# Métodos de categorização de variáveis preditoras em modelos de regressão para variáveis binárias

***Diego Mattozo Bernardes da Silva***

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

DIEGO MATTOZO BERNARDES DA SILVA

# MÉTODOS DE CATEGORIZAÇÃO DE VARIÁVEIS PREDITORAS EM MODELOS DE REGRESSÃO PARA VARIÁVEIS BINÁRIAS

Dissertação apresentada ao Departamento de Estatística – DEs-UFSCar e ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística.

Orientador: Prof. Dr. Gustavo Henrique de Araújo Pereira

**São Carlos**
**Julho de 2017**

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

DIEGO MATTOZO BERNARDES DA SILVA

# CATEGORIZATION METHODS FOR PREDICTOR VARIABLES IN BINARY REGRESSION MODELS

Master dissertation submitted to the Departamento de Estatística – DEs-UFSCar and to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master joint Graduate Program in Statistics.

Advisor: Prof. Dr. Gustavo Henrique de Araújo Pereira

**São Carlos**
**July 2017**

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

## Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Diego Mattozo Bernardes da Silva, realizada em 13/06/2017:

Prof. Dr. Gustavo Henrique de Araujo Pereira
UFSCar

Prof. Dr. Afrânio Márcio Corrêa Vieira
UFSCar

Prof. Dr. Rinaldo Artes
Insper

*À minha mãe Regina Lúcia Mattozo.*

# AGRADECIMENTOS

Ao professor Gustavo, que me ajudou bastante em todas as etapas do mestrado, desde o processo seletivo até a finalização deste trabalho.

Aos meus pais Carlos Alberto e Regina, que sempre me incentivaram a estudar.

À minha namorada Nathalia, que me apoiou na decisão de iniciar o mestrado, mesmo isso significando que iríamos morar em cidades distantes.

Aos meus amigos de mestrado Allan, Juliana, Taís, Eduardo, Fabiano e Nicholas pela companhia e ajuda nos estudos.

*“Para começar, pare de falar e comece a fazer.”*
*(Walt Disney)*

# RESUMO

DA SILVA, D. M. B. **Métodos de categorização de variáveis preditoras em modelos de regressão para variáveis binárias**. 2017. 82 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Modelos de regressão para variáveis resposta binárias são muito comuns em diversas áreas do conhecimento. O modelo mais utilizado nessas situações é o modelo de regressão logística, que assume que o logito da probabilidade de ocorrência de um dos valores da variável resposta é uma função linear das variáveis preditoras. Quando essa suposição não é razoável, algumas possíveis alternativas são: realizar transformação das variáveis preditoras e/ou inserir termos quadráticos ou cúbicos no modelo. O problema dessa abordagem é que ela dificulta bastante a interpretação dos parâmetros do modelo e, em algumas áreas, é fundamental que eles sejam interpretáveis. Assim, uma abordagem muitas vezes utilizada é a categorização das variáveis preditoras quantitativas do modelo. Sendo assim, este trabalho tem como objetivo propor duas novas classes de métodos de categorização de variáveis contínuas em modelos de regressão para variáveis resposta binárias. A primeira classe de métodos é univariada e busca maximizar a associação entre a variável resposta e a covariável categorizada utilizando medidas de associação para variáveis qualitativas. Já a classe de métodos multivariada tenta incorporar a estrutura de dependência entre as covariáveis do modelo através da categorização conjunta de todas as variáveis preditoras. Para avaliar o desempenho, aplicamos as classes de métodos propostas e quatro métodos de categorização existentes em 3 bases de dados relacionadas à área de risco de crédito e a dois cenários de dados simulados. Os resultados nas bases reais sugerem que a classe univariada proposta têm um desempenho superior aos métodos existentes quando comparamos o poder preditivo do modelo de regressão logística. Já os resultados nas bases de dados simuladas sugerem que ambas as classes propostas possuem um desempenho superior aos métodos existentes. Em relação ao desempenho computacional, o método multivariado mostrou-se inferior e o univariado é superior aos métodos existentes.

**Palavras-chave:** Regressão, Risco de Crédito, Categorização de Variáveis Preditoras.

# ABSTRACT

DA SILVA, D. M. B. **Categorization methods for predictor variables in binary regression models**. 2017. 82 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Regression models for binary response variables are very common in several areas of knowledge. The most used model in these situations is the logistic regression model, which assumes that the logit of the probability of a certain event is a linear function of the predictors variables. When this assumption is not reasonable, it is common to make some changes in the model, such as: transformation of predictor variables and/or add quadratic or cubic terms to the model. The problem with this approach is that it hinders parameter interpretation, and in some areas it is fundamental to interpret the parameters. Thus, a common approach is to categorize the quantitative covariates. This work aims to propose two new classes of categorization methods for continuous variables in binary regression models. The first class of methods is univariate and seeks to maximize the association between the response variable and the categorized covariate using measures of association for qualitative variables. The second class of methods is multivariate and incorporates the predictor variables correlation structure through the joint categorization of all covariates. To evaluate the performance, we applied the proposed methods and four existing categorization methods in 3 credit scoring databases and in two simulated cenarios. The results in the real databases suggest that the proposed univariate class of categorization methods performs better than the existing methods when we compare the predictive power of the logistic regression model. The results in the simulated databases suggest that both proposed classes perform better than the existing methods. Regarding computational performance, the multivariate method is inferior and the univariate method is superior to the existing methods.

**Keywords:** Regression, Credit Scoring, Categorization of Predictor Variables.

# LISTA DE ILUSTRAÇÕES

# LISTA DE ALGORITMOS

# LISTA DE TABELAS

# SUMÁRIO

CAPÍTULO

1

# INTRODUÇÃO

Modelos de regressão para variáveis resposta binárias são muito comuns em diversas áreas do conhecimento. Nesses casos, em geral, o interesse é modelar a probabilidade de ocorrência de um dos valores da variável resposta em função de variáveis preditoras. O modelo mais utilizado nessas situações é o modelo de regressão logística, que assume que o logito da probabilidade de ocorrência de determinado evento é uma função linear das variáveis preditoras. Quando essa suposição não é razoável, realizam-se algumas mudanças no modelo, como: transformação das variáveis preditoras e/ou inserir termos quadráticos ou cúbicos no modelo.

Em áreas como na análise de risco de crédito, onde esses modelos são utilizados como ferramental técnico que fornece informações aos gestores para tomada de decisão, essa abordagem se torna problemática, pois dificulta bastante a interpretação dos parâmetros do modelo que, nesse caso, é de vital importância (THOMAS; EDELMAN; CROOK, 2002). Assim, nesse contexto, uma abordagem muitas vezes utilizada é a de categorizar as variáveis preditoras quantitativas do modelo, ou seja, transformar uma variável quantitativa em uma variável qualitativa ordinal com $k$ níveis. Dessa forma, para cada variável originalmente quantitativa são incluídas $k-1$ variáveis indicadoras no modelo de regressão logística. Esse processo também pode ser denominado discretização. Usaremos, neste trabalho, os termos categorização e discretização como sendo equivalentes.

Existem diversos métodos de categorização de variáveis contínuas. Porém, grande parte dos mesmos são derivados da área de aprendizado de máquina, como Kerber (1992), Liu e Setiono (1995), Tay e Shen (2002), Gonzalez-Abril *et al.* (2009), Tsai, Lee e Yang (2008) e Kurgan e Cios (2004), implementados no R pelo pacote *discretization* (KIM, 2012). Muitos desses métodos de categorização usam suposições que, em geral, não são atendidas. Por exemplo, em vários casos, esses métodos utilizam vários testes homogeneidade, sendo que muitos deles envolvem tamanhos de amostras pequenas. Isso não é razoável, pois a estatística desses testes tem distribuição qui-quadrado, sob a hipótese nula, apenas assintoticamente. Sendo assim, é interessante o estudo de métodos de categorização mais razoáveis do ponto de vista teórico.

Este trabalho tem como principal objetivo introduzir duas novas classes de métodos de categorização para variáveis contínuas, para modelos de regressão para variáveis resposta binárias. A primeira classe de métodos é univariada e tem como base medidas de associação entre variáveis qualitativas, com intuito de maximizar a associação entre a variável resposta e a covariável categorizada. Utilizaremos neste trabalho duas medidas diferentes, a saber, coeficiente de correlação *Kendall's Tau-C* (SOMERS, 1962) e a estatística denominada *Information Statistics* (THOMAS; EDELMAN; CROOK, 2002). O método denomina-se univariado pois categoriza uma variável preditora por vez. O método multivariado será baseado na regressão logística e no coeficiente de gini. Denomina-se multivariado pois busca categorizar todas as covariáveis de maneira conjunta, de forma a considerar a estrutura de associação das mesmas.

Os capítulos restantes deste trabalho estão estruturados da forma a seguir. No Capítulo 2, inicialmente descrevemos os Modelos Lineares Generalizados, principalmente no que tange os modelos para dados binários e a regressão logística (Seção 2.1). Em seguida, abordamos o que é *credit scoring*, a importância da regressão logística e de métodos de categorização nessa área (Seção 2.2). No Capítulo 3, descrevemos quatro dos principais métodos de categorização presentes na literatura (Seção 3.1). O primeiro é denominado CAIM (Seção 3.1.1), o segundo é o Ameva (Seção 3.1.2), o terceiro é chamado CACC (Seção 3.1.3), e por último temos o MDL (Seção 3.1.4). Depois apresentamos o método univariado proposto (Seção 3.2) e, em seguida, o método multivariado (Seção 3.3). No Capítulo 4, fazemos uma comparação dos métodos existentes e métodos propostos aplicando os mesmos em três bases de dados fornecidas por uma instituição financeira e através da criação de dois cenários de dados simulados. Por fim, as conclusões e próximos estudos são apresentados no Capítulo 5.

CAPÍTULO

2

# MODELOS DE REGRESSÃO PARA VARIÁVEIS BINÁRIAS

## 2.1 Modelos de Regressão para Variáveis Binárias

Neste capítulo discutimos os modelos de regressão para variáveis binárias a partir da classe de modelos de regressão definida por Nelder e Wedderburn (1972), denominada de Modelos Lineares Generalizados (MLG), que são uma extensão dos modelos lineares clássicos. Essa classe de modelos abrange também alguns modelos lineares utilizados quando se tem variáveis resposta categóricas.

Podemos definir um modelo linear generalizado a partir da especificação de três componentes necessários. O componente aleatório em que determinamos a variável resposta e sua distribuição de probabilidade; um componente sistemático que define quais são as covariáveis que vão entrar no modelo na forma de uma estrutura linear e um componente de ligação que é uma função que tem como característica fazer a ligação entre os componentes sistemático e aleatório.

O Componente aleatório do MLG consiste de um vetor $\mathbf{Y}$ de $n$ observações independentes, $(y_1,...,y_n)$, pertencentes à família exponencial linear de distribuições. Essa família é caracterizada por uma função de probabilidade ou densidade especificada na seguinte forma (PAULA, 2004):

$$f(y_i;\theta_i,\phi) = \exp[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i,\phi)]. \tag{2.1}$$

Algumas distribuições de probabilidade discretas são casos especiais dessa família, como a Poisson e Binomial. Na equação (2.1), $\theta$ e $\phi$ são parâmetros da distribuição e $c$ e $b$ são funções conhecidas.

Temos que $E(Y_i) = \mu_i = b'(\theta_i)$ e $Var(Y_i) = \phi^{-1}V(\mu_i)$, sendo que $V$ é a função de variância e é definida como $V_i = V(\mu_i) = d\mu_i/d\theta_i$. A função de variância apresenta uma característica

importante, pois a partir dela pode-se identificar distribuições de probabilidade específicas. Por exemplo, a função de variância dada por $V(\mu) = \mu(1-\mu)$ especifica a família de distribuições binomial com probabilidade de sucesso $\mu$ (PAULA, 2004).

O componente sistemático do MLG é definido por um vetor $\eta$, $(\eta_1, ..., \eta_n)$, corresponde a uma combinação linear entre as variáveis preditoras. Assim, considerando $x_{ij}$, o valor da variável preditora $j$, $j = 1, ..., p$, para a observação $i$, temos

$$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, ...., n \tag{2.2}$$

Por fim, temos o componente que busca interligar os componentes aleatório e sistemático. Assim, para $\mu_i = E(Y_i)$, $i = 1, ..., n$, definimos a função $g$ que liga $\mu_i$ a $\eta_i$, em que $g$ deve ser uma função monótona e duplamente diferenciável. Desta forma, o modelo linear generalizado é definido por (2.1) e pela equação

$$g(\mu_i) = \eta_i. \tag{2.3}$$

Para utilizarmos o MLG no contexto de variáveis resposta binárias, devemos considerar uma variável aleatória $Y$ que assume valor 1 com probabilidade $\mu$ e o valor 0 com probabilidade $1-\mu$. Dizemos que essa variável aleatória possui distribuição de Bernoulli e sua função de probabilidade é dada por

$$f(y;\mu) = \mu^y (1-\mu)^{1-y}. \tag{2.4}$$

Se tomarmos a exponencial do logaritmo dessa função de probabilidade e reescrevê-la como na equação (2.5) mostramos que essa distribuição pertence à família de distribuições exponencial como definido na equação (2.1), ou seja,

$$f(y;\mu) = \exp\left\{ y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu) \right\}, \tag{2.5}$$

em que $\phi = 1$, $\theta = \log\{\frac{\mu}{1-\mu}\}$, $b(\theta) = \log(1+e^{\theta})$ e $c(y,\phi) = \log(1)$. Além disso, a função de variância $V(\mu)$ fica dada por $\mu(1-\mu)$.

Definida a distribuição de probabilidade que será utilizada no MLG, deve-se primeiro abordar as opções para função de ligação quando temos uma variável resposta binária. Assim, para obter a função de ligação em um modelo linear, poderíamos utilizar o caso mais simples que seria $g(\mu_i) = \mu_i = \sum_{j=1}^{n} \beta_j x_{ij}$. No entanto, essa função de ligação tem a desvantagem de ajustar valores para a variável resposta que podem ser menores que zero ou maiores que um. Como $\mu$ é uma probabilidade isso não é razoável. Então, nesse caso, devemos utilizar funções de ligação que garantam que os valores ajustados de $\mu$ fiquem restritos ao intervalo $[0,1]$.

De acordo com Dobson e Barnett (2008), para garantir essa restrição, geralmente, são usadas funções de ligação baseadas em alguma função de distribuição acumulada

$$\mu = \int_{-\infty}^{t} f(s)ds \tag{2.6}$$

em que $f(s) \geq 0$ e $\int_{-\infty}^{\infty} f(s)ds = 1$. Neste caso $f(s)$ é chamada de distribuição de tolerância.

Quando a função de distribuição normal padrão é utilizada como distribuição de tolerância denomina-se essa função de ligação como probito, em que denotamos $\Phi$ como a sua função de distribuição acumulada. Assim podemos definir a função de ligação probito como a inversa da função acumulada da Normal padrão

$$\Phi^{-1}(\mu_i) = \eta_i \tag{2.7}$$

Outra função de ligação utilizada é a chamada complemento log-log. Nesse caso temos como distribuição de tolerância a distribuição do valor extremo que é definida por

$$f(s) = \exp\{s - exp(s)\}, \tag{2.8}$$

em que $-\infty < s < \infty$. Desse modo, a função de distribuição acumulada é dada por

$$F(s) = 1 - \exp\{-\exp(s)\}, \tag{2.9}$$

e a função de ligação complemento log-log em modelos MLG para variável resposta binária é dada por

$$\log\{-\log(1-\mu_i)\} = \eta_i \tag{2.10}$$

Em alternativa as duas funções citadas acima, também existe a função de ligação logística que tem como distribuição de tolerância a distribuição logística. Sua função de distribuição acumulada é dada por

$$F(s) = \frac{1}{1+e^{-s}} = \frac{e^s}{1+e^s}, \tag{2.11}$$

em que $s \in \mathbb{R}$. Assim, a função de ligação logito é definida como

$$\log\left\{\frac{\mu_i}{1-\mu_i}\right\} = \eta_i \tag{2.12}$$

A utilização da ligação canônica, isto é, quando a função de ligação é igual ao parâmetro canônico $\theta_i$ como ocorre no caso para a função de ligação logito no modelo com variável resposta Bernoulli, garante umas série de vantagens, como por exemplo a concavidade do logaritmo da

função de verossimilhança, o que possibilita a obtenção de certos resultados assintóticos com maior facilidade (PAULA, 2004).

Outra vantagem importante do uso da função de ligação logito é a facilidade na interpretação dos parâmetros. A partir do uso da razão de chances podemos interpretar diretamente o efeito de cada parâmetro no modelo de regressão, o que não acontece com as outras funções de ligações citadas.

Por exemplo, em um modelo com $p$ covariáveis, ou seja, $\log \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 x_{i1} + ,...,+\beta_p x_{ip}$, para $i = 1,...,N$. Assim fixando o valor das $p-1$ covariáveis $(x_{i2},...,x_{ip})$, podemos definir a chance de ocorrer dado evento se o valor de uma covariável binária $x_1$ é igual a zero como

$$\frac{\mu_i}{1-\mu_i} = e^{\beta_0+\beta_2 x_{i2}+,...,+\beta_p x_{ip}}, \tag{2.13}$$

e se $x_1 = 1$

$$\frac{\mu_i}{1-\mu_i} = e^{\beta_0+\beta_1+\beta_2 x_{i2}+,...,+\beta_p x_{ip}}. \tag{2.14}$$

A partir disso, pode-se definir que a razão de chances de ocorrência (OR) de certo evento em relação à presença ou não de certa característica medida por $x_1$ como

$$OR = \frac{e^{\beta_0+\beta_1+\beta_2 x_{i2}+,...,+\beta_p x_{ip}}}{e^{\beta_0+\beta_2 x_{i2}+,...,+\beta_p x_{ip}}} = e^{\beta_1} \tag{2.15}$$

em que $e^{\beta_1}$ é o incremente relativo na chance de ocorrência de um evento para um indivíduo que possui certa característica $x_1$ em relação a um que não possui, mantidas as demais covariáveis constantes. A interpretação para variáveis preditoras contínuas funciona de modo análogo, apenas tratamos da chance de ocorrência dado um incremento em uma unidade nessa variável, mantendo as outras constantes.

Após essa introdução das características básicas do MLG para dados binários, vamos abordar como se dá a estimação dos parâmetros, intervalos de confiança e testes de hipóteses. Pelas vantagens discutidas, utilizaremos neste trabalho a função de ligação logito e por isso consideraremos apenas essa função de ligação nas demais discussões desta seção. O MLG com função de ligação logito é conhecido como regressão logística (JR; LEMESHOW, 2004).

Supondo um vetor aleatório independente $(Y_1, Y_2, ..., Y_n)$ em que $Y_i \sim Bernoulli(\mu_i)$, $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ o vetor de valores observados de $\boldsymbol{Y}$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ o vetor de parâmetros, denotamos a função de verossimilhança $L(\boldsymbol{\mu}; \boldsymbol{y})$, que é dada por

$$L(\boldsymbol{\mu}; \boldsymbol{y}) = \prod_{i=1}^{n} \mu_i^{y_i} (1-\mu_i)^{1-y_i} = \prod_{i=1}^{n} \exp\left\{ y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + \log(1-\mu_i) \right\}. \tag{2.16}$$

Como $\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i$ e $(1-\mu_i) = [1+\exp(\eta_i)]^{-1}$, fazemos a substituição na equação (2.16), para deixar a função de verossimilhança em função de $\boldsymbol{\beta}$

$$L(\boldsymbol{\beta};\boldsymbol{y}) = \prod_{i=1}^{n} \exp\left\{ y_i \sum_j \beta_j x_{ij} - \log(1+e^{\sum_j \beta_j x_{ij}}) \right\}. \tag{2.17}$$

Para obtermos a função score, que é a derivada do logaritmo função de verossimilhança em relação à cada um dos parâmetros do modelo logito, devemos tomar o logaritmo da equação (2.17) obtendo-se

$$l(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n} \left\{ y_i \left( \sum_j \beta_j x_{ij} \right) - \log(1+e^{\sum_j \beta_j x_{ij}}) \right\}. \tag{2.18}$$

A partir disso, podemos escrever o termo da função escore referente à $\beta_j$ como

$$U_{\beta_j}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ y_i x_{ij} - \frac{1}{(1+e^{\sum_j \beta_j x_{ij}})} e^{\sum_j \beta_j x_{ij}} \times x_{ij} \right\}. \tag{2.19}$$

Assim é fácil mostrar que o vetor escore é dado por

$$U_{\beta_j}(\boldsymbol{\theta}) = \boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{\mu}) \tag{2.20}$$

em que $\boldsymbol{X}$ é uma matriz $n \times p$ contendo os valores das variáveis preditoras, $\boldsymbol{y} = (y_1, ..., y_n)^T$ e $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$.

As estimativas de máxima verossimilhança são obtidas por meio da resolução do sistema de equações $U_\beta(\boldsymbol{\theta}) = 0$. No entanto, esse sistema de equações não tem solução algébrica. Portanto, as estimativas de máxima verossimilhança devem ser obtidas por métodos numéricos. Assim, será utilizado no presente trabalho o método numérico denominado *iteratively reweighted least squares* (GREEN, 1984) presente no pacote *glm* do programa estatístico R (R Development Core Team, 2008).

A partir da função escore, podemos obter a matriz de informação de Fisher, que é dada em sua forma matricial por

$$\boldsymbol{K}_{\beta\beta} = \boldsymbol{X^T V X}, \tag{2.21}$$

em que $\boldsymbol{X}$ é uma matriz $n \times p$, $\boldsymbol{y} = (y_1, ..., y_n)^T$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ e $\boldsymbol{V} = diag\{\mu_1(1-\mu_1), ..., \mu_n(1-\mu_n)\}$.

Com base na matriz de informação de Fisher e utilizando as propriedades do estimador de máxima verossimilhança, podemos, sob condições de regularidade, definir intervalos de confiança assintóticos para os parâmetros $\boldsymbol{\beta}$ do modelo. Sendo $\widehat{Var}(\hat{\beta}_r)$ o elemento de posição $(r,r)$ de $(\boldsymbol{X^T \hat{V} X})^{-1}$ e $z_{(1+\gamma)/2}$ o quantil de ordem $(1+\gamma)/2$ da distribuição normal padrão, então o intervalo de confiança para $\beta_r$ com $(\gamma \times 100)\%$ de confiança é dado por (CORDEIRO; DEMÉTRIO, 2008)

$$\hat{\beta}_r \pm z_{(1+\gamma)/2} \widehat{Var}(\hat{\beta}_r)^{1/2}.$$

Podemos realizar testes de hipótese simples para MLG, isto é, quando queremos testar se um conjunto de $m$ parâmetros do modelo são iguais a algum vetor $\boldsymbol{c}$, de forma trivial após a obtenção das expressões do logaritmo da função de verossimilhança, da função escore e da matriz de informação de Fisher. Deste modo, podemos expressar a estatística do teste de razão de verossimilhanças por

$$\xi_{RV} = 2\{l(\hat{\boldsymbol{\beta}};\boldsymbol{y}) - l(\boldsymbol{\beta^0};\boldsymbol{y})\}, \tag{2.22}$$

em que $l(\boldsymbol{\beta^0};\boldsymbol{y})$ e $l(\hat{\boldsymbol{\beta}};\boldsymbol{y})$ são, respectivamente, os logaritmos da função de verossimilhança ajustados sob a hipótese nula e sob as estimativas de máxima verossimilhança.

A estatística do teste de Wald ficar dado por

$$\xi_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^0})^T \widehat{Var}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^0}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^0})^T (\boldsymbol{X^T}\widehat{\boldsymbol{V}}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^0}), \tag{2.23}$$

em que $(\boldsymbol{X^T}\widehat{\boldsymbol{V}}\boldsymbol{X})$ é a matriz de informação de Fisher aplicada em $\hat{\boldsymbol{\beta}}$.

Temos também o teste de escore, que utiliza a função score encontrada na equação (2.19), que pode ser definido para o modelo de regressão logístico como

$$\xi_{SR} = U_\beta(\boldsymbol{\beta^0})^T (\boldsymbol{X^T}\widehat{\boldsymbol{V}}_{\boldsymbol{0}}\boldsymbol{X})^{-1} U_\beta(\boldsymbol{\beta^0}), \tag{2.24}$$

em que a informação de Fisher $(\boldsymbol{X^T}\widehat{\boldsymbol{V}}\boldsymbol{X})$ e a função score são obtidas sob a hipótese nula.

Observamos que assintoticamente e sob a hipótese nula, os três testes expressados pelas equações (2.22, 2.23 e 2.24), tem distribuição qui-quadrado com $m$ graus de liberdade.

## 2.2 Aplicação em Risco de Crédito

O sistema bancário e financeiro de um país pode ser considerado um componente crítico de sua economia. Em alguns casos podemos até dizer que são componentes importantes da economia mundial. Bancos funcionam como intermediários pegando depósitos de clientes, consumidores e empresários de modo a fornecer crédito para uma variedade de clientes e setores da economia. Nesse contexto, a análise de crédito se torna importante como técnica que busca mensurar a idoneidade creditícia dos clientes para gerenciar o processo de decisão de aceitação dos empréstimos nos bancos (MESTER *et al.*, 1997).

No passado essas análises de crédito eram feitas segundo o julgamento individual conduzido pelos funcionários treinados pelo bancos, que efetuavam essas decisões por meio da análise qualitativa das características de cada cliente. No entanto, a necessidade do gerenciamento eficiente do risco e aumento da competitividade do setor fez com que essas instituições começassem a primar por modelos quantitativos e técnicas analíticas para a realização dessas avaliações de risco (THOMAS; EDELMAN; CROOK, 2002).

Modelos quantitativos para avaliação do risco de crédito são denominados *modelos de credit scoring*. Segundo Thomas, Edelman e Crook (2002), modelos de *credit scoring* são um conjunto de

critérios de decisão e técnicas estatísticas que visam auxiliar instituições financeiras na realização de empréstimos. Além disso, essas técnicas podem determinar quem receberá crédito, o seu limite e quais decisões operacionais poderão ser aplicadas para aumentar os lucros dos emprestadores em relação ao conjunto de mutuários disponíveis.

A literatura na área de *credit scoring* é bastante ampla. Podemos, segundo Louzada, Ara e Fernandes (2016), agrupar os trabalhos sobre esse tema em 7 áreas diferentes. Primeiro, temos os trabalhos que propõem novos métodos para mensuração de risco como Lee *et al.* (2002) e Gestel *et al.* (2006). A segunda área é determinada por trabalhos que buscam comparar técnicas tradicionais de *credit scoring* (WEST, 2000; HUANG; CHEN; WANG, 2007). Em terceiro, temos trabalhos que buscam fazer uma discussão conceitual sobre *credit scoring*, como Bardos (1998) e Banasik, Crook e Thomas (1999). A quarta área apresenta trabalhos que buscam abordar métodos de seleção de covariáveis (TSAI, 2009). A quinta área é determinada por trabalhos de revisão de literatura sobre *credit scoring* (HAND; HENLEY, 1997). Na sexta temos trabalhos que visam analisar ou desenvolver medidas de mensuração de performance desses modelos (HAND, 2005). Por fim, temos trabalhos sobre outros temas como seleção de modelos por reamostragem (BIJAK; THOMAS, 2012) e avaliação da segmentação de indivíduos na criação de modelos de crédito (ZIARI; LEATHAM; ELLINGER, 1997). No entanto, segundo os autores, as três primeiras áreas representam cerca de 90% dos artigos na literatura nos últimos anos.

Para o presente trabalho estamos interessados em um subconjunto dos modelos de *credit scoring*, que conferem um escore a um indivíduo ou a um contrato, que é um indicador referente ao risco de crédito associado a ele. Isto é, gera-se uma medida que expressa o risco de um novo indivíduo ou contrato vir a se tornar inadimplente.

Muitos desses subconjuntos de modelos de *credit scoring* são baseados em modelos de regressão para variáveis binárias. Nesse caso podemos definir a variável resposta como

$$y_i = \begin{cases} 1, \text{observação i é um mau contrato} \\ 0, \text{caso contrário} \end{cases} \quad \text{,i = 1, .., n.} \tag{2.25}$$

Na definição da variável resposta, duas questões são importantes. A primeira é a definição do espaço de tempo que os $n$ indivíduos ou contratos serão observados e classificados como bons ou maus. O mais comum é a utilização de um espaço de tempo de um ano (PEREIRA; ARTES, 2016). A outra é a própria definição de mau contrato, sendo comum definir como mau contrato aquele em que o cliente atrasou suas obrigações por mais de 90 dias dentro do espaço de tempo de observação.

A regressão logística surge como o método mais comum na área de avaliação de risco de crédito (THOMAS; EDELMAN; CROOK, 2002). Entre os motivos disso estão a facilidade de aplicação desse método estatístico e a não imposição de suposições restritivas que são encontradas em outros modelos lineares para variável resposta binária, como a análise discriminante. Outro fator importante é a facilidade da interpretação dos parâmetros do modelo aplicado. A partir da razão

de chances podemos avaliar a influência de cada variável preditora na chance dos indivíduos se tornarem maus.

Nesses casos, estamos interessados em ajustar a probabilidade de um contrato se tornar mau em função de variáveis preditoras. Para isso, conforme discutido no Capítulo 1, é comum discretizar todas as variáveis quantitativas do modelo para melhorar a capacidade preditiva do modelo e facilitar a interpretação dos parâmetros.

Na prática, grande parte dos modelos de *credit scoring* são aplicados com variáveis contínuas discretizadas (PEREIRA; ARTES, 2016). No entanto, não temos conhecimento de artigos na literatura de risco de crédito cujo principal objetivo seja descrever ou comparar métodos de categorização de variáveis. Os métodos de categorização existentes foram produzidos principalmente por pesquisadores da área de aprendizado de máquina. Desse modo, no próximo capítulo descreveremos alguns métodos de categorização existentes e introduziremos os novos métodos propostos neste trabalho.

CAPÍTULO

3

# MÉTODOS DE CATEGORIZAÇÃO DE VARIÁVEIS CONTÍNUAS

A categorização de variáveis contínuas é definida por Kerber (1992) como o processo de dividir os valores de uma variável em um número pequeno de intervalos, em que cada intervalo é formado por um símbolo discreto. Já para Fayyad e Irani (1993), categorização é apenas uma condição lógica que se deve elaborar para dividir uma covariável em dois ou mais subconjuntos.

Os métodos de categorização possuem características distintas. Os métodos podem ser diferenciados entre supervisionados ou não supervisionados. Os métodos supervisionados, que é uma nomenclatura utilizada na área de aprendizado de máquina, são aqueles algoritmos que consideram a variável resposta no processo de categorização. Por outro lado, os métodos não supervisionados são aqueles que não levam em consideração a informação contida na variável resposta. Como por exemplo, temos um algoritmo denominado de *equal-frequency algorithm*. Esse algoritmo recebe um parâmetro $k$, que é o número de níveis desejados pelo usuário. A partir disso, o algoritmo ordena os valores da variável, define as $k$ categorias com um número igual de observações e retorna os seus pontos de corte.

Também podemos definir os métodos de categorização entre univariados ou multivariados. No primeiro método, discretizam-se as covariáveis independentemente. No método multivariado busca-se incorporar a dependência entre as covariáveis nesse processo.

Por fim, os algoritmos de categorização podem ser caracterizados pela maneira que são descobertos os pontos de corte para a criação dos níveis de cada variável. Isso é importante pois acaba impactando sua complexidade computacional.

Uma das abordagens é denominada *bottom-up* em que, para cada covariável discretizada, cria-se uma lista de todos os valores contínuos como pontos de corte, e então remove-se vários pontos de corte por meio de algum critério, criando assim os intervalos de categorização. Outra abordagem é chamada de *top-down*, na qual inicia-se a categorização com apenas um intervalo,

e a partir disso, dada alguma regra, tenta-se acrescentar novos pontos de corte a cada passo do algoritmo, de modo a gerar os intervalos de categorização.

O primeiro método, usualmente, necessita de um número de passos superior que o segundo, resultando em custo computacional maior. Segundo Tsai, Lee e Yang (2008), dada uma variável contínua com 1000 valores distintos e supondo que esse atributo será discretizado em 50 níveis, no geral, uma abordagem *top-down* requer somente 50 passos, enquanto a outra abordagem requer 950 passos do algoritmo.

Assim, na próxima seção temos como objetivo descrever quatro métodos de categorização supervisionados e univariados que seguem a abordagem *top-down* ou uma variação dela e estão implementados no pacote *Discretization* do *software* R (R Development Core Team, 2008). A partir disso, vamos introduzir dois novos métodos de categorização supervisionados, um univariado e o outro multivariado.

## 3.1 Métodos Existentes

Os métodos apresentados nessa seção serão apresentados no contexto de modelos para variáveis resposta binárias. Porém, os mesmos foram definidos originalmente para modelos com variáveis resposta com múltiplas categorias, sendo os modelos com variável resposta binária um caso particular.

Podemos definir o problema proposto por esses métodos da seguinte maneira. Vamos considerar um esquema de categorização $D$ para uma determinada covariável contínua $X$, ordenada crescentemente, com $n$ valores $(x_1^\star, x_2^\star, ..., x_n^\star)$ e uma variável resposta binária Y que assume valores $(y_1^\star, y_2^\star)$. Definimos também um vetor de pontos de corte $c$, $(c_0, c_1, ..., c_k)$, que caracteriza os $k$ níveis da covariável $X$.

A partir disso, descrevemos o esquema de categorização $D$ pela tabela de contingência apresentada abaixo

Tabela 1 – Tabela de Contingência para D

| Var. Resposta | Intervalos | | | | Total |
|---|---|---|---|---|---|
| | $[c_0, c_1)$ | $[c_1, c_2)$ | ... | $[c_{k-1}, c_k]$ | |
| $y_1^\star$ | $n_{11}$ | $n_{12}$ | ... | $n_{1k}$ | $\boldsymbol{n_{1+}}$ |
| $y_2^\star$ | $n_{21}$ | $n_{22}$ | ... | $n_{2k}$ | $\boldsymbol{n_{2+}}$ |
| | $\boldsymbol{n_{+1}}$ | $\boldsymbol{n_{+2}}$ | ... | $\boldsymbol{n_{+k}}$ | $\mathbf{n}$ |

Cada elemento da Tabela 1, $n_{ij}$, representa a quantidade de observações na categoria $i$ da variável resposta e no nível $j$ da variável preditora. Já $\boldsymbol{n_{+j}}$ e $\boldsymbol{n_{i+}}$ representam, respectivamente, os números de observações do intervalo $j$ da covariável e da categoria $i$ da variável resposta. Cada valor da covariável $X$ deve pertencer somente a um dos $k$ intervalos.

Para a realização da categorização no contexto definido por $D$, temos que abordar duas questões. Devemos procurar o número de níveis $k$ e definir quais são os melhores pontos de corte, levando sempre em conta a informação contida na variável resposta.

O primeiro problema será resolvido ao utilizar-se a abordagem *top-down* descrita anteriormente. Adicionalmente, através de um método combinatório ou recursivo procura-se o melhor vetor de pontos de corte. No entanto, é necessária a criação de uma medida que avalie quais são os melhores pontos de corte do esquema de categorização. É nesse ponto que entram os métodos que serão abordados a seguir.

### 3.1.1 CAIM

O CAIM (KURGAN; CIOS, 2004) e os dois métodos descritos nas subseções seguintes funcionam de modo similar. Esses métodos buscam o ponto de corte que maximiza uma determinada medida. Isso é feito até que a adição de uma novo ponto de corte reduza a medida considerada. Assim, a cada passo desses métodos muda-se $k$, que é o número de categorias criadas e $c$ que é o vetor de pontos de corte. Logo, se no passo 3, temos $k = 3$ e $c = (0, 20, 30, 50)$, então temos que, nesse passo, a variável está categorizada em 3 níveis, sendo o primeiro contendo valores maiores ou iguais a 0 e valores inferiores a 20, o segundo contendo valores iguais ou superiores a 20 e inferiores a 30 e o terceiro contendo valores iguais ou superiores a 30 e menores ou iguais a 50. Nesse exemplo, de acordo com esses métodos, se calcularia, no passo 4, o valor da medida de interesse para todos os possíveis $c$ com dimensão 5 considerando que quatro das posições de $c$ devem conter os valores 0, 20,30 e 50 que estavam em $c$ no passo anterior. Assim, no passo 4, teríamos $k = 4$ e o valor de $c$ que maximizou o valor da medida de interesse, caso esta tenha aumentado em relação ao passo 3. Em caso contrário, a categorização final da variável terá 3 níveis com $c = (0, 20, 30, 50)$.

O método de discretização CAIM (*Class-Attribute Interdependency Maximization*) proposto por Kurgan e Cios (2004), cria uma medida que tem como objetivo encontrar o esquema de categorização que minimize o número de níveis da covariável $X$ e a perda de informação relativa à associação entre a covariável e a variável resposta em questão.

Os autores criaram um critério heurístico que busca mensurar a associação entre as categorias da variável resposta e a variável discretizada. Dado um esquema de categorização $D$, a medida criada pelos autores é dada por

$$CAIM(c,k) = \frac{\sum_{j=1}^{k} \frac{max_{j,c}^2}{n_{+j,c}}}{k}, \tag{3.1}$$

em que $k$ é o número de categorias, $max_{j,c}$ é o máximo entre o número de sucessos e fracassos na categoria j quando utilizamos o vetor de pontos de corte $c$ e $n_{+j,c}$ é o número total de observações presentes no j-ésimo nível quando utilizamos o vetor de pontos de corte $c$.

Busca-se esquemas de categorização que maximizem esse critério. O mesmo tem algumas

características importantes. Primeiro, ele favorece discretizações em que os níveis possuam todos os seus valores agrupados dentro de uma categoria da variável resposta. Segundo, a divisão de $max^2_{r,c}$ por $n_{+j,c}$ penaliza categorizações que possuam classes que tenham ao mesmo tempo muitos valores tanto de $y_1^\star$ como de $y_2^\star$. E, conforme aumenta-se $k$ o valor do critério decresce, assim favorecendo discretizações com poucos níveis.

A procura pelo máximo global desse critério pode se tornar computacionalmente complexa, para um modelo com um número de observações considerável. Assim, os autores aconselham a utilização de um processo denominado *greedy search* na implementação deste método. Esse algoritmo aproxima o máximo global através da procura de máximos locais. Ele se dá por um processo incremental. Adiciona-se um novo ponto ao vetor de pontos de corte apenas se ele caracteriza um máximo local.

Essa procura pelo máximo global também é utilizada pelos dois próximos métodos abordados, denominados de Ameva e CACC. Devido às similaridades dos métodos como já citamos, descreveremos a seguir um algoritmo que também servirá para o Ameva e CACC. Deve-se apenas trocar a medida para a do método desejado.

Assim, para uma covariável $X$, podemos definir o algoritmo da seguinte forma:

**Algoritmo 1** – Algoritmo Método Univariado Existente

1: **Passo 1**:
2: Ordene X em ordem crescente.
3: $B \leftarrow$ Valores distintos de X.
4: $c \leftarrow [\min(X), \max(X)]$.
5: $CAIMGlobal \leftarrow 0$.
6: $l \leftarrow 2$. # $l$ é o número de categorias da var. resposta.
7: $k \leftarrow 2$.
8: **Passo 2**:
9: *Para todos valores $B_i$ de B que não estão em c faça*:
10: Inclua $B_i$ em $c$.
11: $CAIM_i \leftarrow CAIM(c,k)$
12: Remova $B_i$ de $c$.
13: $CAIM \leftarrow \max(CAIM_1, CAIM_2, ..., CAIM_n)$.
14: $j \leftarrow argmax_{b_i}(CAIM_1, CAIM_2, ..., CAIM_n)$.
15: *Se CAIM > CAIMGlobal ou k < l*:
16: Inclua $j$ em $c$.
17: $CAIMGlobal \leftarrow CAIM$
18: *Caso contrário, vá para linha 20.*
19: $k \leftarrow k+1$ e vá para linha 9.
20: *Retorne vetor c.*

O processo incremental de adesão de novos pontos de corte se dá até que o novo máximo local comece a se manter constante ou decrescer. Além disso, o algoritmo força uma covariável discretizada com no mínimo dois níveis. A característica de forçar a covariável discretizada a ter um número maior ou igual de categorias que a variável resposta é um fator importante deste algoritmo.

Gonzalez-Abril *et al.* (2009) fazem uma crítica importante dessa característica citada, quando tratamos de variáveis resposta com mais de duas categorias. Segundo os autores, sendo $l$ o número de categorias da variável resposta, a razão para a utilização dessa restrição, é que usualmente temos que $CAIM(k) > CAIM(l+1)$ para $k = 1,2,..,l$. Deste modo, ao forçar um número mínimo de níveis para a covariável discretizada, é razoável supor que o algoritmo nem sempre minimiza o número de intervalos de categorização. É nesse ensejo que foi proposto o próximo algoritmo de categorização denominado Ameva. Esse algoritmo visa resolver as mesmas questões que o CAIM porém sem a mesma restrição.

### 3.1.2 Ameva

O método de categorização Ameva proposto por Gonzalez-Abril *et al.* (2009) tem como objetivo encontrar o esquema de categorização que maximiza a associação entre a covariável categorizada e a variável resposta. E, assim como o critério CAIM, busca minimizar o número de níveis para a covariável categorizada. Como veremos a seguir ele também faz isso com a medida criada pelos autores que tem no denominador o número de intervalos do esquema de categorização.

Assim, dado um esquema de categorização $D$, definido na seção 3.1, a medida Ameva é baseada na estatística do teste de homogeneidade de Pearson $(\chi^2)$ envolvendo uma covariável e uma variável resposta qualitativa, cuja distribuição assintótica, sob a hipótese nula de homogeneidade entre os níveis da covariável, é qui-quadrado. A partir dessa estatística, o critério Ameva é dado por

$$Ameva(c,k) = \frac{\chi^2(k,c)}{k(l-1)} = \frac{\chi^2(k,c)}{k}, \tag{3.2}$$

em que $l = 2$ é o número de categorias da variável resposta $Y$, $\chi^2(k,c)$ é o valor da estatística do teste de homogeneidade envolvendo a covariável dividida em $k$ níveis a partir do vetor de pontos de corte $c$.

A definição dessa medida parte de uma característica da estatística do teste de homogeneidade. Temos que para uma tabela de contingência, o valor máximo da estatística é dada por

$$\max \chi^2(k) = n(min\{l,k\} - 1) \tag{3.3}$$

em que $k$ é o numero de categorias da covariável, $n$ é o número de observações do modelo e $l$ o número de categorias da variável resposta.

Dessa forma, temos que $Ameva_{\max}(k) = \max_{n,k,l} Ameva(k) = \frac{n(k-1)}{k(l-1)}$ para $k < l$ e $\frac{n}{k}$ caso contrário. O máximo do coeficiente ameva é portanto uma função crescente quando $k < l$ e decrescente para $k > l$. Então, sob situação ótima, isto é, quando os valores dos intervalos de categorização são pertencentes somente a uma categoria da variável resposta, atinge-se o valor máximo do critério quando $k$ e $l$ são iguais. Logo, não há a necessidade de uma restrição para o nível mínimo de categorias, como no algoritmo definido na seção anterior.

Dada a explicação do Ameva podemos abordar o próximo método, que também busca fazer melhorias em relação ao CAIM.

### 3.1.3 CACC

O método de categorização denominado CACC (*Class-Attribute Contigency Coefficient*) proposto por Tsai, Lee e Yang (2008), tem como proposta criar um critério que leve em conta a associação entre a variável resposta e a covariável na tentativa de maximizar a dependência essas variáveis. Para isso, esse método se baseia em uma modificação do Coeficiente de Contingência (AGRESTI; KATERI, 2011), que é uma medida de associação entre variáveis qualitativas. O CACC é dado por

$$CACC(c,k) = \sqrt{\frac{\frac{\chi^2(k,c)}{log(k)}}{\frac{\chi^2(k,c)}{log(k)} + n}}, \tag{3.4}$$

em que $n$ é o número total de observações e demais medidas como definidas anteriormente. A divisão por $\log(k)$ no CACC tem como objetivo limitar o número de intervalos criado pelo esquema de categorização.

Para Tsai, Lee e Yang (2008), se tomarmos a variável resposta e a covariável discretizada como duas variáveis aleatórias, o Coeficiente de Contingência é um bom critério para medir a associação entre duas variáveis. Já a divisão por $\log(k)$ de dois termos é uma decisão meramente prática, que no trabalho deles se provou razoável por produzir discretizações de variáveis com poucas categorias e que levaram a modelos com bom poder preditivo.

### 3.1.4 MDL

O método proposto por Fayyad e Irani (1993) é geralmente chamado de MDL, pois é baseado em um princípio denominado *Minimum Description Length* (RISSANEN, 1978). O MDL difere dos métodos anteriores em três aspectos. O primeiro é que novos pontos de corte só serão aceitos se uma medida de ganho (equação 3.7) for maior que determinado valor (equação 3.8). Nos três métodos anteriores, se a medida considerada aumentar de um passo para o outro um novo ponto de corte é acrescentado. O segundo é que, só serão avaliados como possíveis pontos de corte os valores das covariáveis definidos como *boundary points*, que serão definidos posteriormente. Nos métodos anteriores esses pontos eram definidos como todos os valores distintos de uma variável. Em terceiro, a definição do vetor de pontos de corte, $c$, que determina o esquema de categorização é encontrado de forma recursiva. Inicia-se o processo de categorização com apenas um intervalo e vai se realizando sucessivos particionamentos de cada subconjunto criado até que o critério de parada seja atingido. Esse processo pode ser observado na Figura 1. Nos métodos anteriores isso era feito pela abordagem *top-down*.

O método é baseado no conceito de entropia (MACKAY, 2003). Entropia pode ser entendida como uma medida da aleatoriedade de determinada variável. Geralmente, árvores de decisão, em que se tem como intuito encontrar a melhor partição de determinada variável, são construídas com base na otimização dessa medida. Sendo assim, os autores apresentam o método como uma extensão de métodos de categorização binários. A ideia por trás desse método é a de encontrar o esquema de categorização que minimize a perda de informação da variável resposta em relação à covariável discretizada.

Podemos definir a entropia para uma variável resposta $Y$ binária como

$$Ent(Y) = -\sum_{i=1}^{2} f(y_i^\star)\log_2\left[f(y_i^\star)\right], \tag{3.5}$$

em que $f(y_i^\star)$ é a proporção de observações pertencentes à classe $y_i^\star$.

O interesse do método MDL é avaliar a entropia da variável resposta após a categorização de $X$ ou de um subconjunto de $X$ em dois intervalos. Isto é, temos um vetor de ponto de cortes $c$, em que $c = (c_0, c_1, c_2)$. E, sendo $X_1$ e $X_2$ os dois subconjuntos de $X$ criados por esse particionamento, a entropia da variável resposta induzida pelo vetor $c$ é dada por

$$E(Y,c) = \frac{n_1}{n} Ent_{X_1}(Y) + \frac{n_2}{n} Ent_{X_2}(Y), \tag{3.6}$$

em que $n$ é o número de observações da amostra, $n_i$ é o número de observações de cada subconjunto de $X$ e $Ent_{X_i}(Y)$ é o valor da entropia de $Y$ no subintervalo $X_i$.

A categorização binária é determinada pelo vetor de pontos de corte $c$, $c = (c_0, c_1, c_2)$, dentre todos os pontos de corte, o qual minimiza o valor da medida definida por $E(Y,c)$. No entanto, segundo Fayyad e Irani (1993), considerar como possíveis pontos de corte todos os valores distintos de $X$ é computacionalmente custoso. Então, os autores consideram apenas um subconjunto desses valores como possíveis pontos de corte, denominados de *boundary points*.

Para uma covariável $X$ e uma variável resposta binária $Y$, $b_i$ é um *boundary point* se, e somente se, na sequência de valores ordenados por X, existem duas observações consecutivas $(x_1^\star, x_2^\star) \in X$, que pertençam a diferentes categorias da variável resposta, tal que $x_1^\star < b_i < x_2^\star$. Ou seja, um *boundary point*, é um ponto $b_i$ que está entre dois valores consecutivos, $x_1^\star$ e $x_2^\star$, dentre os valores ordenados de $X$, e possuem categorias $y_i^\star$ e $y_j^\star$ para $i \neq j$ na variável resposta.

Foi mostrado por Fayyad (1992) que se o vetor de pontos de corte $c$ minimiza a medida $E(Y,c)$, então $c$ é determinado por *boundary points*. Essa característica torna o algoritmo do MDL mais eficiente pois temos que avaliar apenas $q$ *boundary points* como possíveis pontos de corte, sendo que $1 \leq q \leq n-1$.

Definido os possíveis pontos de corte, temos que descrever o critério de aceitação de um novo ponto de corte. Isso define até que ponto o algoritmo irá discretizar recursivamente determinado intervalo ou subintervalo de $X$. É nesse ponto que entra o conceito de *Minimum Description Length*.

Esse conceito define o tamanho da informação, em *bits*, necessária para especificar as categorias, relativas à variável resposta, de determinada covariável (FAYYAD; IRANI, 1993). O MDL é utilizado para estimar as funções custos das hipóteses de aceitação ou não do novo ponto de corte, através de uma função de ganho da informação. O ganho de informação causada pela categorização do atributo pode ser descrita como

$$Gain(Y,c) = Ent(Y) - E(Y,c), \tag{3.7}$$

em que $Ent(Y)$ e $E(Y,c)$ como definidos anteriormente.

Pelo método MDL, um novo ponto corte para X ou para um subconjunto de X, deve ser aceito se, e somente se,

$$Gain(Y,c) > \frac{\log_2(n-1)}{n} + \frac{\Delta(Y,c)}{n}, \tag{3.8}$$

em que $\Delta(Y,c) = \log_2(3^l - 2) - [lEnt(Y) - l_1 Ent_{X_1}(Y) - l_2 Ent_{X_2}(Y)]$ e $l$, $l_1$ e $l_2$ são os números de diferentes categorias da variável resposta presentes no conjunto $X$ e nos subconjuntos $X_1$ e $X_2$.

A implementação desse algoritmo é trivial e não será demonstrada. No entanto, para melhor exemplificar o funcionamento básico e a recursividade do algoritmo temos a Figura 1. No exemplo, temos uma covariável $X$ contínua e uma variável resposta binária, em que os *boundary points* são os pontos médios entre $(4{,}5;5{,}1)$, $(10{,}1;11{,}3)$ e $(15{,}0;16{,}3)$. A partir disso, o método realiza a primeira partição no ponto médio de $(10{,}1;11{,}3)$ e começa recursivamente realizar partições em cada subintervalo criado, levando em conta os *boundary points* definidos, até atingir o critério de parada. Ao final, criou-se uma covariável discretizada com quatro níveis e com pontos de corte dados por $[1{,}2;5{,}1)$, $[5{,}1;11{,}3)$, $[11{,}3;16{,}3)$ e $[16{,}3;25{,}1]$.

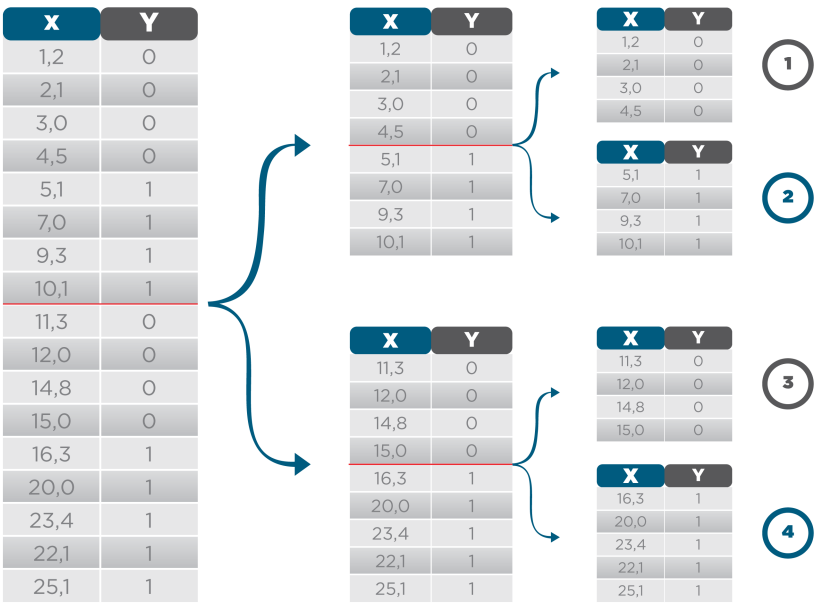


Figura 1 – Funcionamento MDL

## 3.2 Nova Classe de Métodos Univariados

A classe de métodos de categorização univariados proposta neste trabalho segue de certa forma as ideias discutidas anteriormente. Isto é, o método proposto é supervisionado e segue a abordagem *top-down* na definição dos intervalos de categorização. Além disso, também buscamos a maximização de determinada medida para a obtenção do esquema de categorização $D$ (definido anteriormente) ótimo, assim como nos primeiros três métodos existentes. Entretanto, este método difere em relação aos métodos descritos anteriormente em 3 aspectos.

O primeiro aspecto aborda a definição dos pontos de corte avaliados no processo de categorização para cada covariável. Sugerimos uma avaliação dos pontos de corte para cada variável baseada em seus quantis. O propósito disso é reduzir o custo computacional e evitar a criação de categorias com um número muito pequeno de observações, como pode acontecer quando consideramos os *boundary points* no método MDL. O número de quantis considerados não pode ser nem muito pequeno para evitar perda de informação e nem muito grande para evitar a criação de categorias com poucas observações. Assim, aconselhamos a criação de um vetor de possíveis pontos de corte, de forma que se divida a covariável em 200 categorias com o mesmo número de observações. Se o número de observações da base de dados não for grande, propomos a definição desse vetor de forma que o número de observações por categoria fique com pelo menos $m = 30$.

O segundo aspecto que propomos é que o processo de categorização seja baseado em medidas de associação entre variáveis qualitativas. Dessa forma, fica mais claro que o objetivo do processo de categorização é maximizar a associação entre cada covariável e a variável resposta.

Em terceiro, notamos que ao utilizar uma medida de associação entre variáveis qualitativas no processo de categorização, pode-se criar um número muito elevado de categorias por covariável levando a um superajuste do modelo. No caso dos três primeiros métodos discutidos anteriormente, isso era evitado, pois a própria medida dificultava a criação de covariáveis com muitas categorias ao considerar $k$ no denominador da mesma. Nos métodos propostos, para evitar esse problema, sugerimos que se interrompa o processo de categorização quando a inclusão de um novo ponto de corte em $c$ não aumente a medida de associação considerada em pelo menos $(\alpha \times 100)\%$. Propomos que o valor de $\alpha$ seja escolhido a partir de validação cruzada (JAMES *et al.*, 2013). Isso torna desnecessária a mudança de medidas de associação existentes, como ocorre no CACC e Ameva.

Para a utilização da classe de métodos proposta pode ser usada qualquer medida que mensure a associação entre uma variável qualitativa ordinal e uma variável binária. Deste modo, pode-se utilizar diferentes medidas dependendo do problema proposto. Neste trabalho vamos avaliar a performance da classe proposta considerando as medidas de associação *Information Statistics* (IF) (THOMAS; EDELMAN; CROOK, 2002) e *Kendall's Tau-c* ($\tau_c$) (SOMERS, 1962).

A medida de associação IF, também denominada de *information value*, busca identificar o quão diferentes são as probabilidades $p(x|y_1^\star)$ e $p(x|y_2^\star)$ para cada categoria de $X$. Essa medida, no

Tabela 2 – Esquema de categorização

| | $k_1$ | $k_2$ | $k_3$ | $k_4$ |
|---|---|---|---|---|
| f | 25 | 10 | 10 | 60 |
| s | 5 | 40 | 40 | 10 |

caso que a variável resposta assume dois valores, sucesso e fracasso, é definida como

$$IF(c,k) = \sum_{j=1}^{k} \left( \frac{s_{j,c}}{s} - \frac{f_{j,c}}{f} \right) \left[ \log\left( \frac{s_{j,c}}{s} \right) - \log\left( \frac{f_{j,c}}{f} \right) \right], \tag{3.9}$$

em que $k$ é o número de categorias, $s_{j,c}$ e $f_{j,c}$ são, respectivamente, os números de sucessos e fracassos para cada categoria $j$ considerando vetor de pontos de corte $c$; $s$ e $f$ são os números totais de ocorrências de sucessos e fracassos.

A medida de associação $\tau_c$, que é utilizada quando se quer medir a associação entre duas variáveis ordinais, no caso em que a variável resposta tem apenas 2 níveis, é definida como

$$\tau_c(c,k) = \frac{(N_{s,c} - N_{d,c}) \times 4}{n^2}, \tag{3.10}$$

em que $N_{s,c} = \sum_{r=1}^{k-1} \sum_{i=r+1}^{k} f_r \times s_i$ é o número de pares concordantes e $N_{d,c} = \sum_{r=2}^{k} \sum_{i=1}^{r-1} f_r \times s_i$ é o número de pares discordantes e $n$ é o número total de observações. A constante presente no numerador da equação (3.10) garante que a medida de associação assuma valores no intervalo $[-1;1]$. Além disso, no método proposto toma-se o valor absoluto de $\tau_c(c,k)$ para que as associações positivas e negativas sejam consideradas equivalentes.

O cálculo da medida de associação *Kendall's Tau-C* para um esquema de categorização definido pela Tabela 2, é exemplificado abaixo através das equações (3.11), (3.12), e (3.13).

$$N_{s,c} = [25 \times (40+40+10)] + [10 \times (40+10)] + [10 \times (10)] = 2850, \tag{3.11}$$

$$N_{d,c} = [60 \times (40+40+5)] + [10 \times (40+5)] + [10 \times (5)] = 5600, \tag{3.12}$$

$$\tau_c(c,k) = \frac{(2850-5600) \times 4}{200^2} = -0,275. \tag{3.13}$$

Podemos definir o algoritmo para a nova classe de métodos, dado um esquema de categorização $D$, uma covariável $X$ contínua e uma variável resposta $Y$ que assume dois valores $(y_1^\star, y_2^\star)$ como

No algoritmo acima, $B$ define o vetor dos possíveis pontos de corte que serão utilizados para discretizar a covariável, $c$ é o vetor que definirá a discretização ao final do algoritmo e $\alpha$ determina a penalização definida para criação de novos intervalos. Utilizamos como exemplo $\alpha = 0,01$, mas conforme mencionado sugere-se que $\alpha$ seja escolhido por validação cruzada. Utilizamos também a

**Algoritmo 2** – Algoritmo Método Univariado Proposto

1: **Passo 1**:
2: Ordene X em ordem crescente.
3: $B \leftarrow$ Valores dos quantis de X.
4: $c \leftarrow [\min(X), \max(X)]$.
5: $IFGlobal \leftarrow 0$.
6: $k \leftarrow 2$.
7: **Passo 2**:
8: *Para todos valores $B_i$ de B que não estão em c faça*:
9: Inclua $B_i$ em $c$.
10: $IF_i \leftarrow IF(c,k)$
11: Remova $B_i$ de $c$.
12: $IF \leftarrow \max(IF_1, IF_2, ..., IF_n)$.
13: $j \leftarrow argmax_{b_i}(IF_1, IF_2, ..., IF_n)$.
14: *Se IF > IFGlobal*(1 + $\alpha$)*:
15: Inclua $j$ em $c$.
16: $IFGlobal \leftarrow IF$
17: *Caso contrário, vá para linha 19.*
18: $k \leftarrow k+1$ e vá para linha 8.
19: *Retorne vetor c.*

medida *information statistics*, mas como já foi dito, pode-se utilizar qualquer tipo de medida de associação entre uma variável resposta binária e uma covariável qualitativa.

Um problema encontrado nos métodos univariados de categorização, para modelos de regressão, é que os mesmos não consideram a dependência entre as covariáveis no processo de categorização. Isso pode levar a perda de informação e a esquemas de categorização sub-ótimos. Assim, na próxima seção abordaremos o método multivariado proposto. Esse método incorpora no processo de categorização a relação entre todas as covariáveis do modelo para a determinação dos pontos de corte que definirão a categorização das variáveis.

## 3.3 Nova Classe de Métodos Multivariados

Métodos univariados de categorização não levam em conta a correlação entre as variáveis explicativas dos modelos, assim como também falham em capturar padrões que podem acontecer conjuntamente entre algumas variáveis em alguns casos. Isso pode reduzir o poder preditivo do modelo. Assim, propomos um modelo que considere esses problemas.

Há poucos métodos multivariados na literatura e sua grande maioria advém da área de aprendizado de máquina assim como nos métodos univariados. Alguns dos algoritmos observados implementam a categorização multivariada por meio da incorporação do agrupamento das variáveis preditoras como uma nova variável resposta na estrutura de categorização (GUPTA; MEHROTRA; MOHAN, 2010; MONTI; COOPER, 1999). Esses métodos são estáticos. Ou seja, realizam as categorizações como se fosse um método univariado, como os expostos anteriormente, tomando

em consideração duas variáveis resposta, uma real e outra gerada por algum método de análise de agrupamentos (GUPTA; MEHROTRA; MOHAN, 2010). Além disso, não foram encontradas implementações disponíveis dos mesmos para o software *R*. Os métodos propostos na literatura também não apresentaram performance superior que o MDL para todos os bancos de dados considerados. Por todos esses motivos, não descreveremos os métodos existentes na literatura no presente trabalho e nem avaliaremos sua performance com os bancos de dados que utilizaremos.

O método de categorização proposto é dinâmico. Métodos de discretização dinâmicos consideram a associação entre as covariáveis enquanto discretizam as variáveis contínuas no processo de ajuste de algum método de classificação (TSAI; LEE; YANG, 2008). Para realizar esse processo, vamos construir um modelo de regressão logística em que a inclusão e determinação das variáveis preditoras discretizadas será feita por meio de um processo de construção de uma árvore de decisão (MILLER; RANUM, 2011). A cada passo do processo de discretização escolhe-se o melhor ponto de corte para uma determinada covariável considerando os pontos de cortes já criados nas demais covariáveis. Assim, o método proposto leva em consideração a associação entre as variáveis preditoras. Gama, Torgo e Soares (1998) também consideram árvores de decisão em sua proposta de método de categorização multivariado, mas o método proposto por eles é bem diferente do que introduzimos neste trabalho.

Árvores são estrutura de dados hierárquicas, que usualmente são usadas como método de classificação em estatística (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Uma árvore possui diversos níveis ou nós, que são conectados por ramos. A cada nó realiza-se um teste lógico para determinar-se qual o próximo nó descendente. Temos no final um nó terminal, que é único e assume algum valor ou rótulo.

Utilizaremos esse modelo com intuito de obter esquemas de categorização que maximizem a associação entre as variáveis categorizadas e a variável resposta, levando em consideração a estrutura de correlação entre as variáveis explicativas. Podemos dizer então que será um método multivariado e supervisionado.

Na construção de uma árvore devemos seguir certos procedimentos. Primeiro, apresenta-se uma estrutura inicial que define o topo da hierarquia da árvore, chamado de nó raiz, e a partir disso, através de alguma condição lógica, a árvore ramifica-se para algum nó descendente. Esse procedimento é repetido até que um nó terminal é alcançado. Geralmente, aplica-se recursivamente para cada nó criado esse procedimento, dependendo de quantos nós descendentes são permitidos para cada nível. No entanto, no contexto de nosso método, só estamos interessados nas ramificações da árvore que produziram o melhor critério. Portanto, não precisaremos revisitar, no algoritmo, esses nós que não produzem o melhor critério, como fica mostrado na Figura 2.

Para o método proposto, no topo da hierarquia teremos todas as covariáveis categorizadas em apenas um nível, que será representado por um vetor de estados, $a = (a_1, a_2, ..., a_p)$, para um modelo com $p$ covariáveis, $(x_1, x_2, ..., x_p)$. Nesse caso teremos um vetor de tamanho $p$ que será inicializado por $a = (1, 1, ..., 1)$. Além disso, para cada covariável teremos um vetor $b_i$, $(b_1, b_2, ..., b_p)$, em que $b_i$

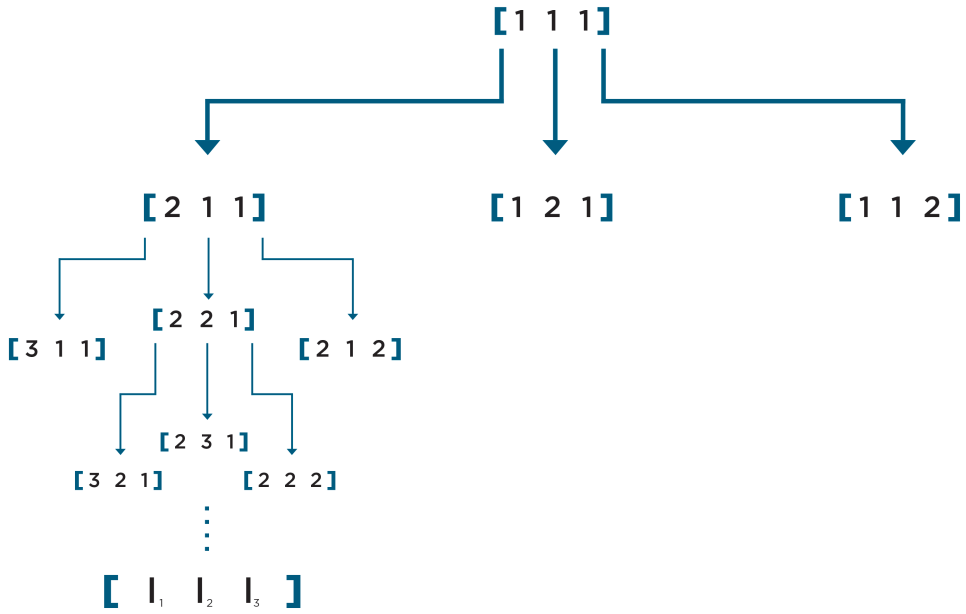


Figura 2 – Algoritmo para um modelo com 3 covariáveis.

são os pontos de corte avaliados pelo método para cada covariável. Esses vetores serão baseados nos quantis de cada variável. Sugerimos, para base de dados grandes, que as variáveis sejam divididas em no máximo 200 quantis. Para bases de dados com poucas observações, sugerimos que cada categoria tenha pelo menos 30 observações.

Para efetuar a avaliação dos vetores de possíveis pontos de corte, a base de dados disponível será divida em base de desenvolvimento e validação. Assim, para cada possível categorização definida por $a_i$ e $c_i$ será ajustado um modelo de regressão logística na base de desenvolvimento e obteremos os valores preditos na base de validação. A partir disso, calcularemos o valor do coeficiente de gini, (THOMAS; EDELMAN; CROOK, 2002) na base de validação para mensurarmos a qualidade do ajuste de cada vetor de estados na estrutura de árvore criada. O coeficiente de gini, que é dado pela a área sob a curva ROC (JR; LEMESHOW, 2004) menos 0,5 multiplicada por 2, variando assim no intervalo $(0,1)$. Ele é muito utilizado em regressão logística para avaliar o poder de discriminação de um modelo, isto é, seu poder de identificar corretamente as observações que são sucesso ou fracasso na variável resposta. Escolheremos assim a categorização que gerou modelo de regressão logística com maior coeficiente de gini.

Tomando como base a Figura 2, podemos exemplificar esse processo. Partindo do nó raiz $a = [1,1,1]$, vamos em um processo incremental, avaliar o seguintes vetores de estados: $[2,1,1]$, $[1,2,1]$ e $[1,1,2]$. O nó raiz nesse caso possui três descendentes. Agora, para cada nó descendente, devemos encontrar o ponto de corte que dividirá essas variáveis em duas categorias. Esse ponto de corte será aquele, para cada covariável, que maximize o valor do coeficiente de gini na base de validação categorizada. A busca pelos pontos de corte dependem do conjunto de vetores dos possíveis pontos de corte $(b_1, b_2, b_3)$. A partir disso, devemos comparar os coeficientes de gini de cada nó no mesmo nível para averiguarmos qual terá descendentes. Segundo a Figura 2, o vetor

$[2,1,1]$ é o que possui o maior coeficiente de gini na base de validação em relação aos vetores $[1,2,1]$ e $[1,1,2]$. Dado esse passo, os pontos de corte que que definem o vetor $[2,1,1]$ são salvos e continuamos o processo de construção da árvore de maneira incremental como é demonstrado na Figura 2. Ao final desse processo o algoritmo retorna os pontos de corte.

Há dois critérios de parada na construção do método proposto. O primeiro é atingido quando o coeficiente de gini na base de validação é igual a 1. Isso quer dizer que o método consegue classificar perfeitamente as observações da base de validação, portanto não há a necessidade de continuar o algoritmo. O segundo critério de parada é atingido caso o coeficiente de gini que determina os novos nós na árvore não aumente $(\alpha \times 100)\%$ por $m$ vezes. Na aplicação consideramos m = 3 e o valor de $\alpha$ foi escolhido da forma que será explicada na Seção 4.1.

Uma característica interessante que podemos avaliar do algoritmo proposto é que nem todas as variáveis são necessariamente discretizadas. Isso pode ocorrer caso algum critério de parada seja atingido antes de que algum novo ponto de corte seja definido para tais variáveis. Isto é, essas variáveis se mantêm no estado inicial, definido anteriormente, com apenas uma categoria e, assim, podem ser eliminadas do modelo. Nesse caso, o método serve também para a seleção de variáveis. Caso seja de interesse categorizar essas variáveis, pode-se aplicar algum dos métodos univariados discutidos.

O método proposto é bastante flexível. Ele pode ser estendido para qualquer problema em que a variável resposta é categórica e deve-se categorizar as covariáveis contínuas. O mesmo não se restringe as variáveis resposta binárias. Deve-se apenas modificar no algoritmo o modelo utilizado e a medida que será usada na base de validação para comparar os diversos nós da árvore criada. No entanto, devido ao escopo deste trabalho, utilizaremos como método somente a regressão logística e bases de dados com variável resposta binária. Além disso, podem-se utilizar outras medidas de performance que não sejam o coeficiente de Gini, como por exemplo o *deviance* (PAULA, 2004) que permite a comparação do ajuste de diversos tipos de modelos de regressão. Por esse motivo, nos referimos a nossa proposta como uma classe de métodos multivariado, pois para cada medida considerada temos um diferente método de categorização.

Uma desvantagem do método multivariado proposto em relação ao univariado é que o mesmo possui um custo computacional maior. Supondo um modelo com $p$ covariáveis e que todas as covariáveis serão categorizadas em $k$ níveis, sendo que cada vetor de possíveis ponto de corte $b_i$ define $Q$ categorias inicias para cada covariável $X_i$. Nesse caso, podemos descrever o número de instruções do método univariado proposto por

$$Int = [Q \times (Q-1) \times ... \times (Q-k+1)] \times p \tag{3.14}$$

Para o modelo multivariado proposto a cada passo do algoritmo temos que reavaliar todos os pontos de corte possíveis e que não foram inclusos no vetor de pontos de corte $c_i$ de cada covariável.

Deste forma, podemos descrever o custo computacional como

$$Int = Q^p + Q^{p-1} \times (Q-1) + ... + (Q-k+1)^p \tag{3.15}$$

Analisando as equações (3.14, 3.15) podemos ver que o custo computacional do método multivariado cresce de forma exponencial quando aumentamos o número de covariáveis, enquanto o método univariado cresce de forma linear para um aumento do número de variáveis preditoras. Como o método multivariado considera a estrutura de associação das covariáveis do modelo, a cada passo do algoritmo são avaliados todas as combinações de possíveis pontos de corte para as variáveis. Essa é a razão do algoritmo ser mais custoso computacionalmente, como veremos na próxima seção em que aplicaremos os métodos abordados neste trabalho em três bases de dados reais.

Uma maneira de diminuir o custo computacional do método multivariado proposto é através da diminuição do número de variáveis preditoras no modelo. Isso pode ser realizado através de métodos de seleção de variáveis como o *Lasso* (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Outra alternativa é dividir as variáveis em alguns grupos de variáveis que são altamente correlacionadas entre si e utilizar o método multivariado separadamente em cada um desses grupos de variáveis.

CAPÍTULO

4

# APLICAÇÃO

A aplicação dos métodos de categorização introduzidos neste trabalho tem como objetivo estudar a eficiência dos mesmos como possíveis métodos de categorização em modelos de regressão para variável resposta binária. Deste modo, neste capítulo, faremos a comparação entre os métodos propostos e os existentes a partir de dois tipos de base de dados: real e simulada.

## 4.1 Aplicação em Dados Reais

Para os dados reais, faremos a comparação entre os métodos propostos e os existentes a partir de três critérios. Avaliaremos o custo computacional dos algoritmos, o número de categorias criadas em cada covariável e o poder preditivo dos modelos que usam as variáveis preditoras categorizadas.

Utilizou-se uma base de dados real (PEREIRA; ARTES, 2016) fornecida por uma instituição financeira para realizar a comparação entre os diversos métodos de categorização. Nessa base de dados há informações sobre três produtos de crédito fornecidos por essa instituição financeira. O primeiro é sobre o cheque especial, o segundo é sobre o cartão de crédito e o terceiro engloba outro tipos de produtos creditícios.

Essa amostra foi retirada de uma população que possuía uma conta no banco, em Dezembro de 2001, e que não tinha nenhum tipo de dívida em atraso com o banco no mês considerado. A variável resposta binária, pra os três produtos considerados, é definida em função da condição de inadimplência ou não do indivíduo em Junho de 2002. Por política de privacidade da instituição financeira não serão divulgados os nomes das variáveis preditoras presentes nessa base de dados. Foram obtidas diversas variáveis contínuas que buscam descrever o comportamento e as características dos consumidores de cada tipo de produto. Além disso, empregou-se na aplicação uma amostra de 17101 observações com 8 variáveis preditoras para o modelo do cheque especial, uma amostra de tamanho 12353 com 6 covariáveis para o modelo do cartão de crédito e uma amostra de 2544

observações com 3 variáveis preditoras no último modelo.

Na Tabela 3 apresentamos a frequência do número de maus nas três bases. Pode-se notar que as três bases são desbalanceadas. Elas possuem um número baixo de clientes que se tornaram inadimplentes no período de análise. Na base cheque apenas 3,3% dos clientes foram marcados como inadimplentes. Já na base de cartão somente 2.7% das observações foram marcadas como mau contrato. Por fim, a base de outros produtos de crédito possui 7.2% de observações marcadas como inadimplentes.

Tabela 3 – Distribuição das Frequências do Número de Bons e Maus

| | Cheque | Cartão | Outros |
|---|---|---|---|
| Mau | 556 / 3,3% | 337 / 2.7% | 183 / 7.2% |
| Bom | 16545 / 96,7% | 12016 / 97,3% | 2361 / 92,8% |

Na Tabela 4 são apresentados os tempos médios (em segundos) do processo de categorização dos métodos existentes, dos métodos existentes com pré-discretização por quantil (excluído o MDL) e dos métodos propostos. Denominamos de pré-discretização por quantil a definição do vetor de possíveis pontos de corte através dos quantis de cada variável. Não fazemos isso com o método MDL pois ele utiliza o vetor de possíveis pontos de corte por meio dos *boundary points*. Os programas foram desenvolvidos no software R e cada algoritmo foi executado 5 vezes. Para os métodos existentes utilizamos as funções presentes no pacote *discretization* (KIM, 2012). Já para os métodos existentes com pré-discretização e para o método univariado proposto modificamos as funções existentes no pacote mencionado. O método multivariado proposto foi integralmente implementado neste trabalho.

A definição do vetor de possíveis pontos de corte, quando baseado em quantis, foram feitas paras as três bases de dados, da maneira que será descrita a seguir. Conforme mencionado na Seção 3.2, para a base do cheque especial e do cartão de crédito definimos o vetor por meio dos valores de 200 quantis igualmente espaçados das variáveis. Para a base de dados dos outros produtos creditícios escolhemos os quantis de forma que cada categoria tenha no mínimo 30 observações. Nota-se na Tabela 4 que o método multivariado proposto apresenta o pior resultado para as três bases de dados usadas, em relação ao tempo médio de discretização. Isso se dá por causa da complexidade computacional do método, como vimos na Seção 3.3. Já os métodos existentes com pré-discretização por quantil são superiores em todas as bases de dados. Podemos ver então que quando incluímos a pré-discretização no processo de categorização, o tempo de execução diminui de maneira considerável. Os métodos univariados propostos também possuem um desempenho interessante, pois são superiores a todos os métodos existentes sem modificação. O melhor método entre os métodos existentes é o MDL. Isso se dá por causa da utilização dos *boundary points* como vetor de possíveis pontos de corte, enquanto os outros métodos utilizam esse vetor como todos os valores distintos de cada covariável. Porém, é razoável supor, pelos resultados apresentados na tabela, que esse vetor de possíveis pontos de corte continua sendo maior que o definido por quantis para todas as bases de dados, causando um maior tempo de execução do MDL.

Tabela 4 – Tempo Médio do Processo de Categorização em Segundos

| Bases de Dados | Métodos Existentes | | | | | |
|---|---|---|---|---|---|---|
| | Ameva | Caim | Cacc | MDL | | |
| CHE | 1185,37 | 1181,29 | 1181,29 | 140,95 | | |
| CAR | 519,39 | 521,26 | 629,85 | 65,14 | | |
| Outros | 8,97 | 13,46 | 684,35 | 0,84 | | |
| Bases de Dados | Métodos Propostos | | | M. E. c/ pré-discretização | | |
| | Information Statistics $\alpha = 0,05$ | Kendalls Tau-C $\alpha = 0,05$ | Multivariado $\alpha = 0,005$ | Ameva | Caim | Cacc |
| CHE | 25,42 | 24,25 | 1936,96 | **15,42** | 15,52 | 15,50 |
| CAR | 22,13 | 22,02 | 1768,32 | 11,75 | **11,66** | 14,3 |
| Outros | 1,65 | 0,6812 | 22,44 | **0,37** | **0,37** | 2,16 |

A Tabela 5 mostra o número médio e máximo do número de categorias criados por cada método nas 3 bases de dados, em apenas uma execução do algoritmo. Esquemas de categorização que geram muitas categorias podem levar a um superajuste do modelo, o que não é desejável e pode levar a um modelo com baixo poder preditivo. Podemos ver que, exceto para o método CACC na última base de dados, os métodos geram um número razoável de níveis para as covariáveis. Dois padrões são observados. O métodos Caim e Ameva, tanto com ou sem pré-discretização, tendem a gerar um número de categorias para as covariáveis iguais ao número de categorias da variável resposta. Já o MDL e os métodos propostos produzem um número de níveis superior ao número de categorias da variável resposta. No entanto, diferente dos outros métodos, com os modelos propostos é possível determinar através do valor da penalização ($\alpha$) um número maior ou menor de categorias para as covariáveis categorizadas. Isso não ocorre nos outros métodos.

Tabela 5 – Média/Máximo de Categorias Criadas Por Método de Categorização

| Bases de Dados | Métodos Existentes | | | | | |
|---|---|---|---|---|---|---|
| | Ameva | Caim | Cacc | MDL | | |
| CHE | 2 / 2 | 2 / 2 | 2 / 2 | 3,62 / 4 | | |
| CAR | 2 / 2 | 2 / 2 | 2,33 / 3 | 3,14 / 4 | | |
| Outros | 2,33 / 3 | 2 / 2 | 41,67 / 121 | 1,67 / 2 | | |
| Bases de Dados | Métodos Propostos | | | M. E. c/ pré-discretização | | |
| | Information Statistics $\alpha = 0,05$ | Kendalls Tau-C $\alpha = 0,05$ | Multivariado $\alpha = 0,005$ | Ameva | Caim | Cacc |
| CHE | 3,12 / 4 | 3 / 4 | 2,28 / 3 | 2 / 2 | 2 / 2 | 2 / 2 |
| CAR | 3,5 / 4 | 3,5/ 4 | 3,75 / 5 | 2 / 2 | 2 / 2 | 2,33 / 3 |
| Outros | 3 / 5 | 3 / 4 | 4 / 7 | 2 / 2 | 2 / 2 | 6,33 / 15 |

Para realizar uma avaliação comparativa da qualidade dos métodos dividimos cada base de dados em base de desenvolvimento, validação e teste. Categorizamos primeiramente a base de desenvolvimento e utilizamos esses pontos de corte para categorizar as bases de validação e teste. A seleção do melhor modelo, para cada método de categorização em determinada base de dados, foi

realizada através do processo denominado *best subset selection* (JAMES *et al.*, 2013) na base de validação, em que selecionamos o modelo com melhor coeficiente de Gini. No método univariado proposto foram selecionados ainda os valores de $\alpha$ que geravam o maior coeficiente de gini na base de validação, entre os componentes do vetor (0,001; 0,01; 0,03; 0,05; 0,1; 0,15; 0,2). Já no método multivariado, os valores de $\alpha$ dentre um vetor com componentes (0,001; 0,003; 0,004; 0,005; 0,007; 0,01; 0,03; 0,05; 0,1) também foram escolhidos usando o coeficiente de Gini na base de validação. No entanto, pelo fato do método multivariado usar a base de validação no processo de categorização, notamos em testes preliminares que a escolha dos valores de $\alpha$ usando o método multivariado sem modificação gerava modelos superajustados. Dessa forma, apenas para a escolha do $\alpha$, o método multivariado foi modificado para não utilizar a base de validação no processo de categorização. Escolhido o $\alpha$, utilizou-se o método multivariado em sua forma original, para a obtenção da categorização final de cada covariável. A partir disso calculou-se os valores preditos dos modelos de regressão logística, usando como covariáveis as variáveis categorizadas pelos diversos métodos citados neste trabalho. Utilizou-se também as variáveis sem modificação (contínuas). Para comparação dos métodos em relação ao poder preditivo, utilizamos como medida o coeficiente de Gini médio (THOMAS; EDELMAN; CROOK, 2002) da base de teste. Os processos descritos foram realizados 25 vezes com sementes de aleatorização diferentes.

Na Tabela 6 apresentamos o coeficiente de gini médio e o desvio padrão na base de teste para todos os métodos. Nas Tabelas 7, 8 e 9 apresentamos os intervalos de confiança com coeficiente de confiança de 95% para a diferença média do coeficiente de gini entre os métodos que estão listados nas linhas em relação aos métodos das colunas. Uma suposição importante na criação de intervalos de confiança em amostras pareadas é a suposição de normalidade das diferenças da variável em estudo nas duas amostras. No Apêndice A apresentamos uma tabela com os valores-p do teste de normalidade de Anderson-Darling (THODE, 2002) para os valores das diferenças do coeficiente de Gini entre os métodos avaliados. Pode-se notar que quase todos os valores-p de testes que envolvem os métodos propostos e os existentes são superiores a 0,05, sugerindo que os intervalos construídos podem ser utilizados para a comparação dos métodos de categorização.

Analisando os coeficientes de Gini médios da Tabela 6, notamos que o método univariado proposto com a medida *Kendall's Tau-C* possui um desempenho superior aos métodos existentes em todos os produtos da instituição financeira. As Tabelas 7, 8 e 9 evidenciam a superioridade desse método em relação aos métodos existentes, pois excetuando o método Ameva na base de dados de outros produtos creditícios, o limite inferior do intervalo de confiança para a diferença média do coeficiente de Gini deste método proposto em relação aos existentes é sempre superior a zero. Isso sugere que o método é superior nessas bases de dados. Já o método univariado proposto com a medida *Information Statistics* apresenta um melhor desempenho relativo nas bases de dados Cheque e Cartão. Analisando os intervalos de confiança para essas duas bases, vemos que o mesmo tem um desempenho superior aos métodos existentes, exceto em relação ao *MDL*. Já o método multivariado é superior aos métodos existentes nas três bases dados dados, exceto em relação ao MDL nos dois primeiros produtos, quando olhamos os ginis médios. Mas quando avaliamos os

Tabela 6 – Média e desvio padrão na base de testes do coeficiente de Gini para os diferentes métodos de categorização

| Bases de Dados | Medidas | | Métodos Propostos | | | Métodos Existentes | | | | M.E. com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Contínuo | Information Statistics | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| CHE | Coef. Gini Médio | 0,795 | 0,791 | **0,800** | 0,784 | 0,766 | 0,766 | 0,766 | 0,792 | 0,766 | 0,766 | 0,766 |
| | D. Padrão Coef. Gini | 0,029 | 0,034 | 0,027 | 0,031 | 0,033 | 0,034 | 0,033 | 0,032 | 0,032 | 0,032 | 0,032 |
| CAR | Coef. Gini Médio | 0,634 | 0,653 | **0,662** | 0,634 | 0,616 | 0,633 | 0,621 | 0,644 | 0,624 | 0,634 | 0,625 |
| | D. Padrão Coef. Gini | 0,040 | 0,055 | 0,047 | 0,056 | 0,056 | 0,054 | 0,053 | 0,045 | 0,054 | 0,057 | 0,054 |
| Outros | Coef. Gini Médio | 0,337 | 0,336 | **0,363** | 0,362 | 0,333 | 0,320 | 0,347 | 0,308 | 0,347 | 0,326 | 0,331 |
| | D. Padrão Coef. Gini | 0,077 | 0,085 | 0,081 | 0,070 | 0,082 | 0,072 | 0,079 | 0,058 | 0,076 | 0,075 | 0,073 |

Tabela 7 – Intervalo de Confiança para a Diferença Média do Coeficiente de Gini na Base Cheque

| | Métodos Propostos | | | Métodos Existentes | | | | M.E. Com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Information Statistics | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% |
| Contínuo | (-0,005;0,013) | (-0,012;0,002) | (0,001;0,021) | (0,020;0,039) | (0,019;0,038) | (0,020;0,039) | (-0,007;0,012) | (0,020;0,039) | (0,020;0,039) | (0,020;0,039) |
| Info.Statistics | | (-0,018;-0,001) | (-0,006; 0,019) | (0,017;0,033) | (0,016;0,032) | (0,017; 0,033) | (-0,01; 0,007) | (0,017; 0,034) | (0,016; 0,033) | (0,017; 0,034) |
| Kendalls Tau-C | | | (0,008; 0,024) | (0,029; 0,040) | (0,028; 0,039) | (0,029; 0,040) | (0,001;0,015) | (0,029; 0,040) | (0,029; 0,040) | (0,029; 0,040) |
| Multivariado | | | | (0,009; 0,027) | (0,009; 0,027) | (0,009; 0,027) | (-0,018; 0,002) | (0,01; 0,027) | (0,010; 0,027) | (0,010; 0,027) |
| Caim | | | | | (-0,002; 0,000) | (0,000; 0,000) | (-0,033; 0,020) | (-0,003; 0,003) | (-0,003; 0,003) | (-0,003; 0,003) |
| Cacc | | | | | | (0,000; 0,002) | (-0,033; -0,019) | (-0,002; 0,004) | (-0,002; 0,004) | (-0,002; 0,004) |
| Ameva | | | | | | | (-0,033; -0,020) | (-0,003; 0,003) | (-0,003; 0,003) | (-0,003; 0,003) |
| MDL | | | | | | | | (0,019; 0,034) | (0,019; 0,034) | (0,019; 0,034) |
| Caim Pré-Disc | | | | | | | | | (0,019; 0,034) | (0,000; 0,000) |
| Cacc Pré-Disc | | | | | | | | | | (0,000; 0,001) |

Tabela 8 – Intervalo de Confiança para a Diferença Média do Coeficiente de Gini da Base Cartão

| | Métodos Propostos | | | Métodos Existentes | | | | M.E. Com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Information Statistics | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C.95% | I.C. 95% | I.C 95% |
| Contínuo | (-0,035;-0,03) | (-0,004;-0,016) | (-0,015;0,015) | (0,001;0,035) | (-0,014;0,016) | (-0,001;0,028) | (-0,022;0,003) | (-0,004; 0,025) | (-0,016; 0,016) | (-0,006;0,023) |
| Info.Statistics | | (-0,021; 0,004) | (-0,002; 0,040) | (0,022; 0,052) | (0,003; 0,037) | (0,016; 0,048) | (-0,004;0,022) | (0,015; 0,043) | (0,001; 0,038) | (0,013; 0,043) |
| Kendalls Tau-C | | | (0,009; 0,046) | (0,029; 0,062) | (0,012; 0,046) | (0,026; 0,056) | (0,007; 0,029) | (0,025; 0,051) | (0,011; 0,045) | (0,023; 0,050) |
| Multivariado | | | | (-0,003; 0,039) | (-0,017; 0,002) | (-0,005; 0,031) | (-0,029; 0,001) | (-0,008; 0,029) | (-0,002; 0,021) | (-0,009; 0,027) |
| Caim | | | | | (-0,027; -0,006) | (-0,012; 0,003) | (-0,045; -0,009) | (-0,017; 0,001) | (-0,005; -0,003) | (-0,019; 0,001) |
| Cacc | | | | | | (0,002; 0,004) | (-0,026; 0,005) | (0,000; 0,018) | (-0,007; 0,005) | (0,000; 0,016) |
| Ameva | | | | | | | (-0,039; -0,007) | (-0,009; 0,002) | (-0,022; -0,004) | (-0,001; 0,001) |
| MDL | | | | | | | | (0,005; 0,034) | (-0,006; 0,026) | (0,004; 0,033) |
| Caim Pré-Disc | | | | | | | | | (-0,018; -0,001) | (-0,003; 0,001) |
| Cacc Pré-Disc | | | | | | | | | | (0,001; 0,016) |

Tabela 9 – Intervalo de Confiança para a Diferença Média do Coeficiente de Gini da Base Outros

| | Métodos Propostos | | | Métodos Existentes | | | | M.E. Com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Information Statistics | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% |
| Contínuo | (-0,031;0,032) | (-0,052;0,000) | (-0,052;0,002) | (-0,027;0,034) | (-0,014;0,048) | (-0,037;0,017) | (0,002; 0,057) | (-0,035;0,016) | (-0,002;0,042) | (-0,002;0,033) |
| Info.Statistics | | (-0,019; 0,003) | (-0,048;-0,003) | (-0,025; 0,03) | (-0,017; 0,005) | (-0,036;0,015) | (-0,002;0,059) | (-0,032;0,011) | (-0,014;0,035) | (-0,019;0,003) |
| Kendalls Tau-C | | | (-0,021; 0,023) | (0,006; 0,053) | (0,002; 0,066) | (-0,002;0,035) | (0,034; 0,076) | (0,004; 0,029) | (0,014; 0,061) | (0,017; 0,048) |
| Multivariado | | | | (0,004; 0,053) | (0,015; 0,068) | (-0,006;0,036) | (0,003; 0,078) | (-0,007;0,037) | (0,016; 0,056) | (0,008; 0,055) |
| Caim | | | | | (-0,008; 0,035) | (-0,022;-0,005) | (0,000; 0,051) | (-0,034;0,008) | (-0,001;0,025) | (-0,023;0,029) |
| Cacc | | | | | | (-0,045; -0,009) | (-0,006;0,003) | (-0,049;-0,005) | (-0,027;0,015) | (-0,031;0,001) |
| Ameva | | | | | | | (0,017; 0,061) | (-0,017; 0,018) | (0,005; 0,037) | (-0,006;0,038) |
| MDL | | | | | | | | (-0,060;-0,018) | (-0,043;0,007) | (-0,005;-0,004) |
| Caim Pré-Disc | | | | | | | | | (-0,002;0,044) | (0,003; 0,029) |
| Cacc Pré-Disc | | | | | | | | | | (-0,029; 0,002) |

intervalos de confiança, no geral, vemos que o método multivariado proposto possui desempenho semelhante ou superior aos métodos existentes.

Quando comparamos os métodos univariados propostos com o modelo ajustado com as variáveis contínuas (sem modificação) vemos que os métodos propostos tem o desempenho superior em uma base de dados e semelhante nas outras duas. Já quando comparamos o desempenho do método multivariado com o modelo com as variáveis contínuas, notamos que o método multivariado é inferior na base do cheque especial e possui desempenho semelhante nas outras duas bases de dados. Quando comparamos os métodos univariados propostos entre si, podemos ver que o método proposto com a medida *Kendall's Tau-C* possui um desempenho superior nas bases de cheque e outros e desempenho semelhante na base de cartão ao *Information Statistics*.

Podemos notar também, avaliando as tabelas dos intervalos de confiança para as três bases de dados avaliadas, que o método multivariado proposto tem um desempenho superior ao método univariado proposto com a medida *Information Statistics* na base de outros produtos creditícios e um desempenho semelhante nas outras duas bases de dados. Já em relação ao método com a medida *Kendall's Tau-C*, o método multivariado apresenta um desempenho semelhante na base de dados de outros produtos de crédito e um desempenho inferior nas outras duas bases.

Observando os desvios padrões, vemos que são em geral pequenos e nenhum método apresenta melhor desempenho nos 3 produtos em relação a esse aspecto. No terceiro produto, devido ao número menor de observações, os desvios padrões, exceto no MDL, aumentam para todos os métodos, atingindo um valor máximo de 0,085 no método univariado proposto com a medida *Information Statistics*.

Por fim, analisando o coeficiente de gini médio e os intervalos de confiança para os métodos Ameva, Caim e Cacc, nota-se que a inclusão da pré-categorização por quantil nesses métodos, em geral, altera pouco a performance preditiva dos métodos, o que torna razoável essa inclusão especialmente em bancos de dados bem grandes, devido ao ganho de custo computacional que vimos na Tabela 4. Além disso, quando comparamos apenas os métodos existentes, observamos que o MDL apresenta o melhor desempenho.

## 4.2 Aplicação em Dados Simulados

A utilização de bases de dados reais para a avaliação de performance dos métodos de categorização descritos é de suma importância. Principalmente quando tratamos do setor financeiro, como foi abordado no presente trabalho. Nesse setor é muito comum a utilização de métodos de categorização em conjunto com o modelo de regressão logística para a definição de políticas de crédito para seus clientes. Assim, neste trabalho, foram utilizadas bases de dados com características de produtos financeiros que são bastante comuns aos bancos brasileiros, como cheque especial e cartão de crédito. A análise dos resultados obtidos na seção anterior sugere que os métodos propostos funcionam bem em bases de dados desse setor. Dado esse contexto, a utilização de dados simulados

também é interessante, pois nos permite avaliar a performance dos métodos em cenários controlados. Isso ajuda avaliar os métodos de categorização propostos em outras situações. Diversos cenários de simulação poderiam ser considerados. No entanto, buscou-se definir dois cenários, um em que as variáveis são altamente correlacionadas e outro em que as variáveis não são correlacionadas. A razão para a escolha desses cenários é que desejamos determinar se o desempenho relativo dos métodos varia muito em função da correlação entre as variáveis preditoras.

Nos dois cenários as bases de dados tem 12000 observações e 6 variáveis explicativas. Essas variáveis foram geradas a partir de uma normal multivariada. No primeiro cenário, que denominaremos de caso correlacionado, as covariáveis foram geradas através de uma normal com vetor de médias $\boldsymbol{\mu}$ = (1000, 1000, 1000, 1000, 1000, 1000), vetor de desvios-padrões $\boldsymbol{\sigma}$ = (250, 250, 250, 250, 250, 250) e com uma matriz de covariâncias $\boldsymbol{\Sigma}$ que foi definida a partir da matriz de correlação descrita a seguir, em que as posições $1,2,3,4,5$ e 6 representam cada uma das 6 variáveis.

$$\rho = \begin{pmatrix} 1 & 0,8 & 0,8 & 0,2 & 0,2 & 0,2 \\ 0,8 & 1 & 0,8 & 0,2 & 0,2 & 0,2 \\ 0,8 & 0,8 & 1 & 0,2 & 0,2 & 0,2 \\ 0,2 & 0,2 & 0,2 & 1 & 0,7 & 0,7 \\ 0,2 & 0,2 & 0,2 & 0,7 & 1 & 0,7 \\ 0,2 & 0,2 & 0,2 & 0,7 & 0,7 & 1 \end{pmatrix}$$

Podemos ver que há dois agrupamentos de variáveis com um alto grau de correlação. As covariáveis $1,2,3$ possuem uma correlação de $0,8$ entre elas e uma correlação baixa com as variáveis $4,5,6$. Já as variáveis $4,5,6$ possuem uma correlação de $0,7$ entre si. Esse cenário será interessante porque poderemos observar o desempenho dos método de categorização propostos em um contexto em que há correlação entre as covariáveis. Isso é importante principalmente no que tange o método multivariado que foi criado para lidar com esse tipo de problema.

Para o segundo cenário, que denominaremos de independente, geramos as 6 variáveis a partir de uma distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}$ = (1000, 1000, 1000, 1000, 1000, 1000) e vetor de desvios-padrões $\boldsymbol{\sigma}$ = (250, 250, 250, 250, 250, 250) e com a matriz de correlação a seguir:

$$\rho = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Analisando a matriz de correlação, podemos ver que no segundo cenário todas as variáveis

são independentes. Cabe ressaltar que correlação zero não implica que duas variáveis são independentes. No entanto, isso é válido para variáveis que seguem a distribuição Normal. Assim, no cenário independente, conseguiremos analisar a performance dos métodos propostos em um cenário em que não há nenhuma relação entre as variáveis.

Já os valores da variável resposta foram determinados de maneira similar para os dois cenários. Esse processo foi feito em duas etapas. Primeiro definimos um vetor $\mu$ com as probabilidades de sucesso e que utiliza o mesmo modelo linear, em relação ao logito da probabilidade de sucesso, para os dois cenários. Cada elemento desse vetor é dado por

$$\mu_i = \frac{e^{1+0{,}005x_{i1}+0{,}001x_{i2}+0x_{i3}-0{,}008x_{i4}+0x_{i5}+0x_{i6}}}{1+e^{1+0{,}005x_{i1}+0{,}001x_{i2}+0x_{i3}-0{,}008x_{i4}+0x_{i5}+0x_{i6}}}, \tag{4.1}$$

em que $i = 1, ..., n$ é o índice de observações e $\mu_i \in (0,1)$.

A partir desse vetor $\mu$ definimos a variável resposta $\mathbf{y}$ através da equação 4.2. A determinação do vetor $\mathbf{y}$ é repetido para cada nova simulação com uma semente de aleatorização diferente.

$$y_i = \begin{cases} 1, \text{se } U(0,1) < \mu_i \\ 0, \text{caso contrário} \end{cases} \quad \text{,i = 1, .., n.} \tag{4.2}$$

Para a avaliação do desempenho dos método de categorização propostos foram realizadas 100 réplicas. O processo de ajuste dos modelos e determinação da performance, através do gini em uma base de testes, foi realizado seguindo os mesmos padrões definidos na Seção 4.1. Por isso, não iremos detalhar novamente como se deu essa dinâmica.

A Tabela 10 apresenta os ginis médios e desvio-padrão dos ginis para as simulações do caso correlacionado e para o caso independente. Olhando os ginis médios, podemos ver que o método univariado proposto com a medida *Kendall's Tau-C* obteve o melhor desempenho no caso correlacionado enquanto o método multivariado proposto apresentou o melhor desempenho no caso independente. Analisando os desvios-padrão percebemos que todos têm valores baixos, mas os métodos novos e o MDL parecem apresentar desvio do coeficiente de Gini inferior aos demais.

Os ginis médios para o modelo de regressão logística ajustado com as variáveis usando os valores contínuos (sem modificação), para os cenários correlacionado e independentes foram, respectivamente, 0,761 e 0,777, pouco superior ao obtido pelos métodos propostos. Isso sugere que mesmo em situações em que o logito da probabilidade de sucesso apresenta uma relação perfeitamente linear com as covariáveis, a perda da capacidade preditiva quando usamos os métodos propostos não é tão grande.

Nas Tabelas 11 e 12 apresentamos os intervalos de confiança com coeficiente de confiança de 95% para a diferença média do coeficiente de gini entre os métodos que estão listados nas linhas em relação aos métodos das colunas. Analisando os intervalos de confiança pode-se ver que o método univariado proposto com a medida *Kendall's Tau-C* realmente tem o melhor resultado para

o cenário correlacionado. Já em relação ao independente, o método multivariado proposto teve um resultado superior a todos os métodos de categorização, exceto ao método univariado *Kendall's Tau-C*, que possui desempenho semelhante.

Ainda analisando os intervalos de confiança, observamos que o MDL tem, assim como na aplicação, o melhor desempenho entre os métodos de categorização existentes. Ele possui desempenho superior aos métodos existentes em ambos os cenários. Quando levamos em conta os métodos existentes em relação aos métodos propostos, o método univariado proposto com as duas medidas e método multivariado proposto apresentam desempenho superior aos métodos existentes nos dois cenários. Já quando avaliamos os métodos propostos observamos que o método univariado com a medida *Information Statistics* possui o pior desempenho.

Nos dois cenários os métodos Caim, Cacc e Ameva apresentam coeficientes de gini médios bem inferiores ao métodos propostos e ao método MDL. Por exemplo, no cenário correlacionado o valor do gini médio para o Caim, Cacc e Ameva é, respectivamente, 0,551, 0,628, 0,613 enquanto o método proposto com medida *Kendall's Tau-C*, o método multivariado e o MDL tem o coeficiente de gini de 0,749, 0,744 e 0,723 respectivamente. A diferença dos ginis médios chega a cerca de 10% nesses casos. O mesmo ocorre para o cenário independente. Isso também fica evidenciado nas Tabelas 11 e 12 em que os limites inferiores dos intervalos de confiança são consideravelmente superiores a 0 quando comparamos esses métodos existentes com os métodos propostos e o MDL. Isso sugere que, dependendo da base de dados, o uso de Caim, Ameva e Cacc pode levar a uma perda substancial de capacidade preditiva em relação ao uso dos métodos propostos e do MDL.

Tabela 10 – Gini Médio e Desvio Padrão dos Dados Simulados

| Simulação | Medidas | Métodos Propostos | | | Métodos Existentes | | | | M.E. com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Information Statistics | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| Var. Correlacionadas | Coef. Gini Médio | 0,736 | **0,749** | 0,744 | 0,551 | 0,628 | 0,613 | 0,723 | 0,556 | 0,628 | 0,612 |
| | D. Padrão Coef. Gini | 0,014 | 0,013 | 0,013 | 0,021 | 0,020 | 0,0,17 | 0,014 | 0,022 | 0,019 | 0,017 |
| Var. Independentes | Coef. Gini Médio | 0,756 | 0,758 | **0,761** | 0,576 | 0,628 | 0,628 | 0,753 | 0,576 | 0,636 | 0,628 |
| | D. PadrãoCoef. Gini | 0,012 | 0,011 | 0,013 | 0,023 | 0,021 | 0,017 | 0,012 | 0,021 | 0,018 | 0,017 |

Tabela 11 – I.C. para as Diferenças Pareadas Usando Dados Simulados Com Variáveis Correlacionadas

| | | | Métodos Existentes | | | | M.E. Com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|
| | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% |
| Info.Statistics | (-0,015;-0,012) | (-0,012;-0,004) | (0,179;0,190) | (0,103;0,113) | (0,118;0,126) | (0,006;0,008) | (0.175;0.184) | (0,104;0,112) | (0,121;0,126) |
| Kendalls Tau-C | | (0,002;0,009) | (0,193;0,203) | (0,116;0,126) | (0,132;0,140) | (0,019;0,023) | (0,189;0,197) | (0,118;0,125) | (0,135;0,139) |
| Multivariado | | | (0,187;0,197) | (0,111;0,112) | (0,126;0,134) | (0,011;0,019) | (0,182;0,193) | (0,111;0,121) | (0,127;0,136) |
| Caim | | | | (-0,082;-0,071) | (-0,066;-0,058) | (-0,183;-0,172) | (-0,010;0,000) | (-0,082;-0,071) | (-0,066;-0,056) |
| Cacc | | | | | (0,010;0,019) | (-0,106;-0,096) | (0,066;0,077) | (-0,006;0,006) | (0,010;0,021) |
| Ameva | | | | | | (-0,119;-0,111) | (0,052;0,062) | (-0,019;-0,009) | (-0,003;0,005) |
| MDL | | | | | | | (0,168;0,177) | (0,097; 0,104) | (0,114;0,119) |
| Caim Pré-Disc | | | | | | | | (-0,076;-0,066) | (-0,060;-0,051) |
| Cacc Pré-Disc | | | | | | | | | (0,013;0,018) |

Tabela 12 – I.C. para as Diferenças Pareadas Usando Dados Simulados Com Variáveis Independentes

| | | | Métodos Existentes | | | | M.E. Com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|
| | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% | I.C 95% |
| Info.Statistics | (-0,003; -0,001) | (-0,008;-0,002) | (0,176;0,185) | (0,124;0,133) | (0,126;0,132) | (0,002;0,004) | (0,177;0,185) | (0,117;0,124) | (0,125;0,132) |
| Kendalls Tau-C | | (-0,007;0,000) | (0,178;0,186) | (0,126;0,135) | (0,128;0,134) | (0,004;0,006) | (0,179;0,186) | (0,119;0,125) | (0,127;0,134) |
| Multivariado | | | (0,180;0,191) | (0,129;0,138) | (0,129;0,138) | (0,005;0,012) | (0,180;0,191) | (0,121;0,130) | (0,129;0,138) |
| Caim | | | | (-0,058;-0,045) | (-0,056;-0,047) | (-0,181;-0,173) | (-0,003;0,004) | (-0,065;-0,055) | (-0,056;-0,047) |
| Cacc | | | | | (-0,005;0,005) | (-0,130;-0,121) | (0,046;0,058) | (-0,014;-0,003) | (-0,005;0,005) |
| Ameva | | | | | | (-0,129;-0,122) | (0,047; 0,056) | (-0,011;-0,005) | (-0,001; 0,001) |
| MDL | | | | | | | (0,173;0,181) | (0,114;0,120) | (0,122;0,129) |
| Caim Pré-Disc | | | | | | | | (-0,065;-0,056) | (-0,057;-0,047) |
| Cacc Pré-Disc | | | | | | | | | (0,005;0,011) |

CAPÍTULO

# 5

# IMPLEMENTAÇÃO DO PACOTE NO R

Os métodos propostos neste trabalho foram implementados em um pacote denominado *multdiscretization* e estão disponíveis para ser instalados através da plataforma Github (2017). Nas seções a seguir vamos descrever o processo de instalação do pacote, assim como suas funções através de exemplos práticos.

## 5.1 Instalação

Para realizar a instalação de pacotes que estão hospedados no *Github* devemos instalar uma biblioteca chamada *devtools*. Essa biblioteca, que geralmente é utilizada na criação de pacotes no *R*, possui uma função chamada *install_github* que nos possibilita a instalação de pacotes presentes no *Github*. Para instalar essa biblioteca devemos digitar o seguinte comando no *R*

```
1 install.packages('devtools')
```

Para instalar o pacote devemos primeiro carregar a biblioteca instalada e depois precisamos chamar a função *install_github* tomando como parâmetro o nome do repositório em que consta o pacote:

```
1 library('devtools')
2 install_github('diegomattozo/categorization')
```

A partir disso, basta carregar o pacote com o comando *library(multdiscretization)* para usá-lo dentro do ambiente de programação *R*.

## 5.2 O Pacote

As funcionalidades desse pacote podem ser divididas em três partes. Primeiro, temos as funções que devem ser utilizadas no contexto do método univariado. Segundo, temos as funções que

devem ser usadas para o método multivariado. Por fim, estão disponíveis algumas funções utilitárias que podem ser utilizadas em conjunto com as funções dos métodos univariados e multivariado ou até mesmo fora do contexto de categorização que é o foco da biblioteca. A Tabela 13 apresenta as funções da biblioteca.

Tabela 13 – Descrição das Funções Presentes no Pacote

| **Método Univariado** | **Descrição** |
|---|---|
| discretize | Função que realiza a categorização univariada para as medidas 1 - Caim, 2 - Cacc, 3 - Ameva, 4 - Info. Stat. e 5 - Kendalls Tau-C. |
| disc_from_cuts | Função que realiza a categorização de uma base não categorizada a partir dos pontos de corte retornados pela função discretize. |
| **Método Multivariado** | **Descrição** |
| multdiscretization | Função que realiza a categorização dos dados usando o método multivariado. |
| cutpoint_discretization | Função que realiza a categorização de uma base não categorizada a partir dos pontos de corte retornados pela função multdiscretization. |
| **Funções Utilitárias** | **Descrição** |
| logistic_reg_giniCoef | Função que calcula o coeficiente de gini da validação a partir de uma base de desenvolvimento e validação. |
| train_test_split | Função que divide aleatoriamente uma base de dados em base de desenvolvimento e validação. |
| quantile_discretization | Função que categoriza por quantil determinada base de dados. |

Para o método univariado temos duas funções. A primeira denominada *discretize* realiza a categorização univariada das variáveis contínuas presentes em uma base de dados. Essa função recebe como primeiro argumento **db** que é a base de dados a ser categorizada. Essa base de dados deve conter as variáveis preditoras contínuas e uma variável resposta binária na última coluna. O segundo argumento **meth** é um número inteiro de 1 a 5 que define a medida a ser utilizada para a categorização: 1 - Caim, 2 - Cacc, 3 - Ameva, 4 - Information Statistics e 5 - Kendalls Tau-C. O terceiro argumento é **alpha** que define a penalização que visa evitar a criação de um número excessivo de categorias de cada variável contínua discretizada. O quarto argumento **n** define o número mínimo de observações por quantil que cada covariável deve ser pré-categorizada. O último argumento **prediscretized**, que é opcional, informa se as variáveis preditoras já foram pré-categorizadas ou não. Isso é útil pois possibilita a utilização de tipos de pré-categorização diferentes da pre-categorização por quantil, que é a utilizada por padrão. Por fim, essa função retornará a base de dados categorizada e uma matriz com os pontos de corte de cada variável preditora categorizada.

A segunda função do método univariado *disc_from_cuts* gera a categorização de uma nova base de dados a partir dos pontos de corte retornados pela função *discretize*. Essa função recebe dois argumentos, **db** que é a base de dados a ser categorizada, em que essa base deve contar a variável resposta binária na última coluna, e **cutpoints** que é a matriz com os pontos de corte de cada variável preditora. No geral, essa função será utilizada em conjunto com duas funções utilitárias *train_test_split* e *quantile_discretization* como demonstraremos a seguir, mas antes vamos abordar os argumentos dessas funções. A função *train_test_split* tem como argumento **db** que é a base a ser

divida em desenvolvimento e validação, **percentual** que é a porcentagem da base de dados que deve ser dedicada a base de validação e **seed** que é a semente de aleatorização dessa divisão. Já a função *quantile_discretization*, que realiza a categorização por quantil de uma base de dados, tem como argumentos **db** que é a base de dados a ser categorizada e **n** que é o número mínimo de observações por quantil.

Para exemplificar o processo de utilização dessas funções abordadas, iremos utilizar a base de dados cheque que abordamos na Aplicação. Agora, vamos supor um cenário em que queremos mensurar a performance do método de categorização univariado com a medida *Kendalls Tau-C* nessa base. Para isso, vamos dividir essa base em desenvolvimento e teste. Depois iremos categorizar a base de desenvolvimento e utilizar os pontos de corte desse processo para categorizar a base de teste. Por fim, iremos calcular o coeficiente de gini da base de teste. Para realizar esse processo, supondo que a base cheque já foi carregada, devemos realizar o particionamento da mesma:

```
1 aux <- train_test_split(cheque, 0.2)
2 cheque_test <- aux$test
3 cheque_desenv <- aux$train
```

O código acima cria uma base de teste que tem 20% do tamanho total da base de dados de cheque. Como podemos ver a função train_test_split retorna uma lista com as duas bases de dados particionadas. Agora, devemos categorizar a base de desenvolvimento e utilizar os pontos de corte definidos nessa categorização para discretizar a base de dados de teste.

```
1 disc_cheque_desenv <- discretize(db=cheque_desenv, meth=5,
2 alpha=0.05, n=100)
3 cuts_desenv <- disc_cheque_desenv$cuts # salvo os pontos de corte
4 disc_cheque_desenv <- disc_cheque_desenv$data
5
6 # Os pontos de corte retornados são para a base pré-categorizada.
7 # Então devo pré-categorizar a base de testes com base nos
    quantis
8 # da base de desenvolvimento.
9
10 quantile_cuts <- quantile_discretization(db= cheque_desenv,
11  n = 100)$cuts
12 quantile_cheque_test <- disc_from_cuts(db=cheque_test,
13 cutpoints=quantile_cuts)
14 disc_cheque_test <- disc_from_cuts(db=quantile_cheque_test,
15 cutpoints=cuts_desenv)
```

Agora que temos as variáveis categorizadas com os nomes *disc_cheque_desenv* e *disc_cheque_test*, precisamos apenas ajustar um modelo de regressão logístico na base de desenvolvimento,

gerar os valores preditos na base de teste e calcular o coeficiente de gini. Isso pode ser feito através da função utilitária *logistic_reg_giniCoef*. Essa função recebe os argumentos **train** que é a base de desenvolvimento, **test** que é a base de teste, **respName** que é o nome da variável resposta nessa base de dados e nos retorna o coeficiente de gini na base de teste.

```
giniCoef <- logistic_reg_giniCoef(train=disc_cheque_desenv,
test=disc_cheque_test, respName="y")
giniCoef
```

Conforme apresentado na Tabela 13, temos também duas funções relacionadas ao método multivariado. A primeira função chamada *multdiscretization* realiza a categorização de uma base dados utilizando o método multivariado. Essa função assume uma base de dados em que a variável resposta binária está presente na última coluna da tabela e tem como argumentos **train** que é a base de desenvolvimento, **test** que é uma base de validação, **alpha** que é a penalização utilizada pelo método para manter um nível pequeno de intervalos para cada covariável categorizada e nos retorna as duas bases categorizadas e uma matriz de pontos de corte. A segunda função denominada *cutpoint_discretization* é utilizada para categorizar uma base de dados com base nos pontos de corte retornados pela função anterior. Essa função recebe como argumentos **db** que é a base a ser categorizada em que a variável resposta está na última coluna da tabela e **cutpoints** que são os pontos de corte retornados pela função *multdiscretization* e retorna a base de dados categorizada. Para exemplificar o processo de utilização dessas funções, vamos usar o mesmo exemplo do caso univariado em que temos a base de cheque da aplicação e queremos avaliar a performance do método de categorização utilizando o coeficiente de gini em uma base de teste. Para isso devemos separar essa base em desenvolvimento, validação e teste.

```
aux <- train_test_split(cheque, 0.5)
validation <- aux$test
desenv <- aux$train
aux2 <- train_test_split(validation, 0.5)
validation <- aux2$train
test <- aux$test
```

No código acima criamos uma base de desenvolvimento com 50% do tamanho de cheque e as bases de validação e teste com 25% do tamanho dessa base. Depois do particionamento da nossa base de dados, já podemos categorizar as bases de dados criadas:

```
aux <- multdiscretization(train=desenv, test=validation,
alpha=0.05)
cutpoints <- aux$cuts
disc_desenv <- aux$train
disc_validation <- aux$test
```

```
6 # Agora devemos categorizar a base de teste com os pontos de
    corte.
7 disc_test <- cutpoint_discretization(test, cutpoints)
```

Com isso temos as três bases de dados categorizadas e, então, podemos ajustar o modelo de regressão logística na base de desenvolvimento, gerar os valores preditos na base de teste e calcular o coeficiente de gini.

```
1 giniCoef <- logistic_reg_giniCoef(train=disc_desenv,
2 test=disc_test, respName="y")
3 giniCoef
```

O argumento **alpha** das duas funções *discretize* e *multdiscretization* pode ser selecionado de modo a gerar uma melhor performance do modelo ajustado. Essa busca pode ser feita de maneira simples através da criação de diversas bases categorizadas com diferentes *alphas*, em que selecionamos o valor desse argumento que gera a melhor performance, isto é, o maior coeficiente de gini em uma base de validação.

Por fim, uma característica importante do método multivariado é que ele faz a seleção de variáveis em alguns casos. Isso pode não ser razoável em certas circunstâncias, como por exemplo quando o usuário do pacote sabe que dada variável é de vital importância para seu modelo. Sendo assim, uma sugestão seria a de categorizar essas variáveis pelo método univariado para a inclusão das mesmas no modelo. Isso pode ser facilmente realizado com o nosso pacote.

CAPÍTULO

6

# CONCLUSÃO

Nesta dissertação foram estudados métodos de categorização para variáveis preditoras contínuas em modelos de regressão para variável resposta binária. A área de *credit scoring* é conhecida por utilizar esses modelos de forma extensiva. Por isso, utilizamos três bases de dados de uma instituição financeira para analisar a performance dos diversos métodos categorização abordados no trabalho. Propomos duas classes de métodos de categorização. A primeira classe, que denominamos de univariada, realiza a categorização de cada covariável de modo a maximizar uma medida de associação entre a variável preditora discretizada e a variável resposta em questão. Utilizamos duas medidas de associação para esse processo a *Information Statistics* e a *Kendall's Tau-C*. A segunda classe denominada de multivariada, realiza a categorização de um conjunto de variáveis preditoras contínuas conjuntamente de modo a considerar a estrutura de correlação das mesmas. Esse processo é realizado através da criação de uma árvore de decisão, onde cada nó desta árvore define um esquema de categorização desse conjunto de variáveis.

Observando os resultados obtidos na aplicação e simulação, o método univariado, com a medida *Kendall's Tau-C*, parece apresentar o melhor desempenho na maioria dos casos tratados. Além disso, o desempenho computacional desse método foi superior aos dos métodos de categorização existentes. Já o método multivariado foi melhor que o *Kendall's Tau-C* no cenário simulado com variáveis preditoras independentes e foi ligeiramente inferior nas bases de dados reais. Assim, sugerimos a utilização desses dois métodos em aplicações que as bases de dados não possuem um número muito grande de covariáveis e uso do método *Kendall's Tau-C* quando o número de variáveis preditoras for grande.

Um fator limitante e que deve ser ressaltado é que devido a complexidade computacional do método multivariado, não foi possível a realização de estudos em bases de dados maiores, como são encontradas na prática. Além disso, o método multivariado proposto deveria, segundo a visão do autor, ter um desempenho superior aos métodos univariados quando temos modelos com covariáveis que possuem correlação, o que não aconteceu na simulação para o cenário correlacionado, pois o

método univariado com a medida *Kendall's Tau-C* teve desempenho superior. Sendo assim, para estudos futuros, sugerimos as mudanças a seguir nos métodos propostos:

- Realização de mudanças no método multivariado para ter uma melhor performance, em relação ao gini, em modelos com variáveis correlacionadas.
- Implementação de uma nova interface para o pacote R, de modo que os métodos propostos comecem a funcionar para problemas em que a variável resposta tem várias categorias e o usuário possa determinar (como argumento) a medida de associação desejada.
- Ajustar o pacote para lidar com valores faltantes (*missing*) que são muito comuns em bases de dados reais.
- Buscar na literatura estrutura de dados mais eficazes para melhorar a performance computacional do método multivariado.
- Implementação do método multivariado na linguagem de programação C++ que é uma linguagem compilada e possui um bom suporte no ambiente de programação R.

Além das melhorias sugeridas, estudos futuros poderiam considerar:

- Comparação dos métodos propostos e dos existentes em outras bases de dados tanto da área de risco de crédito como de outras áreas.
- Comparação dos métodos existentes considerando bases de dados resultantes de outros cenários de simulação.

# REFERÊNCIAS

AGRESTI, A.; KATERI, M. **Categorical data analysis**. New York: Springer, 2011. Citado na página 40.

BANASIK, J.; CROOK, J. N.; THOMAS, L. C. Not if but when will borrowers default. **Journal of the Operational Research Society**, Springer, v. 50, n. 12, p. 1185–1190, 1999. Citado na página 33.

BARDOS, M. Detecting the risk of company failure at the banque de france. **Journal of Banking & Finance**, Elsevier, v. 22, n. 10, p. 1405–1419, 1998. Citado na página 33.

BIJAK, K.; THOMAS, L. C. Does segmentation always improve model performance in credit scoring? **Expert Systems with Applications**, Elsevier, v. 39, n. 3, p. 2433–2442, 2012. Citado na página 33.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. **São Paulo**, 2008. Citado na página 31.

DOBSON, A. J.; BARNETT, A. **An introduction to generalized linear models**. [S.l.]: CRC press, 2008. Citado na página 29.

FAYYAD, U.; IRANI, K. Multi-interval discretization of continuous-valued attributes for classification learning. **In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence**, San Francisco, CA, p. 1022–1027, 1993. Citado nas páginas 35, 40, 41 e 42.

FAYYAD, U. M. **On the Induction of Decision Trees for Multiple Concept Learning**. Tese (Doutorado), Ann Arbor, MI, USA, 1992. UMI Order No. GAX92-08535. Citado na página 41.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics Springer, Berlin, 2001. v. 1. Citado nas páginas 46 e 49.

GAMA, J.; TORGO, L.; SOARES, C. Dynamic discretization of continuous attributes. In: SPRINGER. **Ibero-American Conference on Artificial Intelligence**. [S.l.], 1998. p. 160–169. Citado na página 46.

GESTEL, T. V.; BAESENS, B.; SUYKENS, J. A.; POEL, D. Van den; BAESTAENS, D.-E.; WILLEKENS, M. Bayesian kernel based classification for financial distress detection. **European journal of operational research**, Elsevier, v. 172, n. 3, p. 979–1003, 2006. Citado na página 33.

GITHUB. **Github, The world's leading software development platform**. 2017. [Online; acessado 15-Abril-2017]. Disponível em: <https://github.com/>. Citado na página 67.

GONZALEZ-ABRIL, L.; CUBEROS, F. J.; VELASCO, F.; ORTEGA, J. A. Ameva: An autonomous discretization algorithm. **Expert Systems with Applications**, Elsevier, v. 36, n. 3, p. 5327–5332, 2009. Citado nas páginas 25 e 39.

GREEN, P. J. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 149–192, 1984. Citado na página 31.

GUPTA, A.; MEHROTRA, K. G.; MOHAN, C. A clustering-based discretization for supervised learning. **Statistics & probability letters**, Elsevier, v. 80, n. 9, p. 816–824, 2010. Citado nas páginas 45 e 46.

HAND, D. J. Good practice in retail credit scorecard assessment. **Journal of the Operational Research Society**, Nature Publishing Group, v. 56, n. 9, p. 1109–1117, 2005. Citado na página 33.

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, Wiley Online Library, v. 160, n. 3, p. 523–541, 1997. Citado na página 33.

HUANG, C.-L.; CHEN, M.-C.; WANG, C.-J. Credit scoring with a data mining approach based on support vector machines. **Expert systems with applications**, Elsevier, v. 33, n. 4, p. 847–856, 2007. Citado na página 33.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112. Citado nas páginas 43 e 54.

JR, D. W. H.; LEMESHOW, S. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2004. Citado nas páginas 30 e 47.

KERBER, R. Chimerge: Discretization of numeric attributes. In: AAAI PRESS. **Proceedings of the tenth national conference on Artificial intelligence**. [S.l.], 1992. p. 123–128. Citado nas páginas 25 e 35.

KIM, H. **discretization: Data preprocessing, discretization for classification.** [S.l.], 2012. R package version 1.0-1. Disponível em: <https://CRAN.R-project.org/package=discretization>. Citado nas páginas 25 e 52.

KURGAN, L. A.; CIOS, K. J. Caim discretization algorithm. **Knowledge and Data Engineering, IEEE Transactions on**, IEEE, v. 16, n. 2, p. 145–153, 2004. Citado nas páginas 25 e 37.

LEE, T.-S.; CHIU, C.-C.; LU, C.-J.; CHEN, I.-F. Credit scoring using the hybrid neural discriminant technique. **Expert Systems with applications**, Elsevier, v. 23, n. 3, p. 245–254, 2002. Citado na página 33.

LIU, H.; SETIONO, R. Chi2: Feature selection and discretization of numeric attributes. In: IEEE. **Proceedings of IEEE 24th International Conference on Tools with Artificial Intelligence**. [S.l.], 1995. p. 388. Citado na página 25.

LOUZADA, F.; ARA, A.; FERNANDES, G. B. Classification methods applied to credit scoring: A systematic review and overall comparison. **arXiv preprint arXiv:1602.02137**, 2016. Citado na página 33.

MACKAY, D. J. **Information theory, inference and learning algorithms**. [S.l.]: Cambridge university press, 2003. Citado na página 41.

MESTER, L. J. *et al.* Whats the point of credit scoring? **Business review**, v. 3, p. 3–16, 1997. Citado na página 32.

MILLER, B. N.; RANUM, D. L. **Problem Solving with Algorithms and Data Structures Using Python SECOND EDITION**. [S.l.]: Franklin, Beedle & Associates Inc., 2011. Citado na página 46.

MONTI, S.; COOPER, G. F. A latent variable model for multivariate discretization. In: **AISTATS**. [S.l.: s.n.], 1999. Citado na página 45.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, JSTOR, p. 370–84, 1972. Citado na página 27.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. São Paulo, Brazil: IME-USP, 2004. Citado nas páginas 27, 28, 30 e 48.

PEREIRA, G. H. A.; ARTES, R. A comparison of strategies to develop a customer default scoring model. **Journal of the Operational Research Society**, Springer, v. 67, n. 11, p. 1341–1352, 2016. Citado nas páginas 33, 34 e 51.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2008. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>. Citado nas páginas 31 e 36.

RISSANEN, J. Modeling by shortest data description. **Automatica**, Elsevier, v. 14, n. 5, p. 465–471, 1978. Citado na página 40.

SOMERS, R. H. A new asymmetric measure of association for ordinal variables. **American sociological review**, JSTOR, p. 799–811, 1962. Citado nas páginas 26 e 43.

TAY, F. E.; SHEN, L. A modified chi2 algorithm for discretization. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 14, n. 3, p. 666–670, 2002. Citado na página 25.

THODE, H. C. **Testing for normality**. [S.l.]: CRC press, 2002. v. 164. Citado na página 54.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. **Credit scoring and its applications**. [S.l.]: Siam, 2002. Citado nas páginas 25, 26, 32, 33, 43, 47 e 54.

TSAI, C.-F. Feature selection in bankruptcy prediction. **Knowledge-Based Systems**, Elsevier, v. 22, n. 2, p. 120–127, 2009. Citado na página 33.

TSAI, C.-J.; LEE, C.-I.; YANG, W.-P. A discretization algorithm based on class-attribute contingency coefficient. **Information Sciences**, Elsevier, v. 178, n. 3, p. 714–731, 2008. Citado nas páginas 25, 36, 40 e 46.

WEST, D. Neural network credit scoring models. **Computers & Operations Research**, Elsevier, v. 27, n. 11, p. 1131–1152, 2000. Citado na página 33.

ZIARI, H. A.; LEATHAM, D. J.; ELLINGER, P. N. Development of statistical discriminant mathematical programming model via resampling estimation techniques. **American Journal of Agricultural Economics**, Oxford University Press, v. 79, n. 4, p. 1352–1362, 1997. Citado na página 33.

APÊNDICE

A

# TABELAS DOS TESTES DE NORMALIDADE DE ANDERSON-DARLING

Tabela 14 – Valores-p do Teste de Normalidade para a Base de Dados Cheque

| | Métodos Propostos | | | Métodos Existentes | | | | M.E. Com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Medidas | Information Statistics | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| Contínuo | 0,3208 | 0,2687 | 0,2154 | 0,8535 | 0,7039 | 0,8535 | 0,9283 | 0,5222 | 0,4615 | 0,5222 |
| Info.Statistics | | 0,0918 | 0,5868 | 0,2703 | 0,4428 | 0,2703 | 0,0415 | 0,4275 | 0,5068 | 0,4275 |
| Kendalls Tau-C | | | 0,2673 | 0,1191 | 0,3400 | 0,1191 | 0,9612 | 0,4708 | 0,5286 | 0,4708 |
| Multivariado | | | | 0,7203 | 0,6026 | 0,7203 | 0,9385 | 0,4349 | 0,4264 | 0,4349 |
| Caim | | | | | 0,0000 | 0,0000 | 0,8379 | 0,0006 | 0,0028 | 0,0006 |
| Cacc | | | | | | 0,0000 | 0,8279 | 0,0032 | 0,0083 | 0,0032 |
| Ameva | | | | | | | 0,8379 | 0,0006 | 0,0028 | 0,0006 |
| MDL | | | | | | | | 0,4881 | 0,4484 | 0,4881 |
| Caim Pré-Disc | | | | | | | | | 0,0000 | 0,0000 |
| Cacc Pré-Disc | | | | | | | | | | 0,0000 |

Tabela 15 – Valores-p do Teste de Normalidade para a Base de Dados Cartão

| | Métodos Propostos | | | Métodos Existentes | | | | M.E. Com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Medidas | Information Statistics | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| Contínuo | 0,8019 | 0,0157 | 0,2690 | 0,9029 | 0,9585 | 0,8032 | 0,0665 | 0,9029 | 0,9585 | 0,8032 |
| Info.Statistics | | 0,0498 | 0,2471 | 0,1287 | 0,2499 | 0,3972 | 0,2209 | 0,1463 | 0,8038 | 0,2284 |
| Kendalls Tau-C | | | 0,2216 | 0,0798 | 0,9844 | 0,2826 | 0,1177 | 0,3883 | 0,6760 | 0,2744 |
| Multivariado | | | | 0,5316 | 0,2199 | 0,1376 | 0,0800 | 0,0922 | 0,4610 | 0,1926 |
| Caim | | | | | 0,0291 | 0,0000 | 0,4291 | 0,0000 | 0,0129 | 0,0000 |
| Cacc | | | | | | 0,3023 | 0,0928 | 0,0504 | 0,6081 | 0,2602 |
| Ameva | | | | | | | 0,5393 | 0,0431 | 0,3825 | 0,0191 |
| MDL | | | | | | | | 0,5255 | 0,9143 | 0,5217 |
| Caim Pré-Disc | | | | | | | | | 0,0007 | 0,0000 |
| Cacc Pré-Disc | | | | | | | | | | 0,0006 |

Tabela 16 – Valores-p do Teste de Normalidade para a Base de Dados Outros Produtos

| | Métodos Propostos | | | Métodos Existentes | | | | M.E. Com Pré-Discretização | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Medidas | Information Statistics | Kendalls Tau-C | Multivariado | Caim | Cacc | Ameva | MDL | Caim | Cacc | Ameva |
| Contínuo | 0,0072 | 0,1191 | 0,0002 | 0,0253 | 0,1442 | 0,0005 | 0,6337 | 0,0023 | 0,0000 | 0,0017 |
| Info.Statistics | | 0,2457 | 0,7012 | 0,0958 | 0,0437 | 0,08616 | 0,1410 | 0,3207 | 0,3009 | 0,2936 |
| Kendalls Tau-C | | | 0,8011 | 0,0430 | 0,1327 | 0,3423 | 0,9241 | 0,4623 | 0,4752 | 0,3212 |
| Multivariado | | | | 0,7300 | 0,7791 | 0,7692 | 0,3356 | 0,9766 | 0,4665 | 0,4183 |
| Caim | | | | | 0,2242 | 0,0001 | 0,5953 | 0,0273 | 0,2724 | 0,3240 |
| Cacc | | | | | | 0,0045 | 0,0127 | 0,3057 | 0,1341 | 0,0841 |
| Ameva | | | | | | | 0,0784 | 0,0001 | 0,0096 | 0,0070 |
| MDL | | | | | | | | 0,4590 | 0,6020 | 0,3750 |
| Caim Pré-Disc | | | | | | | | | 0,0588 | 0,0000 |
| Cacc Pré-Disc | | | | | | | | | | 0,0009 |