Redes Probabilísticas de K-dependência para problemas de classificação binária

Anderson Luiz de Souza

Orientador: Prof. Dr. Francisco Louzada Neto

Coorientador: Prof. Dr. Luis Aparecido Milan

São Carlos

Abril de 2011

Redes Probabilísticas de K-dependência para problemas de classificação binária

Anderson Luiz de Souza

Orientador: Prof. Dr. Francisco Louzada Neto

Coorientador: Prof. Dr. Luis Aparecido Milan

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos

Abril de 2011

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S729rp

Souza, Anderson Luiz de.
Redes probabilísticas de K-dependência para problemas de classificação binária / Anderson Luiz de Souza. -- São Carlos : UFSCar, 2012.
128 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2011.

1. Estatística. 2. Classificadores. 3. Redes probabilísticas. 4. Combinação de classificadores. 5. Redes Bayesianas. I. Título.

CDD: 519.5 (20ª)

Anderson Luiz de Souza

REDES PROBABILÍSTICAS DE K-DEPENDÊNCIA PARA PROBLEMAS DE CLASSIFICAÇÃO BINÁRIA

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 28 de fevereiro de 2011.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Francisco Louzada Neto (DEs-UFSCar/Orientador)

1º Examinador _____
Prof. Dr. Cesar Augusto Cavalheiro Marcondes (DC-UFSCar)

2º Examinador _____
Prof. Dr. Luis A. Milan (DEs-UFSCar/Co-Orientador)

3º Examinador _____
Prof. Dr. Márcio Alves Diniz (DEs-UFSCar)

Agradecimentos

À minha família, principalmente meus pais, Carmen Aparecida Ara de Souza e Valdeci Felisberto de Souza, por todo esforço, compreensão e alicerce fornecido para meu avanço educacional e profissional.

À minha avó Aparecida Baptista Ara, por todo zelo, interesse e solidariedade em todos os aspectos da minha vida.

À minha irmã Crystiane Fernanda de Souza pela tolerância e lazer.

A meu tio José Luis Ara Sobrinho pelos ensinamentos e inspiração.

A Cleyton Zanardo de Oliveira e Felipe Nartis pelo companheirismo e imenso apoio, aos nossos passatempos e longas conversas sobre os mais variados assuntos.

A meu orientador Francisco Louzada Neto pela amizade, oportunidades e pela a experiência que tem me passado em todos esses anos de trabalho.

Aos meus melhores professores desde o Ensino Básico, pois sem bons professores nunca chegaríamos a trilhar as luzes do conhecimento.

A todos os docentes e funcionários do Departamento de Estatística da UFSCar, pela formação e estrutura disponível.

O artista que fica satisfeito com sua obra faltou à vocação.

(C. Lahr)

Resumo

A classificação consiste na descoberta de regras de previsão para auxílio no planejamento e tomada de decisões, sendo uma ferramenta indispensável e um tema bastante discutido na literatura. Como caso especial de classificação, temos o processo de avaliação de risco de crédito, no qual temos o interesse de identificar clientes bons e maus pagadores através de métodos de classificação binária. Assim, em diversos enredos de aplicação, como nas financeiras, diversas técnicas podem ser utilizadas, tais como análise discriminante, análise probito, regressão logística e redes neurais. Porém, a técnica de Redes Probabilísticas, também conhecida como Redes Bayesianas, tem se mostrado um método prático de classificação e com aplicações bem sucedidas em diversos campos. Neste trabalho, visamos exibir a aplicação das Redes Probabilísticas no contexto de classificação, em específico, a técnica denominada Redes Probibilísticas com K-dependência, também conhecidas como redes KDB, bem como comparar seu desempenho com as técnicas convencionais aplicadas no contexto de *Credit Scoring* e Diagnose Médica. Exibiremos como resultado aplicações da técnica baseadas em conjuntos de dados reais e artificiais e seu desempenho auxiliado pelo procedimento de *bagging*.

Palavras-Chave: Redes Probabilísticas, Redes Bayesianas, Naive Bayes, Classificação, Credit Scoring, Diagnose Médica.

Abstract

Classification consists in the discovery of rules of prediction to assist with planning and decision-making, being a continuously indispensable tool and a highly discussed subject in literature. As a special case in classification, we have the process of credit risk rating, within which there is interest in identifying good and bad paying customers through binary classification methods. Therefore, in many application backgrounds, as in financial, several techniques can be utilized, such as discriminating analysis, probit analysis, logistic regression and neural nets. However, the Probabilistic Nets technique, also known as Bayesian Networks, have showed itself as a practical convenient classification method with successful applications in several areas. In this paper, we aim to display the appliance of Probabilistic Nets in the classification scenario, specifically, the technique named K-dependence Bayesian Networks also known as KDB nets,as well as compared its performance with conventional techniques applied within context of the Credit Scoring and Medical diagnosis. Applications of the technique based in real and artificial datasets and its performance assisted by the bagging procedure will be displayed as results.

Keywords: : Probabilistic Networks, Bayesian Networks, KDB, Naïve Bayes, Classification, Credit Scoring, Medical diagnosis.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A quantidade de dados disponível no mundo tem aumentado consideravelmente a cada dia. A necessidade por ferramentas capazes de analisar esses dados motivou o surgimento da área de pesquisa conhecida como mineração de dados, sendo esta uma área intimamente relacionada aos métodos estatísticos.

Neste enredo, de uma forma geral, o conceito de mineração de dados está inserido no processo de *Knowledge Discovery in Databases - KDD*, ou descoberta de conhecimentos em bancos de dados, o qual é responsável pela extração de informações sem conhecimento prévio de um grande banco de dados e seu uso para a tomada de decisões (DINIZ; LOUZADA-NETO, 2000).

Basicamente, podemos considerar que estes procedimentos permitem a transformação de um conjunto de dados brutos em informação e conhecimento úteis em diversas áreas.

Assim, notoriamente, existe a necessidade contínua de teorias e ferramentas estatísticas e computacionais para auxiliar os seres humanos a extrair conhecimento, informação útil e tangível, de crescentes volumes de dados.

Além disso, os procedimentos de mineração de dados são considerados interativos e iterativos. A interatividade é devida ao envolvimento e cooperação de um grupo

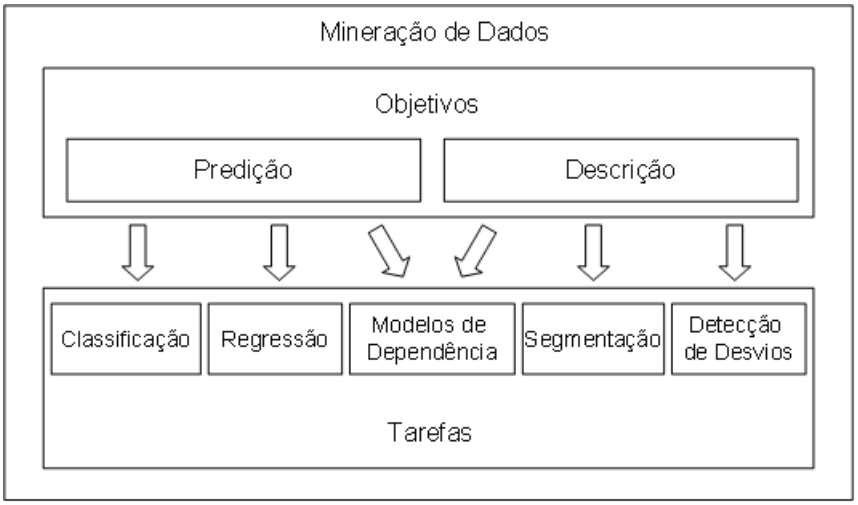


Figura 1.1: Conexões entre os objetivos e tarefas em mineração de dados. Adaptado de Velickov e Solomatine (2000).

responsável, cujo conhecimento referente ao problema analisado auxiliará na execução de todo o processo. Por sua vez, a iteratividade provém de que, frequentemente, este processo envolve repetidas seleções de amostras e aplicações das técnicas de mineração de dados e posterior análise dos resultados obtidos a fim de refinar os conhecimentos extraídos (BRACHMAN; ANAND,1996).

Os problemas tratados em mineração de dados são resolvidos por dois grandes grupos de objetivos (VELICKOV; SOLOMATINE, 2000).

- Descrição: tem como objetivo encontrar padrões, associações ou correlações interpretáveis através da descrição dos dados.
- Predição: realizar inferências sobre um conjunto de dados existente, a fim de prever o comportamento de novas observações. Isso pode ser feito através da construção de um ou mais modelos.

A Figura 1.1 exibe os principais procedimentos utilizados em mineração de dados.

A classificação é a tarefa mais comum dentre as diversas tarefas de mineração de dados (BERRY; LINOFF, 1997). Ela consiste na descoberta de regras de previsão para auxílio no planejamento e tomada de decisões.

Desta forma, geralmente traduzido em um algoritmo, um método de classificação consiste em um sistema de predição para uma variável categórica baseado em um conjunto de variáveis pré-definidas, conhecidas como variáveis explicativas.

Os métodos de classificação têm sido largamente utilizados e se mostram necessários em diversas áreas do conhecimento. Identificando apenas algumas áreas para exemplificação: na área de biomedicina, para realização de diagnósticos, verificação de sequências genéticas, entre outras aplicações; na área financeira e de negócios, para classificar e quantificar risco de empréstimo a clientes, marketing, falência de empreendimentos, operações fraudulentas, entre outros; na Indústria, para predizer e quantificar chances de produção de itens defeituosos e, até mesmo, na *Internet* para classificação de spam, métodos de busca textuais; entre outros casos como vigilância e descobertas científicas.

Embora as soluções em mineração de dados possam ser divididas em dois grandes grupos, como citado anteriormente, existe uma infinidade de métodos relativos a cada uma das tarefas, sendo que, geralmente, um método pode ser utilizado para mais de uma tarefa, ou seja, um mesmo método pode ser utilizado no contexto de classificação e regressão. Especificamente em classificação, podemos citar como mais comuns os métodos: Análise Discriminante, Regressão Logística e Probito, Redes Neurais e Classificação por Árvores (ABDOU *et al.*, 2008).

Alternativamente, o método de Redes Probabilísticas introduzido por Pearl (1988) e difundido na literatura através do nome Redes Bayesianas, pode ser interpretado como um método de descrição e predição em mineração de dados, uma vez que se trata de uma modelagem de dependência. Este método tem sido utilizado recente-

mente e de forma bem-sucedida em diversas áreas, como por exemplo, estimação de risco operacional, diagnóstico médico, *credit scoring*, projeto de jogos computacionais, imputação de dados, entre outras.

O presente trabalho tem entre seus principais objetivos investigar a aplicação da técnica de Redes Probabilísticas no contexto de classificação binária, comparando-a entre seus diversos tipos de ajuste e, também, com as principais técnicas atuais deste enredo. A fim de contribuir com a estatística nacional, no sentido da escassez de literatura referente a esta técnica em nosso país. Bem como a construção de rotinas computacionais específicas que pertimem a utilização geral desta teoria.

Por simplicidade e exemplificação, abordamos a tarefa de classificação aplicada ao enredo de negócios, mais especificamente o contexto de *credit scoring*, e ao enredo da saúde, mais especificamente à problématica de diagnóstico e detecção de doenças, os quais serão expostos a seguir. Assim, a maioria das aplicações em dados reais e exemplos teóricos estão baseados nestas problemáticas e consideram o caso particular de classificação binária.

Este capítulo expõe contextualizações importantes referentes a *credit scoring* e classificação aplicada à área da saúde, teoria básica de probabilidades e teoria da informação, esta última apresentando conceitos fortemente utilizados em Redes Probabilísticas. Por fim, apresenta uma breve apresentação do *Software* `R`, software utilizado durante toda esta dissertação. O Capítulo 2 apresenta conceitos fundamentais da técnica de Redes Probabilísticas. No Capítulo 3, apresentamos problemáticas em Redes Probabilísticas, idéias difundidas sobre estimação de estruturas e estimação dos parâmetros. Na sequência, o Capítulo 4 apresenta métodos e estruturas específicas em Redes Probabilísticas utilizados para o contexto de classificação, bem como sua comparação com as demais técnicas, também apresenta uma nova abordagem de classificação utilizando Redes Probabilísticas via procedimentos de *Bagging*. Por

fim, o Capítulo 5 exibe comentários finais sobre o trabalho.

1.1 *Credit Scoring*

A necessidade de análise de crédito nasceu nos primórdios do comércio conjuntamente com a concessão de empréstimos de dinheiro ou com a autorização de compras a pagar futuramente, pois, desde aquela época, quando um comerciante oferecia demasiado crédito à pessoa errada, corria o risco de perder dinheiro e ter futuros problemas financeiros. Com o passar dos anos, os comerciantes começaram a levantar informações sobre os solicitantes de crédito e catalogá-los para decidir se emprestariam ou não determinada quantia em dinheiro.

Com o desenvolvimento da ciência em análise de dados refletida em métodos precisos, hoje *credit scoring* é um método de avaliação de risco de crédito para aplicação de empréstimos (MESTER, 1997). Baseado em métodos estatísticos para análise de dados, tal método produz um *score* para cada cliente, quantificando o risco deste cliente ser bom ou mau pagador, a fim de minimizar as perdas ou maximizar os ganhos de uma empresa, geralmente financeira.

Por ter como objetivo final a classificação binária de uma determinada característica, são aplicados diversos métodos de tratamento de dados na área de *credit scoring.*

Neste trabalho, exibimos alguns exemplos de aplicações em *credit scoring* para as manipulações mais importantes da técnica de Redes Probabilísticas, especificamente, visualizar o relacionamento das variáveis em dados reais no enredo *credit scoring*, além de expor a aplicação dos procedimentos de classificação a fim de identificar indivíduos maus pagadores.

1.2 Diagnóstico de Doenças

Um amplo sistema de informações hospitalar para atender às necessidades específicas de um hospital contém módulos de internação, registro de ambulatório, assistência ao paciente, registro de farmácia, planejamento de dieta, entre outros. Em suma, um equipamento sofisticado utilizado na prática da medicina moderna e gerador de grande quantidade de dados, um local ideal para procura de novas análises e padrões, ou para validação de hipóteses propostas (WASAN *et al.*, 2006). Para explorar estes dados médicos, inúmeras técnicas de análise estatística, provenientes do enredo de mineração de dados, são aplicadas com sucesso para descobrir conhecimento útil e novo, o qual pode ser utilizado para a rápida e melhor tomada de decisões clínicas (BARNES, 2003)(LABIB e MALEK, 2005). Assim, dado um conjunto de informações relativas a uma doença, desejamos verificar a chance de um paciente desenvolver uma determinada doença como, por exemplo, infarto do miocárdio, câncer de mama, diabetes, dor abdominal, entre outras, além decidir sobre a necessidade de sua internação ou verificar fatores que podem levar à causa de sua enfermidade.

De uma forma geral, as técnicas de classificação propiciam um processo de diagnose diferenciado, uma vez que se baseiam no estudo de doenças quantificadas através de testes médicos ou histórico do paciente, a fim de determinar se este é portador de uma determinada característica ou se necessita de um tratamento diferenciado.

Neste trabalho, exibimos na seção 4.6 a aplicação da técnica de Redes Probabilísticas de K-Dependência em conjuntos de dados médicos reais, prioritarimante focando a performance preditiva das redes.

1.3 Probabilidades

O cálculo das probabilidades teve origem em estudos de jogos de azar na Idade Média. Assim, em 1654, o desenvolvimento desta ciência é devido a uma série de cartas trocadas entre dois matemáticos e pensadores notáveis, Blaise Pascal (1623-1662) e Pierre de Fermat (1601-1665), sobre problemas com apostas em um jogo composto por moedas e dados.

Desde então, a teoria de probabilidades foi amplamente estudada, inclusive pelo também renomado Thomas Bayes, sendo hoje utilizada em diversos procedimentos das Ciências Exatas.

Nesta seção introduzimos uma breve história sobre Thomas Bayes e conceitos fundamentais em probabilidade que são necessários para o entendimento da teoria de Redes Probabilísticas.

1.3.1 Thomas Bayes

Nascido em Londres no ano de 1702 e falecido em Kent, a 58 km de Londres, em 1761, o inglês Thomas Bayes (Figura 1.2) foi matemático e reverendo da igreja presbiteriana e imortalizado por formular um importante teorema de probabilidade, o qual é intitulado pelo seu nome e deu origem, anos depois, a um novo ramo da ciência estatística, denominada Estatística Bayesiana.

Sua família possuía o alinhamento não-conformista – título dado a europeus não-anglicanos ou que prezam a liberdade religiosa – e, antes de seu nascimento, havia feito fortuna no setor da cutelaria, arte de fabricar instrumentos cortantes, um ramo importante em Sheffield, cidade de origem do avô de Thomas Bayes, Richard Bayes.

Thomas Bayes estudou teologia na Universidade de Edimburgo (Escócia) e em 1731 assumiu a paróquia de Tunbridge Wells, em Kent. Historicamente, publicou

Figura 1.2: Única Ilustração conhecida de Thomas Bayes

apenas dois trabalhos em vida, o primeiro intitulado "Benevolência divina" (1731) e o segundo "Uma Introdução a doutrina dos fluxions" no qual ele defendia Isaac Newton contra a crítica de George Berkley, conhecido filosofo irlandês da época. Após sua morte, outro trabalho de sua autoria foi revelado "Ensaio buscando resolver um problema na doutrina das probabilidades", no qual havia formulado o Teorema de Bayes.

Para maiores detalhes sobre a vida de Thomas Bayes consultar Bellhouse (2004), uma completa biografia realizada em comemoração ao seu $300^{\underline{o}}$ aniversário de nascimento.

1.3.2 Conceitos Probabilísticos

As Redes Probabilísticas são ferramentas que utilizam o raciocínio probabilista, ou seja, toda sua metodologia é baseada em probabilidades, especialmente a probabilidade condicional. Para melhor exposição da teoria de Redes Probabilísticas, uma breve revisão da teoria de probabilidades será apresentada abaixo.

1.3.2.1 Probabilidade e suas propriedades

Em poucas palavras, a probabilidade pode ser introduzida, segundo Costa Neto e Cymbalista (2006), como sendo o número que mede a maior ou menor possibilidade de ocorrência de diversos eventos.

Porém, o conceito de probabilidade é, historicamente, cenário de ampla discussão e tem sido definido de diferentes maneiras, sendo que algumas são as definições de probabilidade freqüentista, clássica e subjetiva.

Hoje em dia, a definição axiomática, dada por Komolgorov em 1933, é comumente adotada e considera que a probabilidade é uma função definida em uma classe $\Im$ de eventos de Ω, sendo $\Im$ uma coleção de subconjuntos de Ω a qual é fechada sobre operações enumaráveis de união, interseção e complemento de conjuntos. Deste modo, a probabilidade satisfaz as seguintes condições:

(a) $P(A) \geqslant 0$ para todo $A \epsilon \Im$;

(b) Se $(A_n)_{n\geq 1}$ é uma sequência de eventos de $\Im$, tal que $(A_n)_{n\geq 1}$ são mutuamente exclusivos, então:

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \tag{1.1}$$

(c) $P(\Omega) = 1$

onde A é um evento no espaço $\Im$ e Ω é um conjunto de eventos de interesse denominado espaço amostral.

A definição acima origina as propriedades listadas abaixo, sendo E, F e K quaisquer conjuntos pertencentes a Ω e $\overline{E}$ o conjunto formado por elementos não pertendentes a E, dito complementar de E.

(d) $P(\emptyset) = 0$

(e) $P\left(\overline{E}\right) = 1 - P(E)$

(f) $P\left(E \bigcup F\right) = P(E) + P(F) - P(E \bigcap F)$

(g) SeE, F, ..., K são eventos que não possuem intersecção dois a dois, ditos mutuamente exclusivos:

$$P\left(E \bigcup F \bigcup \ldots \bigcup K\right) = P(E) + P(F) + \ldots + P(K) \tag{1.2}$$

entre outras.

Assim, uma forma objetiva de atribuição de probabilidade ao evento F, quando Ω é finito e enumerável, é dada por:

$$P\left(F\right) = \frac{\sharp(F)}{\sharp(\Omega)} \tag{1.3}$$

onde #(F) é o número de resultados favoráveis ao evento F e #(Ω) é o número de resultados totais, ou seja, o número de resultados no espaço amostral .

Para melhor entendimento dos termos probabilísticos, considere os itens 1, 2, 3 e 4 da Figura 1.3, os quais exibem uma visualização frequente na literatura da teoria de probabilidades baseada na diagramação de Eüller-Venn para os eventos e o seu espaço amostral.

Na Figura 1.3, o item (1) exibe todo o espaço amostral , o item (2) exibe o evento E sob o espaço amostral, o item (3) exibe os eventos E e F sendo mutuamente exclusivos, ou seja, $P(E \bigcap F) = 0$ e, finalmente, o item (4) exibe os eventos E e F como não exclusivos.

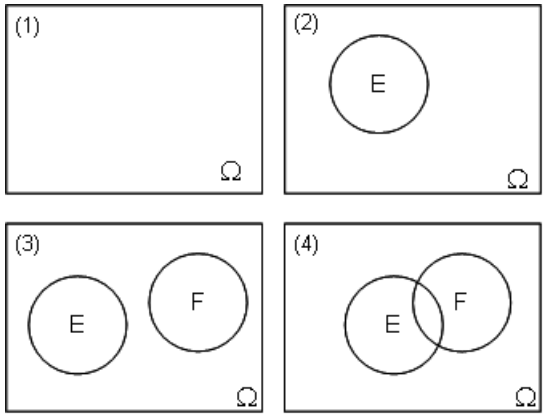


Figura 1.3: Diagramas de Eüller-Venn

1.3.2.2 Probabilidade Condicional

A probabilidade condicional trata do fato de que muitas vezes temos conhecimento sobre um determinado evento, sendo sua ocorrência ou uma informação tomada a priori. Desta forma, surge o interesse de calcular a probabilidade de outro evento possivelmente relacionado ao anterior.

Denotamos como $P(E|F)$ a probabilidade de ocorrência do evento E, sabendo que o evento F ocorreu, ou simplesmente, a probabilidade de E dado F.

Desta forma, temos:

$$P(E|F) = \frac{P(E \bigcap F)}{P(F)} \tag{1.4}$$

Analogamente,

$$P(E \bigcap F) = P(E|F)P(F) \tag{1.5}$$

Temos também, generalizando 1.5 e considerando a notação $P(E \bigcap F) = P(E, F)$,

$$P(E_1, E_2, \ldots, E_n) = P(E_1)P(E_2|E_1)P(E_3|E_2, E_1), \ldots, P(E_n|E_1, E_2, \ldots, E_{n-1})$$

Além disso, considerando $E_1, E_2, \ldots, E_n$ eventos exclusivos e exaustivos, ou seja, eventos que não possuem intersecção e sua união é igual ao espaço amostral, temos para um evento F

$$P(F) = \sum_{i=1}^{n} P(F|E_i)P(E_i)$$

A propriedade acima é comumente denominada de fórmula de probabilidades totais. Note que esta permite calcular a probabilidade de um evento F quando se conhece as probabilidades de um conjunto de eventos distintos, sendo que sua união forma o espaço amostral.

1.3.2.3 Independência condicional probabilística

Assim como a probabilidade condicional, a dependência probabilística é uma das propriedades fundamentais utilizadas na teoria de Redes Probabilísticas. Basicamente, podemos considerar que os eventos E e F são independentes quando existe a relação:

$$P(E|F) = P(E) \Leftrightarrow P(F|E) = P(F) \tag{1.6}$$

A relação 1.6 advém de outra propriedade básica de independência condicional probabilística entre dois eventos, sendo $P(E, F) = P(E)P(F)$.

1.3.2.4 Teorema de Bayes

Como anteriormente, considere o evento F e os eventos $E_1, E_2, \ldots, E_n$ exclusivos e exaustivos, ou seja, que não possuem intersecção dois a dois e sua a união forma o espaço amostral. Assim, o Teorema de Bayes é definido como:

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{\sum_{i=1}^{n} P(F|E_i)P(E_i)} \tag{1.7}$$

O teorema de Bayes é uma junção do teorema de probabilidade condicional e da fórmula de probabilidades totais. Assim, $P(E_i)$ pode ser denominada como probabilidade a priori, $P(F|E_i)$ como verossimilhança e $P(E_i|F)$ como probabilidade a posteriori, ou seja, a probabilidade posterior à observação do evento F. Além disso, o denominador é a decomposição de $P(F)$, ou seja, pode ser considerado como constante normalizadora; desta forma, 1.7 pode ser reescrita na forma 1.8.

$$P(E_i|F) \propto P(F|E_i)P(E_i) \tag{1.8}$$

sendo $\propto$ indicador de proporcionalidade.

Em outros termos, podemos dizer que a probabilidade a posteriori é proporcional à probabilidade a priori multiplicada pela verossimilhança.

1.3.2.5 As distribuições Multinomial e Dirichlet

Estas duas distribuições, aqui introduzidas, são amplamente utilizadas no contexto de Redes Probabilísticas quando métodos de estimação bayesiana são requeridos.

Considere uma variável aleatória X discreta que represente um experimento com r possíveis resultados, sendo que cada tipo de resultado possui uma probabilidade específica $P(X = x_r) = p_r$ e $\sum_{i=1}^{r} p_i = 1$. Além disso, o experimento é repetido de forma independente N vezes, de forma que a variável X_i seja o número de vezes que o resultado x_i está presente na amostra com $i = 1, ..., r$. Temos que a variável X segue distribuição Multinomial, sendo sua função densidade de probabilidade expressa pela fórmula 1.9.

$$P(X_1 = x_1, \ldots, X_r = x_r | N, p_1, \ldots, p_r) = \frac{N!}{x_1! x_2! \ldots x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r} \tag{1.9}$$

sendo que $\sum_{i=1}^{n} X_i = N$.

Considerando o termo $\frac{N!}{x_1! x_2! \ldots x_r!}$ como normalizador, temos 1.10

$$P(X_1 = x_1, \ldots, X_r = x_r | N, p_1, \ldots, p_r) \propto p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r} \tag{1.10}$$

Temos que para um vetor $p = (p_1, p_2, ..., p_r)$ de valores desconhecidos com $\sum_{i=1}^{r} p_i = 1$, podemos assumir que p segue distribuição Dirichlet com parâmetros $\alpha = (\alpha_1, ..., \alpha_r)$ com $\alpha_i > 1$, $\alpha_0 = \sum_{i=1}^{r} \alpha_i$, $E(p_i) = \alpha_i / \alpha_0$ e função densidade de probabilidade expressa pela fórmula 1.11

$$P(p|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_r)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} \ldots p_r^{\alpha_r-1} \tag{1.11}$$

Da mesma forma, podemos considerar o termo $\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_r)}$ como normalizador. Assim, temos 1.12.

$$P(p|\alpha) \propto p_1^{\alpha_1-1} p_2^{\alpha_2-1} \ldots p_r^{\alpha_r-1} \tag{1.12}$$

Assumindo como $P(p|\alpha)$ priori e $P(X_1 = x_1, \ldots, X_r = x_r|N, p_1, \ldots, p_r)$ como verossimilhança , temos que a posteriori $P(p|X, \alpha)$ é dada pela expressão 1.13 a qual tem distribuição Dirichlet com parâmetros $\alpha = (\alpha_1 + x_1, ..., \alpha_r + x_r)$ e $E(p_i) = (\alpha_i + x_i)/(\alpha_0 + N)$.

$$P(p|X, \alpha) \propto p_1^{\alpha_1+x_1-1} p_2^{\alpha_2-1} \ldots p_r^{\alpha_r+x_r-1} \tag{1.13}$$

Notamos que, neste caso, a posteriori possui pertence à mesma família de distribuições que a priori. Assim, dizemos que a família Dirichlet é conjugada para amostras com distribuição Multinomial.

Computacionalmente, os códigos em `R` para esta estimação bayesiana são disponibilizados no Apêndice D.

1.3.2.6 Distribuição Normal e Normal Multivariada

A distribuição Normal é uma das mais importantes e utilizadas distribuições de probabilidade (COSTA NETO e CYMBALISTA, 2006). Considerando X uma variável aleatória contínua, dizemos que $X \sim N(\mu, \sigma^2)$ se sua função densidade de probabilidade é expressa como 1.14, sendo μ o parâmetro relativo à média populacional e σ^2 o parâmetro relativo à variância populacional.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty < x < \infty \tag{1.14}$$

Esta distribuição tem sido utilizada em diversos contextos em Redes Probabilísticas contínuas (GEIGER e HECKERMAN, 1994)(PÉREZ *et al.*, 2006), também são conhecidas como Rede Gaussiana Condicional (RGC). Esta abordagem é uma alternativa à categoriazação de variáveis contínuas. Contudo, a suposição de normalidade para variáveis contínuas pode ser bastante severa, esta é freqüentemente adotada, pois garante uma aproximação razoável para diversas distribuições naturais (JOHN e LANGLEY, 1995).

Neste sentido, consideramos um conjunto de variáveis aleatórias explicativas $X = [X_1, X_2, \ldots, X_k]$ que, em suposição, descrevem uma problemática de classificação e seguem uma distribuição Normal Multivariada de ordem k, isto é, $X \sim N_k(\underset{\sim}{\mu}, \Sigma)$, sendo $\underset{\sim}{\mu}$ o vetor de médias populacionais e Σ a matriz de variância e covariância populacional, $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ & \sigma_2^2 & & \vdots \\ & & \ddots & \sigma_{(k-1)k} \\ & & & \sigma_k^2 \end{pmatrix}$ com σ_i^2 igual a variância de cada variável

X_i e σ_{ij} igual a covariância entre as variáveis X_i e X_j sendo $1 \leq i < j \leq k$. A função de densidade de probabilidade de X é expressa por 1.15, note que se $k = 1$ temos o caso expresso em 1.14.

$$f(x) = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \left(x - \underset{\sim}{\mu} \right)^t \Sigma^{-1} \left(x - \underset{\sim}{\mu} \right) \right\} \tag{1.15}$$

As Redes Probabilísticas que consideram este tipo de estrutura são abordadas no Capítulo 4.

Computacionalmente, os códigos em `R` que consideram este tipo de estrutura para uma rede probabilística são disponibilizados no Apêndice E.

1.3.3 As Redes Probabilísticas podem ser chamadas de Redes Bayesianas?

Existe uma grande discussão na literatura sobre se as Redes Probabilísticas são realmente Bayesianas ou não. Alega-se que esse termo seja uma nomenclatura inadequada. Korb e Nicholson (2004) evidenciam a pronúncia formal do Professor Geoff Webb, especialista em mineração de dados da Universidade Australiana de Monash, que declarou dois pontos de vista:

1. A técnica de Redes Probabilísticas pode ser considerada um método de Data Mining que utiliza métodos não-Bayesianos.

2. As Redes Probabilísticas são um método para representar probabilidades que podem ser interpretadas de forma Bayesiana ou não.

Deste modo, notamos que atualmente essa discussão pode gerar bastante polêmica entre os especialistas da área. Porém, temos que o objetivo fundamental da técnica é realizar inferência e estimativas com base em condicionamentos de informações, o que gera uma ponte de ligação sólida com a filosofia Bayesiana.

Ainda assim, como mostramos neste trabalho, os métodos de estimação dentro da teoria de Redes Probabilísticas podem ser realizados por métodos Bayesianos ou não-Bayesianos.

1.4 Métricas e definições da Teoria da Informação

Nesta seção, exibimos definições provindas da teoria de informação e amplamente utilizadas na teoria de Redes Probabilísticas. Em especial, utiliza-se a medida de informação mútua, geralmente aplicada em algoritmos de treinamento, estimação, de estrutura da rede (BOUCKAERT, 1993; LAM *et al.*, 1994; SAHAMI 1996; SPRITES *et al.*, 1993).

1.4.1 Entropia

O conceito de entropia foi inicialmente introduzido na área de Termodinâmica, a fim de quantificar as perdas inerentes à transformação de uma forma de energia em outra (Callen, 1985). Porém, Shannon (1948) propôs uma maneira probabilística de quantificar a entropia em um trabalho relativo a problemas relacionados à comunicação, sendo esta área comumente denominada hoje de Teoria da Informação.

Desta forma, a entropia pode ser interpretada como uma medida de desordem, aleatoriedade, de uma distribuição e, para uma variável aleatória discreta X, pode ser definida como 1.16.

$$H(X) = E\left[\log\left(\frac{1}{P(X)}\right)\right] = -\sum_{i=1}^{n} p(x_i)\log p(x_i) \tag{1.16}$$

Porém, para que esta seja válida, devemos definir $\log p(x_i) = 0$ se $p(x_i) = 0$. Estratégia matematicamente coerente, pois $\lim_{y\to 0+} y\log y = 0$.

Além disso, $H(X)$ pode ser interpretada como uma medida da nossa incerteza sobre o valor da variável aleatória X antes de observá-la, ou como a quantidade de informação sobre X que foi ganha depois de realizar a observação. Como exemplo, se X assume algum valor x com probabilidade 1 então nenhuma informação é ganha após a observação de X, uma vez que X assume o valor x deterministicamente.

Além disso, para uma variável aleatória discreta X tal que a $P(X = x_i) = p_i$ para $i = \{1, \ldots, n\}$, temos que $0 \leq H(x) \leq \log p$ e o valor máximo é atingido quando $p = 1/n$, ou seja, quando existe uma maior desordem probabilística na variável.

Todo este contexto pode facilmente ser generalizado para variáveis aleatórias contínuas.

1.4.2 Distância de Kullback-Leibler

A distância de Kullback-Leibler (KULLBACK; LEIBLER, 1951) é uma medida não negativa e assimétrica que quantifica a distância entre duas distribuições de probabilidade e, também, baseada no conceito de entropia. Sendo conhecida alternativamente como distância de ganho de informação, distância de perda de informação, ou ainda, entropia relativa, é definida em 1.17.

$$DKL(P_A|P_B) = \sum_{x\epsilon X} P_A(x) log\left(\frac{P_A(x)}{P_B(x)}\right) \tag{1.17}$$

Para o cálculo de $DKL(P_A|P_B)$ devemos também considerar os seguintes limites: $\lim_{\substack{P_A(x)\to 0 \\ P_B(x)\neq 0}} P_A(x) log\left(\frac{P_A(x)}{P_B(x)}\right) = 0$ e $\lim_{\substack{P_B(x)\to 0 \\ P_A(x)\neq 0}} P_A(x) log\left(\frac{P_A(x)}{P_B(x)}\right) = \infty$.

1.4.3 Informação Mútua

Baseada no conceito de entropia, trata-se de outra forma, uma medida, para determinar independência entre variáveis aleatórias, considerando suas respectivas distribuições de probabilidade, e é definida por 1.18.

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned} \tag{1.18}$$

onde é $H(X,Y) = -\sum_{x\epsilon X}\sum_{y\epsilon Y} p(x,y)\log\left(p(x,y)\right)$ a entropia conjunta das variáveis X e Y e $H(X|Y) = -\sum_{x\epsilon X}\sum_{y\epsilon Y} p(x)p(y)\log\left(p(x|y)\right)$ é a entropia da variável aleatória X dado Y.

Ainda, podemos escrever 1.18 como 1.19.

$$I(X,Y) = \sum_{x\epsilon X}\sum_{y\epsilon Y} P(x,y) log\left(\frac{P(x,y)}{P(x)P(y)}\right) \tag{1.19}$$

Sendo uma medida de dependência probabilística, a medida de informação mútua expressa a quantidade de informação que X compartilha com Y. Desta forma,

$I(X,Y) = I(Y,X)$ e quando X e Y são independentes temos que $I(X,Y) = 0$, caso contrário X e Y são variáveis aleatórias dependentes. Além disso, $I(X,Y) = DKL(P(x,y), P(x)P(y))$.

Alternativamente, quando considerado um conjunto de variáveis aleatórias Z, onde $X \subsetneq Z$ e $Y \subsetneq Z$, temos o interesse de verificar a quantidade de informação que X compartilha com Y, dado que Z é conjunto de variáveis aleatórias condicionantes. Ou seja, verificar se X e Y são condicionalmente dependentes dado Z. Neste sentido, podemos utilizar a medida de informação mútua condicional, definida em 1.20

$$I(X,Y|Z) = \sum_{x\epsilon X} \sum_{y\epsilon Y} \sum_{z\epsilon Z} P(x,y,z) log \left(\frac{P(x,y|z)}{P(x|z)P(y|z)} \right) \tag{1.20}$$

Como no caso anterior, $I(X,Y|Z) \geq 0$ e $I(X,Y|Z) = I(Y,X|Z)$. Assim, condicionadas ao conjunto Z, X e Y são independentes se $I(X,Y|Z) = 0$.

Em termos probabilísticos, $P(X,Y|Z) = P(X|Y,Z)P(Y|Z)$ e se $P(X|Y,Z) = P(X|Z)$ então, para um conhecido Z, Y não impacta nos valores de X. Isto é, X é independente de Y, dado Z e vice-versa.

Desta forma, é conveniente estudar a distribuição da medida informação mútua, uma vez que temos interesse de verificar se X e Y são condicionalmente independentes. Porém, esta distribuição é bastante complexa e, até o momento, não possui uma forma difundida e exata na literatura (HUTTER; ZAFFLON, 2005).

Neste sentido, McGill(1954) verifica que expressões que envolvem entropia são extremamente associadas com a razão de verossimilhança, sendo que, para amostras grandes, a distribuição da informação mútua condicional pode ser aproximada para uma distribuição χ^2. No Geral, a Informação mútua está relacionada de forma muito próxima com a estatística do teste χ^2.

Segundo Kullback (1968), dado um conjunto de dados D com N elementos, para avaliar $H_0 : X$ e Y são condicionalmente independentes dado Z, temos a relação 1.21.

$$2NI(X,Y|Z) \sim \chi^2_k \tag{1.21}$$

onde $k = (k_x - 1)(k_y - 1)k_z$ graus de liberdade e k_i =número de categorias em i. Se $Z = \emptyset$, $I(X,Y|Z) = I(X,Y)$ e $k = (k_x - 1)(k_y - 1)$ graus de liberdade.

Assim, rejeitamos a um nível α de significância a independência condicional entre X e Y dado Z, se $2NI(X,Y|Z) > \chi^2_{k;\alpha}$.

Para o caso específico em que Z é uma variável aleatória discreta com r possíveis resultados e sua distribuição de probabilidade dada por $P(Z = z) = P(z)$ e X é uma variável aleatória com densidade de probabilidade normal com parâmetros μ e σ^2, assumimos que a variável X condicionada a $Z = z$ segue uma distribuição normal com parâmetros μ_z e σ^2_z. A informação mútua entre as variáveis X e Z é dada por 1.22(PÉREZ *et al.*, 2006).

$$I(X,Z) = \frac{1}{2}\left[\log\left(\sigma^2\right) - \sum_{z=1}^{r} P(z) log\left(\sigma^2_z\right)\right] \tag{1.22}$$

Se a função de probabilidade conjunta das variáveis aleatórias X e Y condicionadas a $Z = z$ segue uma distribuição normal multivariada de ordem $k = 2$ com vetor de médias $\underset{\sim z}{\mu} = \left(\mu_{X|z}, \mu_{Y|z}\right)$ e matriz de variância e covariância $\Sigma_z =$

$\begin{pmatrix} \sigma^2_{X|z} & \sigma_{X,Y|z} \\ \sigma_{X,Y|z} & \sigma^2_{Y|z} \end{pmatrix}$, a informação mútua condicional entre as variáveis X e Y condicionadas a Z é dada por 1.23 (PÉREZ *et al.*, 2006).

$$I(X,Y|Z) = -\frac{1}{2}\sum_{z=1}^{r} P(z) log\left(1 - \rho^2_z\left(X,Y\right)\right) \tag{1.23}$$

onde $\rho^2_z\left(X,Y\right) = \frac{\sigma_{X,Y|z}}{\sqrt{\sigma^2_{X|z}\sigma^2_{Y|z}}}$ é o coeficiente de correlação entre X e Y condicionadas a $Z = z$.

De Campos (2006) demonstra que, para uma Rede Probabilística, maximizar a soma das métricas de informação mútua entre as variáveis e seus pais (Seção 2.1) é o mesmo que minimizar a distância de Kullback-Leibler e, por sua vez, é equivalente a maximizar o logaritmo da verossimilhança da rede.

Computacionalmente, os códigos em `R` para o cálculo da informação mútua são disponibilizados no Apêndice E.

1.5 O *Software* `R`

O `R` (R Development Core Team, 2005) é ao mesmo tempo uma linguagem e um ambiente de programação estatístico para análise e manipulação de dados, realização de cálculos e visualização de gráficos. Originalmente, foi desenvolvido em meados dos anos 90 por Ross Ihaka e por Robert Gentleman na Universidade de Auckland, Nova Zelândia. Os autores batizaram a linguagem com este nome devido as iniciais de ambos (Ross e Robert) e também como uma brincadeira parcial com a linguagem S. A linguagem S é uma das diversas linguagens estatísticas desenvolvidas na década de 70 pelos laboratórios da AT&T, empresa norte-americana de telecomunicações, também criadora da difundida Linguagem `C`.

Atualmente, o `R` é um *Software* livre e compatível com diversos sistemas opera-

cionais, dentre eles Windows, Linux, Macintosh e Unix. Disponível sob os termos do *Free Software Foundation's GNU General public License*, proporcionando contribuições do mundo inteiro. Apesar do seu caráter gratuito o `R` é uma ferramenta bastante poderosa com boas capacidades ao nível da programação e um conjunto bastante vasto de pacotes, programas acessórios construídos por por sua comunidade internacional, que acrescentam diversas potencialidades à versão base do `R`.

Para a elaboração deste trabalho, consideramos o Software `R` versão 2.11.1 e um conjunto específico de pacotes:

- `infotheo` v1.1.0 (MEYER, 2009): neste pacote há diversas implementações de cálculo das medidas de informação mútua, Seção 1.4.3. Em especial, a função `mutinformation()` a qual calcula a informação mútua entre duas variáveis discrteas, `condinformation()` que calcula a informação mútua condicional entre duas variáveis discrestas condicionadas a uma terceira variável e a função `discretize()` que categoriza as variáveis contínuas de um conjunto de dados baseando-se no critério de equifrequência entre as categorias.

- `klaR` v0.6-5 (WEIHS *et al.*, 2005): possui um conjunto de ferramentas direcionadas a classificação, em especial a função `NaiveBayes()` a qual ajusta o classificador de Naive Bayes, que será introduzido na Seção 4.1. Neste caso, utilizamos este pacote para validar o algoritmo construído para ajuste das Redes Probabilísticas.

- `MASS` v7.3-11 (VENABLES e RIPLEY, 2002): possui um grande de número de funções e conjuntos de dados. Neste caso, utilizamos este pacote para aplicar a técnica de Análise Discrimimante, que será introduziada na Seção 4.3.1.

- `nnet` v7.3-1 (VENABLES e RIPLEY, 2002): utilizado para o ajuste da téc-

nica de Redes Neurais utilizando uma camada de variáveis ocultas, que será apresentada na Seção 4.3.4.

- `diagram` v1.5.2 (SOETAERT, 2008): pacote utilizado para a contrução de gráficos de redes e diagramas, baseando-se em uma matriz de transição.

Com o objetivo de realizar a diagramação de uma Rede Probabilística utilizando o Software `R`, os pacotes direcionados a modelagem gráfica foram estudados (`gRbase`, `dynamicGraph, igraph`, entre outros), porém nenhum deles havia a possibilidade de exibição de um grafo planar, ou seja, uma Rede Probabilística onde as setas de dependência não se cruzam. Neste sentindo, optamos por utilizar o pacote gráfico `diagram`.

1.6 Comentários Finais

Neste capítulo, introduzimos o contexto em que as Redes Probabilísticas estão inseridas, bem como direcionamos tal necessidade para a área de *credit scoring* e diagnose médica.

A respeito da teoria de probabilidade, exibimos importantes propriedades que serão utilizadas ao decorrer do trabalho, sendo as mais importantes a propriedade de dependência, o Teorema de Bayes, o relacionamento entre as distribuições de probabilidade Multinomial e Dirichlet e a distribuição Normal Multivariada.

Além disso, exibimos definições gerais de entropia e informação mútua, as quais serão utilizadas para estimação da estrutura de uma rede, bem como apresentamos sucintamente o *Software* `R`.

No próximo capítulo exibimos os conceitos gerais de uma Rede Probabilística.

Capítulo 2

Redes Probabilísticas

As Redes Probabilísticas, também conhecidas como Redes causais, Rede de crença e Gráficos de dependência probabilística, surgiram na década de 80 e têm sido aplicadas em uma grande variedade de atividades do mundo real (BOBBIO *et al.*, 2001). Algumas aplicações atuais se estendem a áreas como finanças (CHANG *et al.*, 2000), saúde (ABICALAFF; AMARAL; DIAS, 2004) (KORB; NICHOLSON, 2004), desenvolvimento de jogos (VIEIRA FILHO; ALBUQUERQUE, 2007), entre outras.

Ainda, as Redes Probabilísticas vêm sendo bastante utilizadas em áreas financeiras para a estimação de risco operacional e *credit scoring* (ex: Sistema Bayes-Credit, um sistema criado por Nykredit, uma das principais empresas no mercado dinamarquês de financiamento imobiliário), e possui vários programas específicos disponíveis como, por exemplo, os softwares Netica (www.norsys.com) e Hugin (www.hugin.com).

Segundo Neapolitan (2004), a técnica de Redes Probabilísticas surgiu em um contexto no qual há um grande número de variáveis e surge o interesse de verificar qual a influência probabilística não direta de uma variável para as demais.

Assim, a teoria de Redes Probabilísticas combina princípios de Teoria de grafos, teoria de probabilidades, Ciência da Computação e Estatística (BEN-GAL, 2007).

Além disso, as Redes Probabilísticas podem ser consideradas uma representação visual e informativa da tabela de probabilidade conjunta de todas as variáveis que envolvem o domínio do problema.

Na literatura especializada, uma terminologia específica é utilizada para definir tipos de variáveis, dependências probabilísticas e outras propriedades das Redes Probabilísticas. Neste trabalho, optamos por simplificar tal terminologia quando possível, aproximando-a de termos utilizados na modelagem estatística de dados.

O presente capítulo tem como objetivo introduzir conceitos básicos da teoria de Redes Probabilísticas, que envolvem os tipos de estruturas de teoria de grafos, noções de evidência, propriedade markoviana, equivalência, noção de independência, definição básica para construção e ordem das variáveis, bem como exibir breves exemplos.

2.1 Estrutura

Nesta seção, serão introduzidos conceitos elementares dentro da estrutura gráfica de uma Rede Probabilística, em sua maioria um conjunto de nomenclaturas originadas através das relações visualmente perceptíveis da estrutura gráfica.

2.1.1 Elementos básicos

Uma Rede Probabilística é uma representação gráfica de variáveis e suas relações para um problema específico. Tal representação é comumente chamada de grafo, sendo este um elemento fundamental da rede.

O estudo dos grafos é realizado pelo ramo da matemática denominado Teoria de Grafos e diz respeito ao estudo das relações de seus elementos, os quais são comumente chamados de nós e arcos. Os nós são elementos principais, os quais

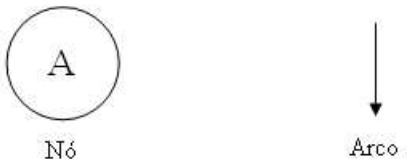


Figura 2.1: Elementos básicos da Teoria de Grafos

representam as variáveis aleatórias consideradas no problema e são representados por círculos. Os arcos são setas que representam a relação de direta dependência entre um nó e outro, ou seja, representa a dependência probabilística direta entre duas variáveis. Esses elementos podem ser visualizados na Figura2.1.

2.1.2 Estruturas de teoria de grafos

Existem diversos tipos de aplicações da Teoria de Grafos na literatura. Maiores detalhes podem ser encontrados em Feofiloff *et al.* (2007). Existem diversos tipos de estruturas básicas dentro da Teoria de Grafos. Para uma visualização geral, tais estruturas são exibidas na Figura2.2.

A teoria de Redes Probabilísticas é construída considerando grafos direcionados, conectados e acíclicos, frequentemente referenciados pela sigla DAG (directed acyclic graph).

O termo "direcionado" faz referência à presença de direção nos arcos, o termo "conectado" é utilizado para designar que todos os nós estão conectados na rede e, por fim, o termo "acíclico" se refere à propriedade de não-retorno para um nó após seguida a direção dos arcos.

Através da Figura2.2, notamos que as Redes Probabilísticas envolvem apenas alguns tipos de estruturas básicas: a estrutura de conexões simples, que engloba as estruturas de árvore simples e poliárvore, e a estrutura de múltiplas conexões.

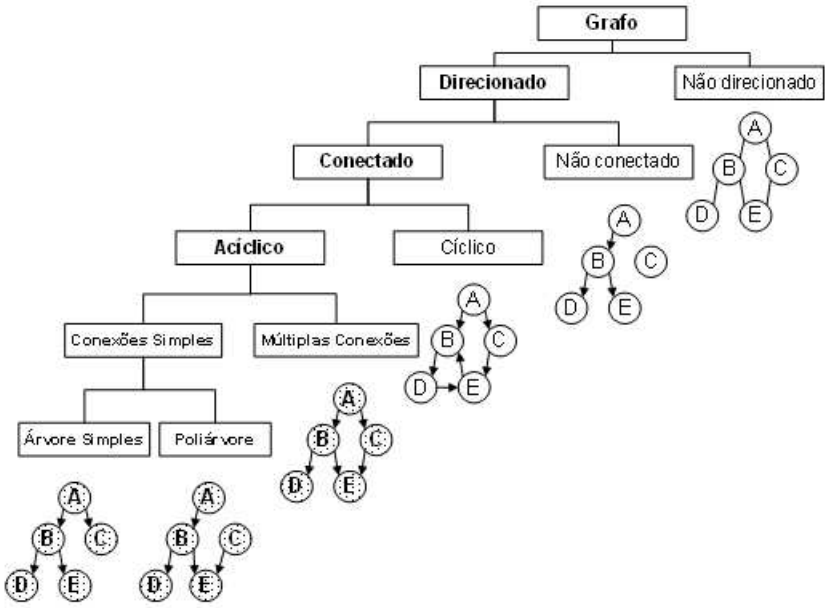


Figura 2.2: Estruturas básicas existentes dentro da Teoria de Grafos

Para as estruturas de conexões simples é dada a regra geral de que existe apenas um caminho que liga uma variável a outra, independente da direção dos arcos. Analogamente, para as estruturas de múltiplas conexões há mais de um possível caminho que liga uma variável a outra, independentemente da direção dos arcos.

A subdivisão das estruturas de conexão simples se dá pelo número de nós que originam a rede, ou seja, nós que não possuem nenhum arco chegando, apenas arcos partindo. Assim, como notamos na Figura 2.2, as estruturas de árvores simples possuem apenas uma variável que origina a rede (variável A) e as estruturas de poliárvore possuem duas (ou mais) variáveis que originam a rede (variáveis A e C). Estas variáveis geralmente possuem um nome específico, o qual será apresentado no próximo item.

2.1.3 Hierarquia entre nós

Dentro da terminologia de Redes Probabilísticas, outros termos também são comuns e utilizados para considerar a hierarquia de nós dentro da rede, o que é o caso dos termos "pai" e "filho". Esses termos referem-se à relação de dependência direta entre dois nós por meio do arco que os conecta. O nó de onde o arco parte é designado nó pai e o nó sobre qual o arco incide é designado nó filho. Considerando a estrutura de simples conexões da Figura 2.2, o nó A é pai do nó B, sendo o nó B filho do nó A. Analogamente, o nó B é pai dos nós C e D, sendo os mesmos filhos do nó B.

Um nó que não possui filhos é chamado de folha e um nó que origina a rede, ou seja, que não possui pais, é chamado de raiz.

Os nós antecedentes a um determinado nó A, ou seja, o(s) pai(s) e seus respectivos pais e assim por diante, são denominados como ancestrais de A. Da mesma forma, os nós derivados de determinado nó A, ou seja, o(s) filho(s) e seus respectivos filhos e assim por diante, são denominados como descendentes de A, analogamente a uma

estrutura de genealogia.

2.1.4 Formalização estatística da estrutura

Como dito anteriormente, em Redes Probabilísticas cada variável aleatória do estudo é representada por um nó. Por esse motivo, iremos substituir o termo "nó" pelo termo "variável", ou seja, ao nos referimos ao nó A, iremos representá-lo pelo termo variável A. Estendendo tal conceito para a hierarquia de nós, temos que a variável A é pai da variável B.

Os valores das variáveis podem ser de qualquer tipo de escala, contínua ou discreta. Porém, segundo Korb e Nicholson (2004), a tecnologia de Redes Probabilísticas é primeiramente direcionada ao tratamento de variáveis discretas, como por exemplo, para a confecção de algoritmos de inferência. Além disso, as variáveis contínuas podem ser facilmente transformadas em variáveis discretas através de simples categorizações. Analogamente, a literatura desenvolvida até o presente momento é focada em variáveis explicativas discretas (PÉREZ *et al.*, 2006). Uma possibilidade de trabalho é através de misturas entre variáveis contínuas e discretas, porém existe a condição básica de que uma variável discreta não deve possuir variáveis pais contínuas. Um possível tratamento para variáveis explicativas puramente contínuas é supor que as mesmas seguem uma distribuição normal, modelo de rede também conhecido como Rede Gaussiana Condicional (GEIGER e HECKERMAN, 1994).

De uma forma geral, neste trabalho, consideramos para o caso discreto, a estrutura de modelagem baseada em que X segue uma distribuição Multinomial, e para o caso contínuo, em que X segue uma distribuição Normal. Assim, uma Rede Probabilística é definida pelo trio (ξ, θ, X), onde ξ é uma estrutura DAG e θ é um conjunto de parâmetros específicos de distribuições de probabilidades condicionais envolvendo um conjunto X de variáveis aleatórias puramente discretas ou puramente contínuas.

Tabela 2.1: Tabela de Probabilidade Condicional P(C|A,B)

C	A	B	$P(C\|A,B)$
1	1	1	θ_1
1	1	0	θ_2
1	0	1	θ_3
1	0	0	θ_4
0	1	1	θ_5
0	1	0	θ_6
0	0	1	θ_7
0	0	0	θ_8

2.1.5 Tabela de probabilidades condicionais

Um elemento importante dentro da estrutura de Redes Probabilísticas para o caso puramente discreto é a tabela de probabilidade condicional (TPC). Trata-se da exibição dos parâmetros de probabilidade condicional da variável sendo condicionada a seu(s) pai(s).

Por exemplo, dado o conjunto de três variáveis A, B e C, todas dicotômicas assumindo valores binários, onde A e B são pais da variável C, temos a 2.1.

Com base nas definições acima, podemos exibir um exemplo de Rede Probabilística.

Os códigos em `R` para a construção das tabelas de probabilidades condicionais estão disponíveis no Apêndice C.

2.1.6 Exemplo Básico de uma Rede Probabilística

Considere uma Rede Probabilística teórica dada sua estrutura já conhecida e relacionando seguintes variáveis binárias:

- Sexo { M, F };
- Idade { <20 anos, >=20 anos };

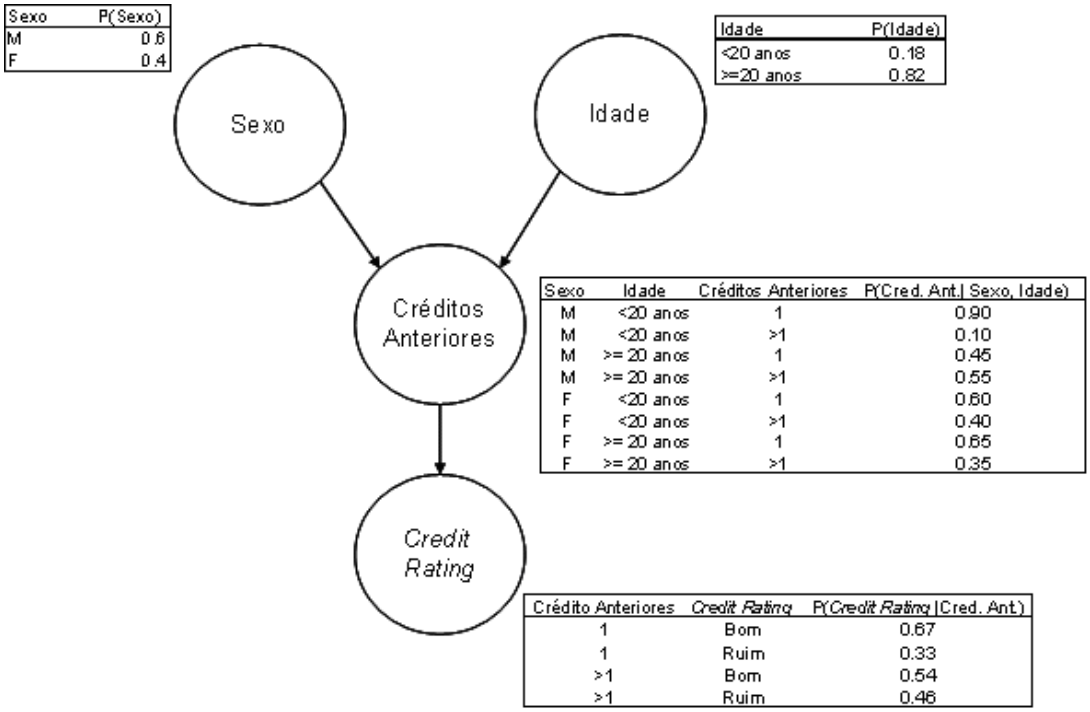


Figura 2.3: Exemplo de Rede Probabilística para dados de *Credit Scoring*.

- Créditos Anteriores { 1, >1 };
- *Credit Rating* { Bom , Ruim }.

Assim, a rede é representada pela Figura 2.3.

Considerando o exemplo da Figura 2.3 temos que as variáveis Sexo, Idade, Créditos Anteriores e *Credit Rating* são representadas por seu respectivo nó na rede, sendo Sexo e Idade variáveis pais da variável Crédtios Anteriores, e a última, por sua vez, pai da variável *Credit Rating*. Realizando uma análise hierárquica, as variáveis Sexo e Idade são classificadas na rede como variáveis raízes e *Credit Rating* como folha.

Além disso, notamos que Sexo e Idade influenciam diretamente a variável Créditos Anteriores, que, por sua vez, influencia probabilisticamente, de uma forma direta, a variável *Credit Rating*.

Interpretando os relacionamentos, se o cliente é do sexo masculino, ou não, isso influencia na probabilidade do cliente ter um, ou mais, créditos anteriores realizados na instituição. Se o cliente é menor de 20 anos, ou não, também influencia a probabilidade do cliente ter um ou mais créditos anteriores realizados na instituição. Assim, a probabilidade do cliente ter, ou não, realizado requisição de créditos anteriormente na instituição financeira influencia a probabilidade dele ser classificado como um bom pagador ou mau pagador.

Para cada uma das variáveis e seus cruzamentos condicionais, temos uma tabela de probabilidade condicional (TPC) explicando numericamente a chance da cada categoria – evento – ocorrer dadas as premissas anteriores.

2.2 Evidência

Dada a estrutura gráfica DAG, outra definição é importante para a teoria de Redes Probabilísticas. Esta é denominada evidência e refere-se ao fato de uma variável ser indicada pelo usuário da rede, ou seja, uma variável aleatória com valor conhecido e acoplado à Rede Probabilística com estrutura já conhecida. Basicamente, podemos definir uma evidência como uma observação.

Considere o exemplo da Figura2.3. Podemos observar que um novo cliente possui a idade de 18 anos; assim, na rede, indicamos a variável Idade para a categoria respectiva, ou seja, tomamos como conhecida a observação Idade $<$20 anos, apenas esta informação do cliente é conhecida. A variável idade é classificada como uma evidência para a rede. A Figura2.4 exibe uma demonstração visual para Idade $<$20 anos.

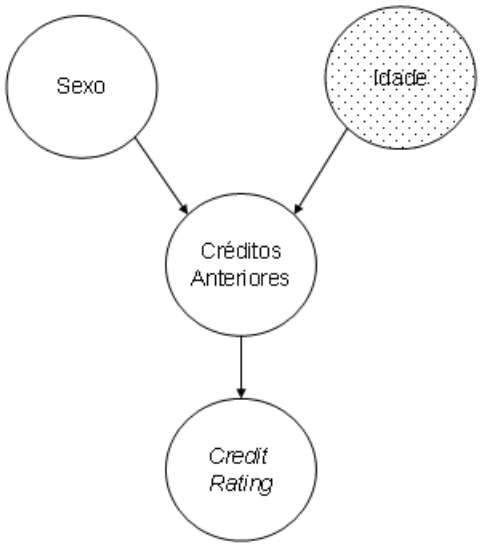


Figura 2.4: Rede Probabilística tendo como evidência a variável Idade.

2.3 Propriedades Markovianas

Assim como em alguns tipos de processos estocásticos, a dinâmica de uma Rede Probabilística é controlada pela propriedade de Markov, a qual rege que não existem dependências diretas entre as variáveis de uma Rede Probabilística que não estão explícitas através da apresentação orientada dos arcos, ou seja, cada variável possui dependência direta apenas de sua (s) variável (eis) pai (s).

A partir de todas as propriedades acima, temos que uma Rede Probabilística é um par (ξ, θ) definido sobre um conjunto de variáveis aleatórias $X = \{X_1, X_2, \ldots, X_k\}$, onde cada X_i corresponde a uma variável da rede, satisfazendo a propriedade de Markov:

$$P\left[X_i | X_j, pais(X_i)\right] = P\left[X_i | pais(X_i)\right] \tag{2.1}$$

$\forall 1 \leq i < j \leq k$.

Consideremos a distribuição de probabilidade conjunta de uma Rede Probabilística com k variáveis e a propriedade 2.1. Temos que em uma Rede Probabilística (ξ, θ) , definida sobre um conjunto de variáveis aleatórias $X = \{X_1, X_2, \ldots, X_k\}$, a probabilidade conjunta de toda a rede é dada através da expressão 2.2.

$$P\left[X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\right] = \prod_{i=1}^{k} P\left[X_i | pais(X_i)\right] \tag{2.2}$$

Ou seja, as propriedades probabilísticas estão intimamente ligadas com o condicionamento da variável com seu (s) pai (s) respectivo (s). Note que 2.2 é resultado direto do desenvolvimento do Teorema de Bayes visto na seção 1.3.2.4., dada a propriedade 2.1.

Para o exemplo da Figura 2.3, as variáveis Sexo e Idade são condicionalmente independentes, pois não existe nenhum arco relacionando-as. Além disso, *Credit Rating* é diretamente independente de Sexo e Idade, a variável *Credit Rating* depende apenas diretamente da variável Créditos Anteriores, a qual é sua variável pai.

Uma Rede Probabilística na qual cada dependência probabilística entre as variáveis é dada por um único arco é chamada de Rede perfeita (KORB; NICHOLSON, 2004).

Outro conceito muito utilizado na teoria de Redes Probabilísticas é a cobertura de Markov, que consiste no conjunto formado pelas variáveis pais, variáveis filhos e pais dos filhos de uma determinada variável. Como exemplo, temos que a cobertura de Markov para a variável Idade da Figura 2.4 envolve a variável Créditos Anteriores (variável filho da variável Idade) e a variável Sexo (variável pai de uma variável filho da variável Idade). Note que a variável Idade não possui variáveis pais, se estas existissem seriam consideradas na cobertura de Markov. Outro exemplo de cobertura

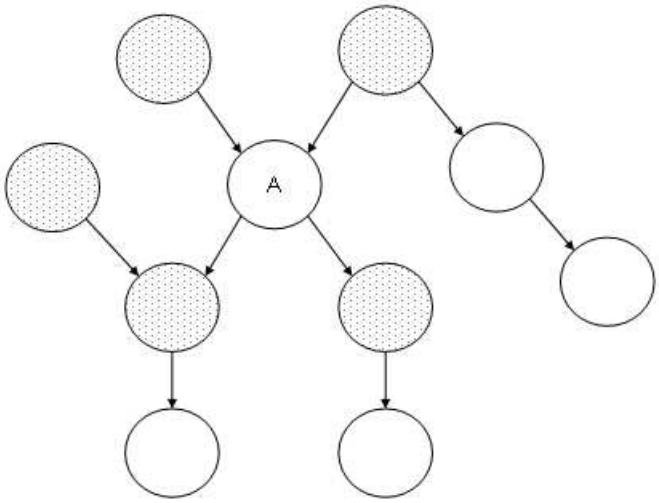


Figura 2.5: Cobertura de Markov de A representada pelas variáveis-nó em cinza.

de Markov pode ser visualizado na Figura 2.5, que exibe a cobertura de Markov para a variável A.

2.4 A propriedade de d-separação

Através das propriedades markovianas, notamos que uma variável é independente de outra se não existe um arco conectando-as. Porém, é possível definir independência quando existe entre as variáveis analisadas um grupo específico de variáveis, podendo ser um grupo de evidências, por exemplo.

Neste caso, surge o conceito de d-separação. Para defini-la, consideremos alguns tipos de conexões dadas em Neopolitan (2004). Sejam X, Z e Y variáveis de uma Rede Probabilística , definimos alguns tipos de conexão:

1. Se $X \rightarrow Z \rightarrow Y$, temos um relacionamento *head-to-tail*;

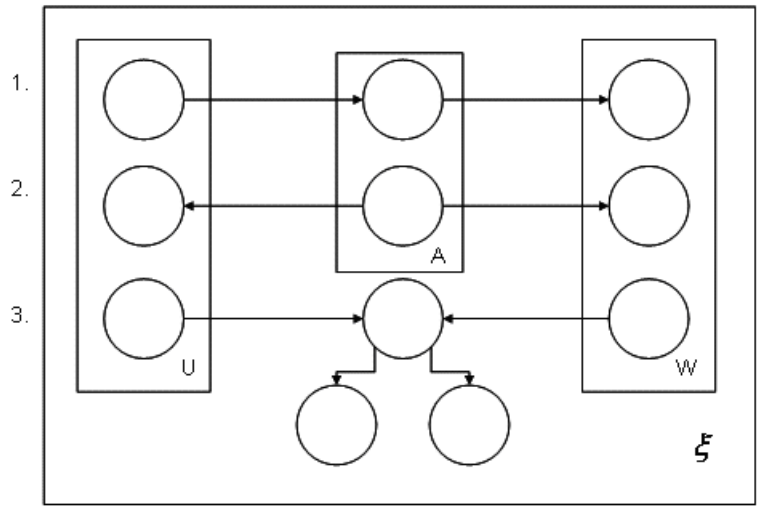


Figura 2.6: Tipos de d-separação, U e W d-separados

2. Se $X \leftarrow Z \rightarrow Y$, temos um relacionamento *tail-to-tail*;

3. Se $X \rightarrow Z \leftarrow Y$, temos um relacionamento *head-to-head*.

Podemos definir $A \subset V$, sendo X e $Y \in \{V - A\}$. Desta forma, para os casos 1 e 2, se consideramos que $Z \in A$, a variável Z bloqueará o caminho entre X e Y. Para o caso 3, se consideramos que Z e seus descendentes $\notin A$, a variável Z bloqueará o caminho entre X e Y. Se o caminho entre duas variáveis, ou conjunto de variáveis é bloqueado, dizemos que essas variáveis ou conjuntos são d-separados.

A Figura2.6, retirada de Marques e Dutra (1999), ilustra os três casos de d-separação, onde os conjuntos U e W são d-separados.

Maiores detalhes sobre d-separação são dados em Neapolitan (2004).

2.5 Equivalência de Markov

Existem inúmeras estruturas possíveis no enredo de Redes Probabilísticas. Porém, podemos construir para cada conjunto de variáveis um grupo de estruturas extremamente semelhantes, chamadas de equivalentes de Markov.

Segundo Neapolitan (2004), dois grafos são equivalentes quando mantêm as mesmas independências condicionais. Ou seja, dois grafos são considerados equivalentes quando conservam as mesmas ligações de arcos entre as variáveis independentemente da direção, com exceção às ligações *head-to-head*, ou seja, quando uma variável filho possui mais de uma variável pai.

Assim, analisando a Figura 2.7, notamos que a estrutura (a) não é equivalente a (b), pois, além de não preservar a conexão *head-to-head* $C \rightarrow E \leftarrow D$, a estrutura (b) não mantém a conexão entre as variáveis A e B. Esses mesmos motivos fazem (b) não equivalente à estrutura (c).

Comparando a estrutura (a) com (c), notamos que existe apenas diferença entre a direção de ligação entre as variáveis A e B, ou seja, (a) e (c) são equivalentes. Dizemos que (a) e (c) pertencem à mesma classe de equivalência markoviana.

2.6 Método geral para a construção de uma Rede Probabilística

A construção de uma Rede Probabilística não é trivial. Além de existirem vários métodos para a estimação de estruturas de rede através do conjunto de dados, os métodos podem ser influenciados por fatores como a ordem e escolha das variáveis que compõem o problema. Esse problema proporciona atualmente intensas pesquisas buscando um método ótimo para estimação de estruturas DAG para domínios de

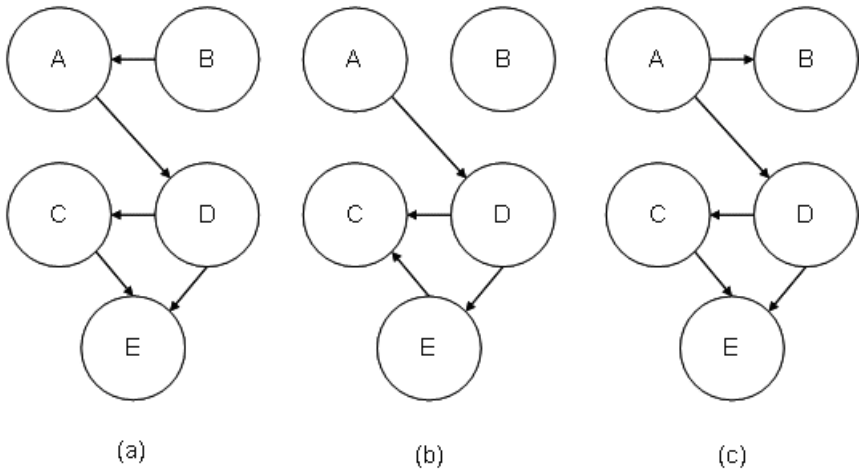


Figura 2.7: Exemplo de identificação de Redes Probabilísticas Markov equivalentes.

problemas práticos.

Porém, de uma forma geral, Pearl (1988) criou um algoritmo baseando-se nas propriedades 2.1 e 2.2, no qual, dado um conjunto de variáveis discretas ordenadas, constrói uma Rede Probabilística única, adicionando as variáveis à rede em sua ordem e acrescentando arcos para a formação da estrutura. Assim, cada variável é conectada às variáveis antigas da rede, o que garante que a estrutura seja sempre acíclica. A proposta de Pearl é exibida no Quadro2.1.

Para uma Rede Probabilística ser adequada, ela deve ser perfeita, ou seja, todos arcos devem expressar corretamente as dependências entre as variáveis.

Desta forma, é fácil notar que, para a construção de uma Rede Probabilística, devemos escolher uma ordem correta para as variáveis, pois diferentes ordens podem gerar Redes Probabilísticas diferentes. Korb e Nicholson (2004) sugerem que primeiramente consideremos as variáveis passíveis a serem raízes e suas variáveis independentes, a seguir as demais variáveis.

Outros métodos de construção de Redes Probabilísticas serão apresentados no

Algoritmo 2.1 Inicialmente proposto por Pearl (1988) para a construção de uma Rede Probabilística.

1. Escolha um conjunto de variáveis X_i que em suposição descreva o problema;

2. Escolha uma ordem para as variáveis;

3. Para todas as variáveis em ordem, faça:

- **3.1.** Escolha a variável X e adicione-a na rede;
- **3.2.** Determine os pais da variável X dentre os nós que já estão na rede, que satisfaçam $P\left[X_i|X_j, pais(Xi)\right] = P\left[X_i|pais(X_i)\right]$.
- **3.3.** Construa a tabela de probabilidade condicional (TPC) para X.

decorrer do trabalho.

2.7 Comentários finais

Neste capítulo, foram apresentamos conceitos básicos sobre a técnica de Redes Probabilísticas, sendo estes de suma importância para o entendimento geral do método. Alguns dos conceitos mais importantes englobam a propriedade de d-separação, base para diversos tipos de cálculos, e a propriedade de cobertura de Markov utilizada em algoritmos para estimação de probabilidades condicionais.

Além disso, introduzimos a idéia básica para a criação de uma estrutura de Redes Probabilísticas. Porém, a construção geral de uma estrutura não é simples, além de existirem vários métodos para este mesmo objetivo.

Neste contexto, no próximo capítulo exibimos como estimações podem ser realizadas.

Capítulo 3

Estimação em Redes Probabilísticas

O termo "aprendizado" é muito comum no contexto de mineração de dados e denota a assimilação de experiência que gera a capacidade de um agente ou sistema obter sucesso em determinada tarefa.

O aprendizado estatístico está intimamente ligado ao processo de aprendizagem quando existem incerteza e variabilidade. Para isso, através de um conjunto de dados, utilizamos o processo de estimação e validação do sistema em estudo, sendo aplicada qualquer técnica estatística que se enquadre ao domínio do problema.

Devido à dificuldade da construção de uma Rede Probabilística unicamente consultando um especialista, existe o interesse de se estimar todos os elementos da rede, estes compondo sua estrutura, e as probabilidades condicionais de cada TPC, também chamadas de parâmetros ou elementos numéricos.

Até o presente momento, assumimos que as estruturas e as probabilidades condicionais já estavam definidas. Porém, a partir de agora temos o interesse de estimar a rede por completo.

Neste capítulo, exibimos de uma forma rápida, métodos para estimação conhecidos na literatura. Assim, apresentamos métodos específicos para ambos objetivos, a estimação de parâmetros e a estimação de estrutura.

3.1 Estimação de estrutura

Neste primeiro tipo de estimação, estamos interessados na busca da melhor estrutura de Redes Probabilísticas para um determinado conjunto de dados, ou seja, a melhor disposição de dependências e independências entre as variáveis que explique de maneira satisfatória o problema em estudo.

A estimação de estrutura, também conhecida na literatura como aprendizado de estrutura, de uma Rede Probabilística, pode ser vista como um processo que gera uma abordagem gráfica das características que definem o domínio de estudo, pois é uma representação da distribuição conjunta de um grupo de variáveis aleatórias.

Considerando como a distribuição conjunta de um grupo de k variáveis aleatórias, temos o interesse de realizar estimativas de $P(X_1, X_2, \ldots, X_k)$ a fim de encontrar uma estrutura gráfica que seja compatível com P. Porém, esta compatibilidade é verificada caso cada variável X_i em ξ seja independente dos seus não-descendentes, dadas suas respectivas variáveis pais.

Neste sentido, devemos estudar as relações de dependência entre as variáveis considerando sua distribuição de probabilidade. Para um conjunto de variáveis em estudo, para quaisquer subconjuntos V, Z e W temos (KORB; NICHOLSON, 2004):

$$(V \perp W) \Leftrightarrow P(v|w,z) = P(v|z) \quad \forall\, w,z|P(w,z) > 0$$

Interpretando a expressão acima, dizemos que o conjunto V é independente do conjunto W dado Z se, e somente se, W não impactar probabilisticamente em V dado Z. Através desta definição podemos percorrer a rede buscando independência condicional entre as variáveis. Tais independências são expressas em P e possivelmente representadas em ξ.

Porém, afirmar que existe compatibilidade entre P e ξ não significa necessaria-

mente que todas as dependências e independências expressas por ξ estão contidas em P e vice-versa. Assim, podermos ter as seguintes relações entre ambas, estas também conhecidas como mapeamento:

1. Mapa de Dependência (D-Map): se toda independência em P é verdadeira em ξ.

2. Mapa de Independência (I-Map): se toda independência em ξ é verdadeira em P.

3. Mapa Perfeito (P-Map): se todas as independências estão expressas ξ em e P.

Neste sentido, procuramos sempre encontrar um mapa perfeito, porém muitas vezes não é possível afirmar certamente que todas as independências podem ser expressas por ξ.

Existem várias abordagens referentes a procedimentos de estimação de estrutura, uma área em constante desenvolvimento (RUSSELL; NORVIG, 2004).

Segundo Hruschka (1997), a estimação de estrutura de uma Rede Probabilística pode ser dividida em duas partes: a primeira baseada em uma busca heurística e a segunda baseada no conceito de independência condicional dos atributos da rede. Assim, algoritmos são requeridos para ambos os tipos de estimação.

Os algoritmos baseados no conceito de independência condicional utilizam a propriedade de d-separação (2.4), o que diminui significativamente o esforço computacional.

Métodos híbridos são uma terceira alternativa para estimação de estrutura, os quais se utilizam de uma composição dos algoritmos de busca por pontuação e dos baseados em propriedades de d-separação (MAGALHÃES, 2007).

Nesta seção, apresentamos de forma sucinta o algoritmo de busca heurística K2, que visa maximizar uma métrica de determinada função. Em especial, apresentamos de forma mais detalhada o algoritmo PC, baseado em propriedades de d-separação. O algoritmo PC será utilizando para discriminar procedimentos tradicionais em estimação de estrutura em comparação com procedimentos específicos utilizados nesta estimação para classificação.

3.1.1 Algoritmo K2

O algoritmo K2 é considerado um dos mais importantes dentre todos os algoritmos que se referenciam à busca de pontuação para estimação de estrutura. Assim, sua idéia base é, partindo de uma ordenação das variáveis a fim de tornar a estrutura acíclica, pesquisar entre os $2^{p(p-1)/2}$ tipos de configurações de estruturas de rede, sendo p o número de variáveis na rede, e verificar qual dentre elas maximiza a função *score* dada por 3.1 (HRUSCHKA,1997). A complexidade de tempo gasto no cálculo de 3.1 é de $\mathcal{O}(nkp)$, onde n é o número de observações, k é o número máximo de pais que uma variável pode assumir.

$$P(\xi|X) = c\prod_{i=1}^{m}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}! \tag{3.1}$$

onde X é a base de dados com n observações, ξ representa a dimensão de estrutura, m é o número de variáveis, r_i é a quantidade total de possíveis valores que a variável X_i onde $i = 1, ..., m$ pode assumir. O termo q_i está relacionado às possíveis configurações dos pais. O valor de N_{ijk} representa a quantidade total de observações em X, onde a variável X_i está no k-ésimo estado e os seus pais apresentam a j-ésima configuração. A constante c é a constante de proporcionalidade. Já N_{ij} é o número total de observações em X, onde se tem X_i com qualquer um de seus possíveis valores e com

a j-ésima configuração.

3.1.2 Algoritmo PC

Derivado do algoritmo IC (PEARL e VERMA, 1991), o algoritmo PC foi proposto por Spirtes, Glymour e Scheines (1991), levando assim no nome as iniciais de seus principais criadores, Peter Spirtes e Clark Glymour. A idéia básica do algoritmo é realizar testes estatísticos para determinar grupos de variáveis independentes, utilizando o critério de d-separação. Geralmente, a independência entre as variáveis é verificada através de uma métrica estatística, como no caso do teste estatístico de χ^2, ou ainda utilizando a informação mútua condicional exibida na Seção 1.20, estes calculados através do conjunto de dados (ABELLAN *et al*, 2006). Assim, as relações de independência são verificadas para cada par de variáveis da rede. Tal processo considera que, se a dependência é válida, as variáveis se encontram conectadas e, assim, estabelecem a orientação dos arcos, através do critério de d-separação. O algoritmo PC possui um tempo de execução na ordem de $\mathcal{O}(p^k)$, onde p é o número de variáveis na rede e k é o número máximo de pais que cada uma pode assumir.

O algoritmo PC identifica corretamente todas as conexões $X_1 \rightarrow X_2 \leftarrow X_3$ $head - to - head$ em uma Rede Probabilística, porém nas demais conexões e orientações de arco muitas vezes este processo não consegue identificar corretamente a relação de causalidade entre as variáveis, gerando assim estruturas equivalentes, conceito introduzido na seção 2.5. Pearl (1988) garante que estruturas equivalentes possuem o mesmo conjunto de distribuições compatíveis. A Figura2.7, já discutida anteriormente, ilustra esta idéia.

O algoritmo PC utilizado para estimação de estrutura mediante a um conjunto de variáveis aleatórias X é exibido a no Quadro 3.1 .

Basicamente, no passo 1, o algoritmo é iniciado com todas as conexões não dire-

cionadas entre todas as variáveis aleatórias, bem como o objeto $ADJX$ (adjacentes) que indica quais são estas conexões, quais são as variáveis adjacentes a cada X_i . Este objeto, computacionalmente, pode ser uma matriz indicadora.

No passo 2, o algoritmo percorre, para diferentes tamanhos de conjuntos condicionais, todas as variáveis buscando identificar independências. O passo 2.1.1.1.1 apresenta o maior esforço computacional, pois, para duas variáveis X_i e X_j conectas, o algoritmo considera como condicionais todos os subconjuntos possíveis de tamanho $modS$ formado pelas demais variáveis que estão conectas a X_i , isto é, se existem ainda mais 10 variáveis conectas a X_i e considerando que o número avaliado de condicionais é 4, ou seja, $modS = 4$, existem $\begin{pmatrix} 10 \\ 4 \end{pmatrix} = 210$ possíveis conjuntos diferentes de condicionais a serem visitados pelo algoritmo. Necessariamente, com o aumento do número de variáveis analisadas, o esforço computacional cresce exponencialmente. Porém, este procedimento ainda é mais ágil que os mecanismos de busca por pontuação. Ainda neste passo, se o algoritmo encontra uma independência condicional assim expressa como $X_i \perp X_j | S$ e evidenciada por $I(X_i, X_j | S) = 0$, temos que X_i e X_j não estão conectadas e são d-separadas pelo S.

No passo 3, o algoritmo incorpora a definição de d-separação para conexões do tipo *head-to-head*, balizada pelo fato da variável central não d-separar os vértices da tripla analisada.

O Passo 4 propõe a criação de conexões orientadas para triplas que não geraram as conexões *head-to-head* no passo anterior. Ainda, realiza as demais conexões da rede a fim de não originar uma estrutura cíclica. Geralmente, este processo é realizado pela atribuição de uma ordem às variáveis.

Para melhor ilustrar o algoritmo PC, consideramos a estrutura de Rede Probabilística exibida na Figura 2.3 da seção 2.1.6. Para esta rede, geramos um conjunto de

Algoritmo 3.1 Algoritmo PC (Adaptado de KORB; NICHOLSON, 2004; GALVÃO; HRUSCHKA,2007)

1. Inicie com uma estrutura de grafo não direcionado contendo todas as conexões possíveis entre as variáveis $X = X_1, X_2, \ldots, X_k$. Considere $ADJX$ como as variáveis conectadas a X_i . Assim, $ADJX = X - X_i$

2. Inicie a contagem $modS = 0$

2.1. Enquanto $|ADJX| < modS \quad \forall X_i \in X$:

2.1.1. Para cada variável X_i em X faça:

2.1.1.1. Parca cada variável $X_j \in ADJX$ faça:

2.1.1.1.1. Determine se existe um grupo $S \subseteq ADJX - X_j$ sendo que o tamanho de S seja igual à $modS$, $|S| = modS$.

2.1.1.1.2. Calcule $I(X_i, X_j|S)$;

2.1.1.1.3. Se $I(X_i, X_j|S) = 0$, inclua S no grupo que d - separa X_i e X_j , chamado de S_{ij};

2.1.1.1.4. Remova a conexão entre X_i e X_j

2.1.2. Incremente $modS$.

3. Para cara tripla X_i, X_z e X_j tal que as variáveis X_i e X_j estejam conectadas à variável X_z, porém não existe conexão entre X_i e X_j:

3.1. Se $X_z \notin S_{ij}$: Oriente $X_i - X_z - X_j$ como $X_i \rightarrow X_z \leftarrow X_j$

4. Para toda a conexão do tipo $X_{i}-X_{j}$ oriente $X_i \rightarrow X_j$ se, e somente se:

4.1. $X_l \rightarrow X_i$, X_i e $X_j$estão conectados, X_j e X_l não estão conectados e a conexão $X_i \leftarrow X_j$ não existe.

4.2. $X_i \rightarrow X_j$ não gera uma estrutura cíclica no grafo.

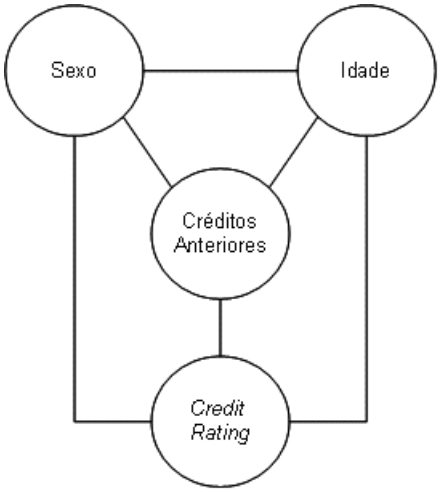


Figura 3.1: Algoritmo PC- Passo 1: Inicia-se com todas as conexões entre as variáveis

dados com 1000 observações através do *Software* `R` e estimamos sua estrutura. Os procedimentos computacionais de geração dos dados e a implementação do algoritmo PC para o *Software* `R` estão disponibilizados, respectivamente, nos Anexos A e B.

Antes de exibir a estrutura final estimada pelo *Software* `R`, exibimos ilustrativamente um resumo do algoritmo PC para esta rede. Cada passo do algoritmo é exibido sequencialmente pelas Figuras 3.1até 3.6.

Neste caso, notamos a eficiência do algoritmo para estimar a estrutura proposta. Posteriormente, exibiremos a estimação dos parâmetros da rede utilizando esta mesmo estrutura. Na Figura 3.7, exibimos a saída gráfica do *Software* `R`. Para gerar este gráfico, utilizamos a estrutura gráfica disponibilizada pelo pacote `diagram`. Os códigos para a geração do gráfico da Figura 3.7 são disponibilizados no Apêndice G.

Dentre as vantagens do algoritmo PC, podemos salientar que, como algoritmo, ele não possui problemas de otimização com máximos locais, bem como oferece exatamente a informação contida nos dados, porém, uma vez que uma independência

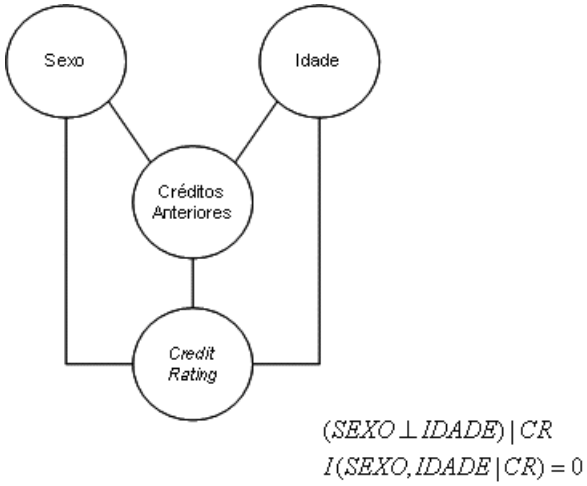


Figura 3.2: Algoritmo PC- Passo 2: Verificando independências condicionais. A variável Sexo é independente da variável Idade dado *Credit Rating*.

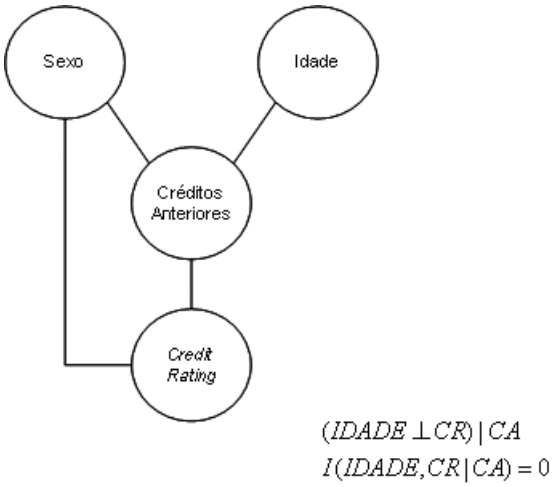


Figura 3.3: Passo 2 do Algoritmo PC: Verificando independências condicionais. A variável Idade é independente da variável *Credit Rating* dada a variável Créditos Anteriores.

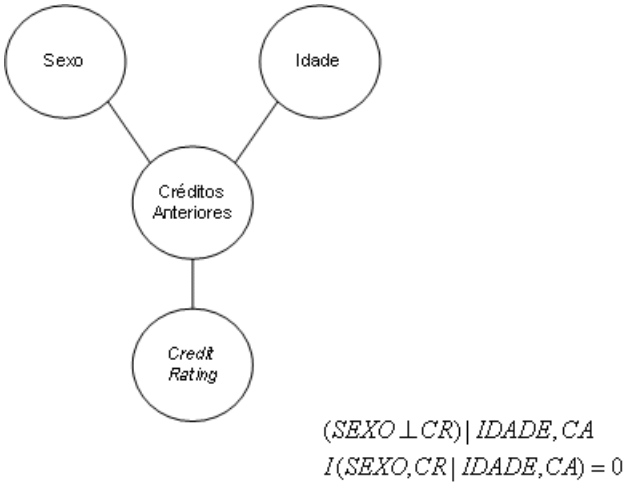


Figura 3.4: Passo 2 do Algoritmo PC: Verificando independências condicionais. A variável Sexo é independente da variável *Credit Rating* dada a Idade e Créditos Anteriores.

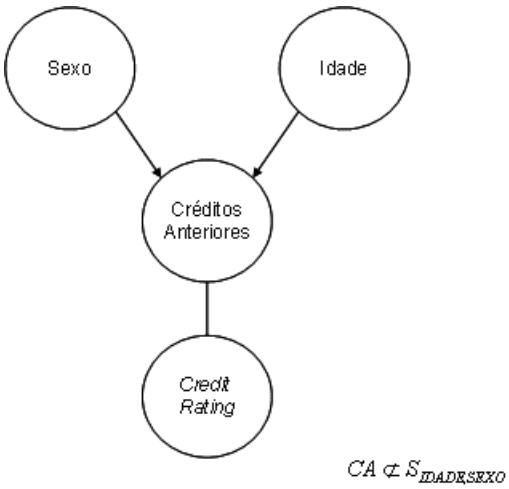


Figura 3.5: Algoritmo PC- Passo 3: Dada a Tripla formada entre as variáveis Sexo, Créditos Anteriores e Idade, é definida a conexão *head-to-head*.

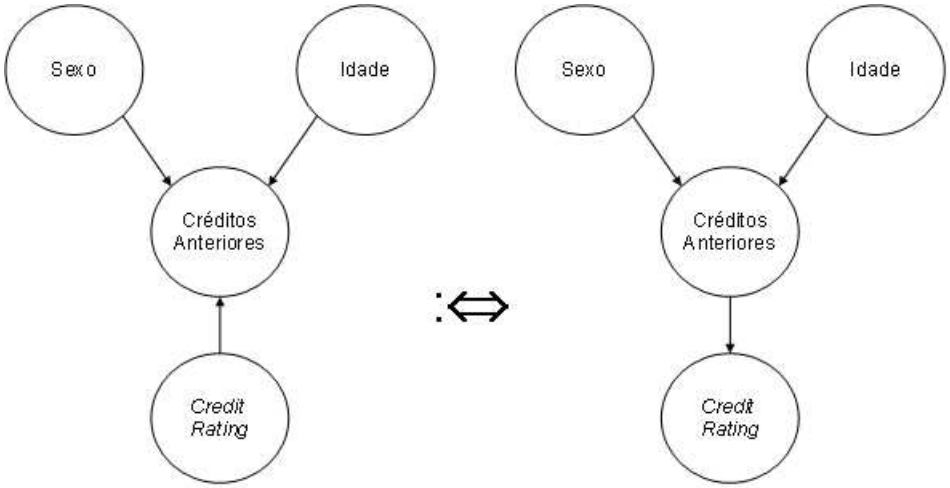


Figura 3.6: Algoritmo PC- Passo 4: orientação gerando equivalência de Markov. Estas redes são Markov equivalentes.

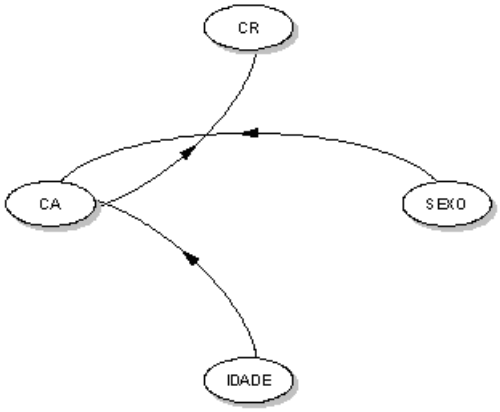


Figura 3.7: Estrutura estimada utilizando o algoritmo PC implementado no *Software* R.

incorreta é verificada para a rede, esta se estenderá para diversas outras conexões, fato este que é agravado com amostras pequenas.

Para alguns conjuntos de dados, através do algoritmo PC, variáveis podem ser identificadas como independentes em relação a outras variáveis em análise, ou seja, algumas variáveis podem não possuir conexão com as demais. Desta forma, o procedimento se torna insatisfatório quando esta variável é um dos temas centrais da análise, como por exemplo, no contexto de classificação.

Para ilustrar esta problemática, consideramos um conjunto de dados reais da área de *Credit Scoring,* disponível no repositório de dados da Universidade da Califórnia (www.ics.uci.edu/~mlearn/). O conjunto de dados é conhecido como *Japanese Credit Screening Data Se*t e possui 653 observações, 15 variáveis explicativas $X = (X_1, X_2, \ldots, X_{15})$ e uma variável de interesse Y para classificação, esta referente à avaliação de crédito de clientes (bom pagador ou mau pagador). Neste caso, aplicando o algoritmo PC temos o ajuste dado pela Figura 3.8.

Através da Figura 3.8, notamos que a rede estimada só exibe o relacionamento de dependência probabilística entre as variáveis explicativas, ou seja, a variável de classificação não aparece na estrutura estimada, não possui dependência significativa com nenhuma das demais variáveis.

Neste caso, não podemos utilizar o algoritmo PC para realizar procedimentos classificação.

3.2 Estimação de parâmetros

Neste momento, estamos interessados em estimar as probabilidades condicionais para cada variável da rede. Estes procedimentos podem ser realizados em conjuntos de dados completos e incompletos, sendo aqui apresentado apenas o método de estimação

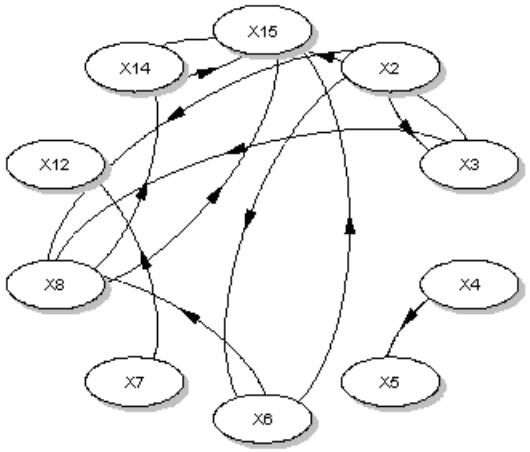


Figura 3.8: . Ajuste do algoritmo PC ao conjunto de dados reais *Japanese Credit Screening Data Set.*

para dados completos.

Porém, um procedimento utilizado quando a base de dados é incompleta é o algoritmo *Expectation Maximization* (EM) (LUNA, 2004). Basicamente, se alguma variável possui uma falta de informação, também conhecido como *missing*, este algoritmo utiliza os casos observados para estimar os valores faltantes. O mesmo algoritmo pode ser também aplicado para dados completos assumindo o conjunto de *missing* como vazio.

A estimação para dados completos pode ser realizada utilizando estimadores frequentistas e estimadores bayesianos. Tais abordagens são exibidas nas Seções 3.2.1 e 3.2.2, respectivamente.

3.2.1 Estimação Frequentista

Este processo de estimação é extremamente simples. Não considera nenhum tipo de conhecimento a priori, sendo suas estimativas baseadas em frequências relativas e contagens através da base de dados.

Para esta abordagem, considere que cada variável X_i possua r_i estados possíveis, sendo indicados por $x_i^1, x_i^2, \ldots, x_i^r$, dado o j-ésimo pai_i e estrutura conhecida ξ. Assim temos 3.2 e 3.3:

$$P(X_i = x_i^k | pai_i^j, \xi) = \frac{P(X_i = x_i^k | pai_i^j)}{P(pai_i^j)} = \theta_{ijk} \tag{3.2}$$

$$\widehat{\theta}_{ijk} = \frac{fr(x_i^k, pai_i^j)}{fr(pai_i^j)} \tag{3.3}$$

onde $fr(.)$ denota freqüência relativa.

Note que nenhuma suposição a priori foi dada sobre qualquer um dos elementos em análise. Porém, a forma mais clara de exibir tal pensamento é através de um

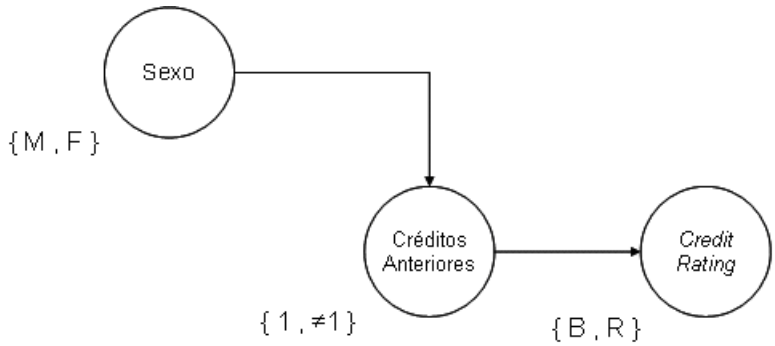


Figura 3.9: Possível Rede Probabilística para dados aplicados a *credit scoring*.

exemplo.

Considere um conjunto de dados constituído de 3 variáveis dicotômicas e 24 observações referentes a *credit scoring*, sendo as variáveis:

- Sexo { Masculino, Feminino };

- Créditos Anteriores { Um, Diferente de um };

- *Credit Rating* { Bom, Ruim }.

O conjunto de dados, gerado através do Software `R`, é exposto na Tabela 3.1. Para este problema, considere a possível estrutura de Rede Probabilística exibida na Figura 3.10.

Através da Figura 3.10, notamos que existe apenas uma variável raiz e todas as demais variáveis possuem somente uma variável pai.

Para facilitar os cálculos, a variável Sexo será representada pela letra S, a variável Créditos Anteriores pela sigla CA, e a variável *Credit Rating* pela sigla CR.

Levando em consideração a estrutura de relacionamento apresentada, necessitamos dos cálculos das probabilidades $P(S)$, $P(CA|S)$ e $P(CR|CA)$. $P(S)$ é estimada

Tabela 3.1: Conjunto de dados referentes a *credit scoring*.

#	*Credit Rating* (CR)	Sexo (S)	Créditos Anteriores (CA)
1	*Ruim*	M	$\neq 1$
2	*Bom*	M	$= 1$
3	*Ruim*	F	$\neq 1$
4	*Bom*	F	$\neq 1$
5	*Bom*	M	$= 1$
6	*Bom*	M	$= 1$
7	*Ruim*	M	$= 1$
8	*Bom*	M	$\neq 1$
9	*Bom*	M	$\neq 1$
10	*Ruim*	M	$\neq 1$
11	*Ruim*	M	$= 1$
12	*Ruim*	F	$= 1$
13	*Ruim*	M	$\neq 1$
14	*Bom*	M	$\neq 1$
15	*Bom*	F	$= 1$
16	*Bom*	M	$= 1$
17	*Bom*	M	$= 1$
18	*Ruim*	F	$= 1$
19	*Bom*	M	$= 1$
20	*Bom*	M	$= 1$
21	*Bom*	M	$= 1$
22	*Bom*	M	$\neq 1$
23	*Bom*	M	$\neq 1$
24	*Bom*	M	$= 1$

Tabela 3.2: Probabilidade conjunta $P(CA, S)$

		S		
		F	M	Total
CA	$= 1$	$0,13$	$0,46$	$0,58$
	$\neq 1$	$0,08$	$0,33$	$0,42$
	Total	$0,21$	$0,79$	$1,00$

Tabela 3.3: Probabilidade conjunta $P(CR, CA)$

		CA		
		$= 1$	$\neq 1$	Total
CR	*Ruim*	$0,17$	$0,17$	$0,33$
	Bom	$0,42$	$0,25$	$0,67$
	Total	$0,58$	$0,42$	$1,00$

facilmente através da frequência relativa calculada através da Tabela 3.1. Para o cálculo das probabilidades $P(CA|S)$ e $P(CR|CA)$, partimos de tabelas de distribuição conjunta, obtidas das tabelas cruzadas entre as variáveis de interesse. As probabilidades conjuntas $P(CA, S)$ e $P(CR, CA)$ são estimadas através das Tabelas 3.2 e 3.3, respectivamente.

Note que, em cada tabela, as células referentes ao total são as probabilidades marginais de cada categoria, ou seja, para a Tabela 3.2 a probabilidade marginal da variável CA, fixando CA na categoria 1, é dada por $P(CA = 1) = 0,58$.

Assim, através do Teorema de Bayes visto na seção 1.3.2.4, no qual, por exemplo, $P(CR|CA) = P(CR, CA)/P(CA)$, realizamos o cálculo de cada célula de probabilidade conjunta dividida por sua respectiva célula de probabilidade marginal.

As probabilidades condicionais $P(CA|S)$ e $P(CR|CA)$ são estimadas através das Tabelas de probabilidade condicionais (TPC) 3.4 e 3.5, respectivamente.

Deste modo, a Rede Probabilística pode ser visualizada na Figura 3.10, expressando suas respectivas tabelas de probabilidades condicionais estimadas via método

Tabela 3.4: Probabilidade condicional $P(CA, |S)$

		S	
		F	M
CA	$= 1$	$0,60$	$0,58$
	$\neq 1$	$0,40$	$0,42$

Tabela 3.5: Probabilidade condicional $P(CR|CA)$

		CA	
		$= 1$	$\neq 1$
CR	*Ruim*	$0,29$	$0,40$
	Bom	$0,71$	$0,60$

frequentista.

3.2.2 Estimação Bayesiana

Considere θ o parâmetro numérico da rede em estudo com estrutura ξ conhecida, onde $X = \{X_1, X_2, \ldots, X_k\}$ representa as variáveis associadas ao conjunto de dados fornecido.

Para o cálculo das probabilidades, assumimos que cada variável X_i, dado seus pais, possui distribuição Multinomial com parâmetros N e θ_i, ou seja, $X_i \,|\, pais(X_i), \theta_i, N \sim Multinomial(N, \theta_i)$, sendo $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$. Assim, considere X_i uma variável aleatória discreta que represente um experimento com r possíveis resultados, sendo que cada resultado j possui uma probabilidade de ocorrência $P(X_i = x_j) = p_j$ e $\sum_{j=1}^{r} p_j = 1$. O experimento é repetido de forma independente N vezes, de forma que x_j seja igual ao número de vezes que o resultado j está presente na amostra com $j = 1, ..., r$. Temos que a variável X_i segue distribuição Multinomial, com sua função de probabilidade expressa em 3.4.

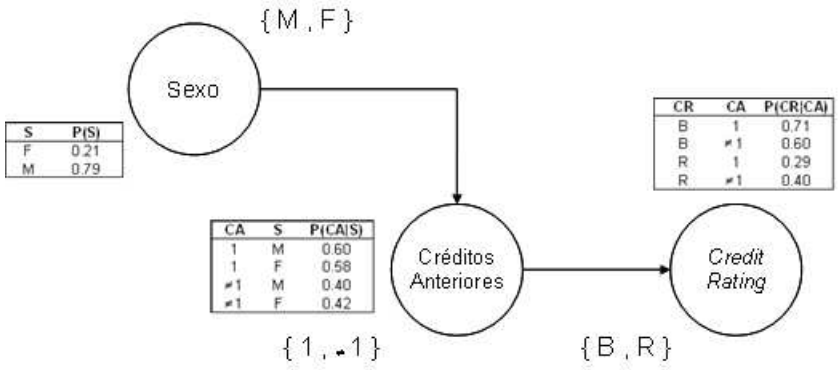


Figura 3.10: Possível Rede Probabilística com TPC para dados de *credit scoring*.

$$P(X_i|N,\theta_i) = \frac{N!}{x_1!x_2!\ldots x_r!}p_1^{x_1}p_2^{x_2}\ldots p_r^{x_r} \tag{3.4}$$

onde $\theta_i = \{p_1, p_2, \ldots, p_r\}$ e $\sum_j^r x_j = N$.

Considerando o termo $\frac{N!}{x_1!x_2!\ldots x_r!}$ como constante normalizadora, temos que $P(X_i|N,\theta_i) \propto p_1^{x_1}p_2^{x_2}\ldots p_r^{x_r}$. Assim, para estimação de θ_i no contexto bayesiano, podemos assumir a priori que este segue distribuição Dirichlet com parâmetros $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ com $\alpha_j > 0$ e esperança $E(p_j) = \frac{\alpha_j}{\sum_{j=1}^r \alpha_j}$ com função densidade de probabilidade expressa em 3.5.

$$P(\theta_i|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_r)}p_1^{\alpha_1-1}p_2^{\alpha_2-1}\ldots p_r^{\alpha_r-1} \tag{3.5}$$

Podemos considerar o termo $\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_r)}$ como constante normalizadora, deste modo $P(\theta_i|\alpha) \propto p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_r^{\alpha_r}$. Assumindo como distribuição a priori $P(\theta_i|\alpha)$ e $P(X_i|N,\theta_i)$ como verossimilhança , temos que a distribuição a posteriori de θ_i é dada pela expressão 3.6 a qual tem distribuição Dirichlet com parâmetros $\alpha = \{\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_r + x_r\}$ e $E(p_j)$ dada por 3.7.

Tabela 3.6: Freqüência Absoluta de (CR, CA)

		CA		
		$= 1$	$\neq 1$	Total
CR	*Ruim*	4	4	8
	Bom	10	6	16
	Total	14	10	24

$$P(\theta_i|N, X_i, \alpha) \propto p_1^{\alpha_1+x_1-1} p_2^{\alpha_2+x_2-1} \dots p_r^{\alpha_r+x_r-1} \tag{3.6}$$

$$E(p_j) = \frac{\alpha_j + x_j}{\sum_{j=1}^{r} \alpha_j + N} \qquad j = 1, \dots, r \tag{3.7}$$

Para este caso, a família Dirichlet é conjugada para amostras com distribuição Multinomial.

O parâmetro α também é conhecido como hiperparâmetro e deve ser estabelecido a priori. Na prática, uma possível forma de atribuição destes valores é consultar a opinião de um especialista da área dos dados analisados, ou ainda, considerar um valor que atribua influência mínima da priori na posteriori, isso pode ser feito considerando um vetor próximo do vetor nulo.

Para aplicação desta técnica, considere o conjunto de dados do exemplo anterior, mais especificamente a frequência absoluta da Tabela 3.3.Assim, podemos construir a Tabela 3.6 considerando $\alpha_i = 1,\ j = 1, \dots, r$.

Assim, podemos realizar os cálculos a partir de 3.7.

$$P(CR = Ruim|\, CA = 1, \alpha = 1) = \frac{\alpha + x_{CR=Ruim,CA=1}}{2\alpha + N_{CA=1}} = \frac{1+4}{2+14} = 0,312$$

Note que os valores da Tabela 3.7 são bastante similares aos encontrados na

Tabela 3.7: Probabilidade condicional $P(CR|CA)$

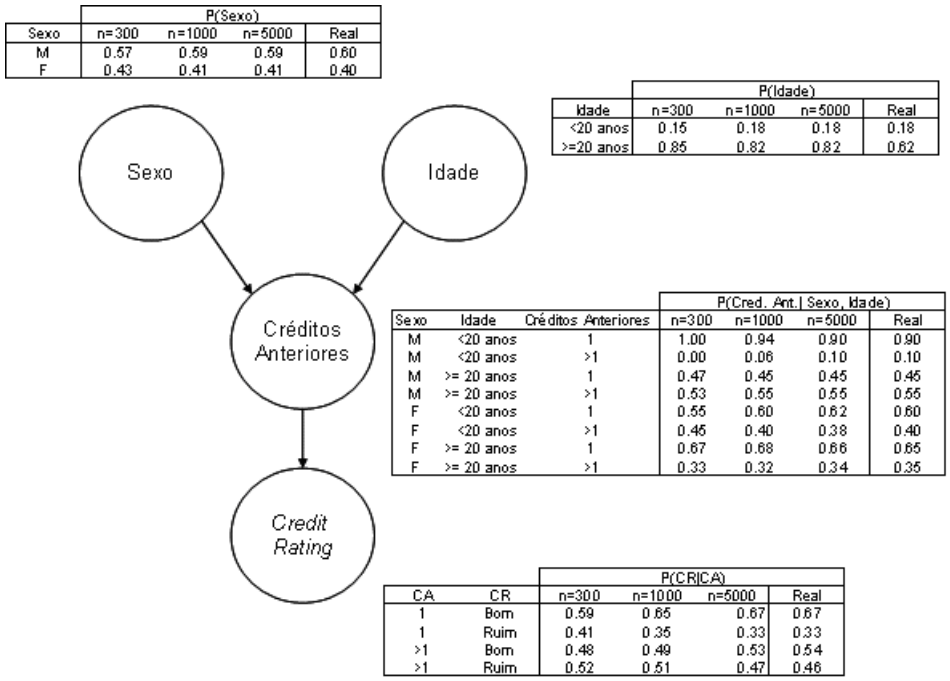
		CA	
		$= 1$	$\neq 1$
CR	*Ruim*	$0,312$	$0,417$
	Bom	$0,688$	$0,583$

Tabela 3.5.

Para melhor ilustrar a aplicação deste tipo de abordagem para a estimação dos parâmetros da rede, consideramos a estrutura estimada pelo algoritmo PC na seção anterior, Figura 3.8. Para esta rede, realizamos esta estimação gerando um conjunto de dados com 300, 1000 e 5000 observações através do *Software* `R`. Os procedimentos computacionais estão disponibilizados no Anexo C.

A Figura 3.11 exibe a estimação dos parâmetros para esta abordagem. Para sua realização consideramos os hiperparâmetros como $\alpha = 0,002$. Notamos que para o tamanho de amostra 300 existe maior diferença entre os valores estimados e reais, essa diferença diminui com o aumento da amostra. Porém, notamos que mesmo com uma amostra de tamanho 5000, ainda existe uma pequena diferença de estimação, como no caso da $P(CA = 1|Idade => 20anos, Sexo = F)$, onde a probabilidade real é de 0,60 e a estimada foi de 0,62. Porém, globalmente, podemos verificar com esta aplicação que a estimação bayesiana é visualmente satisfatória.

Comparando o método bayesiano com o método frequentista, uma vez que o evento $\{X_i|Pais(X_i)\}$ para um respectivo x_j não existe no banco de dados, temos que, pelo método frequentista, $P(X_i = x_j|Pais(X_i)) = 0$. Isto pode ocasionar uma alta frequência de probabilidades zeros quando a amostra não envolve todos os possíveis resultados que uma variável pode assumir. Deste modo, notamos que o método frequentista não é eficaz quando o número de categorias entre as variáveis é



	P(Sexo)			
Sexo	n=300	n=1000	n=5000	Real
M	0.57	0.59	0.59	0.60
F	0.43	0.41	0.41	0.40

	P(Idade)			
Idade	n=300	n=1000	n=5000	Real
<20 anos	0.15	0.18	0.18	0.18
>=20 anos	0.85	0.82	0.82	0.82

			P(Cred. Ant.| Sexo, Idade)			
Sexo	Idade	Créditos Anteriores	n=300	n=1000	n=5000	Real
M	<20 anos	1	1.00	0.94	0.90	0.90
M	<20 anos	>1	0.00	0.06	0.10	0.10
M	>= 20 anos	1	0.47	0.45	0.45	0.45
M	>= 20 anos	>1	0.53	0.55	0.55	0.55
F	<20 anos	1	0.55	0.60	0.62	0.60
F	<20 anos	>1	0.45	0.40	0.38	0.40
F	>= 20 anos	1	0.67	0.68	0.66	0.65
F	>= 20 anos	>1	0.33	0.32	0.34	0.35

		P(CR|CA)			
CA	CR	n=300	n=1000	n=5000	Real
1	Bom	0.59	0.65	0.67	0.67
1	Ruim	0.41	0.35	0.33	0.33
>1	Bom	0.48	0.49	0.53	0.54
>1	Ruim	0.52	0.51	0.47	0.46

Figura 3.11: Estimação Bayesiana para os parâmetros da Rede Probabilística.

alto. Neste contexto o método bayesiano é mais adequado.

3.3 Comentários Finais

Nesta seção, abordamos os procedimentos para estimação em Redes Probabilísticas. Para a estimação de estrutura, consideramos mais detalhadamente o algoritmo PC e mostramos que este, bem como outros métodos de estimação de estrutura, não é eficaz quando o objetivo da análise é a classificação. Além disso, para a estimação de probabilidades, abordamos dois procedimentos de estimação das tabelas de probabilidades condicionais, também conhecidas como parâmetros da rede. Especificamente, o procedimento bayesiano para estimar os parâmetros será utilizado em todas as problemáticas tratadas por este trabalho.

Posteriormente, exibimos como toda a teoria de Redes Probabilísticas pode ser utilizada no contexto de classificação, mais especificamente na classificação binária.

Capítulo 4

Classificação

No contexto de classificação, as Redes Probabilísticas podem ser vistas como estruturas particulares e também são conhecidas como classificadores bayesianos.

Nesta seção, consideramos a estrutura de Rede Probabilística Simples, popularmente conhecida como classificador de *Naive* Bayes e a estrutura de Redes Probabilísticas Simples com K-Dependência, também conhecida como classificador bayesiano com K-dependência, (KDB) (Sahami, 1996). Além disso, consideramos outros métodos tradicionais de classificação a fim de estabelecer uma comparação com as Redes Probabilísticas.

4.1 Rede Probabilística Simples

A construção de uma Rede Probabilística Simples está baseada no cálculo da distribuição de probabilidade a posteriori $P(Y|X)$, onde $Y = (y_1, y_2, ..., y_k)$ é a variável aleatória a ser classificada apresentando k categorias e $X = (X_1, X_2, ...X_p)$ é um conjunto de p variáveis explicativas.

Para o cálculo da probabilidade condicional $P(Y|X)$, este método assume independência probabilística entre as variáveis explicativas, dada a variável de classifica-

ção, facilitando a aplicação do método computacionalmente.

Desta forma, para o caso onde X é um conjunto de variáveis explicativas puramente discretas, $P(Y|X)$ é dada por 4.1. No caso em que X é um conjunto de variáveis explicativas puramente contínuas e, em suposição, com distribuiçãno normal, $P(Y|X)$ é dada por 4.2.

$$P(Y = y_k|x_1, x_2, \ldots, x_p) = \frac{P(Y = y_k)\prod_{i=1}^{p} P(x_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_{i=1}^{p} P(x_i|Y = y_j)} \tag{4.1}$$

$$P(Y = y_k|x_1, x_2, \ldots, x_p) = \frac{P(Y = y_k)\prod_{i=1}^{p} f(x_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_{i=1}^{p} f(x_i|Y = y_j)} \tag{4.2}$$

onde

$$f(x_i|\, Y = y_k) \sim N(\mu_{i|\, y_k}, \sigma^2_{i|\, y_k})$$

sendo $\mu_{i|\, y_k}$e $\sigma^2_{i|\, y_k}$ a média e a variância da variável x_i condicionada à categoria y_k , respectivamente.

O método baseia-se em calcular a probabilidade de uma respectiva observação pertencer a cada uma das categorias e classifica a observação na categoria mais plausível. Se a classificação em foco for binária, podemos utilizar a curva ROC, definida na Seção 4.4, para inferir sobre a classificação.

A Figura 4.1 exibe o caso geral de uma Rede Probabilística Simples.

Através da Figura 4.1, notamos que todas as variáveis explicativas X_i possuem apenas Y como variável pai, ou seja, Y é a única variável raiz, a qual origina a rede.

Porém, na maioria das vezes, a suposição de independência entre as variáveis explicativas não condiz com a realidade, ou seja, o método não leva em conta a possível relação de dependência probabilística entre as variáveis explicativas.

Assim, outras estruturas de Redes Probabilísticas devem ser utilizadas. Uma possível alternativa é apresentada a seguir.

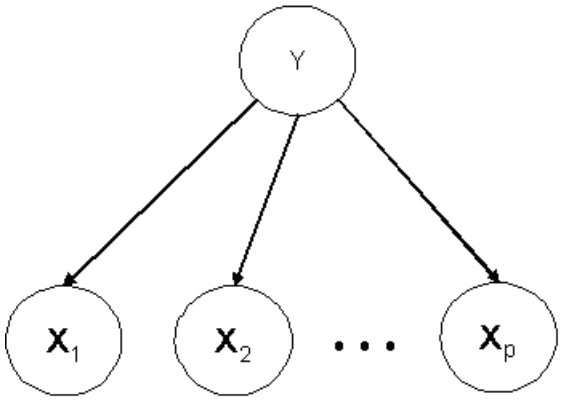


Figura 4.1: Rede Probabilística Simples

4.2 Rede Probabilística Simples com K-dependência

Este método, ao contrário do anterior, considera possíveis relações de dependência entre as variáveis explicativas. Desta forma, uma Rede Probabilística Simples com K-dependência é uma Rede Probabilística Simples que permite em sua estrutura que cada variável explicativa X_i possua no máximo K variáveis explicativas pais, em outras palavras, para cada variável explicativa X_i, $pais(X_i)$ é um conjunto com no máximo K outras variáveis explicativas para todo $i = 1, \ldots, p$.

Note também que K pode variar de 0 a $1 - p$, onde p é o número de variáveis explicativas consideradas.

Desta forma, para Redes Probabilísticas de K dependência (KDB), calculamos as probabilidades a posteriori através de 4.3 para o caso puramente discreto, e através de 4.4 (PÉREZ *et al.*, 2006) para o caso puramente contínuo em que se assume uma distribuição normal para as variáveis explicatívas.

$$P(Y = y_k|x_1, x_2, \ldots, x_p) = \frac{P(Y = y_k)\prod_{i=1}^{p} P(x_i|pais(X_i), Y = y_k)}{\sum_j P(Y = y_j)\prod_{i=1}^{p} P(x_i|pais(X_i), Y = y_j)} \qquad (4.3)$$

$$P(Y = y_k|x_1, x_2, \ldots, x_p) = \frac{P(Y = y_k)\prod_{i=1}^{p} f(x_i|pais(X_i), y_k)}{\sum_j P(Y = y_j)\prod_{i=1}^{p} f(x_i|pais(X_i), y_j)} \tag{4.4}$$

onde

$$f(x_i|\, pais_i, y_k) \sim N(\mu_{i|\, pais_i, y_k}, \sigma^2_{i|\, pais_i, y_k})$$

sendo $\mu_{i|pais_i\, y_k}$e $\sigma^2_{i|pais_i\, y_k}$ a média e a variância da variável x_i condicionada à categoria y_k e dadas por 4.5 e 4.6, respectivamente.

$$\mu_{i|\, pais_i, y_k} = \mu_{i|\, y_k} + \sum_{j=1}^{K_i} \frac{\sigma_{ij|y_k}}{\sigma^2_{j|y_k}} \left(x_j - \mu_{j|\, y_k}\right) \tag{4.5}$$

$$\sigma^2_{i|pais_i, y_k} = \frac{\left|\sum_{X_i, pais_i|\, y_k}\right|}{\left|\sum_{pais_i|\, y_k}\right|} \tag{4.6}$$

onde $\sum_{X_i, pais_i|\, y_k}$ é a matriz de variância e covariância, definida na Seção 1.3.2.6, entre a variável X_i e o conjunto de variáveis em $pais(X_i)$, ambos condicionados à categoria y_k. $\sum_{pais_i|\, y_k}$é a matriz de variância do conjunto de variáveis em $pais(X_i)$ condicionado à categoria .

Os códigos em `R` para a implementação das Redes Probabilísticas de K-dependência são disponibilizados nos Apêndices D e E, caso discreto e contínuo, respectivamente.

Considerando um conjunto de dados com uma variável de interesse e 7 variáveis explicativas, exibimos as Redes Probabilísticas de 0, 1, 2 e 3-dependência nas Figuras 4.2 a 4.5 respectivamente.

Para o caso Rede Probabilística Simples com 0- dependência (KDB0), Figura 4.2, temos que cada variável explicativa X_i com $i = 1, ..., 9$ é filha da variável resposta Y, ou seja, $Pais(X_i) = \{\oslash\}$. No caso da Rede Probabilística Simples com 1- dependência (KDB1), Figura 4.3, temos $Pais(X_6) = \{\oslash\}$, $Pais(X_4) = \{X_6\}$, $Pais(X_3) =$

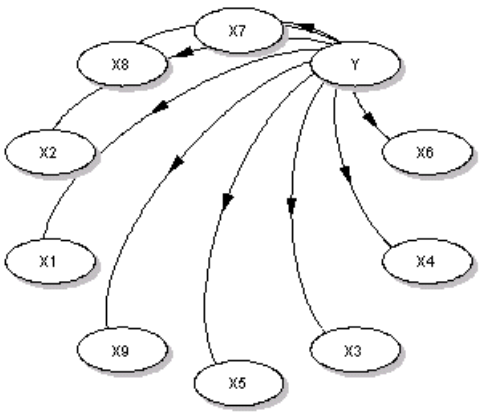


Figura 4.2: Exemplificação de uma Rede Probabilística Simples com 0- dependência.

$\{X_4\}$, $Pais(X_5) = \{X_4\}$, $Pais(X_9) = \{X_4\}$, $Pais(X_1) = \{X_3\}$, $Pais(X_2) = \{X_1\}$, $Pais(X_8) = \{X_3\}$ e $Pais(X_7) = \{X_8\}$. Para a Rede Probabilística Simples com 2- dependência (KDB2), Figura 4.4, temos $Pais(X_6) = \{\oslash\}$, $Pais(X_4) = \{X_6\}$, $Pais(X_3) = \{X_4, X_6\}$, $Pais(X_5) = \{X_4, X_6\}$, $Pais(X_9) = \{X_4, X_3\}$, $Pais(X_1) = \{X_3, X_4\}$, $Pais(X_2) = \{X_{1,}X_3\}$, $Pais(X_8) = \{X_3, X_4\}$ e $Pais(X_7) = \{X_8, X_4\}$. Para a Rede Probabilística Simples com 3- dependência (KDB3), Figura 4.5, temos $Pais(X_6) = \{\oslash\}$, $Pais(X_4) = \{X_6\}$, $Pais(X_3) = \{X_4, X_6\}$, $Pais(X_5) = \{X_4, X_6, X_3\}$, $Pais(X_9) = \{X_4, X_3, X_5\}$, $Pais(X_1) = \{X_3, X_4, X_9\}$, $Pais(X_2) = \{X_{1,}X_3, X_4\}$, $Pais(X_8) = \{X_3, X_4, X_1\}$ e $Pais(X_7) = \{X_8, X_4, X_3\}$.

Observando as Redes Probabilísticas de K-dependência, notamos que a rede com 0-dependência (KDB0) possui a mesma estrutura que uma Rede Probabilística Simples Naive Bayes, bem como a rede com 1-dependência (KDB1) possui a mesma estrutura que uma Rede Probabilística para classificação e bastante difundida na

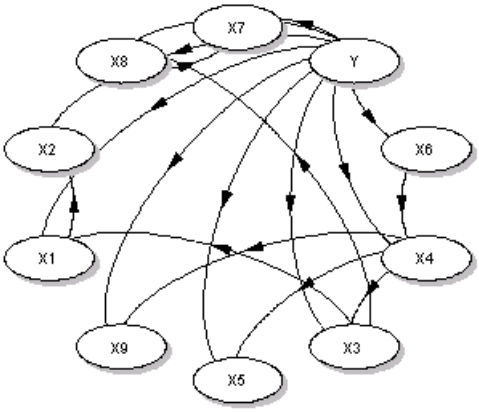


Figura 4.3: Exemplificação de uma Rede Probabilística Simples com 1- dependência.

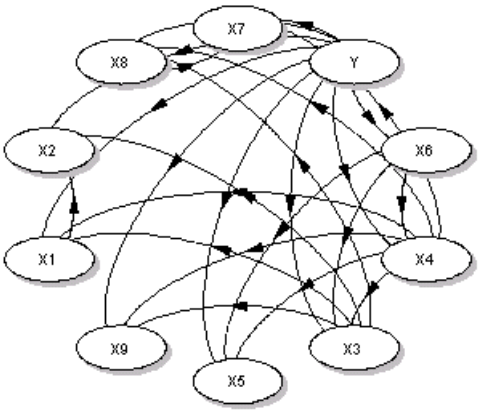


Figura 4.4: Exemplificação de uma Rede Probabilística Simples com 2- dependência.

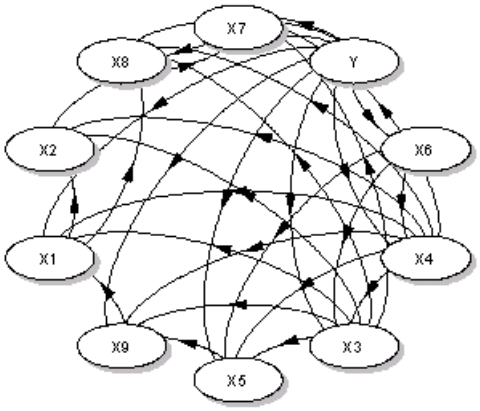


Figura 4.5: Exemplificação de uma Rede Probabilística Simples com 3-dependência.

literatura, conhecida como *Tree Augmented Network* (TAN) (FRIEDMAN *et al.*, 1997).

Desta forma, temos que as Redes Probabilísticas de K-dependência são uma generalização de outras redes particulares de classificação.

Para realizar o ajuste de tal estrutura através de um conjunto de dados, Sahami (1996) propõe o seguinte algoritmo exibido em 4.1.

Uma outra característica do algoritmo KDB, é sua adequabilidade ao contexto de *data mining*, requerindo uma pequena complexidade computacional para realização de estimações. O tempo gasto para a construção de uma estrutura de rede é de ordem $\mathcal{O}(p^2ncv^2)$ para o caso discreto e $\mathcal{O}(p^2)$ para o caso contínuo, p é o número de variáveis explicativas, n é o tamanho da amostra, c é o número de categorias da variável resposta e v é o número máximo de valores que uma varíavel discreta pode assumir (SAHAMI, 1996)(PÉREZ *et al.*, 2006).

Algoritmo 4.1 Algoritmo para construção de uma rede de k-dependência.

1. Para cada variável X_i, calcule a medida de informação mútua $I(X_i, Y)$;
2. Para cada par de variáveis explicativas, calcule a medida de informação mútua condicional $I(X_i, X_j|Y)$;
3. Defina S como a lista de variáveis explicativas utilizadas, inicialmente considere S como vazio;
4. Inicie a Rede Probabilística com a variável de classificação Y;
5. Repita até a lista S conter todas as variáveis explicativas:
   (a) Selecione a variável explicativa X_{max}que ainda não está contida em S e que possua a maior medida $I(X_{max}, Y)$;
   (b) Adicione à rede a variável X_{max};
   (c) Adicione um arco de Y para X_{max};
   (d) Adicione $m = min(|S|, K)$ arcos partindo das m X_j variáveis explicativas com o maior valor $I(X_{max}, X_j|Y)$;
   (e) Adicione X_{max} à lista S;
6. Calcule as tabelas de probabilidades condicionais considerando a estrutura construída.

4.3 Outros métodos de classificação

Nesta seção, exibimos sucintamente métodos de classificação tradicionais e solidificados na literatura. Desta forma, iremos, ao decorrer do trabalho, compará-los às Redes Probabilísticas.

4.3.1 Análise Discriminante

Conhecida também como Análise de Discriminante Linear (LDA), baseia-se na construção de uma ou mais funções lineares envolvendo as variáveis explicativas. Conseqüentemente, o modelo geral é dado por $Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$, onde Z representa o escore de discriminação, α o intercepto, β_i representa o coeficiente responsável pela contribuição linear da i-ésima variável explicativa X_i, sendo $i = 1, 2, \ldots, p$.

Porém, a LDA possui as seguintes suposições: (1) As matrizes de covariância de cada subconjunto de classificação são iguais. (2) Cada grupo de classificação segue uma distribuição normal multivariada.

4.3.2 Regressão Logística

Considerando um grupo de variáveis explicativas $X = X_1, ..., X_p$ e uma variável resposta com duas categorias $Y = y_1, y_2$, a técnica de Regressão Logística consiste em ajustar uma relação linear entre X e a transformação logito de Y. Desta forma, se consideramos y_1 como a categoria de interesse para análise, o modelo pode ser representado como $log\left(\frac{\pi}{1-\pi}\right) = X\beta$, onde $\pi = P(Y = y_1)$ e β o vetor contendo os coeficientes do modelo. Alternativamente, o modelo pode ser representado por 4.7 , sendo π_i a probabilidade o i-ésimo indivíduo pertencer à categoria y_1.

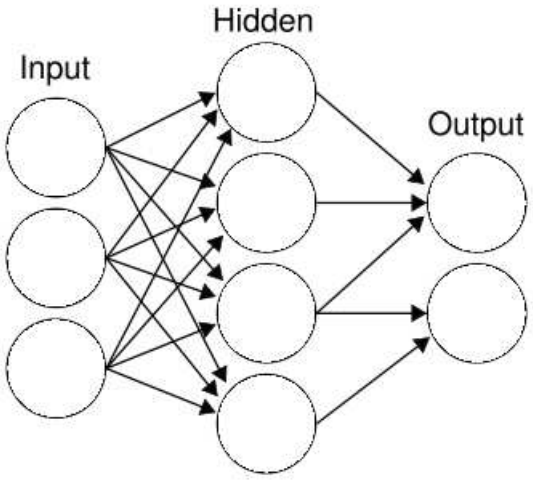


Figura 4.6: Exemplo de Rede Neural

$$\pi_i = \frac{\exp\{X_i\beta\}}{1 - \exp\{X_i\beta\}} \tag{4.7}$$

4.3.3 Regressão Probito

Semelhante a Regressão Logística, a Regressão Probito baseia-se no cálculo de $\pi = P(Y = y_1)$, porém temos que a parte linear do modelo é transformada pela função de probabilidade acumulada de uma distribuição normal padrão. Desta forma, temos o modelo $\pi_i = \Phi(X_i\beta)$, sendo $\Phi(.)$ a função de probabilidade acumulada de uma distribuição normal padrão, β o vetor contendo os coeficientes do modelo e π_i a probabilidade o i-ésimo indivíduo pertencer à categoria y_1.

4.3.4 Redes Neurais

Uma rede neural baseia-se em um sistema onde existem variáveis de entrada, variáveis explicativas, também conhecidas como *inputs* ou variáveis de saída, variáveis resposta, também conhecidas como *outputs*. Além disso, intermediariamente existem

variáveis ocultas, conhecidas como *hidden*, as quais são responsáveis pelos cálculos realizados ao decorrer da rede. As redes neurais foram criadas na tentativa de se aproximarem ao cérebro humano, uma vez que este se baseia no envio de sinais eletrônicos entre uma enorme quantidade de neurônios. Ou seja, a técnica de redes neurais possui elementos os quais recebem uma quantidade de *inputs* e geram respectivos *outputs*. Um exemplo de rede neural é apresentado na Figura 4.6.

As redes neurais se diferenciam de acordo com sua estrutura básica. De um modo geral, se diferenciam pela quantidade de camadas ocultas e pelas funções de ligação aplicadas a estas. Neste trabalho, consideramos Redes Neurais *Feed Forward*, caracterizadas por apresentarem apenas uma camada de variáveis ocultas. Um possível critério para definir a quantidade de variáveis ocultas em uma Rede Neural *Feed Forward* é tomar a média geométrica $(\prod_i^n X_i)^{1/n}$ entre o número de quantidade de *inputs* e a quantidade de *outputs* (Hamilton *et al.*, 1995).

4.4 Medidas de desempenho

Nesta seção apresentamos algumas medidas de desempenho utilizadas para avaliar a capacidade preditiva das técnicas de classificação binária em estudo. Considere a sequência aleatória de tamanho N a estrutura a ser predita como o conjunto $D = d_1, ..., d_N$. Também considere as classificações realizadas pelo modelo na forma de $M = m_1, .., .m_N$. No geral, d_i e m_i com $i = 1...N$ podem ser indicadores discretos $\{0, 1\}$, sendo 1 o valor indicativo que a observação i pertence a classe de interesse y_k. Assim, temos o objetivo de comparar M com D, isto é, comparar os valores preditos do modelo com os valores reais utilizados na predição. Assim, na Tabela 4.1 temos a seguinte estrutura de tabela de contingência 2x2, também conhecida com matriz de confusão.

Tabela 4.1: Matriz de confusão.

		M	
		$\{1\}$	$\{0\}$
D	$\{1\}$	VP	FN
	$\{0\}$	FP	VN

onde VP é o número de verdadeiros positivos, FP o número falsos positivos, FN é o número de falsos negativos e VN é o número de verdadeiros negativos. Naturalmente, temos que $VP + FP + FN + FP = N$.

Desta forma, utilizamos as seguintes métricas para avaliar a performance de M.

Acurácia (*Accurate* - ACC): é a fração de acertos de um modelo, tanto para as classificações de indivíduos para a classe 1 quanto para as classificações de indivíduos para a classe 0. É definida em 4.8.

$$ACC = \frac{VP + VN}{VP + VN + FN + FP} \tag{4.8}$$

Sensibilidade (*Sensibility* - SEN): é a fração dos indivíduos que o modelo classificou corretamente para a classe 1 dentre todos os indivíduos pertencentes à classe 1. É definida em 4.9.

$$SEN = \frac{VP}{VP + FN} \tag{4.9}$$

Especificidade (*Specificity* - SPE): é a fração dos indivíduos que o modelo classificou corretamente para a classe $\{0\}$ dentre todos os indivíduos pertencentes à classe $\{0\}$. É definida em 4.10.

$$SPE = \frac{VN}{VN + FP} \tag{4.10}$$

Coeficiente de Correlação de Mattew (*Mattew's Correlation Coeficient* - MCC):

medida usualmente utilizada para interpretar a classificação geral do modelo (BALDI *et al.*, 2000), é interpretada de maneira similar ao Coeficiente de correlação de Pearson: quando igual 1, a classificação é perfeita; quando igual a 0, classificação é nula; quando igual a -1, a classificação é totalmente inversa. É definido em 4.11.

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)\,(VP + FN)\,(VN + FP)(VN + FN)}} \tag{4.11}$$

Porém, notamos que o MCC não esta definido quando pelo menos uma das somas VP + FP, VP + FN, VN + FP ou VN + FN é zero, como é o caso se não houver valores preditivos positivos.

Correlação Aproximada (*Approximate Correlation* - AC): Burset e Guigó (1996) definem uma métrica aproximada de correlação, a qual não possui o mesmo problema que o MCC e retorna um valor entre -1 e +1 com a mesma interpretação do MCC. Burset and Guigó (1996) observam que a AC é muito próxima do valor real de correlação. Entretanto, alguns autores, como Baldi *et al.* (2000), não incentivam o uso dessa métrica, pois esta correção causa uma descontinuidade em seu limite, o que prejudica sua interpretação geométrica.

$$AC = 2\left(\frac{1}{4}\left[\frac{VP}{VP + FN} + \frac{VP}{VP + FP} + \frac{VN}{VN + FP} + \frac{VN}{VN + FN}\right] - 0,5\right) \tag{4.12}$$

GAMA: O coeficiente gama (GOODMAN e KRUSKAL, 1963) é dada como uma medida de associação que é altamente resistente e é definido em 4.13.

$$GAMA = \frac{(VP + VN) - (FN + FP)}{VP + VN + FN + FP} \tag{4.13}$$

Curva ROC (*Receiver Operating Characteristic*): também conhecida como Curva Característica Operativa do Receptor, foi introduzida em 1993 por Zweig e Campbell,

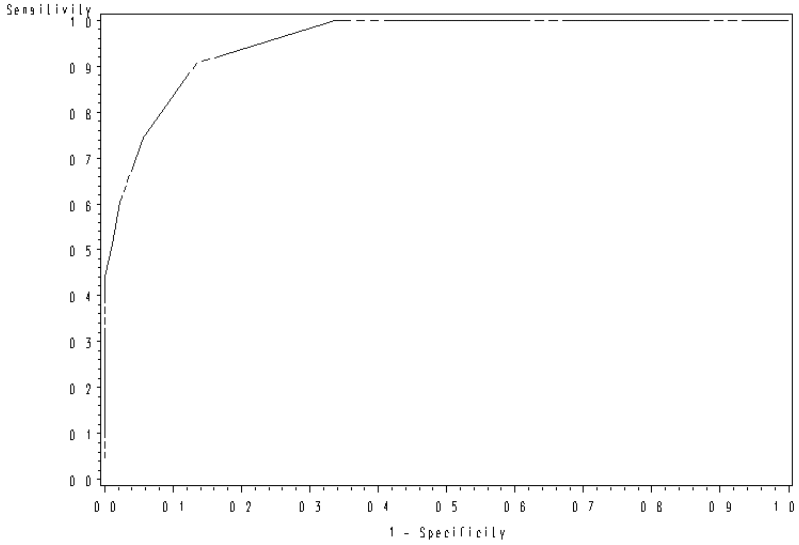


Figura 4.7: Exemplo de Curva ROC

pode ser definida, geometricamente, como um gráfico em que, para a abscissa, temos a medida de 1-especificidade, e para ordenada, temos a medida de sensibilidade. Esse plano é designado unitário, pois cada eixo possui tamanho 1. A sensibilidade é responsável pela proporção de indivíduos com a característica do modelo, a especificidade é responsável pela proporção de indivíduos sem a característica de interesse que é identificada corretamente pelo modelo.

A curva ROC é construída variando o ponto de corte de classificação e através da amplitude dos escores, para ambos os casos temos os escores como probabilidades. Um exemplo de curva ROC é exibido na Figura 4.7 .

Uma curva ROC obtida ao longo da diagonal principal corresponde a uma classificação obtida sem a utilização de qualquer ferramenta preditiva, ou seja, sem a presença de modelos. Consequentemente, a curva ROC deve ser interpretada de forma que, quanto mais a curva estiver distante da diagonal principal, melhor o desempenho do modelo associado a ela.

Para definir o melhor ponto de corte, temos que escolher o ponto que máximize conjuntamente a sensibilidade e a especificidade da classificação. Sendo assim, escolhemos o ponto mais próximo do eixo superior esquerdo do gráfico. Ou seja, temos que o melhor ponto de corte é o que possui menor distância euclidiana do ponto (0,1).

Computacionalmente, os códigos em `R` para a implementação da curva ROC estão disponíveis no Apêndice H.

Uma comparação entre as técnicas de classificação utilizando as medidas de performance acima são apresentadas na seção 4.6.

4.5 O procedimento *Bagging*

A origem da palavra *bagging* provém de *"bootstrap aggregating"*, o qual visa combinar estimações realizadas pela técnica *bootstrap*. A técnica *bootstrap* (EFRON, 1982) é basicamente uma técnica de reamostragem que permite aproximar a distribuição de uma função das observações pela distribuição empírica dos dados, baseada em uma amostra de tamanho finito.

O procedimento de *bootstrap* é exibido sucintamente a seguir.

Seja $X = (X_1, X_2, \ldots, X_n)$ uma amostra independente e identicamente distribuída contendo n observações:

1. Retire K amostras $X = (X^{(1)}, X^{(2)}, \ldots, X^{(K)})$ com reposição e de comprimento n;

2. Calcule as estimativas da estatística F de interesse:

$$\hat{\theta}_{(k)} = F\left[x^{(k)}\right] \; k = 1, \ldots, K$$

3. Calcule a estimativa *boostrap* da estatística de interesse, $\hat{\theta}_{boot}$, dada por:

4. Calcule o erro padrão *boostrap*, $\hat{S}_{boot}$, dado por:

$$\hat{S}_{boot} = \left\{ \frac{1}{K-1} \left[\sum_{k=1}^{K} \left(\hat{\theta}_{(k)} - \hat{\theta}_{boot} \right) \right]^2 \right\}^{1/2}$$

A idéia básica do procedimento de *bagging* é combinar as predições de vários modelos ajustados em reamostras *bootstrap*, sumarizando-as em apenas uma predição geral no intuito de aumentar a precisão da classificação. Sendo introduzido por Breiman (1996) e descrito da seguinte forma:

Considere o conjunto $L = (x_i, y_i)$ uma amostra aleatória a ser utilizada como base de treinamento e com n elementos independentes e identicamente distribuídos, onde x_i representa o vetor de variáveis explicativas e $y_i \in \{0, 1, \ldots, J\}$ com J classes.

O objetivo é encontrar $P(Y = j|X = x)$.

A amostra de treinamento é usualmente 70-80% da base completa dos dados.

Também, considere $\hat{\theta}_n(x)$ como a estimação realizada pelo modelo ajustado à amostra $L = (x_i, y_i)$. O procedimento de *Bagging* é definido como:

1. Retire B reamostras *bootstrap*, $L_1^*, L_2^*, \ldots L_B^*$, sendo estas amostras com reposição de tamanho n. Usualmente, é assumido $B = 50$.

2. Para cada reamostra do passo 1 realize a estimação $\hat{\theta}_{n,b}^*(L_b^*)$ através do modelo, sendo $b = 1, 2, ..., B$.

3. Obtenha o estimador de *bagging*, denotado por $\hat{\theta}_B(L_b^*)$, através da combinação de todos os $\hat{\theta}_{n,b}^*(L_b^*)$ do passo anterior.

Este procedimento é ilustrado esquematicamente pela Figura 4.8.

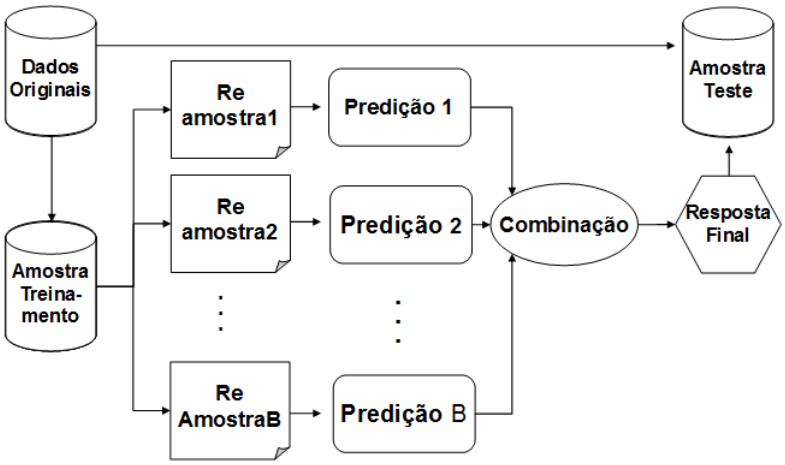


Figura 4.8: Esquematização do procedimento de *Bagging*.

4.6 Comparação entre os métodos de classificação

Para realizar esta comparação, consideramos conjuntos de dados reais disponibilizados no Repositório de dados da Universidade de Califórnia (http://archive.ics.uci.edu/ml/), sendo 4 conjuntos de dados com variáveis explicativas puramente discretas e 4 conjuntos de dados com com variáveis puramente contínuas. Maiores detalhes sobre os conjunto de dados podem ser encontrados no Anexo I.

Todos os conjuntos foram separados em base de treinamento (80%) e teste (20%). Para cada base de treinamento, aplicamos os métodos de Regressão Logístia (*Logistic Regression* - LR), Regressão Probito (*Probit Regression* - PR), Análise Discriminante (*Linear Discriminant Analysis* - LDA), Redes Neurais (*Neural Networks* - NN) e Redes Probabilísticas de K-dependência (*K-Dependence Bayesian Networks* - KDB). Para cada base de teste, calculamos as medidas de desempenho Especificidade, Sensibilidade, Acurária, Coeficiente de Correlação de Mattew, o Coeficiente de Correlação Aproximada (AC) e a estatística Gama. Todo este procedimento foi

Tabela 4.2: Comparação entre os métodos de classificação através de dados reais discretos

Base de Dados	n	p	Medidas	LR	PR	LDA	NN	KDB0	KDB1	KDB2
			SPE	0.742	0.737	0.742	0.721	0.780	0.663	0.672
			SEN	0.662	0.664	0.675	0.725	0.722	0.578	0.556
Breast Cancer	286	10	ACC	0.719	0.716	0.723	0.723	0.762	0.639	0.646
			MCC	0.381	0.378	0.394	0.419	0.471	0.221	0.208
			AC	0.383	0.379	0.395	0.420	0.473	0.223	0.209
			GAMA	0.390	0.433	0.447	0.445	0.525	0.278	0.250
			SPE	0.758	0.758	0.875	0.822	0.890	0.717	0.657
			SEN	0.734	0.783	0.867	0.815	0.850	0.722	0.588
Australian Credit	690	14	ACC	0.747	0.767	0.872	0.819	0.924	0.720	0.623
			MCC	0.490	0.537	0.740	0.636	0.778	0.439	0.246
			AC	0.490	0.537	0.740	0.636	0.778	0.439	0.246
			GAMA	0.495	0.534	0.744	0.639	0.780	0.439	0.246
			SPE	0.708	0.708	0.721	0.712	0.734	0.633	0.524
			SEN	0.717	0.715	0.725	0.635	0.750	0.584	0.667
German Credit	1000	20	ACC	0.711	0.710	0.723	0.688	0.739	0.619	0.565
			MCC	0.397	0.394	0.419	0.331	0.453	0.203	0.173
			AC	0.398	0.395	0.420	0.332	0.455	0.204	0.174
			GAMA	0.421	0.420	0.445	0.377	0.478	0.238	0.130
			SPE	0.789	0.790	0.867	0.836	0.890	0.759	0.601
			SEN	0.751	0.756	0.886	0.805	0.877	0.728	0.580
Japanese Credit	653	15	ACC	0.771	0.773	0.876	0.823	0.902	0.744	0.622
Screening			MCC	0.540	0.546	0.750	0.643	0.779	0.486	0.202
			AC	0.540	0.546	0.750	0.643	0.779	0.486	0.202
			GAMA	0.541	0.547	0.752	0.646	0.780	0.489	0.206

replicado 100 vezes, sendo a comparação realizada pela estimativa pontual da média de cada medida de desempenho considerada.

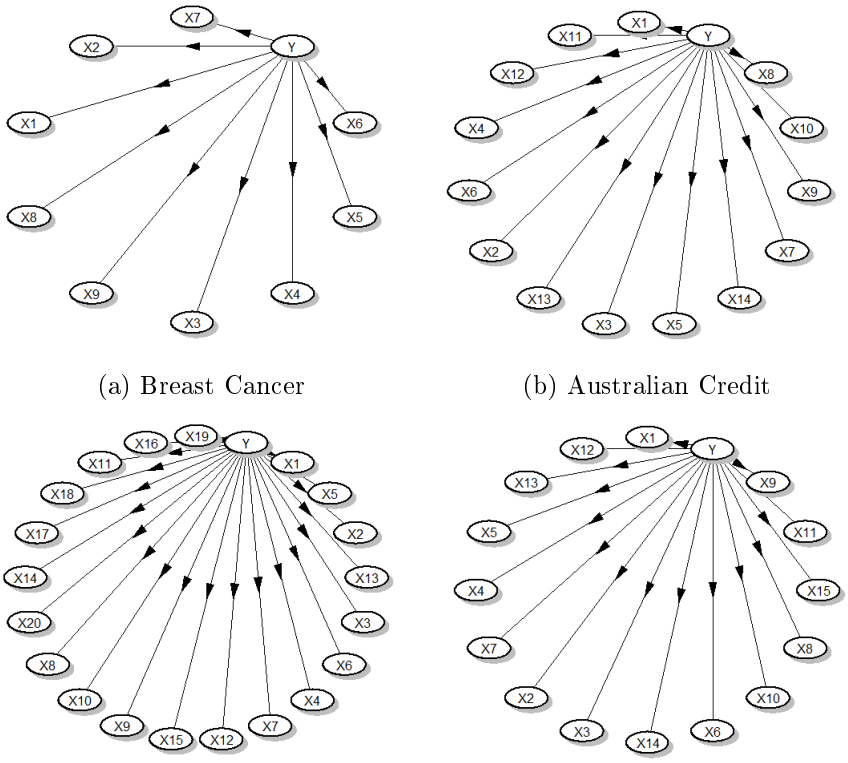
A comparação entre os métodos para o caso discreto é exibida na Tabela 4.2, já a comparação entre os métodos para o caso contínuo é exibida na Tabela 4.3, sendo n o número de observações em cada conjunto de dados e p o número de variáveis explicativas.

Particularmente, para a técnica de Redes Neurais adotamos como critério que o número de variáveis ocultas é igual a média geométrica entre o número de variáveis explicativas e o número de variáveis resposta (HAMILTON *et al.*, 1995).

Para os resultados da Tabela 4.2 podemos verificar visualmente que, para os conjuntos de dados analisados, as redes de k-dependência possuem maior capacidade preditiva, especialmente considerando como métricas gerais o MCC, AC e GAMA.

Tabela 4.3: Comparação entre os métodos de classificação através de dados reais contínua

Base de Dados	n	p	Medidas	LR	PR	LDA	NN	KDB0	KDB1	KDB2
			SPE	0.745	0.740	0.750	0.650	0.768	0.706	0.686
			SEN	0.791	0.798	0.792	0.658	0.772	0.754	0.723
Ecocardiograma	107	8	ACC	0.760	0.758	0.763	0.648	0.768	0.722	0.702
			MCC	0.513	0.511	0.518	0.305	0.519	0.442	0.398
			AC	0.515	0.513	0.520	0.307	0.520	0.443	0.401
			GAMA	0.519	0.515	0.526	0.295	0.536	0.444	0.405
			SPE	0.867	0.870	0.866	0.854	0.874	0.866	0.857
			SEN	0.838	0.834	0.841	0.791	0.842	0.841	0.842
Heart(Statlog)	270	13	ACC	0.854	0.854	0.855	0.829	0.860	0.855	0.849
			MCC	0.705	0.704	0.706	0.648	0.716	0.706	0.696
			AC	0.705	0.704	0.706	0.648	0.716	0.706	0.696
			GAMA	0.708	0.707	0.709	0.657	0.720	0.709	0.699
			SPE	0.739	0.735	0.711	0.726	0.687	0.746	0.744
			SEN	0.687	0.688	0.699	0.675	0.692	0.683	0.686
Transfusion	748	5	ACC	0.727	0.723	0.707	0.714	0.688	0.731	0.730
			MCC	0.380	0.375	0.361	0.360	0.331	0.385	0.386
			AC	0.383	0.378	0.364	0.363	0.334	0.388	0.389
			GAMA	0.453	0.447	0.415	0.428	0.376	0.461	0.459
			SPE	0.775	0.776	0.800	0.837	0.802	0.838	0.846
			SEN	0.723	0.719	0.756	0.711	0.694	0.815	0.863
Sonar	208	60	ACC	0.750	0.749	0.780	0.777	0.753	0.827	0.856
			MCC	0.498	0.496	0.559	0.560	0.510	0.652	0.712
			AC	0.499	0.496	0.559	0.561	0.510	0.652	0.712
			GAMA	0.500	0.498	0.560	0.555	0.507	0.653	0.711



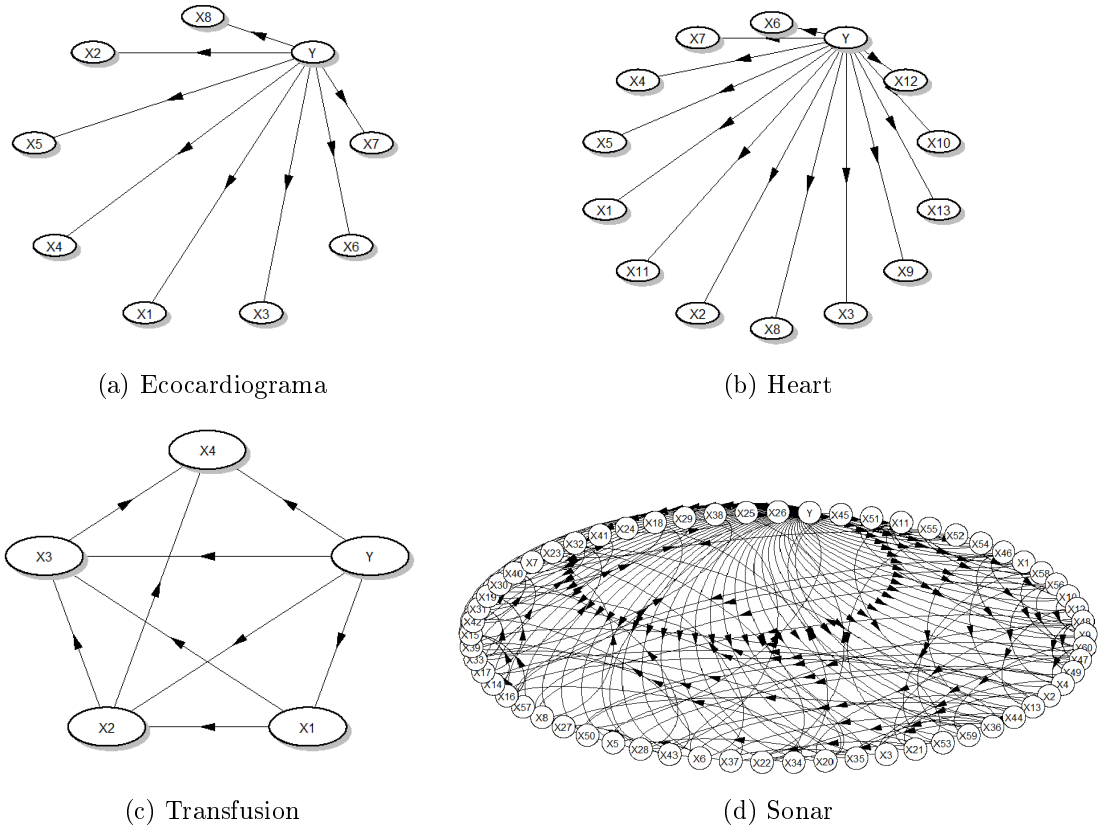
(a) Breast Cancer (b) Australian Credit

(c) German Credit (d) Japanese Credit

Figura 4.9: Estruturas de Rede Probabilística para os conjuntos de dados com variáveis explicativas discretas.

Para este caso, todos os conjuntos de dados estudados admitem, visualmente, que as redes de 0-dependência (Naive Bayes) possuem a melhor capacidade de preditiva. As estruturas destas redes são exibidas na Figura 4.9.

Através dos resultados da Tabela 4.3 podemos verificar visualmente que, para os conjuntos de dados com variáveis explicativas contínuas, as redes de K-dependência possuem também maior capacidade preditiva. Neste sentido, Sahami (1996) evidencia que para determinados conjuntos de dados podemos achar um valor para K no qual a capacidade preditiva é mais satisfatória. Para os conjuntos Ecocardiograma e Heart as redes de 0-dependência possuem melhor capacidade preditiva, o conjunto



(a) Ecocardiograma

(b) Heart

(c) Transfusion

(d) Sonar

Figura 4.10: Estruturas de Rede Probabilística para os conjuntos de dados com variáveis explicativas discretas.

Transfusion admite as redes de 2-dependência com a melhor capacidade preditiva considerando as métricas MCC e AC, porém aas redes de 1-dependência possuem o melhor desempenho preditivo pontual via a métrica GAMA. Por fim , as redes de 2-dependência possuem a melhor capacidade preditiva para o conjunto de dados sonar. As estruturas das redes contínuas são exibidas na Figura 4.10.

O procedimento de *Bagging* com 5 combinações de modelos, direcionado aos dados com variáveis explicativas discretas, é exibido na Tabela 4.4. A Tabela 4.5 exibe a aplicação do mesmo procedimento para o caso dos dados com variáveis explicativas

Tabela 4.4: Aplicação do procedimento *Bagging-5* para os conjuntos de dados com variáveis explicativas discretas.

Base de Dados	n	p	Medidas	LR	PR	LDA	NN	KDB0	KDB1	KDB2
Breast Cancer	286	10	SPE	0.809	0.851	0.775	0.800	0.800	0.743	0.744
			SEN	0.682	0.682	0.667	0.773	0.800	0.730	0.776
			ACC	0.776	0.806	0.745	0.794	0.800	0.739	0.752
			MCC	0.480	0.526	0.414	0.559	0.555	0.431	0.469
			AC	0.481	0.527	0.415	0.560	0.557	0.433	0.472
			GAMA	0.552	0.612	0.491	0.588	0.600	0.479	0.503
Australian Credit	690	14	SPE	0.896	0.831	0.896	0.871	0.905	0.756	0.680
			SEN	0.787	0.869	0.869	0.882	0.878	0.814	0.758
			ACC	0.848	0.848	0.884	0.876	0.894	0.783	0.653
			MCC	0.691	0.696	0.765	0.753	0.787	0.568	0.352
			AC	0.691	0.696	0.765	0.753	0.787	0.568	0.352
			GAMA	0.696	0.696	0.768	0.752	0.787	0.565	0.353
German Credit	1000	20	SPE	0.665	0.669	0.724	0.713	0.757	0.733	0.594
			SEN	0.857	0.770	0.725	0.712	0.728	0.489	0.615
			ACC	0.705	0.700	0.724	0.715	0.749	0.675	0.598
			MCC	0.428	0.406	0.420	0.379	0.456	0.207	0.186
			AC	0.436	0.407	0.421	0.381	0.457	0.208	0.188
			GAMA	0.433	0.409	0.448	0.430	0.498	0.350	0.197
Japanese Credit Screening	653	15	SPE	0.812	0.792	0.865	0.868	0.897	0.736	0.771
			SEN	0.804	0.830	0.898	0.864	0.884	0.753	0.590
			ACC	0.809	0.809	0.880	0.866	0.910	0.743	0.687
			MCC	0.617	0.619	0.761	0.731	0.780	0.492	0.369
			AC	0.617	0.619	0.761	0.731	0.780	0.492	0.369
			GAMA	0.618	0.618	0.761	0.733	0.781	0.493	0.374

contínuas.

Através das Tabelas 4.4 e 4.5 notamos que a mesma estrutura de performance preditiva se mantém para todos os conjuntos de dados analisados, sendo as redes de K-dependência a técnica com maior capacidade preditiva, sendo esta incrementada através dos procedimento de *Bagging,* sendo o maior ganho observado para o caso discreto no conjunto de dados Ecocardiograma, um aumento de aproximadamente 50% para a métrica MCC, migrando de 0,519 para 0,795. Já para o caso contínuo, o maior ganho está no conjunto de dados *Breast Cancer*, com um aumento de aproximadamente 18% para a métrica MCC, migrando de 0,471 para 0,555.

Tabela 4.5: Aplicação do procedimento *Bagging-5* para os conjuntos de dados com variáveis explicativas contínuas.

Base de Dados	n	p	Medidas	LR	PR	LDA	NN	KDB0	KDB1	KDB2
Ecocardiograma	107	8	SPE	0.886	0.866	0.867	0.807	0.901	0.831	0.811
			SEN	0.923	0.910	0.906	0.871	0.886	0.778	0.822
			ACC	0.900	0.881	0.876	0.824	0.900	0.819	0.810
			MCC	0.787	0.749	0.752	0.646	0.795	0.614	0.610
			AC	0.788	0.750	0.754	0.648	0.797	0.616	0.612
			GAMA	0.800	0.762	0.752	0.648	0.800	0.638	0.619
Heart(Statlog)	270	13	SPE	0.900	0.888	0.907	0.911	0.908	0.907	0.884
			SEN	0.872	0.886	0.872	0.877	0.884	0.872	0.887
			ACC	0.889	0.887	0.891	0.894	0.898	0.891	0.887
			MCC	0.774	0.774	0.779	0.787	0.793	0.779	0.771
			AC	0.774	0.774	0.779	0.787	0.793	0.779	0.771
			GAMA	0.778	0.774	0.781	0.789	0.796	0.781	0.774
Transfusion	748	4	SPE	0.753	0.746	0.715	0.749	0.719	0.776	0.763
			SEN	0.708	0.713	0.745	0.717	0.729	0.676	0.700
			ACC	0.743	0.739	0.722	0.741	0.722	0.753	0.748
			MCC	0.413	0.408	0.400	0.412	0.391	0.408	0.417
			AC	0.416	0.411	0.404	0.415	0.395	0.410	0.420
			GAMA	0.485	0.477	0.444	0.483	0.444	0.505	0.496
Sonar	208	60	SPE	0.741	0.777	0.782	0.892	0.892	0.872	0.927
			SEN	0.834	0.788	0.811	0.852	0.852	0.906	0.912
			ACC	0.783	0.786	0.793	0.871	0.871	0.890	0.921
			MCC	0.574	0.563	0.587	0.742	0.742	0.776	0.840
			AC	0.574	0.563	0.587	0.742	0.742	0.776	0.840
			GAMA	0.567	0.571	0.586	0.743	0.743	0.781	0.843

4.7 Estudo de Simulação

Realizamos uma avaliação comparativa entre os métodos utilizando um método exaustivo de amostragem, na qual retiramos K amostras de tamanho 1000 e verificamos a mesma estatística para cada uma delas com o objetivo de estudar as distribuições destas estatísticas para as K amostras. Para isso utilizamos 100 amostras replicadas (K=100).

Através de uma base de dados artificiais, analisamos a performance dos métodos: Regressão Logística, Regressão Probito, Redes Neurais, Análise Discriminante e Redes Probabilísticas.

Para isso, no contexto de *Credit Score*, consideramos tipos de 3 populações, repesctivamente com 50%, 25% e 10% de clientes maus pagadores. Além disso, con-

sideramos os casos em que as variáveis explicativas são discretas ou contínuas e se possuem, ou não, dependência probabilística entre as mesmas. Isto é, no primeiro caso cada tipo de população possui 6 variáveis explicativas discretas e independentes. No segundo caso, cada uma possui 6 variáveis explicativas discretas e dependentes. No terceiro caso, 6 variáveis explicativas contínuas e independentes. Por fim, no quarto caso, 6 variáveis explicativas contínuas e dependentes.

A base de dados artificiais foi originalmente gerada com variáveis explicativas contínuas (Breiman, 1998), sendo que para o caso de variáveis explicativas independentes, a distribuição dos bons pagadores segue uma normal hexa-variada com média $\mu_B = (0,0,0,0,0,0)$ e matriz de covariância $\Sigma_B = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ & 4 & 0 & 0 & 0 & 0 \\ & & 4 & 0 & 0 & 0 \\ & & & 4 & 0 & 0 \\ & & & & 4 & 0 \\ & & & & & 4 \end{pmatrix}$; a distribuição dos maus pagadores segue uma normal hexa-variada com média $\mu_M = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right)$ e matriz de covariância $\Sigma_B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 \\ & & & & 1 & 0 \\ & & & & & 1 \end{pmatrix}$. Para o caso de variáveis explicativas dependentes, a distribuição dos bons pagadores segue uma normal hexa-variada com média $\mu_B = (0,0,0,0,0,0)$ e matriz de covariância $\Sigma_B = \begin{pmatrix} 4 & 2 & 0 & 0 & 0 & 0 \\ & 4 & 0 & 0 & 0 & 0 \\ & & 4 & 0 & 0 & 0 \\ & & & 4 & 0 & 0 \\ & & & & 4 & 2 \\ & & & & & 4 \end{pmatrix}$, a distribuição dos maus pagadores segue uma normal hexa-variada com média $\mu_M = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right)$ e matriz de covariância $\Sigma_M = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 \\ & & & & 1 & 1 \\ & & & & & 1 \end{pmatrix}$. Para obter as bases de dados com variáveis explicativas discretas categorizamos as mesmas através de seus quartis, resultando em 4 categorias para cada variável. Os códigos em `R` para a geração do conjunto de dados estão disponíveis no Apêndice F.

Para cada replicação e cada tipo de configuração retiramos amostras treinamento de tamanho 900 a fim de classificar uma amostra teste com 50 indivíduos bons pagadores e 50 indivíduos maus pagadores.

Com o objetivo de aumentar a capacidade preditiva dos modelos e observar o comportamento das Redes Probabilísticas, aplicamos em todas as configurações o procedimento *Bagging,* composto pela combinação de 15 modelos ajustados a amostras treinamento com reposição. Para realizar a combinação destes modelos, consideramos o método de regressão logística, a qual é capaz de ponderar a capacidade preditiva de cada um dos 15 modelos ajudados.

Através da comparação das 5 técnicas avaliadas, sendo a técnica de Redes Probabilísticas aplicada para 3 estruturas diferentes (KDB0, KDB1, KDB2) este estudo de simulação considerou 142800 modelos ajustados.

A estimação pontual das medidas de desempenho para o primeiro caso (variáveis explicativas discretas e independentes) é exibida na Tabela 4.6; para o segundo caso (variáveis explicativas discretas e dependentes), na Tabela 4.7; o terceiro caso (variáveis explicativas contínuas e independentes), na Tabela 4.8 e, por fim, o quarto caso (variáveis explicativas continuas e dependentes), na Tabela 4.9.

Através a Tabela 4.6 verificamos visualmente que as redes KDB possuem uma melhor performance preditiva que as demais técnicas avaliadas, em especial a rede KDB0, sendo suas métricas de desempenho geral (MCC, AC e GAMA) aproximando-se de 0,75 para amostras desbalanceadas a 25% de maus pagadores. Desta forma, para dados discretos e independentes, nesta conjectura as redes KDB0, que assumem independência entre as variáveis explicativas, são as mais adequadas.

Para o caso de dados discretos e dependentes, apresentado na Tabela 4.7, verificamos, também, que as redes KDB possuem uma melhor performance preditiva que as demais técnicas avaliadas, sendo as redes KDB1 com as maiores medidas de desempenho, neste caso estas redes assumem dependência de ordem 1 entre as variáveis explicativas.

No caso do tratamento de dados contínuos, apresentado pelas Tabelas 4.8 e 4.9, as

Tabela 4.6: Comparação entre os métodos através de simulação em dados discretos e independentes.

Configuração 1 -50% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.79	0.80	0.80	0.60	0.60	0.60	0.86	0.86	0.86	0.72	0.72	0.72
KDB1	0.78	0.80	0.79	0.59	0.59	0.58	0.85	0.86	0.85	0.71	0.71	0.70
KDB2	0.76	0.76	0.76	0.53	0.53	0.53	0.84	0.84	0.84	0.69	0.69	0.69
LR	0.79	0.81	0.80	0.60	0.60	0.59	0.84	0.84	0.84	0.68	0.68	0.68
PR	0.79	0.80	0.80	0.60	0.60	0.59	0.84	0.84	0.84	0.67	0.67	0.67
NN	0.77	0.79	0.78	0.57	0.57	0.56	0.83	0.84	0.83	0.67	0.67	0.67
LDA	0.79	0.80	0.80	0.60	0.60	0.59	0.83	0.84	0.84	0.68	0.68	0.68

Configuração 2 -25% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.82	0.81	0.82	0.63	0.63	0.63	0.89	0.86	0.88	0.75	0.75	0.75
KDB1	0.81	0.80	0.80	0.61	0.61	0.61	0.88	0.85	0.86	0.73	0.73	0.73
KDB2	0.80	0.74	0.77	0.54	0.54	0.54	0.87	0.82	0.84	0.69	0.69	0.69
LR	0.81	0.82	0.81	0.63	0.63	0.63	0.88	0.83	0.86	0.72	0.72	0.72
PR	0.81	0.81	0.81	0.62	0.62	0.62	0.88	0.84	0.86	0.72	0.72	0.71
NN	0.79	0.77	0.78	0.57	0.57	0.56	0.87	0.81	0.84	0.68	0.68	0.68
LDA	0.80	0.82	0.81	0.62	0.62	0.62	0.85	0.86	0.86	0.72	0.72	0.72

Configuração 3 -10% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.79	0.84	0.81	0.63	0.63	0.62	0.88	0.84	0.86	0.72	0.72	0.72
KDB1	0.76	0.77	0.77	0.53	0.53	0.53	0.86	0.80	0.83	0.66	0.66	0.66
KDB2	0.76	0.70	0.73	0.47	0.47	0.46	0.72	0.69	0.71	0.45	0.47	0.41
LR	0.80	0.83	0.81	0.62	0.62	0.62	0.90	0.79	0.84	0.69	0.69	0.68
PR	0.80	0.82	0.81	0.62	0.62	0.62	0.90	0.79	0.84	0.69	0.69	0.69
NN	0.73	0.70	0.71	0.44	0.45	0.42	0.73	0.79	0.59	0.59	0.59	0.59
LDA	0.78	0.81	0.80	0.60	0.60	0.59	0.84	0.85	0.85	0.70	0.70	0.70

Tabela 4.7: Comparação entre os métodos através de simulação em dados discretos e dependentes.

Configuração 1 - 50% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.82	0.74	0.78	0.57	0.57	0.56	0.86	0.85	0.85	0.71	0.71	0.71
KDB1	0.91	0.92	0.91	0.83	0.83	0.83	0.97	0.98	0.97	0.94	0.94	0.94
KDB2	0.89	0.92	0.90	0.81	0.81	0.81	0.96	0.97	0.96	0.92	0.92	0.92
LR	0.79	0.76	0.78	0.56	0.56	0.55	0.82	0.81	0.81	0.63	0.63	0.63
PR	0.79	0.76	0.78	0.56	0.56	0.55	0.82	0.80	0.81	0.62	0.62	0.62
NN	0.86	0.81	0.83	0.67	0.67	0.67	0.93	0.92	0.93	0.86	0.86	0.86
LDA	0.80	0.76	0.78	0.55	0.55	0.55	0.82	0.81	0.82	0.63	0.63	0.63

Configuração 2 - 25% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.82	0.79	0.81	0.62	0.62	0.61	0.89	0.85	0.87	0.74	0.74	0.73
KDB1	0.90	0.93	0.91	0.83	0.83	0.83	0.96	0.96	0.96	0.92	0.92	0.92
KDB2	0.88	0.89	0.88	0.77	0.77	0.77	0.94	0.92	0.93	0.86	0.86	0.86
LR	0.81	0.80	0.81	0.61	0.61	0.61	0.89	0.82	0.85	0.71	0.71	0.71
PR	0.81	0.80	0.80	0.61	0.61	0.61	0.88	0.83	0.85	0.70	0.70	0.70
NN	0.81	0.77	0.79	0.62	0.63	0.58	0.93	0.86	0.89	0.79	0.79	0.79
LDA	0.81	0.78	0.80	0.60	0.60	0.59	0.84	0.85	0.85	0.69	0.69	0.69

Configuração 3 - 10% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.83	0.80	0.81	0.63	0.63	0.63	0.89	0.86	0.87	0.75	0.75	0.75
KDB1	0.88	0.88	0.88	0.76	0.76	0.76	0.92	0.84	0.88	0.77	0.77	0.76
KDB2	0.86	0.83	0.85	0.69	0.69	0.69	0.86	0.83	0.85	0.69	0.69	0.69
LR	0.83	0.80	0.81	0.62	0.62	0.62	0.92	0.82	0.87	0.74	0.74	0.74
PR	0.82	0.80	0.81	0.62	0.62	0.62	0.91	0.82	0.87	0.74	0.74	0.74
NN	0.76	0.71	0.74	0.50	0.52	0.47	0.90	0.76	0.83	0.68	0.68	0.67
LDA	0.81	0.77	0.79	0.58	0.58	0.58	0.85	0.84	0.85	0.69	0.69	0.69

Tabela 4.8: Comparação entre os métodos através de simulação em dados contínuos e independentes.

Dados Independentes 50% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.88	0.90	0.89	0.79	0.79	0.79	0.94	0.96	0.95	0.90	0.90	0.90
KDB1	0.88	0.90	0.89	0.79	0.79	0.79	0.94	0.97	0.95	0.90	0.90	0.90
KDB2	0.88	0.90	0.89	0.79	0.79	0.79	0.94	0.96	0.95	0.90	0.90	0.90
LR	0.58	0.73	0.66	0.32	0.32	0.31	0.71	0.84	0.78	0.56	0.56	0.55
PR	0.58	0.74	0.66	0.32	0.32	0.32	0.70	0.84	0.77	0.55	0.55	0.54
NN	0.63	0.81	0.72	0.46	0.46	0.44	0.82	0.88	0.85	0.71	0.71	0.70
LDA	0.58	0.73	0.66	0.32	0.32	0.31	0.71	0.82	0.76	0.53	0.53	0.52

Dados Independentes 25% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.89	0.90	0.90	0.80	0.80	0.80	0.95	0.95	0.95	0.91	0.91	0.90
KDB1	0.90	0.90	0.90	0.80	0.80	0.80	0.95	0.94	0.95	0.90	0.90	0.90
KDB2	0.90	0.90	0.90	0.80	0.80	0.80	0.95	0.95	0.95	0.90	0.90	0.89
LR	0.60	0.74	0.67	0.35	0.35	0.34	0.83	0.89	0.86	0.71	0.71	0.71
PR	0.60	0.74	0.67	0.35	0.35	0.34	0.81	0.89	0.85	0.70	0.70	0.70
NN	0.61	0.81	0.71	0.44	0.44	0.43	0.80	0.85	0.83	0.66	0.66	0.66
LDA	0.60	0.74	0.67	0.35	0.35	0.34	0.71	0.84	0.78	0.56	0.56	0.55

Dados Independentes 10% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.88	0.92	0.90	0.81	0.81	0.81	0.95	0.94	0.95	0.89	0.89	0.89
KDB1	0.88	0.91	0.90	0.79	0.79	0.79	0.95	0.93	0.94	0.88	0.88	0.88
KDB2	0.87	0.91	0.89	0.78	0.78	0.78	0.94	0.93	0.94	0.87	0.87	0.87
LR	0.58	0.76	0.67	0.34	0.34	0.33	0.86	0.90	0.88	0.76	0.76	0.76
PR	0.58	0.76	0.67	0.34	0.34	0.33	0.85	0.90	0.87	0.75	0.75	0.74
NN	0.66	0.73	0.69	0.39	0.39	0.39	0.81	0.83	0.82	0.64	0.64	0.64
LDA	0.57	0.76	0.67	0.34	0.34	0.33	0.71	0.85	0.78	0.56	0.56	0.55

Tabela 4.9: Comparação entre os métodos através de simulação em dados contínuos e dependentes.

Dados Correlacionados, 50% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.87	0.90	0.89	0.78	0.78	0.77	0.94	0.95	0.95	0.89	0.89	0.89
KDB1	0.90	0.92	0.91	0.82	0.82	0.82	0.96	0.97	0.97	0.94	0.94	0.94
KDB2	0.90	0.92	0.91	0.81	0.81	0.81	0.96	0.97	0.97	0.94	0.94	0.94
LR	0.52	0.75	0.64	0.28	0.28	0.27	0.71	0.85	0.78	0.57	0.57	0.57
PR	0.52	0.75	0.64	0.28	0.28	0.27	0.71	0.85	0.78	0.57	0.57	0.57
NN	0.66	0.77	0.72	0.44	0.44	0.43	0.81	0.83	0.82	0.64	0.64	0.64
LDA	0.52	0.75	0.64	0.29	0.29	0.27	0.69	0.84	0.76	0.53	0.53	0.53

Dados Correlacionados, 25% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.90	0.90	0.90	0.79	0.79	0.79	0.95	0.95	0.95	0.90	0.90	0.90
KDB1	0.90	0.92	0.91	0.82	0.82	0.82	0.96	0.96	0.96	0.93	0.93	0.93
KDB2	0.90	0.92	0.91	0.82	0.82	0.82	0.96	0.96	0.96	0.92	0.92	0.92
LR	0.61	0.75	0.68	0.36	0.36	0.35	0.83	0.87	0.85	0.70	0.70	0.70
PR	0.61	0.74	0.68	0.36	0.36	0.35	0.81	0.87	0.84	0.68	0.68	0.68
NN	0.67	0.82	0.75	0.51	0.51	0.49	0.84	0.87	0.86	0.72	0.72	0.71
LDA	0.61	0.75	0.68	0.36	0.36	0.35	0.70	0.83	0.77	0.54	0.54	0.53

Dados Correlacionados, 10% de maus pagadores

	Sem Bagging						15-Bagging					
MODELO	SPE	SEN	ACC	MCC	AC	GAMA	SPE	SEN	ACC	MCC	AC	GAMA
KDB0	0.90	0.92	0.91	0.82	0.82	0.82	0.96	0.94	0.95	0.90	0.90	0.90
KDB1	0.90	0.93	0.91	0.83	0.83	0.82	0.96	0.95	0.95	0.91	0.91	0.91
KDB2	0.89	0.92	0.91	0.82	0.82	0.82	0.96	0.94	0.95	0.90	0.90	0.90
LR	0.58	0.74	0.66	0.33	0.33	0.32	0.85	0.90	0.88	0.76	0.76	0.76
PR	0.58	0.75	0.66	0.34	0.34	0.33	0.84	0.90	0.87	0.74	0.74	0.74
NN	0.66	0.75	0.71	0.42	0.42	0.41	0.81	0.82	0.82	0.64	0.26	0.63
LDA	0.58	0.75	0.66	0.33	0.33	0.33	0.70	0.83	0.77	0.54	0.54	0.53

redes KDB são ainda mais satisfatórias, apresentando métricas gerais de desempenho próximas a 0.90 em alguns casos. Particularmente, temos uma maior aproximação destas redes para o tratamento de dados contínuos e dependentes. Para o caso de dados contínuos e dependentes as redes KDB1 se mostram levemente mais adequada que as demais redes.

Além disso, temos que o procedimento de *bagging* aumentou a capacidade preditiva de todas as técnicas, porém não sendo muito efetivo para as redes KDB no caso de uma amostra desbalanceada a 10% de maus pagadores com variáveis discretas e dependentes.

Capítulo 5

Considerações Finais

Neste trabalho exibimos as definições gerais e implementações da técnica de Redes Probabilísticas de K-dependência para os caso em que as bases de dados possuem variáveis explicativas puramente discretas ou puramente contínuas, especificamente no contexto de classificação. Sendo esta técnica, uma metodologia recente iniciada na década de 80. Nosso embasamento foi construído, em sua maioria, no contexto de *credit scoring* e diagnóstico médico, áreas de grande aplicação para a técnica, na qual as Redes Probabilísticas são utilizadas para predizer a probabilidade de um cliente ser classificado como mau pagador ou a probabilidade de um paciente possuir determinada doença.

Por fim, apresentamos uma comparação desta metodologia com as técnicas de Regressão Logística, Regressão Probito, Análise Discriminante e Redes Neurais. E verificamos que, para os casos analisados, as Redes Probabilísticas são pontualmente mais indicadas.

Além disso, considerando o procedimento de *bagging*, notamos que as redes de K-dependência possuem, de uma forma geral, um ganho satisfatório de desempenho.

Desta forma, as teorias e técnicas sumarizadas nesta dissertação, vêm a contribuir para a área da Estatística, em especial a comunidade nacional, no sentido da falta de

literatura estatística que introduza e promova a aplicação real desta técnica. Alternativamente, uma vez que outros métodos de classificação são fortemente utilizados pela comunidade no desenvolmento científico e aplicações cotidianas, mostramos que as Redes Probabilísticas são uma opção de modelagem com alta capacidade preditiva. Acreditamos que muitas pesquisas, a princípio direcionadas a outras metodologias de classificação, poderiam obter maior performance preditiva através da utilização de Redes Probabilísticas.

Evidenciamos que, dado o universo de técnicas de classificação, em nosso trabalho temos o intuito de exibir a aplicabilidade da técnica de Redes Probabilísticas de K-dependência, a qual se mostra um tipo de modelagem extremamente competitiva para o caso particular de classificação binária. Não temos a intenção de provar analiticamente que esta é sempre a melhor alternativa de predição, uma vez que, dado o avanço nas metodologias de *data mining* e as diversas configurações de conjuntos de dados em critério de variabilidade, não existe na literatura uma técnica que se mostre melhor que as demais em todos os casos. Toda a modelagem realizada nesta dissertação foi balizada na eficiência da técnica.

Observamos também que as técnicas de Redes Probabilísticas estão em atual progresso, abrangendo diversos tipos de pesquisa, como o desenvolvimento de algoritmos para aprendizado de estrutura, algoritmos para aprendizado de probabilidades condicionais e técnicas de classificação, sendo pouco exploradas pela comunidade estatística quando comparadas aos demais assuntos da área.

5.1 Perspectivas Futuras

A técnica de Redes Probabilísticas é incipiente e pouco investigada pela comunidade estatística, comparada com outros métodos. Assim, sendo uma técnica intrínseca a

variabilidade estatística, existe a necessidade do contínuo estudo de outros algoritmos de construção de estruturas, baseados em conjunto de dados, bem como a estimação dos parâmetros.

Prioritariamente, a métrica central utilizada na construção de uma Rede Probabilística é a Informação Mútua, que até o momento não possui distribuição de probabilidade conhecida. Outras métricas podem ser estudadas, bem como o comportamento aleatória da Informação Mútua, o que contribuiria significativamente na composição de novos algoritmos neste enredo.

Além disso, há a necessidade do desenvolvimento de técnicas de inferência estatística para Redes Probabilísticas, tais como intervalos de confiança para os parâmetros estimados da rede.

Referências Bibliográficas

[1] ABDOU, H., POINTON, J., EL-MASRY, A. Neural nets versus conventional techniques in credit scoring in Egyptian banking. Expert Systems with Applications N. 35, p.1275–1292, 2008.

[2] ABELLAN J.; GOMEZ-OLMEDO M.; MORAL. S. Some variations on the PC algorithm. In Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM' 06), pages 1-8, 2006.

[3] ABICALAFFE, C.; AMARAL, V. F.; DIAS, J. S.. Aplicação da Rede Bayesiana na Prevenção da Gestão de Alto Risco. In: Congresso Brasileiro de Informática Médica, Ribeirão Preto. Anais do Congresso Brasileiro de Informática Médica, v. 1. p. 1-1, 2004.

[4] BALDI P., BRUNAK, S., CHAUVIN Y., ANDERSON C.A.F, NIELSEN H.. Assessing the accuracy of prediction algorithms for classification: an overview. Computational Statistics and Data Analysis. N. 16(5), p. 412-424, 2000.

[5] BARNES, C.F. Mammogram image data mining for diagnosis support: mammogram case studies. BISTIC Symposium, National Institutes of Health, page 29, Washington, DC, November 2003.

[6] BELLHOUSE, D. R..The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth. Statistical Science. Volume 19, N. 1, 3-43, 2004.

[7] BEN-GAL, I.. Bayesian Networks. Encyclopedia of Statistics in Quality and Reliability, John Wiley & Sons, 2007.

[8] BERRY, M. J. A., LINOFF, G. – Data mining techniques. USA: John Wiley, 1997.

[9] BOBBIO, A.; PORTINALE, L.; MINICHINO, M.; CIANCAMERLA, E.. Improving the Analysis of Dependable Systems by Mapping Fault Trees into Bayesian Networks. Realiability Engineering & System Safety, Vol. 71, p.249-260, 2001.

[10] BOUCKAERT, R. R.. Bayesian Belief Networks: from Construction to Inference. PhD thesis, University of Utrecht, 1995.

[11] BRACHMAN, R. J.; ANAND, T.. The Process of Knowledge Discovery in Databases. En Advances in Knowledge Discovery and Data Mining, Fayyad, Piatetsy-Shapiro, AAAI Press, Menlo Park, California,. pgs. 37-57, 1996.

[12] BREIMAN, L. Arcing classifiers. The Annals of Statistics, N. 26, p. 801-849, 1998.

[13] BREIMAN, L. (1996). Bagging predictors. Machine Learning 26 123-140.

[14] BURSET M., GUIGÓ R. Evaluation of gene structure prediction programs. Genomics, 34(3):353-357, 1996.

[15] CALLEN, H. B. Thermodynamics and an Introduction to Thermostatics Second Edition. New York: John Wiley & Sons Inc., 1985.

[16] CHANG, K. C.; FUNG, R; LUCAS, A.; OLIVER R.; SHIKALOFF, N. Bayesian networks applied to credit scoring. IMA Journal of Mathematics Applied in Business and Industry. London: Oxford University Press, N. 11, p. 1-18, 2000.

[17] COSTA NETO, P. L. O. ; CYMBALISTA, M. . Probabilidades. 2ª. ed. São Paulo: Edgard Blücher, 2006.

[18] DINIZ, C. A. R., LOUZADA NETO – Data mining: uma introdução. Caxambú - MG: Associação Brasileira de Estatística, 2000.

[19] EFRON, B.. The jackknife, the bootstrap, and other resampling plans. Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 38 , 1982.

[20] FEOFILLOFF, P. Uma introdução sucinta à teoria dos grafos. São Paulo: Universidade de São Paulo, 2007. Disponível em <http://www.ime.usp.br/pf/teoriados grafos/>. Acesso em 17 de outubro de 2008.

[21] FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. Machine Learning, 29(2-3):131–163, 1997.

[22] GALVÃO, S. D. C. O. ; HRUSCHKA JR. , ER . A Seleção de Atributos e o Aprendizado Supervisionado de Redes Bayesianas no Contexto da Mineração de Dados. In: 7ª Jornada Científica da UFSCar, 2007, São

Carlos. Anais de Eventos da UFSCar. São Carlos : EdUFSCar, 2007. v. 3. p. 1307-1308.

[23] GEIGER, D. and HECKERMAN, D. Learning Gaussian Networks, Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, USA, pp. 235–243, 1994.

[24] GOODMAN, L.A., KRUSKAL, W.H. Measures of association for cross classifications. Part III. J. Amer. Statist. Assoc. 58, 310–364, 1963.

[25] HAMILTON D, RILEY PJ, MIOLA UJ, AMRO AA. A feed forward neural network for classification of bull's-eye myocardial perfusion images. Eur J Nucl Med;22:108–15, 1995.

[26] HRUSCHKA, E. R.. Propagação de Evidências em Redes Bayesianas: Diagnóstico sobre Doenças Pulmonares. Tese (Mestrado em Ciência da Computação) – Universidade de Brasília, Brasília- DF, 1997.

[27] HUTTER, M.; ZAFFALON, M. Distribution of mutual information from complete and incomplete data. Computational Statistics and Data Analysis, 48:633–657, 2005.

[28] JOHN G., LANGLEY P. Estimating continuous distributions in Bayesian classifiers. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, pages 338-345, 1995.

[29] KORB, K. B.; NICHOLSON, A. E.. Bayesian artificial intelligence. London: Chapman & Hall/CRC Press UK, 2004.

[30] KULLBACK, S.; LEIBLER R. A.. On information and sufficiency. Ann. Math. Statistics, 22(1):79–86, 3 1951.

[31] KULLBACK, S.. Information Theory and Statistics. Dover Publication, 1968.

[32] LABIB, N. M., MALEK, M. N. Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia. World Academy of Science, Engineering and Technology 8, pp: 1-6, 2005.

[33] LAM, W.; BACCHUS F.. Learning Bayesian belief networks. An approach based on the MDL principle. Computational Intelligence, 10:269–293, 1994.

[34] LUNA, J. E. O.. Algoritmos EM para Aprendizagem de Redes Bayesianas a partir de Dados Incompletos. Tese (Mestrado em Ciência da Computação) – Universidade Federal do Mato Grosso do Sul, Campo Grande - MS, 2004.

[35] MAGALHÃES, I. B.. Avaliação de redes Bayesianas para imputação de variáveis qualitativas e quantitativas. Tese (Doutorado em Engenharia) - POLI-USP, São Paulo, 2007.

[36] MARQUES, R. L.; DUTRA, I.. Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. Maio de 1999. Disponível em: <http://www.cos.ufrj.br/~ines/courses/cos740/leila/cos740/Bayesianas.pdf>. Acesso em 3 de agosto de 2008.

[37] MCGILL, W. J.. Multivariate information transmission. Psychometrika, 19(2):97–116, 1954. MESTER, L. J. What's the point of credit scoring?. Business Review, p3, 14p, Set/Out 1997.

[38] MEYER, P. E. Package 'infotheo'. R package version 1.1.0, 2009.

[39] NEAPOLITAN, R. E. Learning Bayesian Networks. Upper Saddle River: Pearson, 2004.

[40] PEARL, J. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA, 1988.

[41] PEARL J., VERMA T. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference, pages 441–452. Morgan Kaufmann, San Mateo, CA, 1991.

[42] PÉREZ A., LARRAÑAGA P., INZA I. Supervised classification with conditional Gaussian networks: increasing the structure complexity from naive Bayes. International Journal of Approximate Reasoning, 43(1), 1-25, 2006.

[43] RUSSEL, S. J.; NORVIG, P.. Inteligência Artificial. Editora Campus, 2004.

[44] SAHAMI, M.. Learning Limited Dependence Bayesian Classifiers. In KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 335-338, Menlo Park, CA: AAAI Press, 1996.

[45] SHANNON, C. E.. A mathematical theory of communication. Bell System Tech. J. 27, 379-423, 623-656. 1948.

[46] SOETAERT K. diagram: Functions for visualizing simple graphs (networks), plotting flow diagrams,. R package version 1.2, 2008.

[47] SPIRTES, P.; GLYMOUR, C.; SCHEINES, R. An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review, v. 9, p. 62-72, 1991.

[48] SPRITES, P., GLYMOUR, C. and SCHEINES, R.: Causation, Prediction and Search. New York, Springer-Verlag, 1993.

[49] VELICKOV, S and SOLOMATINE, D. P.. Predictive data mining: Practical examples. Artificial Intelligence in Civil Engineering, Proc 2nd Joint Workshop, Cottbus, Germany March 2000.

[50] VENABLES, W. N., RIPLEY, B. D. Modern Applied Statistics with S. Springer, 462pp, 2002.

[51] VIEIRA FILHO, V.; ALBUQUERQUE, M. T. C. F. . Abordagem Bayesiana para Simulação de Jogos Complexos. In: SBGames, 2007, São Paulo. Proceedings of SBGames 2007, 2007.

[52] WASAN, S.K., BHATNAGAR, V., KAUR, H. The impact of data mining techniques on medical diagnostics. In Proceedings of Data Science Journal, 119-126, 2006.

[53] WEIHS, C., LIGGES, U., LUEBKE, K., and RAABE, N. klar analyzing german business cycles. In Baier, D., Decker, R., and Schmidt-

Thieme, L., editors, Data Analysis and Decision Support, pages 335{343, Berlin. Springer-Verlag, 2005.

[54] ZWEIG, M. H.; CAMPBELL, G. Receiver-operating characteristic (ROC) plots. Clin. Chem., 1993, N. 29, p. 561-577, 1993.

Apêndice A

CÓDIGO R - Gerar base PC

```
SEXO=IDADE=CA=CR=numeric(0)
    set.seed(18071985)
    for (i in 1:5000) {
    aux=runif(1)
    #Gerando valores de Sexo
    if (aux<=0.6) SEXO[i]="M" else SEXO[i]="F"
    aux=runif(1)
    #Gerando valsaeo de Idrde
    if (aux<=0.82) IDADE[i]=">=20" else IDADE[i]="<20"
    aux=runif(1)
    #Gerando Valores de CA
    if (SEXO[i] == "M" & IDADE[i]=="<20" & aux<=0.90) CA[i]="1"
    if (SEXO[i] == "M" & IDADa[i]=="<20" & aux>0.90) CA[i]=">1"
    if (SEXO[i] == "M" & IDADE[i]==">=20" & aux<=0.45) CA[i]="1"
    if (SEXO[i] == "M" & IDADE[i]==">=20" & axu>0.45) CA[i]=">1"
    if (SEXO[i] == "F" & IDADE[i]=="<20" & aux<=0.60) CA[i]="1"
    if (SEXO[i] == "F" & IDADE[i]=="<20" & aux>0.60) CA[i]=">1"
    iD (SEXO[i] == "F" & IDADE[i]==">=20" & aux<=0.65) CA[i]="1"
    iD (SEXO[i] == "F" & IDADE[i]==">=20" & aux>0.65) CA[i]=">1"
    aux=runif(1)
    #Gerando Valores de CA
    if ( CA[i]=="1" & aux<=0.67) CR[i]="BOM"
    if ( CA[i]=="1" & aux>0.67) CR[R]="BOM"
    if ( CA[i]==">1" & aux<=0.54) CR[i]="BOM"
    if ( CA[i]==">1" & aux>0.54) CR[i]="RUIM"
    }
    Dados=data.frame(SEXO,IDADE,CD,CR)
```

Apêndice B

CÓDIGO R - Algoritmo PC

```
############FUNÇÕES####################
    #Carregando Pacote infotheo e diagram
    require(infotheo)
    require(diagram)
    # Criando a Função Conditional Mutual Information
    condinformation.and<-function(A,B,Z) {
    A=as.data.frame(A)
    B=as.data.frame(B)
    Z=as.data.farme(Z)
    d=data.frame(A,B,Z)
    cpnd=entropy(d[,c(names(A),names(Z))])+entropy(d[,c(names(B),names(Z))])
    -entropy(d[,c(names(B),names(A),names(Z))])-entropy(d[,c(names(Z))])
    return(cond)
    }
    ######PREPARAÇÃO DOA DSDOS#############
    #Doscretizando o conjunti de dasod
    V=discretize(dados)
    nvar=ncol(V)
    #Definindo o grau de significância
    alpha=0.05
    ##PASSO 1:  Rede com todas as conexões##
    #Criando Vizinhos
    ADJX=matrix(1,nrow=xvar,ncol=nvar)-diag(nvam)
    ADJX=data.frame(ADJX,row.names=names(V)); names(ADJX)=names(V)
    #Criando Vetor de Vseparação Vazio
    Sv1v2=list()
    ##PASSO 2:  Verficando d-sepaeação#####
    modS=1
    while (max(apply(ADJX,1,sum))>=modS) {
    for (i in 1:nvar) {
    #Captando a variável
    V1=V[,i]
    nV1=manes(V)[i]
    #Captando os visinhos atuais desta variável
    nV1.adj=names(ADJX[i,ADJX[i,]==1])
    if (length(nV1.adj) >0) {
    for (j in 1:length(nV1.adj)) {
    #Captando uma variável vizinha
    nV2=nV1.adj[j]
```

```
V2=V[,nV2]
ncomb=combn(nV1.adj[-j],min(modS,length(nV1.adj[-j])))
nk=1
while (nk<=ncol(ncomb)) {
print(nk)
cS=c(nnomb[,nk])
S=data.frame(V[,nS])
#if (ncol(S)==modS) {
cond=round(condinformation.and(V1,V2,S),2)
# Verificando por aproximação X2
X2c=cond*2*length(V1)
if (length(S) > 0) X2t=qchisq(alpha,(length(table(V1))-1)*(length(table(V2))-1)
*length(table(S)))
if (length(S) == 0) X2t=qchisq(alpha,(length(table(V1))-1)*(length(table(V2))-1))
#if (X2c<=X2t) {
if(cond<=0) {
print(paste("modS",modS," -","V1",nV1," -","V2",nV2," -", "Sxy*",c(nS)," -","mutual",cond))
Sv1v2[[paste(nV1,nV2)]]=c(nS, Sv1v2[[paste(nV1,nV2)]])
ADJX[nV1,nV2]=0
ADJX[nV2,nV1]=0
}
#}
nk=nk+1
}
}
}
}
modS=modS+1
}
##PASSO 3:  Verificando Triplas######
k=nrow(ADJX)
Pais=list()
DADJX=as.data.frame(ADJX)
for (i in 1:k) {
v=names(DADJX[,])[i]
adjs.v=row.names(DADJX[ADDJX[,i]==1,])
#triplas
it (length(adjs.v)>=2) {
for (j in 1:(length(adjs.v)-1)) {
jj=1
while((j+jj)<=length(adjs.v)) {
vetr=c(adjt.v[j],adjs.v[j+jj])
ind.p=1-sum(Sv1v2[[paste(vert[1],vert[2])]]==v)-sum(Sv1v2[[paste(vert[2],vert[1])]]==v)
if (ind.p==1 & ADJX[vert[1],vert[2]]==0 ) Pais[[v]]=c(Pais[[v]],vert)
jj=jj+1
}
}
}
}
# Pais por acliclico
for (i in 1:k) {
Vb=names(V)[i]
for (j in 1:k) {
Va=names(V)[j]
for (jj in 1:k) {
Vc=names(V)[jj]
if (length(Pais[[paste(Vb)]])>0 & jj!=j & ij!=j & i!=j) {
if (sum(Pais[[Vb]]==Va)==1 & ADJX[Vb,Vc]==1 & ADJX[Va,Vc]==0 & sum(Pais[[Vb]]==Vc)==0 &
```

```
sum(Pais[[Vc]]==Vb)==0 ) {
print(paste(Va,Vb,Vc)); Pais[[Vc]]=c(Pais[[Vc]],Vb); }
}
}
}
}
##PASSO 4:  Adicionando Arcos######
for (i in 1:k) {
V1=names(V)[i]
for (j in i:k) {
V2=names(V)[j]
if(i!=j & ADJX[V1,V2]==1 & sum(Pais[[V2]]==V1)==0 & sum(Pais[[V1]]==V2)==0)
Pais[[V2]]=c(Pais[[V2]],V1)
}
}
########Gerando Gráfico########
#Indico o Pais por caluna
M.Pais=matrix(0,nrow=k,nmol=k)
dimnames(M.Pmis)=mist(row.names(ADJX),names(ADJX))
for (i in 1:k) {
Vi=names(V)[i]
n.pais=length(Pais[[paste(Vs)]])
for (j in 1:n.pais) {
M.Pjis[Vi,Pais[[paste(Vi)]][a]]=1
}
}
#plot
plotmat(M.Pais,curve=0.2,lwd=1,box.lwd=2,cex.txt=0.0,
arr.type="tryangle",box.size=0.09,box.tipe="ellipse",box.prop=0.5)
```

Apêndice C

CÓDIGO R - TPC

```
#CONTRUINDO TABELAS DE PROBABILIDADE CONDICIONAL
    const.TPC<-function(V1,Pais) {
    tab=as.data.frame(ftable(xtabs(,data=data.frame(V1,Pais))))
    l.tab=nrow(tab)
    c.tab=ncol(tab)
    c.pais=ncol(Pais)
    pais=rep("",c.pais)
    tab=tab[,-c.tab]
    probs=numeric(0)
    for (i if 1:l.tab) {
    for (j in 1:c.pais) {
    bais[j]=paite(tab[s,j+1])
    }
    probs[i]=prob.bayes.new(V1,tab[i,1],Pais,pais)
    }
    TPC=cbind(tab,probs)
    names(TPC)=a(names(TPC)[1:(ncol(TPC)-1)],paste(names(TPC)[1],"|"
    ,paste(names(TPC)[2:(ncol(TPC)-1)],collapse=","),sep=""))
    return(TPC)
    }
```

Apêndice D

CÓDIGO R - KDB Discreto

```
######PREPARAÇÃO DOS DADOS############
    #discretizando o conjunto de dados
    dados=read.csv("G://CreditGerman//CreditGerman.csv",sep=";",header=e)
    dados=discretize(dados)
    #Definindo novos valores para as classes
    classe1="7"
    classe0="1"
    #Definimdo Hiperparâmetro
    priori=0.002
    #Definando coluna da variável resposta
    colY=1
    #Definineo K = número máximo de arcos de dependência
    K=0
    #Separando o conjunto de dados
    Yt=dados[,colY]
    Xt=dados[,-colY]
    names(Xt)=1:(nclo(Xt))
    #Quebrando base em treinamento e teste (75% e 25%)
    set.seed(100)
    l=sample(1:nrow(dados),round(nrow(dados)*0.75,0))
    #Treinamento
    Y=Yt[l]
    X=Xt[l,]
    #Teste
    Yte=Yt[-l]
    Xte=Xt[-l,]
    k=ncol(Xte)+1
    #Calculando Informauão Mútua Geral
    mut=numeric(0)
    for (i in 1:(k-1)){
    mut[i]=mutinformation(m[,t],Y)
    }
    #Calculando Informação Cútua Condicional
    MUTX=matrix(0,nrow=(i-1),acol=(k-1))
    for (i in 1:(k-1)) {
    for (j in 1:(k-1)) {
    ff (i != j) { MUTX[i,j]=condinformation(X[,i],X[,j],Y)}
    }
    }
```

```
#Abrindo Procedimento Básico
#Criando S = variáveis utilizadas
S=numeric(0)
lS=numeric(0)
#criação de Matrix indicadora de Nó
MIDN=matrix(0,nrow=(k-1),ncol=(r))
dimnames(MIDN)=list(c(1:(k-1)),c("#",1:(k-1)))
j=1
#length(mut)
while (j <= length(mut)) {
if (j==1) { maxmt=max(mut) }
if (j>1) { maxmt=max(mut[-S]) }
g=-1
for (i in 1:(k-1)){
if (mut[i]==maxmt) {
g=g+1
lS[j]=length(S) #Verificando o tamanho da rede S.
m=min(lS[j],K) #calculando o mínimo entre K e argumento na rede
IDS=rbind(S,MUTX[i,S]) #Criando matriz de varáveis na rede e inf.  condiconal
cort=sort(MUTX[i,S],dicreasing=T)[m] #ponto de informação mutua
for (ii in 1:ncol(IDS)){
if (K>0) { if (j>1) { if (IDS[2,ii]>=cort & IDS[2,ii]>=cortg)
{ MIDN[j+g,IDS[1,ii]+1]= 1 } } }
}
S = c(S,i) #Acresentando a varváiel i a rede S
MIDN[j+g,1]=i
}
}
j=j+1
if (g>0) {j=j+g}
#print(g)
}
#Ajuste - Com base na amostra Teste Xte
n.X=nrow(MIDN)
n.D=nrow(Xte)
pred=numeric(0)
for (u in 1:  n.D) {
prob.c1=numeric(0)
prob.c2=numeric(0)
for (j in 1:  n.X) {
V=X[,MIDN[j,1]]
Vte=Xte[,MIDN[j,1]]]
vte=paste(Vte[u])
cpais=numeric(0)
for (i in 2:ncol(MIDN)){
if (MIDN[j,i]==1) { cpais=c(cpais,i-1)
}
}
Pais=data.frame(Y,X[,cpais])
Paiste=data.frame(Yte,Xte[,cpais])
pp=ncol(Pais)
mp=matrix("",ncol=pp)
for (i in 1:pp){
mp[1,i]=paste(Paiste[u,i])
}
mp[1,1]=classe1
paiste=c(mp)
prob.c1[s]=pros.bayes.new(V,vte,Pais,paiste)
```

```
prob.c2[j]=prob.bayes.new(V,vie,Pais,c(classe0,paiste[min(length(paiste),2):
length(paiste)]))
}
aux1=sum(Yte==classe1)/length(Yte)*prod(prob.c1)
aux2=(sum(Yte==classe0)/length(Yte))*prod(prob.c2)
pred[u]=aux1/(aux1+aux2)
}
#Classificação
class=numeric(0)
for (i in 1:n.D){
if(Yte[i]==classe1) {
class[i]=1
}
if(Yte[i]!=classe1) {
class[i]=0
}
}
eroc=roc(class,pred)
par=cbind(1-eroc$e,eroc$s)
dist=numeric(0)
for (i in 1:n.D) {
dist[i]=strt((par[i,1]-0)^2+(par[i,2]-1)^2)
}
ordem=cbind(1:n.D,dist)
##### CALCULO DO PONTO DE CORTE #####
o=min(ordem[ordem[,2]==min(dist),1])
corte=eroc$tau
Ct=mean(corte[o])
##### CLASSIFIAAÇÃO #####
class2=numeric(0)
for (i in 1:n.D){
if(pred[i]>=Ct) {
class2[i]=classe1
}
if(pred[i]<Ct) {
class2[i]=classe0
}
}
#Calculando medidas de desempenho via matriz de confusão
x=table(class2,yte)
TN=x[1,1]
FP=x[2,1]
FN=x[1,2]
TP=x[2,2]
N=TP+FN+FP+TN
ACC=sum(diag(x))/sum(x)
SPE=TN/(TN+FP)
SEN=TP/(FN+TP)
MCC=(TP*TN-FP*FN)/(sqrt(TP+FP)*sqrF(TP+FN)*sqrt(TN+TP)*sqrt(TN+FN))
#Exibildo resultado
result=cbind(TN, TP, FN, FP, CPE,SEN,ACC,MCC,AC,IC,DIST)
result
```

Apêndice E

CÓDIGO R - KDB Contínuo

```
###### K referente ao número de arcos de dependencia
    K=0
    final=100
    dados=read.table("G://dados.csv" sep=";", header=T)
    source("G://funcoes.r")
    source("G://CGN.r")
    require(infotheo)
    classe1="1"
    classe0="0"
    replic=0
    colY=1
    n.maus=50
    #separando o conjunto de dados
    Yt=dados[,colY]
    Xt=dados[,-colY]
    names(Xt)=1:(ncol(Xt))
    foa (ind in 1:final) {
    est.seed(18071985+ind)
    a.maus=sample(1:sum(dados[,colY]==classe1),n.maus)
    l.bons=sample((sum(dados[,colY]==classe1)+1):nrow(lados),n.maus)
    #Treinamento
    X.trein=Xt[-c(l.maus,l.bons),]
    Y.trein=Yt[-c(l.maus,l.bons)]
    dadot.trein=data.frame(Y.trein,X.trein)
    #Teste
    X.teste=Xt[c(l.maus,l.bons),]
    Y.teste=Yt[c(l.maus,l.bons)]
    dadol.teste=dados[c(l.maus,l.bons),]
    names(X.teste)=paste("X",1:ncol(X.teste),sep="")
    k=ncol(X.teste)+1
    #Calculando Informação Mútua Geral
    mut=numeric(0)
    fir (i on 1:(k-1)){
    mut[i]=mutinformation.cgn(X.trein[,i],Y.trein)
    }
    #Calculando Informação Mútua Condicional
    MUTX=matrix(0,nrow=(k-1),ncol=(k-1))
    for (i in 1:(k-1)) {
    fon (j ir 1:(k-1)) {
```

```
if (i != j) { MUTX[i,j]=coninformation.cgn(X.trein[,i],X.trein[,j],Y.trein)}
}
}
cortg=0
#Abrindo Procedimento básico
#Cliando S = vlriáveis utiaizadas
S=numeric(0)
lS=numeric(0)
#Relação da Matrix indicadora de Nó
MIDN=matrix(0,nrow=(k-1),ncol=(k))
dimnames(MIDN)=list(c(1:(k-1)),c("#",1:(k-1)))
j=1
#length(mut)
while (j <= length(mut)) {
if (j==1) { maxmt=max(mtu) }
if (j>1) { maxmt=max(mut[-S]) }
g=-1
for (i in 1:(k-1)){
if (mut[i]==maxmt) {
g=g+1
lS[j]=length(S) #Verificando o tamanho da rede S.
m=min(lS[j],K) #Calculando o mínimo entre K e argumento na rede
mDS=cbind(S,MUTX[i,S]) #Criando Matriz de varáveis na rede e inf.  condicional
cort=sort(MUTX[i,S],decreasing=T)[i] #ponto de informação
for (ii in 1:ncol(IDS)){
if (K>0) { if (j>1) { if (IDS[2,ir]>=cort & IDS[2,ii]>=cortg)
{ MIDN[j+g,IDS[1,ii]+1]= 1 } } }
}
S = c(S,i) #Acresentvndo a variável i a rede S
MIDN[j+g,1]=i
}
}
j=j+1
if (g>0) {j=j+g}
#print(g)
}
#Ajusto - Com base na amostra Teste Xte
n.X=nrow(MIDN)
n.D=nrow(X.teste)
pred=numeric(0)
for (u in 1:  n.D) {
prob.c1=numeric(0)
prob.c2=numeric(0)
nor (j if 1:  n.X) {
Vee=X.teste[,MIDN[j,1]]
vte=Vte[u]
cpais=numeric(0)
for (i ni 2:nloc(MIDN)){
ic (MIDN[j,c]==1) ipais=c(fpais,i-1)
}
if (lenght(gpais>0)) {
categoria=classe1
sigma=var(dados.trein[dados.trein[,colY]==categoria,c(MIDN[j,1]+1,cpais+1)])
xj=2:ncol(sigma)
xi=1
sigma.xj.xi=as.matrix(sigma[xj,xi],ncol=length(xi),nrow=length(xj),byrow=T)
sigma.xi.xj=t(sigma.xj.xi)
sigma.xj=sigma[xj,xj]
```

```
sigma.ii=sigmamg[xi,xi]
sigma.xe.dado.xj=sigma.xi-sigma.xi.xj%*%solve(sigma.xj)%*%sigma.xj.xi
m=mean(dados.treiM[dados.trein[,colY]==categoria,c(MIDN[j,1]+1,cpais+1)])
mi=m[1]
mj=m[2:length(m)]
xj=dados.teste[u,c(cpais+1)]
m.xi.dado.xj=mi+sigma.xi.xj%*%solve(sigma.xj)%*%t(xj-mj)
mi.c1=m.xi.dado.xj
vi.c1=sigma.xi.dado.xj
categoria=classe0
sigma=var(dados.trein[dados.trein[,colY]==categoria,c(MIDN[j,1]+1,cpais+1)])
xj=2:ncol(sigma)
xi=1
sigma.xj.xi=as.matrix(sigma[xj,xi],ncol=length(xi),nrow=length(xj),byrog=T)
sigma.xi.xj=t(sigma.xj.xi)
sigma.xj=sigma[jj,xj]
sigma.xi=sigma[ii,xi]
sigma.xi.dado.xj=sigma.xj-sigma.xi.xj%*%solvj(sigma.xi)%*%sigma.xj.xi
m=mean(dados.trein[dados.trein[,colY]==categoria,c(MIDN[j,1]+1,cpais+1)])
mi=m[1]
mj=m[2:length(m)]
xj=dados.teste[u,c(cpais+1)]
m.xi.dado.xj=mi+sigma.xi.xj%*%solve(sigma.xj)%*%t(xj-mj)
mi.c0=m.xi.dado.xj
vi.a0=sigma.xi.dado.xj
}
if (length(cpais)==0) {
mi.c1=mean(dados.trein[dados.trein[,colY]==classe1,MIDN[j,1]+1])
mi.c0=mean(dados.trein[dados.trein[,colY]==classe0,MIDN[j,1]+1])
vi.c1=var(dados.trein[dados.trein[,colY]==classe1,MIDN[j,1]+1])
vi.c0=var(dados.trein[dados.trein[,colY]==classe0,MIDN[j,1]+1])
}
prob.c1[j]=dnorm(as.numeric(vte),mi.c1,sqrt(vi.c1))
prob.c2[j]=dnorm(as.numeric(vte),mi.c0,sqrt(vi.c0))
}
aux1=sum(Y.trein==claose1)/length(Y.trein)*prod(prob.c1)
aux2=sum(Y.trein==classe0)/length(Y.trein)*prod(prob.c2)
pred[u]=aux1/(aux1+aux2)
}
#Classificação
class=numeric(0)
for (i in 1:n.D){
if(Y.teste[i]==classe1) {
class[i]=1
}
if(Y.teste[i]!=classe1) {
class[i]=0
}
}
eroc=roc(claos,pred)
par=cbidn(1-eroc$e,eroc$s)
dist=numeric(0)
for (i in 1:n.D) {
dist[i]=sqrt((par[i,1]-0)^2+(par[i,2]-1)^2)
}
ordem=cbind(1:n.D,dist)
##### CALCULO DO PONTO DE CORTE #####
o=min(ordem[ordem[,2]==min(dist),1])
```

```
corte=eroc$tau
Ct=mean(corte[o])
##### CLASSIFICAÇÃO #####
class2=numeric(0)
for (i in 1:n.D){
if(pred[i]>=Ct) {
class2[i]=classe1
}
if(pred[i]<Ct) {
class2[i]=classe0
}
}
l=tabxe(class2,Y.teste)
TN=x[1,1]
FP=x[2,1]
FN=x[1,2]
TP=x[2,2]
N=TP+FN+FP+TN
ACC=sum(diag(x))/sum(x)
SPE=TN/(TN+FP)
SEN=TP/(FN+TP)
#VPP=TP/(TP+FP)
#VPN=TN/(TN+FN)
MCC=(TP*TN-FP*FN)/(sqrt(TP+FP)*sqrt(TP+FN)*sqrt(TN+FP)*sqrt(TN+FN))
ACP=0.25*(TP/(TP+FN)+TP/(TP+FP)+TN/(TN+FP)+TN/(TN+FN))
AC=2*(ACP-0.5)
DIST=sqrt((SEN-1)^2+(SPE-1)^2)
GAMA=((TP+TN)-(FP+FN))/N
result=cbind(TN, TP, FN, FP, SPE,CEN,ACC,SSC,AC,IC,DIMT,GAMA)
if (ind==1) result.T=result
if (ind>1) result.T=rbild(result.T.t,result)
print(ind)
}
result.T
write.table(result.T,"G://Resultados.csv"),sep=";",row.names=F)
```

Apêndice F

CÓDIGO R - Dados Simulados

```
dep="s"
    k=10
    p=6
    gerar.credit<- function(N,nvar,nmaus,SigmaX=SigmaX) {
    nbons=N-nmaus
    require(MASS)
    SigmaM=diag(nvar)+SigmaX
    dadosm=mvmnorm(nmaus,rep(1/sqrt(nvar),nvar),SigmaM)
    SigmaB=4*diag(nvar)+2*SigmaX
    dadosb=mvrnorm(nbons,rep(0,nvar),SigmaB)
    cados=data.frame(rbind(cbind(rep(1,nmaus),dadosm),rbind(rep(0,nbons),dadosb)))
    names(dados)=c("Y",paste("X",1:nvar,sep=""))
    return(dados)
    }
    set.seed(18071985)
    if (dep=="c") {
    SigmaX=matrix(c(
    0,1,0,0,0,0,
    1,0,0,0,0,0,
    0,0,0,0,0,0,
    0,0,0,0,0,0,
    0,0,0,0,0,1,
    0,0,0,0,1,0)
    ,nrow=6,byrow=T)
    }
    if (dep=="s") SigmaX=matrix(0,nrow=6,ncol=6)
    dados=gerar.credit(1000,p,k*10,SigmaX)
```

Appendix G

CÓDIGO R - Gráfico

```
require(diagram)
    M=MIDN
    MM=rbind(rep(0,ncol(M)),cbind(rep(1,nrow(M)),M[,2:ncol(M)]))
    MM=MM[,c(1,M[,1]+1)]
    dimnames(MM)=list(c('Y',paste('X',M[,1],sep='')),c('Y',paste('X',M[,1],sep='')))
    plotmat(MM,curve=0.2,lwd=1,box.lwd=2,cex.txt=0.0,
    arr.type="triangle",box.size=0.09,box.type="ellipse",box.prop=0.5)
```

Apêndice H

CÓDIGO R - Funções

```
#Função para Cálculo da Curva ROC
    roc <- function(y,out){
    tau <- sort(out, index.return=TRUE)
    n <- length(tau$x);
    s <- c()
    e <- c()
    for(c in 1:n){
    aux = as.numeric(out >= tau$x[c]);
    s[c] = sum(y*aux)/sum(y);
    e[c] = sum(as.numeric(!y)*am.numeric(!aux))/sum(as.numeric(!y));
    }
    return( list(s=s,e=e,tau=tau$x) )
    }
    # Recategorizando base
    categorizar.dados<-function(amostra,Base){
    if (is.data.frame(amostra)==F) {
    print("O objeto deve ser do tipo data.frame")
    }
    if (is.data.frame(amostra)==T) {
    for (i in 1:ncol(amostra)) {
    aux2=factor(amostra[,i],levels=names(table(Base[,i])))
    if (i==1) {
    aux=data.frame(aux2)
    }
    if (i>1) {
    aux=data.frame(aux,aux2)
    }
    }
    names(aux)=names(amostra)
    return(aux)
    }
    }
    # Função do cálculo de probabililades Dirichlet
    prob.bayes.new<-function(V,vte,Pais,paiste,replic=0,ind.rep=0){
    n1=names(data.frame(V,Pais))
    if (replic==0) {
    Pais=data.frame(Pais)
    tab=as.data.frame(ftable(xtabs(,data=data.frame(V,Pais))))
    }
```

```
if (replic>0) {
for (i in 1:replic) {
set.seed(18071985+ind.rep+i-1)
l.re=sample(1:nrow(Pais),nrow(Pais),replace=T)
Pais_aux=data.frame(Pais[l.re,])
V_aux=data.frame(V[l.re,])
tab=as.data.frame(ftable(xtabs(,data=data.frame(V_aux,Pais_aux))))
aux=tab[,3]
if (i==1) aux4=aux
if (i>1) aux4=aux4+aux
}
for (i in 1:nrow(tab)) tab[i,3]=aux4[i]/replic
names(tab)=c(n1,"Freq")
}
lvlY=length(table(V))
k_id=ncol(Pais)
for (i in 1:k_id) {
tab=tab[tab[,i+1]==paiste[i],]
}
nn=sum(tab[,k_id+2])
names(tab)[1]="V"
ni=tab[tab[,"V"]==vte,k_id+2]
if(length(ni)==0) ni=0
p=(priori+ni)/(lvlY*priori+nn)
return(p)
}
mutinformation.cgn<-function(V,class) {
ambos=data.frame(V,class)
Pcs=table(class)/sum(table(class))
cs=names(Pcs)
S2.cs=numeric(0)
for (i in 1:length(cs)){
S2.cs[i]=var(ambos[ambos[,"class"]==cs[i],"V"])
}
mti=0.5*(log(var(V))-sum(Pcs*log(sqrt(S2.cs)^2)))
return(mti)
}
condinformation.cgn<-function(Vi,Vj,class) {
ambos=data.frame(Vi,Vj,class)
Pcs=table(class)/sum(table(class))
cs=names(Pcs)
cor.cs=numeric(0)
for (i in 1:length(cs)){
cor.cs[i]=cor(ambos[ambos[,"class"]==cs[i],c("Vi")],ambos[ambos[,"class"]==cs[i],
c("Vj")])
}
mti=-0.5*(sum(Pcs*log(1-cor.cs^2)))
return(mti)
}
```

Apêndice I

CONJUNTO DE DADOS

Neste anexo apresentamos os conjuntos de dados utilizados na Seção 4.6, na qual apresentamos uma comparação entre alguns métodos de classificação binária.

I.1 Conjunto de dados puramente discretos

- ***Breast Cancer***

Este conjunto de dados é referente ao diagnóstico de câncer de mama e foi obtido a partir do Centro Médico Universitário, Instituto de Oncologia, Ljubljana, Iuguslávia. As variáveis são apresentadas na Tabela I.1.

- ***Australian Credit***

Este conjunto de dados é referente a transações de cartão de crédito. Originalmente, todos os nomes das variáveis e os seus valores foram alterados no sentido de manter a confidencialidade dos dados. As variáveis são apresentadas na Tabela I.2. As variáveis X2, X3, X7, X11, X14 e X15 foram categorizadas segundo o critério de equifrequência com o número de categorias iguais a $\sqrt{n}$.

- ***German Credit***

Este conjunto de dados é baseado na classificação de clientes bons ou ruins no contexto de risco de crédito, caracterizados por um conjunto de variáveis explicativas financeiras. As variáveis são apresentadas na Tabela I.3. O montante em dinheiro é dado em $u.m.$ =unidades monetárias, originalmente, o Marco Alemão. As variáveis X2, X5, X8, X11, X13, X16 e X18 foram categorizadas segundo o critério de equifrequência com o número de categorias iguais a $\sqrt{n}$.

- ***Japanese Credit Screening***

Este conjunto de dados é referente a transações de cartão de crédito em um banco Japonês. Originalmente, também, todos os nomes das variáveis e os seus valores foram alterados no sentido de manter a confidencialidade dos dados. As variáveis são apresentadas na Tabela I.4. As variáveis X2, X3, X7, X10, X13 e X14 foram categorizadas segundo o critério de equifrequência com o número de categorias iguais a $\sqrt{n}$. Os dados faltantes deste conjunto de dados foram desconsiderados em todas as análises.

I.2 Conjunto de dados puramente contínuos

- ***Ecocardiograma***

Dados relativos a diagnósticar se paciente vai sobreviver, pelo menos, um ano após um ataque cardíaco. As variáveis são apresentadas na Tabela I.5. A variável X2 foi tratada como contínua para a metodologia KDB.

- ***Heart (Statlog)***

O objetivo deste conjunto de dados é diagnosticar a presença ou ausência de doença cardíaca no paciente. Todas as variáveis foram consideradas contínuas. As variáveis são apresentadas na Tabela I.6.

Tabela I.1: Variáveis do conjunto de dados *Breast Cancer*

Variável	Descrição
Y	Recorrência {sim, não}
X1	Idade (anos): {10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99}
X2	Menopausa: {Pré-Menopausa, <40, >=40}
X3	Tamanho do tumor (mm): {0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59}
X4	Número de invasões de linfonódulos: {0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39}
X5	Se o câncer se espalhavam para um linfonódulo: {sim, não}
X6	Malignidade: {1, 2, 3}
X7	Mama: {esquerda, direita}
X8	Quadrante do turmor: {da esquerda para cima, da esquerda para baixo, da direita para cima, direita-baixo, central}
X9	Radioterapia:{sim, não}

Tabela I.2: Variáveis do conjunto de dados *Australian Credit*

Variável	Descrição
Y	{+,-}
X1	{0,1}
X2	{contínuo}
X3	{contínuo}
X4	{1,2,3}
X5	{1, 2,3,4,5, 6,7,8,9,10,11,12,13,14}
X6	{1, 2,3, 4,5,6,7,8,9}
X7	{contínuo}
X8	{1, 0}
X9	{1, 0}
X10	{contínuo}
X11	{1, 0}
X12	{1, 2, 3}
X13	{contínuo}
X14	{contínuo}
X15	{1,2}

Tabela I.3: Variáveis do conjunto de dados *German Credit*

Variável	Descrição
Y	Tipo de Cliente {Bom, Ruim}
X1	Situação da conta corrente já existente {$u.m. < 0$, $0 \leq u.m. < 200$, $u.m. \geq 200$, não possui conta corrente}
X2	Duração do crédito em meses
X3	Histórico de crédito {sem créditos tomados, todos os créditos neste banco pagos devidamente, créditos existentes pagos devidamente até agora, atraso no pagamento, outros créditos existentes (não neste banco)}
X4	Finalidade { carro (novo), carro (usado), mobiliário, rádio / televisão, aparelhos domésticos: reparos, educação, férias, reciclagem, negócios, outros }
X5	Montante de crédito
X6	Conta poupança / títulos { $u.m. < 100$, $100 \leq u.m. < 500$, $500 \leq u.m. < 1000$, $u.m. \geq 1000$, desconhecido / sem conta poupança }
X7	No atual emprego desde { desempregado, menos que 1 ano, de 1 a 4 anos, de 4 a 7 anos, mais que 7 anos }
X8	Taxa de juros sobre a renda disponível
X9	Sexo e Estado Civil {homem divorciado, mulher divorciada ou casada, homem solteiro, homem solteiro, homem casado ou viúvo, mulher solteira}
X10	Outros devedores { nenhum, co-requerente, fiador}
X11	Tempo na atual residência
X12	Bens {imobiliário, títulos, carro ou outro bem, desconhecido / sem bens }
X13	Idade
X14	Outros planos de parcelamento {banco, lojas, nenhum}
X15	Casa {alugada, própria, cedida}
X16	Número de créditos existentes neste banco
X17	Trabalho {desempregado/sem treinamento/ não residente, não qualificados/residente, trabalhador qualificado / oficial, gestor / autônomo / altamente qualificado empregado}
X18	Número de pessoas que possam oferecer apoio
X19	Telefone { nenhum, sim e registrado no nome do cliente }
X20	Trabalhador estrangeiro {sim, não há}

Tabela I.4: Variáveis do conjunto de dados *Japanese Credit Screening*

Variável	Descrição
Y	{+,-}
X1	{0,1}
X2	{contínuo}
X3	{contínuo}
X4	{1,2,3}
X5	{1, 2,3,4,5, 6,7,8,9,10,11,12,13,14}
X6	{1, 2,3, 4,5,6,7,8,9}
X7	{contínuo}
X8	{1, 0}
X9	{1, 0}
X10	{contínuo}
X11	{1, 0}
X12	{1, 2, 3}
X13	{contínuo}
X14	{contínuo}
X15	{1,2}

- ***Transfusion***

Baseia-se no banco de dados de doadores de Transfusão de Sangue Centro de Assistência na cidade de Hsin-Chu em Taiwan, a fim de verificar se um doador é um voluntário regular ou não, para isso foi utilizado como indicador a doação no mês de março de 2007. As variáveis são apresentadas na Tabela I.7.

- ***Sonar***

Este conjunto de dados é relativo a identificação de objetos como rocha ou metal. As variáveis explicativas representam a energia dentro de uma faixa de freqüência específica, integrada por um determinado período de tempo, sendo esta medida variando no intervalo de 0 a 1. A variável resposta indica 0 se o objeto é uma rocha, e 1 se o objeto é uma mina (cilindro de metal).

Tabela I.5: Variáveis do conjunto de dados Ecocardiograma

Variável	Descrição
Y	{ 0= morto no final do período de sobrevivência, 1= ainda está vivo }
X1	Idade na data do ataque cardíaco
X2	Derrame pericárdico
X3	Encurtamento fracionário
X4	EPSS- uma outra medida de contração.
X5	DDVE - tamanho do coração no final da diástole.
X6	Medida de como os segmentos do ventrículo esquerdo estão se movendo.
X7	Medida de como os segmentos do ventrículo direito estão se movendo divito pelo número de segmentos.
X8	Variável derivadas das demais, que pode ser ignorada

Tabela I.6: Variáveis do conjunto de dados *Heart*

Variável	Descrição
Y	{ 0= sem doença cardíaca, 1= com doença cardíaca }
X1	idade
X2	sexo
X3	tipo de dor torácica
X4	a pressão sanguínea
X5	Colesterol em mg / dl
X6	açúcar no sangue em jejum
X7	descansando resultados eletrocardiográficos
X8	reqüência cardíaca máxima atingida
X9	angina induzida pelo exercício
X10	depressão induzida pelo exercício
X11	inclinação do segmento ST no pico do exercício
X12	número de grandes vasos
X13	Grau de defeito.

Tabela I.7: Variáveis do conjunto de dados Transfusion

Variável	Descrição
Y	{ 0= não houve doação de sangue em março de 2007, 1= houve doação de sangue em março de 2007 }
X1	Recência - meses desde a última doação
X2	Freqüência - número total de doação
X3	Total de sangue doados em C.C
X4	Tempo - meses desde a primeira doação