Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Métodos de aprendizado ativo

Luben Miguel Cruz Cabezas

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Métodos de aprendizado ativo

Luben Miguel Cruz Cabezas
Orientador: Rafael Izbicki

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Abril de 2022

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Active learning methods

Luben Miguel Cruz Cabezas
Advisor: Rafael Izbicki

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
November 2022

Luben Miguel Cruz Cabezas

Métodos de aprendizado ativo

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Luben Miguel Cruz Cabezas e aprovado pela banca examinadora.

Aprovado em 12 de abril de 2022

Banca Examinadora:

- Rafael Izbicki (orientador)
- Andressa Cerqueira
- Ricardo Felipe Ferreira

A meus avôs, Jesus Cruz e Ildefonso Cabezas

Agradecimentos

Agradeço a meu pai Luben Cabezas Gómez e minha mãe Isara Lourdes Cruz Hernandez por me darem a vida, amor incondicional, ensinamentos e grande incentivo aos estudos, ao meu irmão Alejandro Miguel, por todo o carinho, consideração e amizade, às minhas avós Estela e Mirta por todo o carinho e companhia, a todos meus amigos de faculdade, André Paulino, Luiz Piccin, Reinaldo, Vitor Schiavone, Vitor Ramos, Vinícius Hideki e Vinícius Souza, por todos os momentos passados na universidade (e fora dela), a meus amigos pessoais de São Carlos David, João Vitor e Gabriel por todos os anos de amizade, união e diversão, aos meus amigos de Belo Horizonte André Sá Alves e Samuel por todos os anos de amizade (desde a infância) e por fim, a meus dois orientadores Rafael Izbicki e Rafael Stern por me guiarem durante todo o caminho da graduação e serem grandes referências e verdadeiros mestres na minha formação.

"Você tem poder sobre sua mente – não sobre eventos externos.
Perceba isso e você encontrará sua força."
(Marco Aurélio)

Resumo

No campo de aprendizado supervisionado, a boa performance de um modelo de predição geralmente está atrelada à presença de um conjunto de treinamento rotulado grande. Existem, porém, situações em que a obtenção dos rótulos é custosa por motivos financeiros, de trabalho e/ou dificuldade, sendo proibitivo rotular todas as observações. Para essas situações, o uso do *aprendizado ativo* é fundamental. O *aprendizado ativo* se caracteriza por, a partir de diferentes métodos, escolher e incluir observações mais informativas ao conjunto de treinamento do modelo de predição, de forma que este tenha um bom desempenho com menos observações. Neste trabalho, estudamos alguns métodos de aprendizado ativo, como os de amostragem por incerteza, consulta por comitê, redução de erro esperada, redução de variância, ponderação por densidade e consulta por mini-lotes. Também propomos um novo método de regressão ativa com uma abordagem distinta das demais metodologias presentes na literatura. Como parte dos resultados, apresentamos um estudo de simulação para ilustrar como se dá o viés na amostragem em um algoritmo de aprendizado ativo. Finalmente, exploramos a nova metodologia de regressão ativa proposta, comparando-a à outras metodologias de regressão ativa e ao aprendizado passivo.

Palavras-chave: *aprendizado supervisionado, aprendizado ativo, rotulação, regressão ativa.*

Abstract

In the field of supervised learning, the good performance of a prediction model is generally tied to the presence of a large labeled training set. However, there are many situations where labeling an instance is expensive for financial, work and/or difficulty reasons, being prohibitive to label all observations. The use of *active learning* is crucial for these situations. *Active learning* is characterized by, through different methods, selecting and adding more informative instances to the training set of a prediction model, so that it performs well with fewer instances. In this work, we studied some active learning methods such as uncertainty sampling, query by committee, expected error reduction, variance reduction, density weighted methods and batch-mode active learning. We also propose a new active regression method with an approach different from other methods in the literature. As part of the results, we present an simulation study to illustrate how sampling bias occurs in active learning algorithms. Finally, we explore our new active regression methodology, comparing it to other active learning methodologies and passive learning.

Keywords: *supervised learning, active learning, labeling, active regression.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O aprendizado de máquinas (*Machine learning*) é um extenso campo pertencente à área da inteligencia artificial com o objetivo geral de aprender padrões com base em dados, tendo alta interseção com a área da estatística (Izbicki e dos Santos, 2020). Dentre os diferentes contextos e tipos de aprendizados, o *aprendizado supervisionado* é um dos mais comuns da área, tendo como principal enfoque a predição de um rótulo (variável dependente) com base em um conjunto de atributos (ou variáveis independentes) através de uma função $g(\cdot)$, com todos os rótulos sendo observados (Izbicki e dos Santos, 2020; Hastie *et al.*, 2009). O problema pode ser de regressão (rótulo quantitativo) ou de classificação (rótulo categórico), tendo diferentes modelos usados para cada caso (James *et al.*, 2013; Hastie *et al.*, 2009).

Geralmente, a boa performance de algoritmos de aprendizado supervisionado está atrelada à presença de um grande número de observações, com centenas ou até mesmo milhares de instâncias rotuladas (Settles, 2009). Tais rótulos usualmente têm um pequeno custo de obtenção, sendo geralmente possível a utilização de todas as observações para o treinamento e teste de diferentes modelos. Porém, existem muitas situações em que a obtenção dos rótulos é custosa por motivos financeiros, de trabalho e/ou dificuldade, como em problemas de reconhecimento de fala, extração de informação e classificação e filtragem de documentos (Zhu, 2005; Settles, 2009; Settles *et al.*, 2008). Para estes problemas, o uso do *aprendizado ativo* é fundamental.

O aprendizado ativo se caracteriza basicamente por, a partir de diferentes métodos, escolher quais observações incluir no treinamento do modelo de forma que este tenha um bom desempenho com um conjunto de treinamento menor (Settles, 2009). Ou seja, ao invés de sortearmos alguma nova observação para ser rotulada (aprendizado passivo),

escolhemos uma instância mais estratégica com base em diferentes medidas de informatividade.

1.1 Objetivos

Assim, na busca de explorar os métodos e procedimentos associados a essa área para entender suas diferentes aplicações, surge o objetivo principal deste trabalho:

- Estudar Aprendizado ativo em problemas de regressão e classificação, como uma forma de abordagem alternativa para certos problemas na área de aprendizado supervisionado que tenham possíveis entraves na rotulação.

Para atingir o objetivo fundamental do trabalho, se propõem os seguintes objetivos específicos desse trabalho:

- Comparar diferentes métodos presentes na literatura de aprendizado ativo, entendendo a vantagem e desvantagem de cada um em diferentes situações.
- Comparar também o aprendizado passivo e o aprendizado ativo no contexto do aprendizado ativo, identificando benefícios e malefícios de cada aplicação nesse contexto.
- Aplicar e comparar métodos de aprendizado ativo no contexto de regressão.

Capítulo 2

Revisão bibliográfica

Nesta seção, serão apresentadas informações sobre as principais metodologias e procedimentos utilizados em aprendizado ativo. Todas as metodologias expostas a seguir e outros aspectos interessantes sobre aprendizado ativo são explorados em Settles (2009, 2012) e nos basearemos em tais trabalhos para o estudo do aprendizado ativo. Primeiramente, a estrutura básica de um problema de aprendizado supervisionado se caracteriza pela presença de um conjunto de dados com $n \in \mathbb{N}^*$ observações i.i.d (independentes e identicamente distribuídos) $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, tal que $\boldsymbol{X}_i \in \mathbb{R}^p$, em que se tem o interesse de obter uma função $g : \mathbb{R}^d \to \mathbb{R}$ com bom poder preditivo, ou seja, dada novas observações i.i.d $(\boldsymbol{X}_{n+1}, Y_{n+1}), \ldots, (\boldsymbol{X}_{n+m}, Y_{n+m})$ para $m > n$, $m \in \mathbb{N}^*$, tenhamos (Izbicki e dos Santos, 2020):

$$g(\boldsymbol{x}_{n+1}) \approx y_{n+1}, \ldots, g(\boldsymbol{x}_{n+m}) \approx y_{n+m}\,.$$

Para medir o desempenho da função de predição $g(\cdot)$, define-se uma função de perda $L(g(\boldsymbol{X}; Y))$ e o risco $R(g) = \mathbb{E}\left[L(g(\boldsymbol{X}; Y))\right]$, buscando-se a função de predição g que minimize o risco de interesse $R(g)$ (Izbicki e dos Santos, 2020). Para a seleção de modelos, usualmente divide-se o conjunto de dados em treinamento e teste, estimando k diferentes regressores ou classificadores $g_l(\cdot)$, $l \in \{1, \ldots, k\}$ no mesmo conjunto de treinamento $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_s, Y_s)$ para $s < n$, $s \in \mathbb{N}^*$, e utilizando o conjunto de teste $(\boldsymbol{X}_{s+1}, Y_{s+1}), \ldots, (\boldsymbol{X}_n, Y_n)$ para estimar o risco definido através do estimador:

$$\hat{R}(g_l) \coloneqq \frac{1}{n-s} \sum_{i=s+1}^{n} L\left(g_l(\boldsymbol{X}_i; Y_i)\right)\,,$$

para validação do modelo. A função de perda, por exemplo, pode ser definida pela perda quadrática $L(g(\boldsymbol{X}; Y)) \coloneqq (g(\boldsymbol{X}) - Y)^2$ no caso de regressão ou pela perda 0-1 $L(g(\boldsymbol{X}; Y)) \coloneqq \mathbb{I}(Y \neq g(\boldsymbol{X}))$ no caso de classificação, tendo diversas outras funções de perda para cada contexto. Alternativamente ao *data spliting*, pode-se utilizar outras técnicas de validação cruzada, como *K-fold* ou *leave-one-out* para a validação de cada modelo. Salienta-se que nesse contexto, todas as observações são rotuladas.

Para um problema de aprendizado ativo, considera-se a presença de um pequeno conjunto de observações $\mathcal{L}$ já rotulados para treinamento inicial e um conjunto $\mathcal{T}$ também rotulado fixado para teste (em certas abordagens apenas $\mathcal{L}$ está presente). A utilização de um conjunto de teste ao invés da aplicação direta de validação cruzada no conjunto $\mathcal{L}$ é para se evitar um possível *covariate shift*, ou seja, as observações de teste não serem mais i.i.d em relação à distribuição de interesse, visto que cada adição de novas observações rotuladas a $\mathcal{L}$ modifica sempre o conjunto de teste e sua distribuição. O conjunto não rotulado $\mathcal{U}$ também denominado de *pool* é o alvo das principais técnicas e cenários de amostragem do aprendizado ativo. Tal conjunto é abordado em dois principais cenários (Settles, 2009):

- **Amostragem seletiva sequencial** (*stream-based*) (Cohn *et al.*, 1994): Nesse cenário, cada instância não rotulada é sequencialmente selecionada, e é decidido através de alguma estratégia de consulta se a observação é rotulada ou descartada. As observações podem ser amostradas de forma "online", ou seja, retiradas continuamente de uma fonte de dados não estática (não necessariamente $\mathcal{U}$ é um conjunto estático nesse cenário). A decisão de rotular cada observação é orientada de acordo com medidas de "informatividade", de forma que as observações amostradas para a rotulação sejam o mais informativas possíveis (Dagan e Engelson, 1995; Settles, 2009).

- **Aprendizado ativo *pool-based*** (Lewis e Gale, 1994): O cenário *pool-based* é direcionado principalmente para cenários de grande bases de dados já coletadas e estáticas, apesar de não haver total necessidade dessa última propriedade. Nessa abordagem, a amostragem das instâncias para consulta é feita de uma maneira mais ambiciosa, avaliando-se e ranqueando todas as instâncias de $\mathcal{U}$ ao mesmo tempo de acordo com alguma medida de informatividade. Subsequentemente seleciona-se a "melhor consulta" do conjunto $\mathcal{U}$, rotulando-se tal instância. Tal procedimento

avaliativo é repetido de forma iterativa de acordo com um limiar ou número de consultas fixas.

Em geral, o cenário *stream-based* percorre e rotula o banco de dados de forma sequencial, decidindo a rotulação (ou não) das instâncias de forma individual, enquanto o *pool-based* avalia e ranqueia todas as instâncias ao mesmo tempo antes de selecionar a melhor consulta de forma iterativa (Settles, 2009). As medidas de informatividade descritas em cada cenário são um dos principais tópicos de pesquisa na área de aprendizado ativo, buscando a partir de medidas de incerteza e/ou ambiguidade selecionar observações mais decisivas para o treinamento dos modelos, de forma que ainda se evite o super ajuste. Tais estratégias podem variar de acordo com a natureza do problema, caso seja tanto classificação quanto regressão. A seguir serão detalhadas as principais estratégias de consulta presentes na literatura de aprendizado ativo. Para cada estratégia de consulta, definimos $\boldsymbol{x}_a^*$ como a observação consultada pelo método a, e a probabilidade $P(y = c|\boldsymbol{x}; \hat{g})$ como a probabilidade estimada pelo modelo $\hat{g}$ de que $y = c$ dado $\boldsymbol{x}$, tal que $c \in \mathcal{A}$, com $\mathcal{A}$ sendo o conjunto de rótulos definido para cada problema.

2.1 Amostragem por incerteza

A estratégia de amostragem por incerteza (Lewis e Catlett, 1994) tem como principal premissa consultar instâncias com maior incerteza associada a rotulação, evitando consultar observações mais "óbvias"que se tem mais certeza sobre sua rotulação. Em outras palavras, tem-se interesse nas observações mais próxima da fronteira de decisão (Settles, 2012). Traduze-se tal intuição de forma matemática em um problema de classificação binária ($y \in \{c_1, c_2\}$) como a seguir: seja $\hat{g}$ um classificador binário probabilístico arbitrário, nosso maior interesse seria encontrar a instância $\boldsymbol{x}$ tal que a probabilidade posterior estimada $P(y = c_1|\boldsymbol{x}; \hat{g})$ se aproxime de 0.5, ou seja:

$$\boldsymbol{x}_b^* = \arg\min_{\boldsymbol{x}} |0.5 - P(y = c_1|\boldsymbol{x}; \hat{g})| \ . \tag{2.1}$$

Assim, para classificação binária, o quão perto $\boldsymbol{x}$ está da fronteira de decisão é uma estratégia boa para a seleção de instâncias informativas ao modelo (Settles, 2012, 2009). A Figura 2.1 ilustra a estratégia de amostragem por incerteza no caso binário para um exemplo simulado parecido ao mostrado em Settles (2009, 2012).

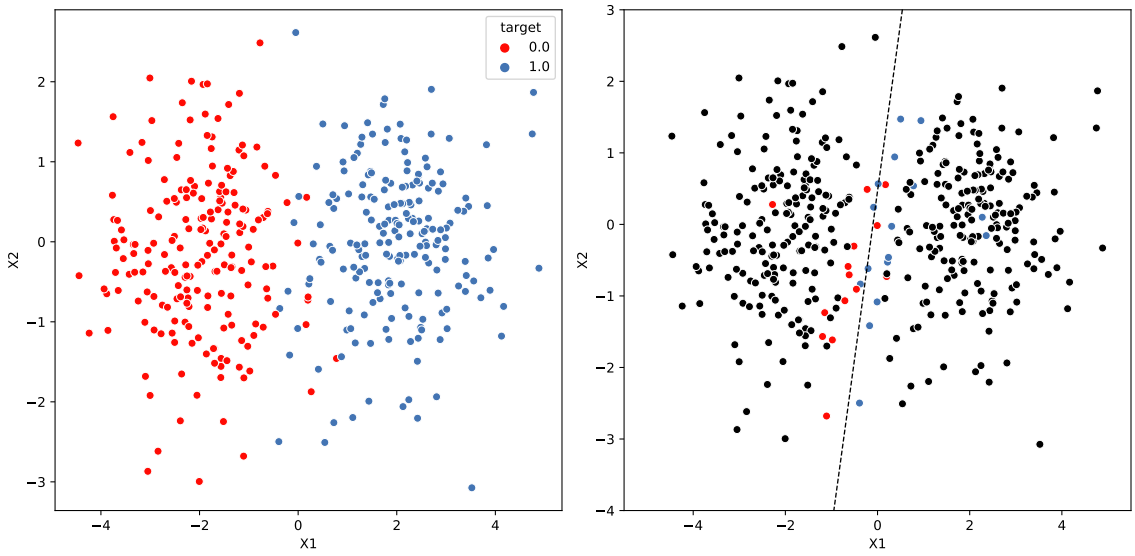


Figura 2.1: Amostragem por incerteza em um banco de dados simulado por 2 normais bivariadas de tamanho 200. A esquerda temos a distribuição dos dados simulados e a separação entre as classes. À direita temos classificadas apenas as observações selecionadas pelo amostrador. Nota-se que as observações amostradas em geral estão muito próximas da fronteira de decisão e são as mais ambíguas dentre todas as demais observações

Porém, para casos de classificação multi-classe ou de classes estruturadas há a necessidade de se introduzir medidas de incerteza generalizadas. Uma primeira abordagem de amostragem por incerteza mais básica é dada pela seleção da observação cuja classe predita é a **menos confiante** (Settles, 2009):

$$\boldsymbol{x}^*_{mc} = \arg\min_{\boldsymbol{x}} P(y^*|\boldsymbol{x}; \hat{g}) \,, \tag{2.2}$$

em que $y^* = \arg\max_y P(y|\boldsymbol{x}; \hat{g})$ é a classe com maior probabilidade estimada sob o modelo de classificação estimado $\hat{g}$. Em palavras, tal estratégia tende a amostrar a instância cujo rótulo mais provável é o menos provável em comparação aos demais preditos para o restante das instâncias não rotuladas. A Equação (2.2) pode também ser escrita como uma maximização:

$$\boldsymbol{x}^*_{mc} = \arg\max_{\boldsymbol{x}} \left[1 - P(y^*|\boldsymbol{x}; \hat{g})\right] \,. \tag{2.3}$$

Apesar da simplicidade e intuitividade de tal medida, esta apenas considera a melhor predição de cada instancia de $\mathcal{U}$, potencialmente perdendo muita informação sobre as outras possíveis rotulações e suas probabilidades. Respondendo a tal limitação, outra

estratégia interessante de se utilizar se baseia na **margem** das possíveis predições:

$$\boldsymbol{x}_m^* = \arg\min_{\boldsymbol{x}} \left[P(y_1^*|\boldsymbol{x}; \hat{g}) - P(y_2^*|\boldsymbol{x}; \hat{g}) \right] , \tag{2.4}$$

em que, seguindo a notação usada na Equação (2.2), y_1^* e y_2^* são respectivamente a primeira e segunda mais prováveis classes sob o modelo estimado $\hat{g}$. Em palavras, a amostragem pela **margem** seleciona a instância com menor margem entre a classe mais provável e a segunda mais provável, tendo tais observações como ambíguas, visto que podem ser facilmente rotuláveis tanto como uma classe quanto pela outra classe. Semelhantemente ao método do menos confiante, pode-se também escrever a Equação (2.4) como uma maximização:

$$\boldsymbol{x}_m^* = \arg\max_{\boldsymbol{x}} \left[P(y_2^*|\boldsymbol{x}; \hat{g}) - P(y_1^*|\boldsymbol{x}; \hat{g}) \right] . \tag{2.5}$$

Por incluir a segunda classe mais provável no cálculo da incerteza sobre $\boldsymbol{x}$, tal abordagem trata bem da limitação do método do **menos confiante**, mas possui problemas caso o número de classes seja relativamente grande, pois se ignora a probabilidade associada a maior parte das classes. Para se obter uma medida ainda mais genérica, pode-se utilizar a amostragem por **entropia** (Shannon, 1948), fazendo uso de toda a distribuição de probabilidade estimada para as classes:

$$\boldsymbol{x}_e^* = \arg\max_{\boldsymbol{x}} \left(- \sum_i P(y_i|\boldsymbol{x}; \hat{g}) \log \left(P(y_i|\boldsymbol{x}; \hat{g}) \right) \right) , \tag{2.6}$$

tal que y_i varia de acordo com todos seus possíveis rótulos. A expressão $-\sum_i P(y_i|\boldsymbol{x}; \hat{g}) \log \left(P(y_i|\boldsymbol{x}; \hat{g}) \right)$ também pode ser escrito como a função de entropia $H(Y|\boldsymbol{x}; \hat{g})$. Tal medida de importância pode ser pensada como uma medida de impureza associada a observação $\boldsymbol{x}$ (Settles, 2012). É interessante notar que para um problema de classificação binário, todas as medidas descritas se resumem a consultar a instância mais próxima da fronteira entre as duas classes como mostrado pela equação (2.1). As diferenças entre as diferentes abordagens é mais clara em problemas multi-classe ou de classe estruturada, tendo uma diferença na distribuição das regiões de incerteza. Pode-se enxergar tais diferenças em um problema com três classes através da configuração de cada função em gráficos ternários como apresentada na Figura 2.2 (adaptada de Settles (2012)), fixando-se em dois eixos as

possíveis probabilidades associadas as classes 1 e 2, com a limitação de que $p_1 + p_2 \leq 1$ e sabendo também que $p_3 = 1 - (p_1 + p_2)$.

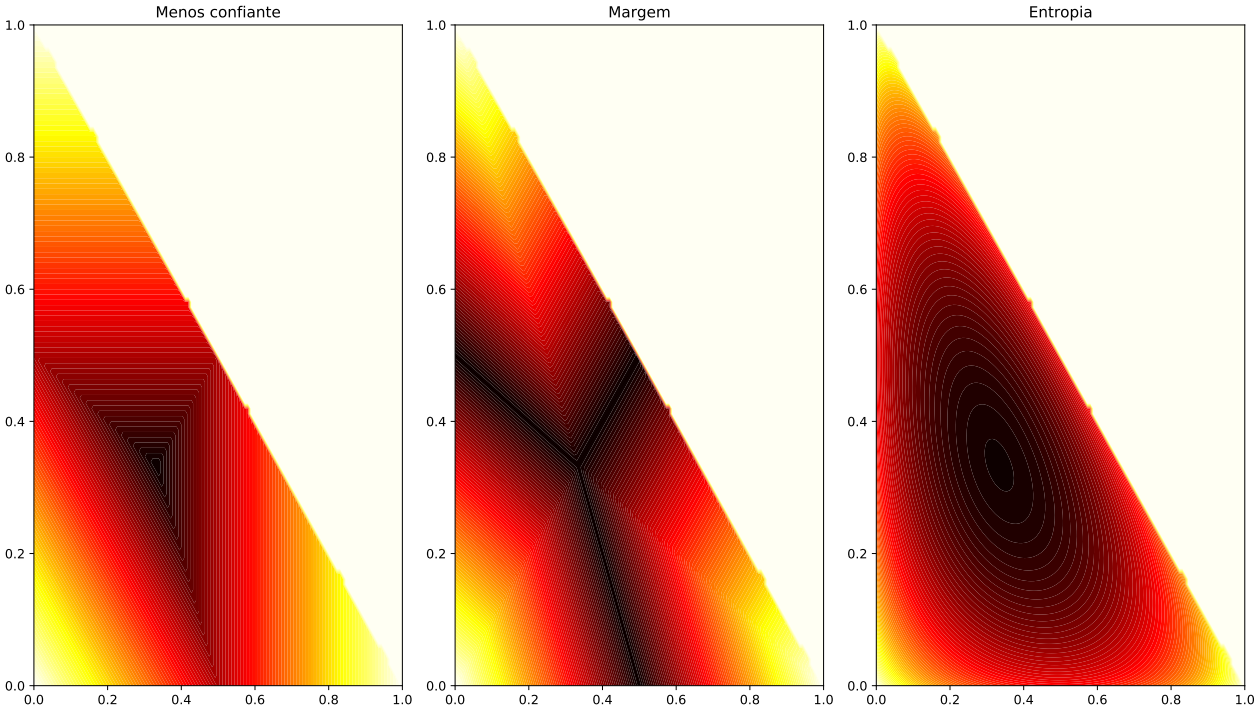


Figura 2.2: Gráficos ternários para cada método. A cor representa o valor de cada função a ser maximizada em cada método. Quanto mais escura, maior o valor da função, logo mais informativa a instancia.

Percebe-se pela Figura 2.2 que as regiões de informatividade de cada método têm a instancia mais informativa localizada mais ao centro de cada região triangular, que é onde a distribuição de probabilidade dos rótulos é mais uniforme (Settles, 2012), enquanto que as instâncias menos informativas localizam-se nas pontas dos triângulos que são onde uma das classes tem alta probabilidade associada. As maiores diferenças em geral se dão pelo fato da entropia ter uma região mais suave, não penalizando tanto a região próxima ao centro do triângulo, enquanto o menos confiante tem uma região mais centralizada na área com maior incerteza, com alta penalização nas vizinhanças do centro do triangulo. O método da margem forma uma região de alta incerteza em 3 linhas, cada uma delas tendo ao menos duas classes com probabilidades próximas, com certa penalização fora de cada linha.

Um grande problema da amostragem por incerteza, que será melhor discutido na Seção 2.7, é o fato da consulta e amostragem de observações ser muito dependente do modelo escolhido, e muitas vezes tal modelo pode não se encaixar bem aos dados. Assim, há um enviesamento da amostragem de observações de acordo com o modelo escolhido, com a

obtenção de observações consideradas ambíguas de acordo com fronteiras erroneamente ajustadas em comparação ao comportamento real das classes.

2.1.1 Adaptação para regressão

Em geral, as estratégias de incerteza até agora discutidas apenas abordavam problemas de classificação. A amostragem por incerteza também pode ser aplicada em problemas de regressão ao amostrar a instancia com maior variância de predição em um ajuste por exemplo, processos gaussianos, em que é possível obter uma estimativa em fórmula fechada da variância de cada predição (Settles, 2012). Outros modelos podem ser usados desde que seja possível tal estimativa. Sob uma suposição gaussiana, a entropia de uma variável aleatória é uma função monótona de sua variância (Settles, 2012), ou seja, essa abordagem é de fato uma extensão da amostragem por incerteza baseada em entropia para um problema de regressão. A Figura 2.3 ilustra a amostragem por incerteza aplicada em problemas de regressão em dados simulados.

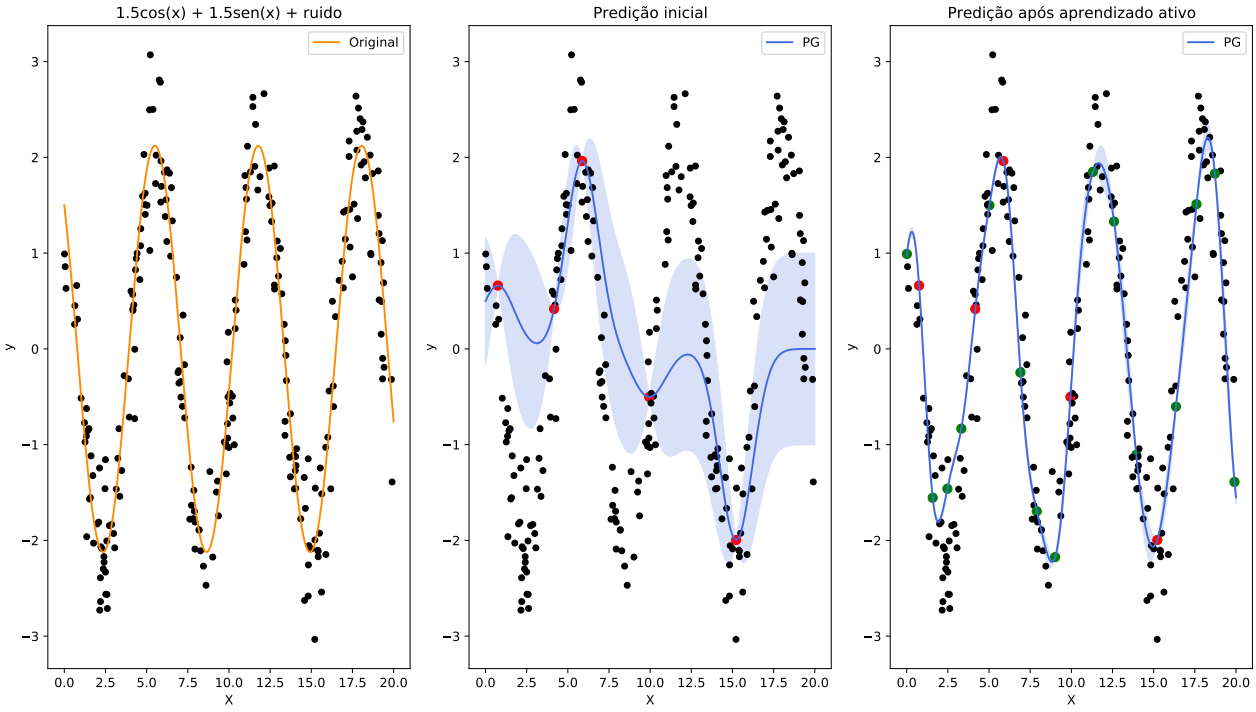


Figura 2.3: Aprendizado ativo em regressão usando um processo gaussiano. Os pontos coloridos de vermelho são a amostra de treinamento inicial, enquanto os pontos verdes são os pontos selecionados pelo aprendizado ativo com um total de 15 consultas.

É interessante notar a alta variância do processo gaussiano ajustado inicialmente e sua consequente diminuição ao amostrar algumas observações através de uma regressão ativa. Um problema dessa abordagem de regressão ativa é o fato de a estimativa das variâncias

de cada predição sem informações da variável resposta y_i ser apenas possível de se obter em uma pequena coleção de modelos (como por exemplo, para processos gaussianos e redes neurais) e ser relativamente complexa. Na Seção 2.2 apresentaremos um método de regressão ativa com tática de *ensemble* que generaliza mais as possíveis classes de modelo a serem empregadas em uma regressão ativa.

2.1.2 Algoritmo

Em geral, a amostragem por incerteza em um contexto **pool-based** (ou seja, com consultas feitas em $\mathcal{U}$ fixo iterativamente) é feita a partir do Algoritmo 1 (Settles, 2012).

Algoritmo 1 Amostragem por incerteza

1: $\mathcal{U}$ = conjunto de instâncias não rotuladas $\{x_i\}_{i\in\mathcal{U}}$
2: $\mathcal{L}$ = conjunto inicial de instâncias rotuladas $\{(x_i, y_i)\}_{i=1}^{n}$
3: k = Número escolhido de consultas
4: **for** $t = 1, 2, \ldots, k$ **do**
4: g = **ajuste**($\mathcal{L}$)
4: selecione $x^* \in \mathcal{U}$ a instância mais incerta de acordo com o modelo g
4: obtenha y^* do oráculo
4: adicione (x^*, y^*) a $\mathcal{L}$
4: remova x^* de $\mathcal{U}$
5: **end for**

Pode-se também avaliar dentro do laço *for* o desempenho do modelo a cada iteração, ajustando as covariáveis de teste ao modelo treinado g e utilizando diferentes métricas (acurácia, precisão, revocação, F_1, etc) para a validação do modelo, criando-se um histórico de desempenho do modelo.

2.1.3 Cenário *stream-based*

A abordagem de amostragem por incerteza também é comumente empregada em um cenário *stream based* com uma leve alteração em comparação ao cenário mais comum *pool based*, selecionando instâncias não rotuladas uma por vez de um conjunto de dados não necessariamente estático $\mathcal{D}$.A maneira mais simples de se utilizar tal procedimento em uma aplicação *stream-based* é definir um limiar para a medida de incerteza pela qual se definira uma **região de incerteza** (Settles, 2009).Assim, sequencialmente, uma instância é selecionada se tiver uma incerteza superior ao limiar fixado (ou seja, dentro da região de incerteza) e descartada caso contrário, tendo o modelo re-treinado para cada nova instância consultada e adicionada ao conjunto $\mathcal{L}$. Em tal abordagem pode-se pré-fixar

um número de instâncias a serem consultadas (assim como em uma abordagem *pool-based*) ou um limiar de certa medida de performance (acurácia, revocação, F_1, entre outros).

A Figura 2.4 mostra um exemplo de aprendizado ativo em um cenário *stream based* com um exemplo de uma figura dada por um quadrado preto. Percebe-se que após 15 consultas, o formato da imagem previsto já se assemelha ao formato inicial, com uma região de incerteza relativamente curta.

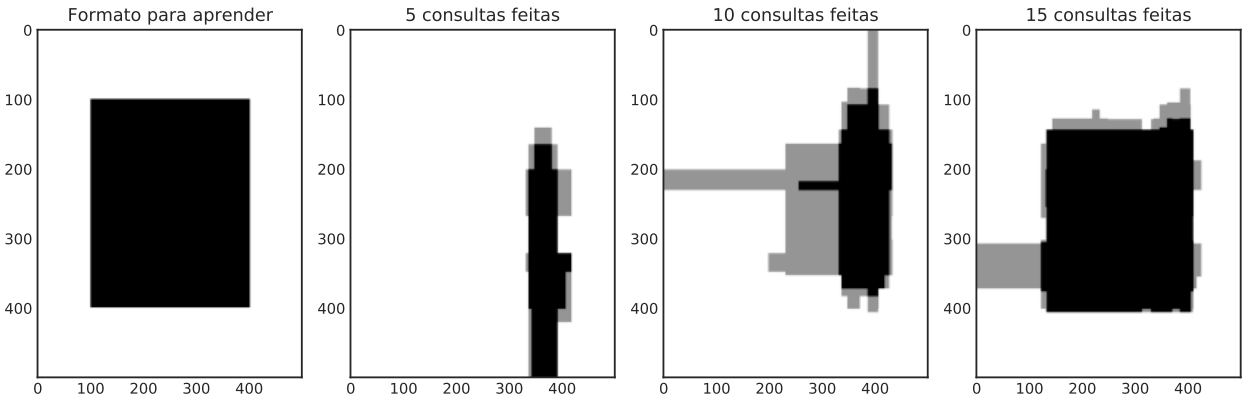


Figura 2.4: Aprendizado ativo em um cenário *stream based* para prever o formato do primeiro gráfico utilizando uma floresta aleatória e um limiar fixo de 0.4 (ou seja, se a incerteza associada a instância selecionada estiver acima de 0.4, tal instância é selecionada para consulta). A região acinzentada é a região de incerteza associada ao modelo em cada etapa correspondente do aprendizado ativo.

Em geral, assume-se que D é um conjunto representativo da distribuição das instâncias $\boldsymbol{X}$, tal que a seleção de cada instância $\boldsymbol{x}$ para consulta ou descarte é feita de forma i.i.d a tal distribuição (Settles, 2012).

Um grande problema da amostragem por incerteza, que será melhor discutido na Seção 2.7, é o fato da consulta e amostragem de observações ser muito dependente do modelo escolhido, e muitas vezes tal modelo pode não se encaixar bem aos dados. Assim, há um enviesamento da amostragem de observações de acordo com o modelo escolhido, com a obtenção de observações consideradas ambíguas de acordo com fronteiras erroneamente ajustadas em comparação ao comportamento real das classes.

2.2 Consulta por comitê

Uma estratégia de aprendizado ativo mais robusta ao viés de um único modelo de predição é a consulta por comitê (Seung *et al.*, 1992), que consiste em manter um comitê de diferentes modelos $\mathcal{C} = \{\hat{g}_1, \ldots, \hat{g}_k\}$ treinados no conjunto $\mathcal{L}$. Na prática há uma

grande variedade de modelos possíveis a serem utilizados em conjunto (Settles, 2012), empregando-se mais comumente algoritmos de *ensemble* genéricos devido ao comportamento de comitê já naturalmente presente nesses modelos (Abe, 1998). Salienta-se que não existe regra sobre o número de membros do comitê, variando comumente de 2 a 15 (Settles, 2012). A ideia desse procedimento é de que cada modelo g_i do comitê vota em um dado rótulo para cada instância de $\mathcal{U}$, e a observação mais informativa é aquela pela qual os modelos do comitê mais discordam quanto a rotulação.

Para se atribuir uma medida ao nível de discordância entre os membros do comitê pode-se utilizar de diferentes heurísticas disponíveis. Uma das mais comuns é a **entropia de voto** (Dagan e Engelson, 1995), que é uma generalização da medida de incerteza baseada em entropia, definida da seguinte forma:

$$\boldsymbol{x}^*_{EV} = \arg\max_{\boldsymbol{x}} \left(-\sum_i \frac{V_{\mathcal{C}}(y_i, \boldsymbol{x})}{|\mathcal{C}|} \log \frac{V_{\mathcal{C}}(y_i, \boldsymbol{x})}{|\mathcal{C}|} \right), \tag{2.7}$$

com y_i variando de acordo com todos seus possíveis rótulos, tal que $V_{\mathcal{C}} \coloneqq \sum_{g \in \mathcal{C}} \mathbb{I}(\hat{g}(\boldsymbol{x}) = y_i)$ é o número de votos que o rótulo y_i recebe para $\boldsymbol{x}$ fixado entre todas as hipóteses do comitê $\mathcal{C}$ e $|\mathcal{C}|$ é o tamanho do comitê. Essa formulação é uma entropia de voto mais "dura"(Settles, 2012), visto que é feita uma contabilização de votos para cada rótulo através do número de membros que têm tal rótulo com maior probabilidade associada. Uma definição mais "flexibilizada" de **entropia de voto**, a **entropia de consenso** é definida como a seguir:

$$\boldsymbol{x}^*_{EC} = \arg\max_{\boldsymbol{x}} \left(-\sum_i P_{\mathcal{C}}(y_i|\boldsymbol{x}) \log \left(P_{\mathcal{C}}(y_i|\boldsymbol{x}) \right) \right), \tag{2.8}$$

tal que $P_{\mathcal{C}}(y_i|\boldsymbol{x}) = \frac{1}{|\mathcal{C}|} \sum_{g \in \mathcal{C}} P(y_i|\boldsymbol{x}; \hat{g})$ é a probabilidade média ou de "consenso"de que y_i é a rotulação correta de acordo com o comitê (Settles, 2012). Em palavras, a instância mais ambígua é selecionada pela primeira medida de acordo com a distribuição dos votos para cada classe dada pelo comitê (entropia de voto mais "dura"), e pela segunda através da média das probabilidades associadas a cada classe obtidas por cada modelo do comitê (entropia de voto mais "flexível"). Outra medida de discordância também comumente utilizada é baseada na divergência de Kullback-Leibler (Kullback e Leibler, 1951), uma medida de informação que quantifica a diferença entre duas distribuições de probabilidade. Nesse caso, temos interesse em quantificar a divergência média da predição de cada

membro do comitê $\hat{g}_i$ com relação a predição do consenso $\mathcal{C}$ (Settles, 2012) e em escolher a observação que nos dê o máximo de discordância possível:

$$\boldsymbol{x}^*_{KL} = \arg\max_{\boldsymbol{x}} \frac{1}{|\mathcal{C}|} \sum_{g \in \mathcal{C}} KL\left(P(Y|\boldsymbol{x}; \hat{g}) || P(Y|\boldsymbol{x}; \mathcal{C})\right) , \tag{2.9}$$

em que a divergência KL é definida como:

$$KL\left(P(Y|\boldsymbol{x}; \hat{g}) || P(Y|\boldsymbol{x}; \mathcal{C})\right) = \sum_i P(y_i|\boldsymbol{x}; \hat{g}) \log\left(\frac{P(y_i|\boldsymbol{x}; \hat{g})}{P(y_i|\boldsymbol{x}; \mathcal{C})}\right) . \tag{2.10}$$

Pode-se detalhar as diferenças em como a entropia de voto e a divergência KL quantificam a discordância a partir do exemplo dado pela Figura 2.5 utilizado por Settles (2012).

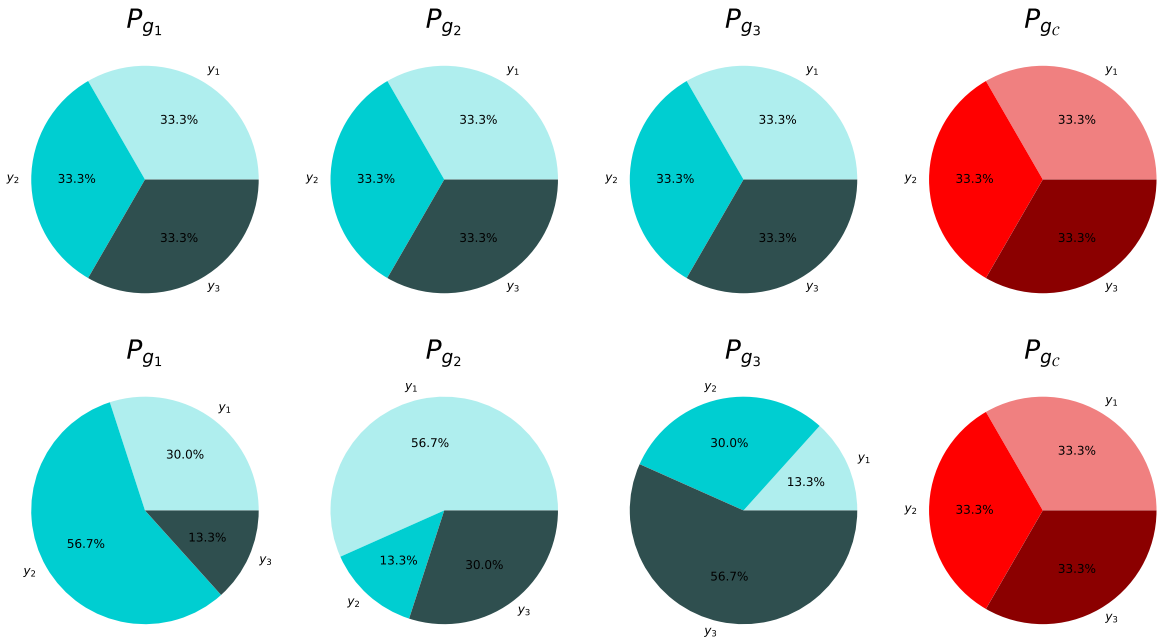


Figura 2.5: Exemplos de distribuições P_{g_i} do comitê e a distribuição $P_{\mathcal{C}}$ de consenso resultante nos rótulos hipotéticos (y_1, y_2, y_3) em dois casos diferentes, primeira linha com todas as probabilidade homogêneas, e segunda linha com homogeneidade na distribuição de consenso e heterogeneidade nas distribuições dos modelos g_i do comitê.

Percebemos primeiramente uma homogeneidade nas distribuições de probabilidade dada pelo caso da primeira linha, já que como as distribuições de cada modelo g_i são uniformes, a distribuição de consenso também será uniforme. Já a segunda linha ilustra um contexto diferente, em que cada modelo g_i do comitê possuem distribuições diferentes e não uniformes, cada uma com certa preferência para certo rótulo, mas a distribuição de consenso é uniforme para todos os rótulos. A entropia de consenso (Equação 2.8)

consideraria diretamente a probabilidade $P_{\mathcal{C}}$ que é homogênea e portanto não distinguiria as duas situações apresentadas. É interessante notar porém que os modelos da primeira linha todos têm ampla concordância quanto a rotulação de y ser incerta, dado que todos atribuem probabilidade uniforme para cada possível rotulação. Sendo assim, esse contexto não se encaixa bem na premissa da consulta por comitê: selecionar observações com base na ampla discordância dos modelos do comitê. A divergência KL (Equação (2.9)) favoreceria mais o contexto da segunda linha: adicionalmente ao fato do consenso ser incerto, cada modelo tem grande variedade de possíveis predições, o que se encaixa mais na natureza da consulta por comitê (Settles, 2012). Assim, a divergência KL é uma heurística mais sensível às distribuições de cada modelo do comitê do que as medidas de entropia de voto, apesar ser um pouco mais complexa. Um exemplo mais concreto e geral de comparação entre diferentes heurísticas de aprendizado ativo também utilizado e idealizado por Settles (2012) é dado pela Figura 2.6.

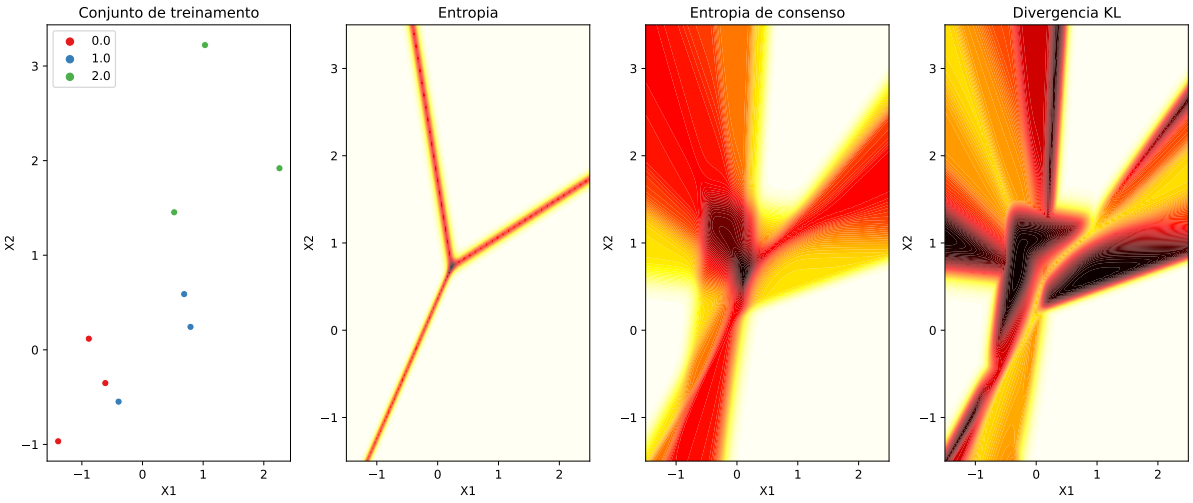


Figura 2.6: Visualizando diferentes heurísticas de aprendizado ativo, para um conjunto de treinamento pequeno simulado com 3 classes. Para a amostragem por incerteza baseada em entropia usou-se um único classificador de regressão logística multinomial e para as entropias de consenso e divergência KL utilizou-se um comitê de 10 classificadores de regressão logística ajustados com *bagging* (Breiman, 1996). Cada mapa de calor mostra as regiões de incerteza de acordo com cada heurística, quanto mais escura a região, mais incerta ela é.

Ajustando-se um único modelo ao conjunto de treinamento mostrado e medindo a incerteza do classificador através da entropia (Equação (2.6)) observa-se grande confiança do modelo na maioria do espaço de atributos, com uma região de incerteza mais estreita em formato de "Y"localizada nas fronteiras entre cada classe. O terceiro e quarto gráfico nos mostram as regiões de incerteza obtidas para o comitê especificado na Figura 2.6 ao

utilizar entropia de consenso (Equação (2.8)) ou divergência KL (Equação (2.9)). Nota-se nos dois gráficos o mesmo formato de "Y"da região de maior incerteza, tendo porém uma maior suavidade na vizinhança e uma menor confiança sobre o espaço de atributos em comparação a amostragem por incerteza baseada em entropia.

Destaca-se também que, apesar de os dados de treinamento e o comitê criado serem os mesmos, os gráficos das medidas de incerteza para a consulta por comitê têm comportamentos razoavelmente diferentes: a região da divergência KL tem trechos de grande incerteza fora do centro e nas fronteiras da região de incerteza, enquanto a entropia de consenso tem grande incerteza mais ao centro da região de incerteza, com valores de incerteza mais atenuados no restante da região.

2.2.1 Adaptação para regressão

Além dessas duas medidas, Settles (2012) destaca outras medidas de discordância para a consulta por comitê existentes na literatura, como por exemplo a divergência de **Jensen-Shannon** (Melville *et al.*, 2005) (uma versão simétrica e mais suavizada da divergência KL).

Assim como para amostragem por incerteza, a consulta por comitê também pode ser aplicada em problemas de regressão, medindo a discordância entre os membros através da variância entre as predições obtidas por cada modelo do comitê. A Figura 2.7 nos mostra os dados simulados a serem modelados pela regressão ativa por comitê.

Pode-se ajustar um comitê formado por três processos gaussianos com *bagging* em um conjunto de treinamento inicial de tamanho 10 e posteriormente consultando 10 observações. Os ajustes separados de cada modelo do comitê e do comitê inteiro para o conjunto de treino inicial e após 10 consultas é dado pelo Figura 2.8. Observa-se uma grande diminuição de variância no modelo conjunto após as 10 consultas, tendo os 3 membros do comitê mais "em acordo"com relação ao formato do oráculo $|X|$ após as 10 consultas em comparação ao conjunto de treinamento inicial. Um problema existente nesse tipo de abordagem é que caso a classe de modelos especificada é impropriadamente viesada ou má especificada (Settles, 2012) a amostragem aleatória será preferível à consulta por comitê baseada em variância da combinação de modelos (Burbidge *et al.*, 2007).

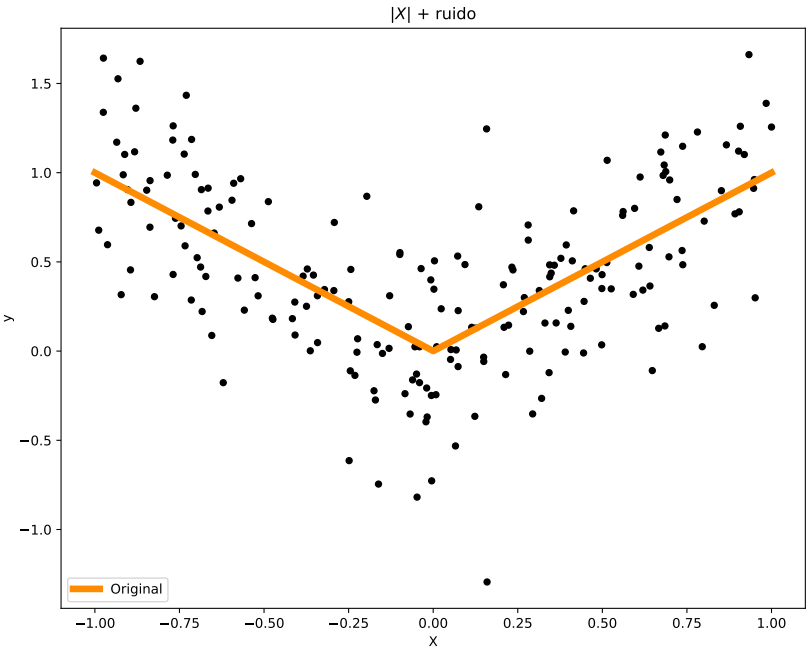


Figura 2.7: Dados simulados para ajuste de uma regressão ativa com $E[Y|X] = |X|$

2.2.2 Algoritmo

O algoritmo para a consulta por comitê em um cenário *pool-based* é semelhante ao algoritmo de consulta por amostragem por incerteza, com as maiores diferenças se dando no ajuste de um conjunto de modelos $\mathcal{C}$ no conjunto de treinamento $\mathcal{L}$ e na seleção de $\boldsymbol{x}^*$ através das heurísticas de discordância abordadas nessa seção. O Algoritmo 2 ilustra essas modificações, verificando-se a necessidade de outro laço *for* para percorrer o comitê de modelo $\mathcal{C}$.

Algoritmo 2 Consulta por comitê

1: $\mathcal{U}$ = conjunto de instâncias não rotuladas $\{\boldsymbol{x}_i\}_{i\in\mathcal{U}}$
2: $\mathcal{L}$ = conjunto inicial de instâncias rotuladas $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$
3: k = Número escolhido de consultas
4: G = Vetor de tamanho $\mathcal{C}$ para armazenar o ajuste de cada modelo g_i
5: **for** $t = 1, 2, \ldots, k$ **do**
6: **for** $c = 1, 2, \ldots, |\mathcal{C}|$ **do**
6: G[c] = **ajuste**($\mathcal{L}$, modelo = g_c)
7: **end for**
7: selecione $\boldsymbol{x}^* \in \mathcal{U}$ a instância mais incerta de acordo com o comitê G
7: obtenha y^* do oráculo
7: adicione $(\boldsymbol{x}^*, y^*)$ a $\mathcal{L}$
7: remova $\boldsymbol{x}^*$ de $\mathcal{U}$
8: **end for**

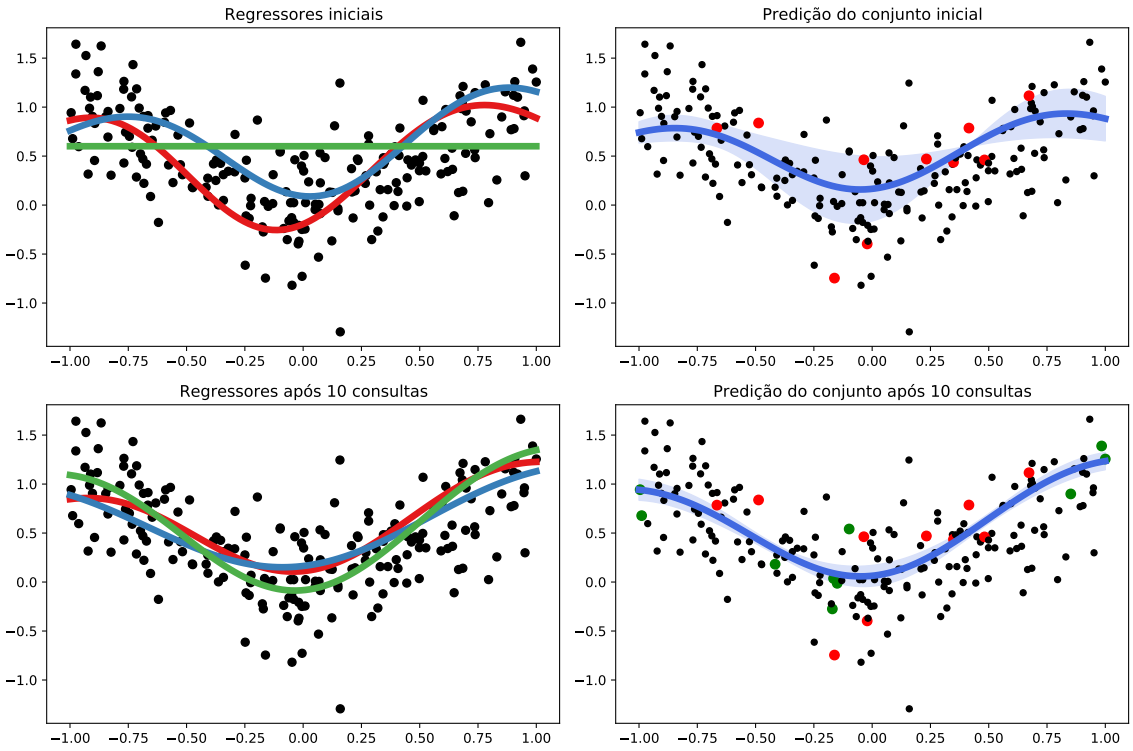


Figura 2.8: Regressão ativa por comitê (PG´s individuais do comitê ajustados ao lado esquerdo, comitê completo ajustado ao lado direito). Os pontos em vermelho são as observações de treinamento inicial e os verdes são as observações amostradas e consultadas durante o aprendizado ativo.

2.3 Redução de erro esperada e redução de variância

Até agora, nos dois métodos analisados, identificávamos a melhor instância para consulta com base no quão incertas e disruptivas estas eram com relação a predição executada pelo modelo através de diferentes heurísticas exploradas. Uma definição alternativa de "melhor instância" a ser consultada é aquela que aumentaria consideravelmente o poder preditivo do modelo ao ser amostrada. Ou seja, podemos nos atentar mais a quão bem o modelo realiza predições do que a quão certo ou incerto o modelo está quanto a sua tarefa (Settles, 2012; Roy e McCallum, 2001). Em uma analogia feita por Settles (2012), o paradigma abordado pelos métodos anteriores (baseados em incerteza) seria pensar sobre o que o modelo pensa com relação as observações **agora**, enquanto esse novo paradigma pensaria que tipo de decisões o modelo faria no **futuro** buscando em particular uma instância tal que, ao observar seu rótulo, provavelmente reduzirá erros futuros do modelo.

O maior problema quanto a esse paradigma é que o modelo não sabe o rótulo de cada observação antes de consultá-la. Um meio de contornar esse problema é de ao invés de tentar reduzir o erro diretamente (que não é possível visto que não conhecemos o rótulo

da instância de interesse), reduze-se seu valor **esperado**. Em outras palavras, utilizando-se de uma interpretação de teoria da decisão, a ideia desse tipo de abordagem é de, ao termos um leque de ações possíveis de serem executadas, podemos identificar todos os resultados possíveis, determinar suas probabilidades e calcular uma soma ponderada para dar um valor esperado para cada ação (Settles, 2012), tendo a decisão racional como aquela que resulte no melhor valor esperado, ou seja, no menor erro esperado. Em termos práticos, para computar o **erro esperado**, precisamos de duas distribuições de probabilidade (Settles, 2012):

1. A probabilidade do rótulo ser y_i ao consultar a observação $\boldsymbol{x}$

2. A probabilidade do modelo errar em outra instância $\boldsymbol{x}'$ quando a resposta sobre $\boldsymbol{x}$ é conhecida

Nenhuma dessas funções é exatamente conhecida (dado que não temos informação sobre o oráculo) e por isso, utiliza-se das distribuições de probabilidade estimadas pelos modelos como uma boa aproximação. Além disso, se tivermos um conjunto não rotulado $\mathcal{U}$ grande, o modelo pode tentar minimizar o erro esperado sobre este, tomando tal conjunto como um conjunto de validação (ou seja, assumimos uma semelhança entre a distribuição de $\mathcal{U}$ e $\mathcal{T}$). Pensando inicialmente na minimização do erro esperado sob a perda 0-1 sobre o conjunto $\mathcal{U}$, a observação com menor erro esperado é dada por:

$$\begin{aligned} \boldsymbol{x}^*_{RE} &= \arg\min_{\boldsymbol{x}} \mathbb{E}_{Y|\hat{g},\boldsymbol{x}} \left[\sum_{\boldsymbol{x}' \in \mathcal{U}} \mathbb{E}_{Y|\hat{g}^+,\boldsymbol{x}'} [\mathbb{I}\left(y \neq \hat{y}\right)] \right] \\ &= \arg\min_{\boldsymbol{x}} \sum_i P(y_i|\boldsymbol{x},\hat{g}) \left[\sum_{\boldsymbol{x}' \in \mathcal{U}} 1 - P\left(\hat{y}|\boldsymbol{x}';\hat{g}^+\right) \right], \end{aligned} \tag{2.11}$$

tal que $\hat{g}^+$ é o modelo aumentado com a adição $\boldsymbol{x}$ e sua suposta rotulação y_i ao modelo e $\hat{y}$ tem o mesmo significado visto na Equação 2.2, ou seja, $\hat{y}$ é o rótulo mais provável de acordo com o modelo g.

Em palavras, fixa-se inicialmente um rótulo verdadeiro para $\boldsymbol{x}$, y_i, e a partir da atribuição desse rótulo, adiciona-se $(\boldsymbol{x}, y_i)$ ao modelo (ou seja ao conjunto $\mathcal{L}$), e utiliza-se o modelo aumentado g^+ no cálculo da soma da probabilidade de se errar a classificação para cada $\boldsymbol{x} \in \mathcal{U}$, ponderando-se essa soma de erros para cada y_i fixado como rótulo verdadeiro de $\boldsymbol{x}$ pela probabilidade de y_i ser a classe verdadeira de fato. Assim, o objetivo nesse caso é reduzir o número esperado de predições incorretas com uma penalização grande

para probabilidades não assertivas (como se poder ver na Figura 2.2 no método do menos confiante). Uma função de perda que torna a redução de erro mais suave é a log-perda, minimizando portanto o seguinte erro esperado:

$$\begin{aligned} \boldsymbol{x}^*_{LP} &= \arg\min_{\boldsymbol{x}} \mathbb{E}_{Y|\hat{g},\boldsymbol{x}} \left[\sum_{\boldsymbol{x}' \in \mathcal{U}} \mathbb{E}_{Y|\hat{g}^+,\boldsymbol{x}'} \left[-\log P(y|\boldsymbol{x}'; \hat{g}^+) \right] \right] \\ &= \arg\min_{\boldsymbol{x}} \sum_i P(y_i|\boldsymbol{x}, \hat{g}) \left[\sum_{\boldsymbol{x}' \in \mathcal{U}} - \sum_i P\left(y'_i|\boldsymbol{x}'; \hat{g}^+\right) \cdot \log P(y'_i|\boldsymbol{x}'; \hat{g}^+) \right] , \end{aligned} \tag{2.12}$$

que é uma extensão da redução de erro esperado proposta pela Equação 2.11 utilizando entropia como medida de erro esperado futuro. Tal metodologia foi proposta por Roy e McCallum (2001) num contexto de classificação de texto utilizando Naive Bayes. Em seu trabalho, essa metodologia produziu melhores curvas de aprendizado (em termos de acurácia) em comparação as estratégias anteriores (consulta por comitê e amostragem por incerteza).

A maior desvantagem da redução de erro esperada é o fato de ela ser muito complexa computacionalmente, pois além de se estimar o erro esperado sobre $\mathcal{U}$ para cada observação $\boldsymbol{x}'$ com o modelo aumentado, também deve se reajustar o modelo para cada rotulação possível em $\boldsymbol{x}$ e repetir o processo para todas as observações de $\mathcal{U}$. Para os classificadores como o naive bayes usado por Roy e McCallum (2001) e certos modelos não paramétricos como processos gaussianos e vizinhos mais próximos, o custo de re-treinamento é relativamente baixo, tornando a aplicação desses modelos mais prática. Porém, para os vários outros modelos e algoritmos de classificação, o custo computacional tende a aumentar. Por exemplo, para uma regressão logística binária, o algoritmo tem uma complexidade $O(|\mathcal{U}| \cdot |\mathcal{L}| \cdot G)$ para amostrar cada observação, tal que $|\cdot|$ é o operador que nos dá o tamanho de cada subconjunto (*pool* e de treino) e G o número de cálculos de gradientes requeridos para a convergência do modelo.

Se aumentarmos o número de rótulos para três ou mais classes o custo computacional tende a aumentar ainda mais de acordo com o numero de rótulo M. Por exemplo, o uso de um modelo de máxima entropia (Berger *et al.*, 1996) nesse caso tem uma complexidade de $O(M^2 \cdot |\mathcal{U}| \cdot |\mathcal{L}| \cdot G)$. Assim, como essa abordagem é relativamente imprática e complexa, realiza-se comumente uma redução do tamanho de $\mathcal{U}$ por uma re-amostragem para que o termo $|\mathcal{U}|$ diminua na medida de complexidade, buscando um balanço entre reduzir a complexidade do algoritmo e a perda de informação sobre $\mathcal{U}$. Salienta-se, por fim, o

fato de as funções de perda utilizada na redução de erro também pode ser muito variada, podendo-se escolher qualquer medida de performance de interesse (Settles, 2012), como por exemplo precisão, revocação, medida F_1, área sobre a curva **ROC** (AUC) e assim por diante, tendo novamente como maior limitante a complexidade computacional do algoritmo de redução de erro esperada.

Em casos de problemas de regressão, não se minimiza a função de erro esperado diretamente, mas sim através da variância em casos de solução com fórmula fechada, o que justamente caracteriza o método de **Redução de variância**, proposto por Cohn *et al.* (1996). Tomando a perda quadrática $L(g(\boldsymbol{X}); Y) = (g(\boldsymbol{X}) - Y)^2$, podemos decompor o erro esperado de um modelo da seguinte maneira (Geman *et al.*, 1992; Izbicki e dos Santos, 2020), neste caso tomando $\hat{g}$ como um estimador para o oráculo $\mathbb{E}[Y|\boldsymbol{X}]$:

$$\mathbb{E}\left[(\hat{g}(\boldsymbol{X}) - Y)^2 | \boldsymbol{X}\right] = \mathbb{V}\left[Y|\boldsymbol{X}\right] + \left(\mathbb{E}_{\mathcal{L}}[\hat{g}(\boldsymbol{X})] - \mathbb{E}[Y|\boldsymbol{X}]\right)^2 + \mathbb{V}_{\mathcal{L}}\left[(\hat{g}(\boldsymbol{X})\right] , \tag{2.13}$$

em que os operadores com sobrescrito $\mathcal{L}$ se referem a operadores sobre o conjunto de treino $\mathcal{L}$, ou seja, nesses casos o conjunto de treinamento é visto como um vetor aleatório i.i.d $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$. Assim a esperança $\mathbb{E}[(\hat{g}(\boldsymbol{X}) - Y)^2|\boldsymbol{X}]$ é computada sobre o conjunto de treinamento aleatório $\mathcal{L}$ e também sobre as variáveis $\boldsymbol{X}$ e Y, tendo Y condicionado em $\boldsymbol{X}$.

Ou seja, o modelo $\hat{g}$ é um componente aleatório. O primeiro termo à direita é visto como um ruído ou variância intrínseca (Izbicki e dos Santos, 2020) da variável resposta que não depende de $\hat{g}$ nem do conjunto de treinamento $\mathcal{L}$. O segundo termo é o viés do modelo em comparação ao oráculo $\mathbb{E}[y|\boldsymbol{x}]$ que pode se dever a classe de modelo usada, como usar um modelo linear quando $\mathbb{E}[y|\boldsymbol{x}]$ tem um formato não linear. Cohn *et al.* (1996) assume que esse componente é invariante dado uma classe fixa de modelos. O último termo, mais importante para essa metodologia, é a variância do estimador $\hat{g}$, que deve ser minimizada para que o erro esperado também seja minimizado. Assim, podemos minimizar o erro quadrático utilizando o mesmo truque da redução de erro (re-treinar o modelo no conjunto $\mathcal{L} \cup \{(\boldsymbol{x}, y)\}$ e minimizar a variância sobre todo o conjunto $\mathcal{U}$):

$$\boldsymbol{x}^*_{RV} = \arg\min_{\boldsymbol{x}} \sum_{\boldsymbol{x}' \in \mathcal{U}} \mathbb{V}_{\hat{g}^+}[Y|\boldsymbol{x}'] , \tag{2.14}$$

com $\hat{g}^+$ sendo novamente o modelo aumentado. Uma questão a ser respondida é como

computar esse valor de forma mais eficiente ao invés de minimizar o erro esperado diretamente por re-treino, como feito pela redução de erro.

Uma possível abordagem atrelada a tarefa de regressão é a utilização da teoria relacionada ao *delineamento ótimo de experimentos* (Fedorov, 2013), que utiliza a *informação de fisher* como uma parte essencial da minimização da variância. Ou seja, uma abordagem mais paramétrica, que supõe certa distribuição (ou classe de distribuições) para a variável resposta Y. Assim, sendo Y uma variável aleatória e θ um parâmetro do qual a a verossimilhança de Y, $f(Y;\theta)$ depende, definimos o *score de fisher* como a seguir (Settles, 2012):

$$\nabla_\theta \boldsymbol{x} = \frac{\partial}{\partial \theta} \log f(Y|\boldsymbol{x};\theta)\,,$$

tal que $\nabla_\theta \boldsymbol{x}$ denota o score fixada a observação $\boldsymbol{x}$, da qual Y também depende. É interessante notar que o score de fisher depende apenas da distribuição de Y sob os parâmetros θ, enquanto $\boldsymbol{x}$ é tido como um valor fixo nesses casos. Como exemplo de distribuição de Y e parâmetros θ, podemos pensar na suposição básica de uma regressão linear,em que supõe-se $Y \sim \text{Normal}(X\boldsymbol{\beta}, 1)$, tal que $\theta = \boldsymbol{\beta}$, tendo o *score de fisher* $\nabla \boldsymbol{x}_\theta = (1, x_1, \ldots, x_p)$. Assim, a **informação de fisher** nada mais é que a variância do score, obtida da seguinte maneira:

$$\begin{aligned}
F(\theta) &= \mathbb{E}\left[\left(\frac{\partial}{\partial \theta}\log f(Y|\boldsymbol{x};\theta)\right)^2\right] - \mathbb{E}\left[\frac{\partial}{\partial \theta}\log f(Y|\boldsymbol{x};\theta)\right]^2 \\
&= \mathbb{E}\left[\left(\frac{\partial}{\partial \theta}\log f(Y|\boldsymbol{x};\theta)\right)^2\right] \quad (\text{por definição } \mathbb{E}[\nabla \boldsymbol{x}_\theta] = 0) \\
&= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2}\log f(Y|\boldsymbol{x};\theta)\right]\,,
\end{aligned}$$

tal que a esperança é tomada tanto na distribuição $f(Y|\boldsymbol{x},\theta)$ quanto $f(\boldsymbol{x})$, ou seja, $\boldsymbol{x}$ é considerado aleatório sobre certa distribuição $f(x)$. $F(\theta)$ resulta em uma matriz $K \times K$ caso o modelo dependa de K parâmetros. O inverso da informação de fisher $F^{-1}(\theta)$ define um limite inferior da variância dos estimadores não enviesados de θ pela desigualdade de Cramér-Rao (Cover e Thomas, 2006). Assim, para minimizar a variância do estimador de θ, o modelo deve selecionar a instância que minimiza o inverso da matriz de informação de fisher (ou maximizar a matriz de informação de fihser). Como a informação de Fisher tem um formato matricial ao termos K parâmetros, a otimização do estimador de θ pela

seleção da instância com menor variância não é exatamente clara, havendo três maneiras de se minimizar a matriz (3 tipos de delineamentos ótimos de experimentos) (Settles, 2012):

- $D-$ *ótimo* minimiza o *determinante* da inversa da matriz de informação de fisher;
- $E-$ *ótimo* maximiza o *autovalor* da matriz de informação de fisher;
- $A-$ *ótimo* minimiza o traço da inversa da matriz de informação de fisher. Esse critério minimiza a variância *média* dos estimadores do conjunto de parâmetros $\boldsymbol{\theta}$

Nesse contexto, não há uma forma analítica de se obter a instância E-ótima. Assim, podemos selecionar a instância $\boldsymbol{x}$ que minimiza a variância de forma $D-$ótima da seguinte maneira (Chaloner e Verdinelli, 1995):

$$\boldsymbol{x}_D^* = \arg\min_{\boldsymbol{x}} \boldsymbol{det}\left([F_{\mathcal{L}}(\theta) + \nabla_\theta \boldsymbol{x} \nabla_\theta \boldsymbol{x}^t]^{-1}\right) , \tag{2.15}$$

em que $F_{\mathcal{L}}(\theta)$ é a matriz de informação de fisher fixada no conjunto de treinamento $\mathcal{L}$. Ao somarmos $F_{\mathcal{L}}(\theta)$ e $\nabla_\theta \boldsymbol{x} \nabla_\theta \boldsymbol{x}^t$ estamos adicionando a informação de $\boldsymbol{x}$ as informações do conjunto $\mathcal{L}$, e ao minimizarmos o determinante da matriz inversa dessa soma obtemos a instância D-ótima com relação a θ. Para o delineamento A-ótimo, relativamente mais popular, minimiza-se comumente o traço $\boldsymbol{tr}(AF_{\mathcal{L}}(\theta)^{-1})$, com A sendo uma matriz simétrica. Em particular podemos tomar $A_{\boldsymbol{x}} = \nabla_\theta \boldsymbol{x} \nabla_\theta \boldsymbol{x}^t$, tendo portanto o traço $\boldsymbol{tr}(\nabla_\theta \boldsymbol{x}^t F_{\mathcal{L}}^{-1}(\theta) \nabla_\theta \boldsymbol{x})$ para cada instância $\boldsymbol{x}$ (Schervish, 2012). Ou seja, minimiza-se a variância da predição considerando-se uma única instância $\boldsymbol{x}$. Porém, como temos interesse em minimizar a variância sobre toda a distribuição de atributos $f(\boldsymbol{x})$, a matriz A deve conter informação sobre toda a distribuição de atributos em $\mathcal{U}$. Assim, utilizando um delineamento A-ótimo, obtemos a instância $\boldsymbol{x}$ que minimiza a variância a partir do método *razão de informação de fisher* como segue:

$$\boldsymbol{x}_{RIF}^* = \arg\min_{\boldsymbol{x}} \sum_{\boldsymbol{x}' \in \mathcal{U}} \boldsymbol{tr}\left(A_{\boldsymbol{x}'}(\theta)[F_{\mathcal{L}}(\theta) + \nabla_\theta \boldsymbol{x} \nabla_\theta \boldsymbol{x}^t]^{-1}\right) \tag{2.16}$$

$$= \arg\min_{\boldsymbol{x}} \boldsymbol{tr}\left(F_{\mathcal{U}}(\theta)[F_{\mathcal{L}}(\theta) + \nabla_\theta \boldsymbol{x} \nabla_\theta \boldsymbol{x}^t]^{-1}\right) , \tag{2.17}$$

interpretando-se tal razão como sendo a variância do estimativa do modelo por toda a distribuição de atributos em $\mathcal{U}$ que não pode ser explicada pelas observações no conjunto

aumentando $\mathcal{L} \cup \{\boldsymbol{x}\}$ (Settles, 2012). A desvantagem principal da redução de variância é semelhante a da redução de de erro esperada: complexidade computacional. Ao estimarmos a variância para cada instância, devemos invertes e multiplicar matrizes $K \times K$, o que tem uma complexidade de operação $O(K^3)$, tendo portanto uma complexidade final de $O(|\mathcal{U}|K^3)$ para consultar a próxima instância, já que percorremos toda o conjunto $\mathcal{U}$ para cada consulta. Além disso, esse tipo de abordagem não é tão flexível quanto os métodos vistos anteriormente (amostragem por incerteza e consulta por comitê), utilizando uma abordagem mais paramétrica, apesar de haver certa adaptação para redes neurais (Mckay, 1992), sendo aplicada de forma mais natural para modelos de regressão linear, não linear e logística. Assim, neste trabalho, daremos mais foco para a redução de erro esperada, técnicas de amostragem de incerteza, consulta por comitê e outras metodologias apresentadas.

2.4 Métodos ponderados por densidade

As estratégias de consulta de comitê e amostragem por incerteza vistas anteriormente nas Subseções 2.1 e 2.2 no cenário *pool-based* tem o problema de serem muito míopes com relação a seleção de observações, visto que a medida de informatividade de cada observação é obtida de forma individual, ou seja, uma instância por vez, sem considerar possíveis similaridades entre observações. Como consequência, muitas vezes podemos acabar selecionando outliers ao invés de observações mais representativas do conjunto de dados, como ilustra a Figura 2.9, que mostra um caso em que a instância é próxima da fronteira de decisão, mas tem baixa representatividade da distribuição das observações. Assim, rotular esse tipo de instância pode nos levar a uma melhora pobre na performance do modelo. Os métodos de redução de erro esperado não possuem esse mesmo problema pois levam em conta a distribuição das covariáveis no cálculo de seus scores ao tentar reduzir o erro esperado futuro sobre todos as observações pertencentes a *pool*, utilizando $\mathcal{U}$ para estimar os erros futuros ou variância. Assim, a redução de erro esperada é relativamente resistente a *outliers* ou ruídos em comparação aos métodos de amostragem por incerteza e consulta por comitê, tendo porém um *trade-off* entre consistência na consulta de instâncias e o custo computacional elevado para sua execução.

Essa "miopia"dos métodos de amostragem por incerteza e consulta por comitê pode ser concertada ao incluir certa influência da estrutura dos dados sobre as heurísticas de

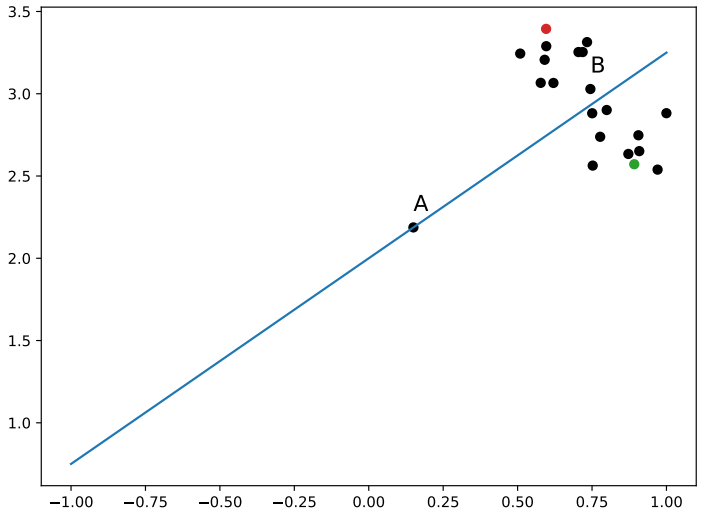


Figura 2.9: Um exemplo de dados simulados que mostram quando amostragem por incerteza e consulta por comitê podem ser estratégias ruins de consulta. Os pontos coloridos são as observações que temos no conjunto $\mathcal{L}$ enquanto o restante das observações (em preto) pertencem ao conjunto $\mathcal{U}$. Como a observação A está exatamente na fronteira de decisão, ela seria tomada como a mais incerta. Porém, a observação B ou outras próximas a essa provavelmente nos dariam informações mais consistentes sobre a distribuição dos rótulos.

cada método. Para se fazer isso, pode-se ponderar tais heurísticas pela **densidade de informação**, cuja intuição é de que observações informativas não apenas devem carregar boa informação para o ajuste do modelo como também devem ser representativas da distribuição da qual fazem parte (Settles, 2012). Para se quantificar o grau de representatividade de uma instância pode-se utilizar medidas de **similaridade** entre a instância atual e as demais instâncias pertencentes ao conjunto de interesse. A heurística genérica resultante de tal ponderação, formulada por Settles e Craven (2008), é dada pela fórmula a seguir:

$$\boldsymbol{x}^*_{DI} = \arg\max_{\boldsymbol{x}} \phi_A(\boldsymbol{x}) \cdot \left(\frac{1}{|\mathcal{U}|} \sum_{\boldsymbol{x}' \in \mathcal{U}} sim(\boldsymbol{x}, \boldsymbol{x}') \right)^{\lambda}, \tag{2.18}$$

tal que $\phi_A(\cdot)$ representa a medida de incerteza sobre $\boldsymbol{x}$ de acordo com certa estratégia de consulta A, enquanto o operador $sim(\cdot, \cdot)$ é o operador de similaridade entre $\boldsymbol{x}$ e $\boldsymbol{x}'$, computando-se uma similaridade média entre $\boldsymbol{x}$ e as demais instâncias pertencentes a $\mathcal{U}$. Assim o segundo termo pondera a informatividade de $\boldsymbol{x}$ pela similaridade média de $\boldsymbol{x}$ com relação a todas as outras observações pertencentes a distribuição de atributos, para o qual

se assume que $\mathcal{U}$ é representativo (Settles, 2012), tendo λ como um hiper-parâmetro que controla a importância do termo relativo a densidade. Dentre as similaridades possíveis de serem usadas, podem ser utilizadas a distância euclidiana, similaridade por cosseno, coeficiente de correlação de Pearson, correlação de Spearman, entre outras. Tomando o gráfico analisado anteriormente na Figura 2.9, a Figura 2.10 nos dá a densidade de informação euclidiana para cada instância.

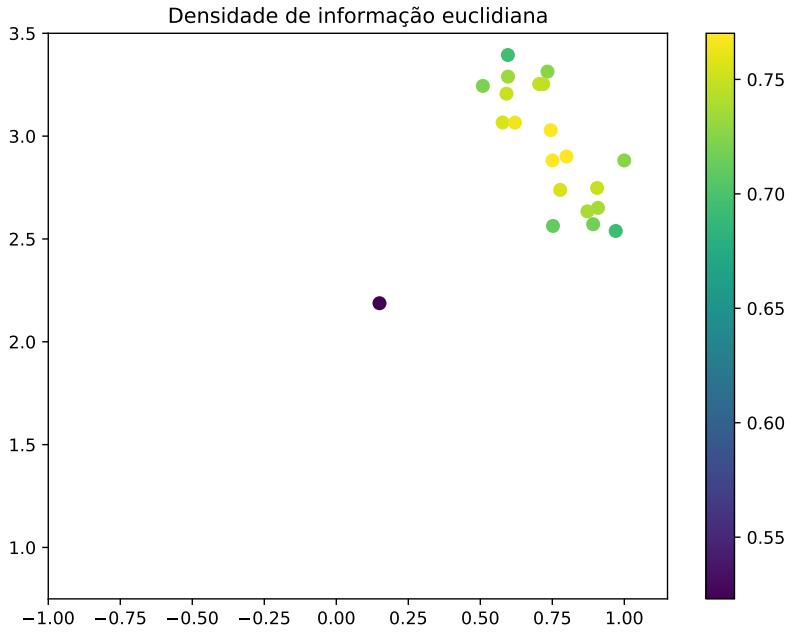


Figura 2.10: Similaridade média de cada observação com relação as demais. É interessante notar a penalização grande feita sobre a observação A que possivelmente será deixada de em favor das outras observações em uma consulta que utilize uma heurística ponderada por densidade.

2.5 Consulta por minilotes (*batches*)

Até o momento, os métodos explorados até a seção anterior empregaram um modo de seleção de observações baseado no cenário *pool-based*, ou seja, um cenário em que a consulta é feita de forma serial (uma observação consultada e removida de $\mathcal{U}$ por vez). Porém em muitas situações, o tempo necessário para se obter uma amostra rotulada para o modelo de interesse é grande e custoso, em casos, por exemplo, de modelos de *ensemble* em consultas, ou modelos complexos para tarefas de classificação de dados estruturados (para alguns casos de genética). Apesar de em muitos casos se ter uma convergência razoável do conjunto de treinamento em uma boa quantidade de tempo, algoritmos de redução de erro esperada chegam a ter um custo proibitivo para toda a consulta (Settles, 2012). Além disso, em muitos contextos, a rotulação pode ser feita de forma distribuída, tendo por exemplo, múltiplos oráculos.

Em casos como esses, pode ser interessante o uso de uma estratégia de consulta que faça uso de um *paralelismo* na consulta, consultando observações em grupos (ou minilotes). É nesse cenário que comumente se utiliza o aprendizado ativo em minilotes (ou no inglês *batch-mode active learning*, *BMAL*), que permite a consulta de instâncias em grupo (Settles, 2012). A ideia principal por trás da consulta por minilotes é de se obter um conjunto de consultas $\mathcal{Q}$ que terão suas instâncias consultadas simultaneamente (fazendo uso de um rotulação paralela). Para se obter tal conjunto, há diversos desafios, visto que caso simplesmente consultarmos as Q-melhores consultas de acordo com certa estratégia de consulta desconsideraremos as possíveis interseções de informação entre as melhores instâncias selecionadas, amostrando observações não tão informativas para o modelo e desperdiçando dessa forma tempo na rotulação de instâncias redundantes.

Settles (2012) discute mais especificamente a abordagem de redução de variância e erro esperada sobre a configuração de minilote, elucidando o problema de que a amostragem aleatória em minilotes é geralmente superior a consulta por minilotes, com exceção da classe de metodologia citada. Para tal abordagem, faz-se uso de funções submodulares (Nemhauser *et al.*, 1978). Cardoso *et al.* (2017) introduzem uma nova abordagem de *BAML* mais flexível denominada **aprendizado ativo ranqueado por minilotes**, cujas consultas geram uma lista ranqueada otimizada de instâncias a serem rotuladas (Cardoso *et al.*, 2017). Inicializando com $\mathcal{Q}$ vazio, o procedimento de tal método tem as seguintes etapas (Cardoso *et al.*, 2017):

1. Treinar o modelo no conjunto $\mathcal{L}$ e em seguida estimar a incerteza associada $\phi(\boldsymbol{x})$ para cada $\boldsymbol{x} \in \mathcal{U}$. $\phi(\boldsymbol{x})$ pode ter uma variedade de formatos, mais comumente os apresentados nas subseções 2.1 (amostragem por incerteza) e 2.2 (consulta por comitê). Em, seguida inicia-se um laço até se obter um conjunto $\mathcal{Q}$ para ser rotulado

2. Crie/atualize os *scores* de similaridades entre as observações pertencentes ao conjunto $\mathcal{U}$ com relação ao conjunto $\mathcal{L}$, ou seja, computar para cada $\boldsymbol{x} \in \mathcal{U}$:

$$sim(\boldsymbol{x}, \mathcal{L}) \coloneqq \frac{1}{|\mathcal{L}|} \sum_{\boldsymbol{x}' \in \mathcal{L}} sim(\boldsymbol{x}, \boldsymbol{x}') \,.$$

3. Dado o seguinte score:

$$S(\boldsymbol{x}, \mathcal{U}, \mathcal{L}) = \alpha \cdot (1 - sim(\boldsymbol{x}, \mathcal{L})) + (1 - \alpha) \cdot \phi(\boldsymbol{x}) \,, \tag{2.19}$$

   α sendo o parâmetro de peso para o impacto da similaridade com relação a incerteza geralmente tomado como $\alpha = \frac{|\mathcal{U}|}{|\mathcal{U}|+|\mathcal{L}|}$. Busca-se a instância que maximize esse *score*, ou seja, uma observação tal que:

$$\boldsymbol{x}^*_{RBMAL} = \arg\max_{\boldsymbol{x}} \; S(\boldsymbol{x}) \,, \tag{2.20}$$

   após seleção, insere-se $\boldsymbol{x}^*_{RBMAL}$ no conjunto $\mathcal{Q}$ e remove-se tal observação de $\mathcal{U}$. Além disso, cada instância adicionada a $\mathcal{Q}$ também é adicionada a $\mathcal{L}$ porém sem rotulação feita. O item 2 e 3 são repetidos até se esvaziar o conjunto $\mathcal{U}$ (ou seja, $|\mathcal{U}| = 0$) ou até que uma quantidade pré definida de instâncias para cada minilote seja adicionada a $\mathcal{Q}$.

4. Apresentar o conjunto $\mathcal{Q}$ a um ou mais oráculos para rotulação, incorporando a rotulação em $\mathcal{L}$ e esvaziando-se novamente $\mathcal{Q}$. Após isso, reinicia-se todo o processo do item 1 ao 4, até se atingir uma quantidade pré determinada de minilotes.

Através de tal procedimento elabora-se o Algoritmo 3, formulado originalmente por Cardoso *et al.* (2017). Em geral, a proposta desse método é evitar o problema de selecionar observações com possíveis interseções na informação fornecida ao ponderar a incerteza em cada etapa pela similaridade da observação com relação as observações de $\mathcal{L}$, oferecendo ao mesmo tempo um algoritmo flexível e eficiente para uma configuração de busca *batch-mode*.

Algoritmo 3 Aprendizado ativo ranqueado por minilotes

$\mathcal{U}$ = conjunto de instâncias não rotuladas $\{\boldsymbol{x}_i\}_{i\in\mathcal{U}}$
$\mathcal{L}$ = conjunto de instâncias não rotuladas $\{\boldsymbol{x}_i\}_{i=1}^{n}$
n_{bach} = Número escolhido de minilotes
l_{bach} = Número de instâncias para cada minilote (tamanho do minilote)
$\mathcal{Q}$ = ListaVazia()
for $n = 1, 2, \ldots, n_{bach}$ **do**
    $\mathcal{L}_{temp} = \mathcal{L}$ (conjunto das instâncias rotuladas e ranqueadas sem rotulação)
    g = **ajuste**($\mathcal{L}$)
    CalculeIncertezas(g, $\mathcal{U}$)
    for $l = 1, 2, \ldots, l_{bach}$ **do**
        s = selecione $\boldsymbol{x}^* \in \mathcal{U}$ com maior *score* $S(\boldsymbol{x}^*, \mathcal{U}, \mathcal{L}_{temp})$
        $\mathcal{L}_{temp} = \mathcal{L}_{temp} \cup s$
        $\mathcal{U} = \mathcal{U} - s$
        AdicionarALista($\mathcal{Q}$, s)
    end for
    $\mathcal{L}_{novo}$ = **Rotular**($\mathcal{Q}$)
    $\mathcal{L} = \mathcal{L} \cup \mathcal{L}_{novo}$
end for

2.6 Outras abordagens

Na literatura de aprendizado ativo existe uma gigantesca diversidade de métodos para consulta e rotulação de observações nos cenários *stream-based* e *pool-based*, para as mais diversas tarefas, como por exemplo aprendizado ativo para rótulos estruturados (Cohn *et al.*, 1994; Settles e Craven, 2008) e aprendizado ativo com custos (Culotta e McCallum, 2005), com diversas proposições de algoritmos com incorporação de diferentes propriedades teóricas como por exemplo, algoritmos de consulta que lidam bem com ruídos (Algoritmo A^2 (Hanneke, 2007), Aprendizado ativo ponderado por importância (Beygelzimer *et al.*, 2009) e Aprendizado ativo baseado em marginais (Balcan *et al.*, 2007)). O próprio aprendizado ativo por minilotes (Cardoso *et al.*, 2017; Settles, 2012) compreende uma área de aplicação específica, quando se tem a presença de diversos oráculos em paralelo, apesar de uma capacidade de generalização maior que muitos dos métodos citados anteriormente (que em muitas situações tem custo computacional proibitivo).

De forma geral, os métodos descritos e trabalhados nessa seção compreendem uma classe de métodos mais comumente utilizada no cotidiano de aprendizado ativo, nas variantes de problemas mais comuns, envolvendo problemas de classificação e regressão e dos quais muitos métodos mais específicos são derivados com variações nos modelos ou heurísticas e as vezes com maiores abstrações, sendo tais métodos muito personalizáveis (Settles, 2012). Há, porém, duas classes de métodos também empregadas cotidianamente

que são interessantes de se detalhar nessa Seção: **amostragem hierárquica** (Dasgupta e Hsu, 2008) e **aprendizado semi-supervisionado** (Zhu e Goldberg, 2009).

A **amostragem hierárquica** consiste de retirar informações sobre a distribuição dos atributos a partir de métodos de agrupamentos hierárquicos, associando tais *clusterings* a uma rotulação. Ou seja, tal método consiste de selecionar consultas e montar o conjunto $\mathcal{L}$ com base unicamente na estrutura dos agrupamentos e na rotulação associada a estes, de forma que a consulta seja agnóstica ao modelo escolhido (não se utiliza o modelo para a seleção de instâncias). Para tal, Dasgupta e Hsu (2008) propõe um algoritmo cuja ideia básica é realizar um agrupamento inicial grosseiro, consultar instâncias aleatoriamente selecionadas em cada agrupamento e iterativamente refinar (dividir) os agrupamentos até que eles fiquem mais puros, focando a consulta e divisão em agrupamentos impuros. Mais detalhes sobre o algoritmo são discutidos em Settles (2012) e Dasgupta e Hsu (2008).

Caso a estrutura do agrupamento não possua tanta correlação com a distribuição dos rótulos, a amostragem hierárquica se torna basicamente uma amostragem aleatória (todos os agrupamentos são impuros), mas caso haja certa relação e exista uma boa segmentação dos rótulos no *clustering*, explora-se e amostrasse principalmente de agrupamentos mais impuros. Ao finalizar o algoritmo, utiliza-se o conjunto de treinamento final obtido $\mathcal{L}$ para treinar algum algoritmo de aprendizado supervisionado de escolha.

O **aprendizado semi-supervisionado** (Zhu, 2005; Zhu e Goldberg, 2009) é uma área semelhante ao **aprendizado ativo** que também tem como cenário um banco de dados com algumas instâncias rotuladas e uma grande quantidade de instâncias não rotuladas, tendo-se interesse de usar o conjunto de dados não rotulados para de alguma forma melhorar o desempenho do modelo de interesse. Porém, a abordagem do aprendizado semi-supervisionado é diferente em comparação ao aprendizado ativo. Ao invés de buscar e rotular as instâncias mais informativas para o modelo como se faz em geral em aprendizado ativo, no aprendizado semi-supervisionado o próprio modelo "se ensina"a partir de extrapolações do modelo sobre as instâncias não rotuladas, explorando estruturas latentes dos dados para melhorar a qualidade do modelo.

Dessa forma aprendizado ativo e aprendizado semi-supervisionado possuem certas interseções por trabalharem no mesmo tipo de problema (ausência de rótulos) tentando melhorar o desempenho do modelo a partir da exploração do conjunto sem rótulo. Três técnicas de aprendizado semi-supervisionado que são semelhantes a diferentes técnicas de aprendizado ativo exploradas são: *auto-treinamento* proposta por Yarowsky (1995) (seme-

lhante a amostragem por incerteza), *co-treinamento* proposta por Blum e Mitchell (1998) (semelhante a consulta por comitê) e *regularização por entropia* proposta por Grandvalet *et al.* (2005) (semelhante a redução de erro esperada). Resumindo, no *auto-treinamento* adiciona-se iterativamente a instância não rotulada **mais** confiante e sua predição a $\mathcal{L}$, enquanto que o *co-treinamento* consiste de tomar cada modelo separadamente de um *ensemble* para classificar os dados não rotulados e ajustar o restante dos modelos com a instância da qual o modelo fixado é o mais confiante e enfim, a *regularização por entropia* é uma abordagem com a intuição de que o melhor modelo é aquele que faz as predições mais confiantes dos dados não rotuladas.

Em suma, vemos que o **aprendizado ativo** e o **aprendizado semi-supervisionado** atacam o mesmo problema porém de maneiras complementares: enquanto o **aprendizado semi-supervisionado** explora aquilo que o modelo sabe sobre os dados não rotulados, os métodos de **aprendizado ativo** estudam mais o que tais modelos não sabem (Settles, 2012). Dessa forma, é interessante em diversos contextos unir ambos métodos, como feito por exemplo por McCallum *et al.* (1998) ao combinar uma consulta por comitê ponderada por densidade com o algoritmo EM (*expectation-maximization*), tido como uma espécie de auto-treinamento.

2.7 Discussão sobre os métodos e outros detalhes do aprendizado ativo

Ao estudar e trabalhar com diversos problemas em que aprendizado ativo é aplicável para rotulação ou possível redução do conjunto treinamento, o primeiro questionamento que se faz é qual a o melhor método/algoritmo a ser utilizado. Porém, assim como para escolha de modelo e seleção de variáveis em diferentes problemas de aprendizado supervisionado, não há uma resposta certa a priori para cada situação (Settles, 2012). De fato, não há uma heurística ou método de consulta "vencedor", havendo muita dependência no contexto do problema ou interesse da pesquisa/estudo com relação a algum modelo específico ou abordagem. Inclusive, em diversas situações, por má escolha de modelo ou de método de aprendizado ativo/heurística, o aprendizado ativo pode ter desempenho pior ou comparável à amostragem aleatória, especialmente ao tomarmos a amostragem por incerteza como estratégia de consulta (Schütze *et al.*, 2006; Tomanek *et al.*, 2009; Wallace *et al.*, 2010; Settles, 2012). Logo, é necessário atenção e entendimento redobrado

sobre o problema que se tem interesse. Se detalha e compara as vantagens, desvantagens e aspectos gerais dos diferentes métodos através da Tabela 5.1 originalmente elaborada por Settles (2012).

Tabela 2.1: Comparação entre as estratégias de aprendizado ativo abordadas

Estratégia de consulta	**Vantagens**	**Desvantagens**
Amostragem por incerteza	Abordagem simples e intuitiva, fácil de se implementar e rápida. Além disso pode ser usado com qualquer modelo probabilístico.	Essa abordagem é facilmente viesada pelo modelo que está associada, tendo como resultado um viés alto na amostragem e uma confiança excessiva sobre regiões que na verdade podem ser relativamente incertas (miopia).
Consulta por comitê	Abordagem também simples, melhor aplicada em algoritmos de *ensemble* e mais robusta ao viés do modelo com relação a amostragem	Pode ser custoso de se treinar e manter multiplos modelos em cômite, além de esse modelo ainda ter certa miopia na consulta de observações mais informativas.
Redução de erro/variância	Seleciona instâncias com base na redução direta do erro esperado do estimador, tendo resultados empiricos satisfatórios	Computacionalmente caro, dificil de se implementar, limitado ao cenário *pool-based*. Além disso, a redução de variância se atém a modelos parâmétricos de regressão.
Ponderação de densidade	Melhora a miopia da heurísticas anteriores com relação à distribuição dos atributos, captando melhor outliers ou ruídos e evitando consulta-los. Facilmente implementavel e rápido.	A distribuição dos atributos podem não ter relação com os rótulos e herda certas limitações das heurítiscas passadas.

Primeiramente a abordagem de **amostragem por incerteza** é a mais popular entre as demais abordadas pela simplicidade e intuitividade mencionadas na Tabela 5.1, sendo um método com alta flexibilidade na escolha de heurísticas que representem a incerteza de cada modelo, podendo ser também empregado em um cenário *stream-based* como se pode ver na Seção 2.1. Além disso, para modelos probabilísticos muito pesados, com treinamento caro, o uso desse tipo de método é ideal pelo seu custo computacional baixo (Settles, 2012). O maior problema desse tipo de abordagem é ela basear-se unicamente em um modelo, que geralmente é treinado em um conjunto de treinamento inicial $\mathcal{L}$ muito pequeno. Assim, surge um alto **viés na amostragem** tendo a amostra consultada muito correlacionada ao modelo de escolha. Esse viés pode nos levar a resultados indesejáveis se não muito controlado, estando atrelado tanto a uma amostra inicial de treinamento quanto a possivelmente um modelo escolhido ruim (veja Capítulo 4 para exemplos). Tal

viés pode ser reduzido ao utilizar-se a **consulta por comitê**, que utiliza a discordância entre incertezas de diferentes modelos para escolher observações ao invés de depender da decisão de apenas um único modelo, aumentando dessa forma a região de incerteza no espaço de atributos e diminuindo uma miopia muito presente no método de amostragem por incerteza.

Apesar disso, ainda existe certa miopia (região de incerteza não muito correspondente com a situação real dos dados ou muito confiante em regiões de alta incerteza) associada a tal modelo, principalmente com a presença de outliers e/ou ruídos no banco de dados. O método de **ponderação por densidade** corrige melhor tal miopia e diminui a sensibilidade à ruídos e outliers ao adicionar mais informações sobre a distribuição dos atributos $\boldsymbol{X}$, ponderando as heurísticas de incerteza pela similaridade média de cada observação pertencente a $\mathcal{U}$ com relação a todas as observações restantes de $\mathcal{U}$. Porém, muitas vezes, essa correção não melhora tanto o desempenho do classificador pelo fato de a distribuição dos atributos ter pouca relação com a distribuição dos rótulos, havendo um custo computacional adicional atrelado ao cálculo das medidas de similaridade para cada observação $\boldsymbol{x} \in \mathcal{U}$. Para obter melhores performances com menos observações de treinamento, os métodos de **redução de erro esperada** e **redução de variância**, ao reduzirem o risco futuro esperado do modelo, consultam observações que melhoram consideravelmente o desempenho do modelo em uso, porém com um custo computacional relativamente mais caro e uma ampla dificuldade de generalização para modelos mais complexos.

Alternativamente, pode-se também misturar os métodos comumente utilizados em **aprendizado ativo** com métodos em **aprendizado semi-supervisionado**, para garantir uma exploração mais profunda da estrutura do banco de dados ao combinar duas metodologias distintas para dados semi-rotulados, ou utilizar a **amostragem hierárquica** para reduzir ainda mais o viés na amostragem ao utilizar um método agnóstico ao modelo utilizado e consultar observações com base na estrutura do agrupamento hierárquico. Apesar de tais vantagens, ambos métodos ainda possuem desvantagens: alto custo computacional para computar o agrupamento hierárquico (e também para armazenar o dendrograma) para a amostragem hierárquica e o fato de ser herdar problemas tanto da área de **aprendizado ativo** quanto **semi supervisionado** quando se utiliza **aprendizado ativo** e **semi-supervisionado** (Settles, 2012). Assim, observa-se que há diversos prós e contras para cada abordagem, com cada método se complementando ou simplificando de acordo com o contexto do aprendizado ativo, analisando as observações não rotuladas sob

diferentes óticas.

Enfatiza-se também que o problema de **viés na amostragem** destacado principalmente para a **amostragem por incerteza**, é um aspecto em comum a todos os métodos presentes na Tabela 5.1. De fato, por mais que se possa diminuir tal viés, a amostra rotulada $\mathcal{L}$ sempre estará atrelada ao modelo escolhido, o que torna o conjunto de treinamento não *iid*. Portanto, o reuso do conjunto de treinamento $\mathcal{L}$ obtido por aprendizado ativo em outros modelos é cercado de debates e pouco há de teoria quanto a habilidade de utilizar um conjunto rotulado obtido por um modelo em outro (Settles, 2012). Certos estudos mostraram resultados positivos quanto ao reuso de dados em situações e modelos específicos (Lewis e Gale, 1994; Tomanek *et al.*, 2007; Hwa, 2001), porém em geral, Tomanek e Morik (2011) mostram a partir de um amplo estudo empírico que a reutilização de amostras em um contexto de classificação possuem resultados inconclusivos. Todos esses aspectos levantados em geral, mostram a importância de se escolher bem a classe de modelos a ser utilizada para aprendizado ativo, de forma que, o aprendizado ativo é mais seguro de se utilizar quando se tem certa ideia do quão bom o modelo é no contexto de aplicação. Settles (2012) propõe que caso não se saiba que classe de modelos aplicar para o problema em especifico, utilizar a amostragem aleatória pode ser mais seguro e garantir melhores resultados do que ajustar modelos "no escuro"que podem ser inadequados. Pode-se inclusive elaborar um estudo piloto sobre a tarefa de interesse a partir de tal amostra, conhecendo-se melhor do contexto de aplicação e ganhando intuição sobre que tipo de modelo utilizar para o aprendizado ativo.

Por fim, outro aspecto interessante de se destacar sobre aprendizado ativo é o fato da literatura desse tipo de abordagem estar mais voltada para problemas de classificação ou variantes, tendo como principais métodos para regressão ativa as adaptações de consulta por comitê (Burbidge *et al.*, 2007) e amostragem por incerteza (Settles, 2012) juntamente com a redução de variância (Cohn *et al.*, 1996) que é voltada para modelos paramétricos. Assim, em geral, esse tipo de tarefa é relativamente ignorada no contexto de aprendizado ativo. No capítulo 3 desse trabalho propõe-se um novo tipo de abordagem mais flexível específica para modelos de regressão. Esse procedimento será futuramente testado e comparado às demais metodologias de regressão ativa.

Capítulo 3

Método proposto para regressão ativa

Na seção 2.3 a redução de variância tinha interesse principal em reduzir o risco esperado $\mathbb{E}[(\hat{g}(\boldsymbol{X}) - Y)^2|\boldsymbol{X}]$ ao consultar a instância que minimiza-se o termo da variância associada ao modelo $\hat{g} : \mathbb{R}^d \to \mathbb{R}$ na Equação (2.13), considerando $\hat{g}(\cdot)$ como aleatório. Nossa proposta também utiliza da decomposição dada pela Equação (2.13), mas tem uma formulação diferente da redução de variância. Primeiramente, considerando $\hat{g}(\boldsymbol{X})$ como fixo, obtemos como nova decomposição do risco esperado:

$$R(\hat{g}, \boldsymbol{x}) = \mathbb{E}[(\hat{g}(\boldsymbol{x}) - Y)^2|\boldsymbol{X} = \boldsymbol{x}] = (\mathbb{E}[Y|\boldsymbol{x}] - \hat{g}(\boldsymbol{x}))^2 + \mathbb{V}[Y|\boldsymbol{x}]\,. \tag{3.1}$$

Para fins de simplificação no desenvolvimento da metodologia, consideraremos neste trabalho a variância intrínseca constante $\mathbb{V}[Y|\boldsymbol{x}] = \sigma^2$, ou seja, supõe-se homocedasticidade dos dados. A ideia principal dessa nova abordagem é de ao invés de encontrar $\boldsymbol{x}$ que minimize o risco esperado, buscaremos a instância $\boldsymbol{x}$ que maximiza a distância entre o oráculo $r(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{x}]$ e o modelo de interesse $\hat{g}(\boldsymbol{x})$. Isso é baseado na intuição de que ao consultar a instância $\boldsymbol{x}$ mais má ajustada pelo modelo (mais distante do oráculo) há maiores chances de melhora na performance do modelo $\hat{g}(\cdot)$ (em outras palavras, diminuir a distância de $\hat{g}(\boldsymbol{x})$ para $r(\boldsymbol{x})$). Com tal proposta em mente, a Equação (3.1) pode ser reformulada da seguinte maneira:

$$(r(\boldsymbol{x}) - \hat{g}(\boldsymbol{x}))^2 = \mathbb{E}[Z|\boldsymbol{x}] - \sigma^2\,, \tag{3.2}$$

em que $Z = (Y - \hat{g}(\boldsymbol{x}))^2$. Assim, tomando o conjunto de treinamento inicial $\mathcal{L}$, podemos dividir $\mathcal{L}$ em $\mathcal{L}_T$ e $\mathcal{L}_V$, ajustando o modelo $\hat{g}$ sobre o conjunto de treinamento $\mathcal{L}_T$ e posteriormente, ajustar um modelo $\hat{h}$ considerando a variável resposta $Z_i = (Y_i - \hat{g}(\boldsymbol{x}_i))^2$ sobre o conjunto de validação $\mathcal{L}_V$. Ou seja, enquanto na primeira etapa consideramos os dados $(\boldsymbol{x}_i, y_i)_{i \in \mathcal{L}_T}$ para ajustar o modelo $\hat{g}$ na segunda etapa consideramos os dados $(\boldsymbol{x}_i, z_i)_{i \in \mathcal{L}_V}$ para estimar $\mathbb{E}[Z|\boldsymbol{x}]$ através do ajuste de um **modelo de validação** $\hat{h}$, que pode ser diferente do modelo original $\hat{g}(\cdot)$. Logo, com $\mathbb{E}[Z|\boldsymbol{x}]$ estimado pelo conjunto de validação $\mathcal{L}$, queremos encontrar no conjunto $\mathcal{U}$:

$$\boldsymbol{x}^*_{RV} = \arg\max_{\boldsymbol{x} \in \mathcal{U}} \hat{\mathbb{E}}_{\mathcal{L}_V}[Z|\boldsymbol{x}]\,. \tag{3.3}$$

Um problema que surge nessa abordagem é como dividir o conjunto de treinamento $\mathcal{L}$ em treino e validação ao adicionar uma nova observação $\boldsymbol{x}^*_{RV}$, visto que estaríamos perdendo informação sobre o conjunto de dados tanto para o conjunto de treino $\mathcal{L}_T$ quanto de validação $\mathcal{L}_V$ se adicionarmos permanentemente tal observação ao outro conjunto, havendo grande chance de se escolher uma nova observação que não necessariamente leve a diminuição da distância de $\hat{g} : \mathbb{R}^d \to \mathbb{R}$ para o oráculo $r : \mathbb{R}^d \to \mathbb{R}$. Para evitar esse tipo de arbitrariedade, podemos considerar todas as possíveis combinações de conjuntos de treino $\mathcal{L}_T$ e validação $\mathcal{L}_V$ através do *K-fold*. Dessa forma, a estimação de $\mathbb{E}[Z|\boldsymbol{x}]$ é feita pela média dos erros preditos obtidos em cada combinação de conjunto de treino e validação, tendo a observação a ser consultada dada pela fórmula:

$$\boldsymbol{x}^*_{RV} = \arg\max_{\boldsymbol{x} \in \mathcal{U}} \frac{1}{k} \sum_{i=1}^{k} \hat{\mathbb{E}}_{\mathcal{L}_{V_i}}[Z|\boldsymbol{x}]\,, \tag{3.4}$$

sendo k o número fixado de folds. O Algoritmo 4 ilustra o algoritmo de aprendizado ativo para nosso novo método. Ao contrário dos algoritmos mais comuns no contexto de regressão ativa (Burbidge *et al.*, 2007; Cohn *et al.*, 1996; Wu, 2018), nosso algoritmo suporta modelos genéricos para $g(\boldsymbol{x})$, não limitando este a processos gaussianos, modelos lineares ou modelos de *ensemble*, visto que dependemos apenas das predições de $g(\boldsymbol{x})$ para selecionar as observações. Porém, ao mesmo tempo, também requeremos do usuário um modelo de validação $h(\boldsymbol{x})$ que pode ser relativamente arbitrário de se escolher.

Ressalta-se também que o modelo de validação $h : \mathbb{R}^d \to \mathbb{R}$ pode ser diferente de

g para potencialmente evitar maior enviesamento do aprendizado ativo e captar melhor possíveis comportamentos dos erros Z que possivelmente o modelo $g : \mathbb{R}^d \to \mathbb{R}$ original não captaria, como é o caso do exemplo da Figura 3.1 em que os dados são não lineares e o erro também é relativamente não linear, e o modelo de validação de floresta aleatória consegue captar isso ao contrário do modelo $g(\boldsymbol{x})$ de regressão linear utilizado.

Algoritmo 4 Redução de viés

$\mathcal{U}$ = conjunto de instâncias não rotuladas $\{\boldsymbol{x}_i\}_{i\in\mathcal{U}}$
$\mathcal{L}$ = conjunto inicial de instâncias rotuladas $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$
k = Número escolhido de folds
n = Número escolhido de consultas
for $t = 1, 2, \ldots, n$ **do**
    E = Matriz de tamanho $|\mathcal{U}| \times k$ para armazenar os erros de cada combinação de *fold* para cada observação em $\mathcal{U}$
    $\mathcal{K}$ = **atribuirfolds**($\mathcal{L}$, k) (Vetor do indice de folds)
    for $i = k, \ldots, 1$ **do**
        $\mathcal{L}_V$ = Selecionar observações com $\mathcal{K} = i$
        $\mathcal{L}_T$ = Selecionar o restante das observações
        g = **ajuste**($\mathcal{L}_T$, resposta = Y)
        $Z = (\textbf{predizer}(g, \mathcal{L}_V) - \textbf{selecionar}(\mathcal{L}_V, Y))^2$
        h = **ajuste**($\mathcal{L}_V$, resposta = Z)
        E[,i] = **predizer**(h, $\mathcal{U}$)
    end for
    média = **MediaPorLinha**(E)
    selecione $\boldsymbol{x}^* \in \mathcal{U}$ a instância com maior média
    obtenha y^* do oráculo
    adicione $(\boldsymbol{x}^*, y^*)$ a $\mathcal{L}$
    remova $\boldsymbol{x}^*$ de $\mathcal{U}$
end for

Na mesma Figura 3.1 notamos em geral como funciona o método de Redução de Viés. Vemos que para o conjunto de treino inicial ($n = 0$), as maiores estimativas de erro estão nos menores valores de X, e logo após isso, com $n = 5, 15$ e 30 consultas, vemos que o erro tende a ser maior em regiões antes não muito exploradas, que possuem certa não linearidade não captada pela amostra inicial em vermelho, tendo a estimativa do erro um formato relativamente flexível. Ao final, em $n = 30$, vemos que a relação não linear entre Y e X é praticamente explicitada pelo aprendizado ativo, tendo uma maior concentração de pontos consultados no pico da curva não linear entre Y e X. Em geral, vemos exemplificado pela Figura 3.1 uma maneira de quantificar incertezas para qualquer modelo de regressão ativa predizendo erros para isso.

A Figura 3.2 nos mostra o mesmo processo mas utilizando um knn como modelo de

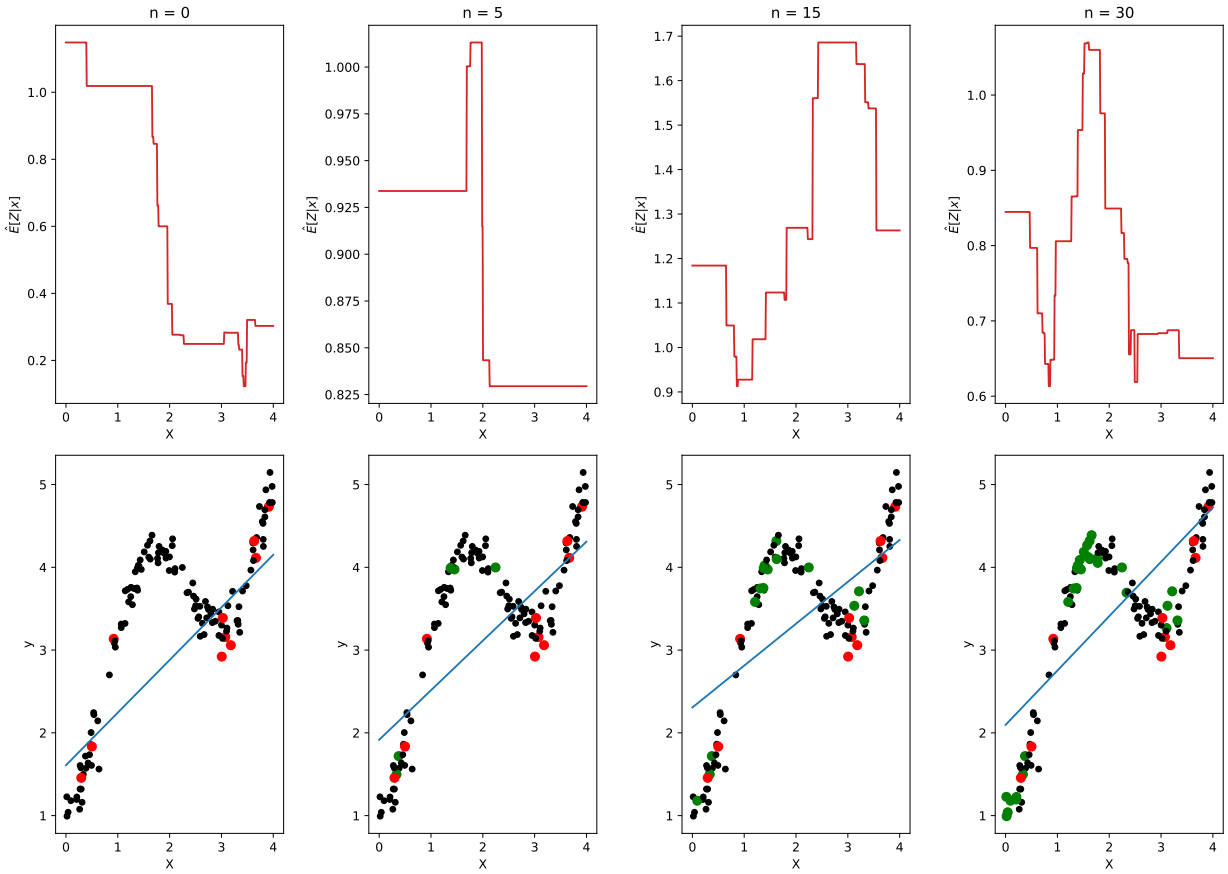


Figura 3.1: Acima: $\hat{\mathbb{E}}[Z|\boldsymbol{x}]$ estimado para cada x em diferentes números de consulta. Abaixo: Pontos amostrados em cada iteração do aprendizado ativo (em verde) e amostra inicial de treinamento (em vermelho). O modelo para ajuste dos dados escolhido foi a regressão linear e o modelo para validação foi a floresta aleatória.

validação. Nota-se certa mudança quanto ao formato dos erros preditos em comparação a Figura 3.1, tendo no inicio erros iguais para todos X, e depois, para $n = 5$, erros maiores para $X < 2$. Já, para $n = 15, 30$, nota-se maiores incertezas em regiões de não linearidade, havendo novamente uma grande concentração de consultas no pico da curva não linear entre Y e X. Em particular, vemos que para $n = 30$ temos dois picos em $\hat{\mathbb{E}}[Z|x]$ concentrados tanto no pico quanto no vale não linear entre Y e X. Logo, apesar de existirem diferenças entre as estimativas de $\hat{\mathbb{E}}[Z|x]$ tomando o modelo de validação como um KNN ou floresta aleatória, em ambos casos nota-se que as consultas feitas são relativamente similares, com estas concentrando-se em regiões mais não lineares (pico e vale) ou com X muito pequeno (próximo de zero).

Nas próximas seções, desenvolveremos diversos experimentos utilizando dados simulados e reais, comparando o método de Redução de Viés tanto com o aprendizado passivo (amostragem aleatória) quanto com métodos de regressão ativa conhecidos como a

amostragem por incerteza e a consulta por comitê (Burbidge *et al.*, 2007). Salienta-se que estamos considerando uma versão relativamente crua de nosso método, tomando a variância intrínseca $\mathbb{V}[Y|\boldsymbol{X}]$ como constante e o modelo de validação como entrada do usuário.

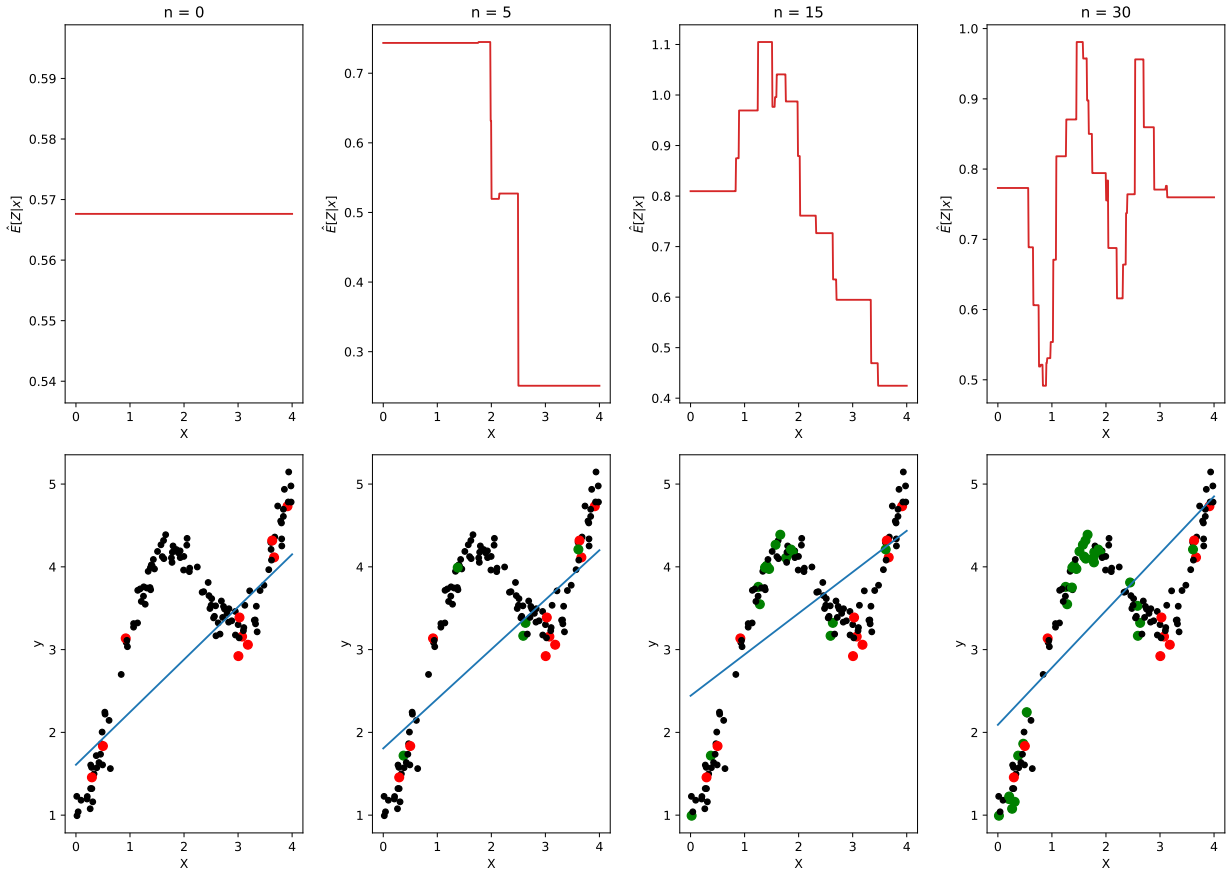


Figura 3.2: Encima: $\hat{\mathbb{E}}[Z|\boldsymbol{x}]$ estimado para cada x em diferentes números de consulta. Embaixo: Pontos amostrados em cada iteração do aprendizado ativo (em verde) e amostra inicial de treinamento (em vermelho). O modelo para ajuste dos dados escolhido foi a regressão linear e o modelo para validação foi o knn.

Capítulo 4

Resultados e Experimentos

Neste trabalho, serão exploradas algumas das diferentes metodologias de aprendizado ativo expostas no Capítulo 2, focando-se principalmente nas técnicas e no contexto associado a regressão. Para o contexto de classificação especificamente será aplicado e discutido o método de amostragem por incerteza (Lewis e Gale, 1994) em um problema binário simulado, com o intuito de se ilustrar como se dá o viés na amostragem e como o aprendizado ativo pode ser melhor ou pior que o aprendizado passivo. Já, no contexto de regressão, será explorada a nova metodologia proposta no Capítulo 3, comparando-a ao aprendizado passivo e às metodologias de regressão ativa já existentes: amostragem por incerteza e consulta por comitê (Burbidge *et al.*, 2007), buscando entender o potencial desse novo método para regressão ativa. O método de redução de variância não será utilizado pela complexidade de seu algoritmo. Todos os experimentos de aprendizado ativo serão implementados em *python*, utilizando-se as bibliotecas *scikit-learn* (Pedregosa *et al.*, 2011) para a implementação dos diferentes modelos de aprendizado de maquinas e *modAL* (Danka e Horvath, 2018) para a implementação do aprendizado ativo.

4.1 Aprendizado ativo na classificação binária

Desenvolveremos a seguir o caso mais básico de aprendizado ativo, analisando-se as situações que o aprendizado passivo pode ser melhor que o aprendizado ativo, e como podemos aplicar o aprendizado ativo de forma satisfatória em um exemplo binário

4.1.1 Um exemplo simulado

Para uma experimentação inicial mais simples, entendendo mais claramente as vantagens do aprendizado ativo e sua relação com o passivo, podemos inicialmente gerar um banco de dados simulado como mostrado pela Figura 4.1 em que a classe 1 tem um formato de disco interno a classe 0. Temos interesse em averiguar a convergência do

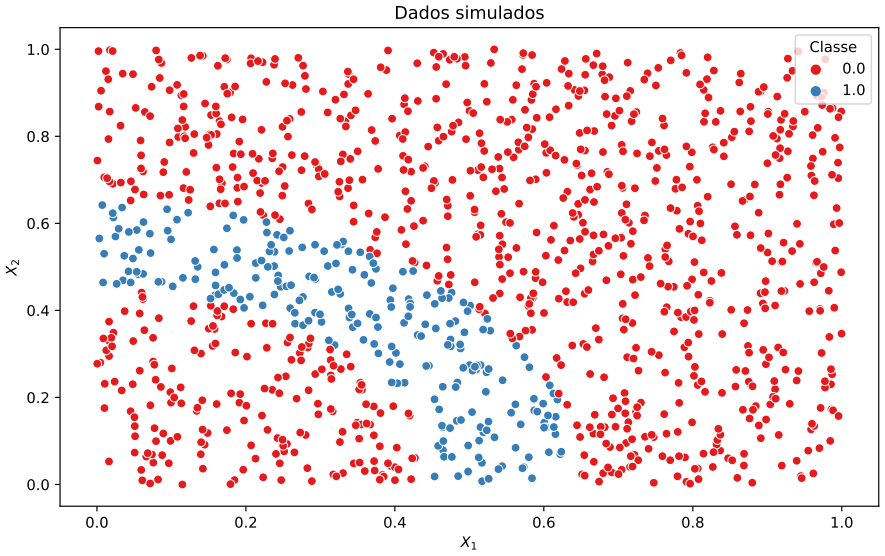


Figura 4.1: Dados simulados pela função $\mathbb{I}(X_{i,1}^2 + X_{i,2}^2 \leq (0.65)^2) \cdot \mathbb{I}(X_{i,1}^2 + X_{i,2}^2 \geq (0.45)^2)$ com $n = 1200$ observações

aprendizado ativo para tal banco de dados simulado, analisando-se o desempenho de tal método com relação ao aprendizado passivo. Para esse formato de disco, é adequado de se ajustar modelos não paramétricos que captem bem padrões não lineares. Para tal, inicialmente, podemos ajustar um *Support Vector Machine* com Kernel gaussiano (RBF) com hiper-parâmetro de regularização $C = 1$ a um conjunto de treino fixo de tamanho 20 mostrado pela Figura 4.2.

É interessante notar que apesar da pouca informação sobre a não linearidade das distribuições das classes, já descarta-se modelos lineares como, por exemplo, a regressão logística e o *SVM* com Kernel linear dado uma separação curva entre os rótulos. Ainda assim, com base no formato dos dados de treinamento, poderíamos equivocadamente escolher um Kernel polinomial ou sigmoide que não modelariam muito bem o formato visualizado na Figura 4.1. Isso deixa claro que o fato de a amostra de treinamento ser relativamente pequena pode levar a escolha errada de modelos e um possível enviesamento sobre a região de incerteza.

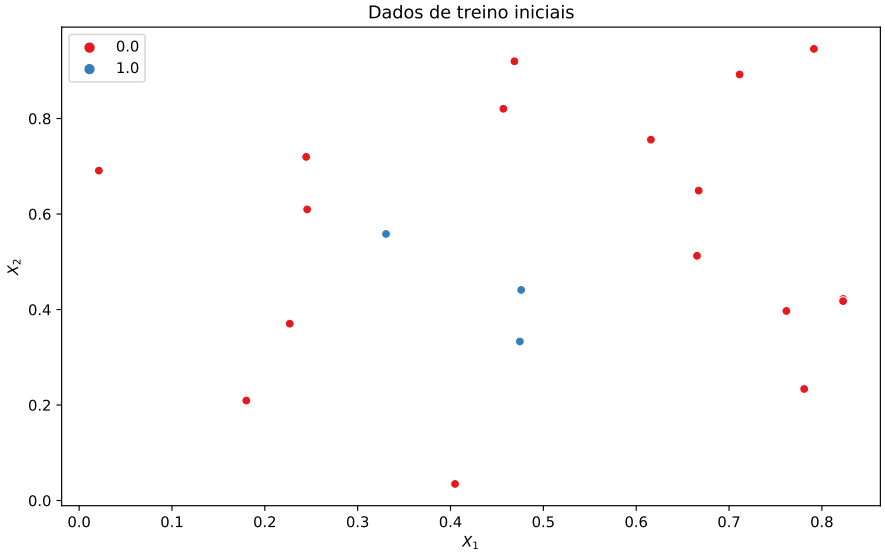
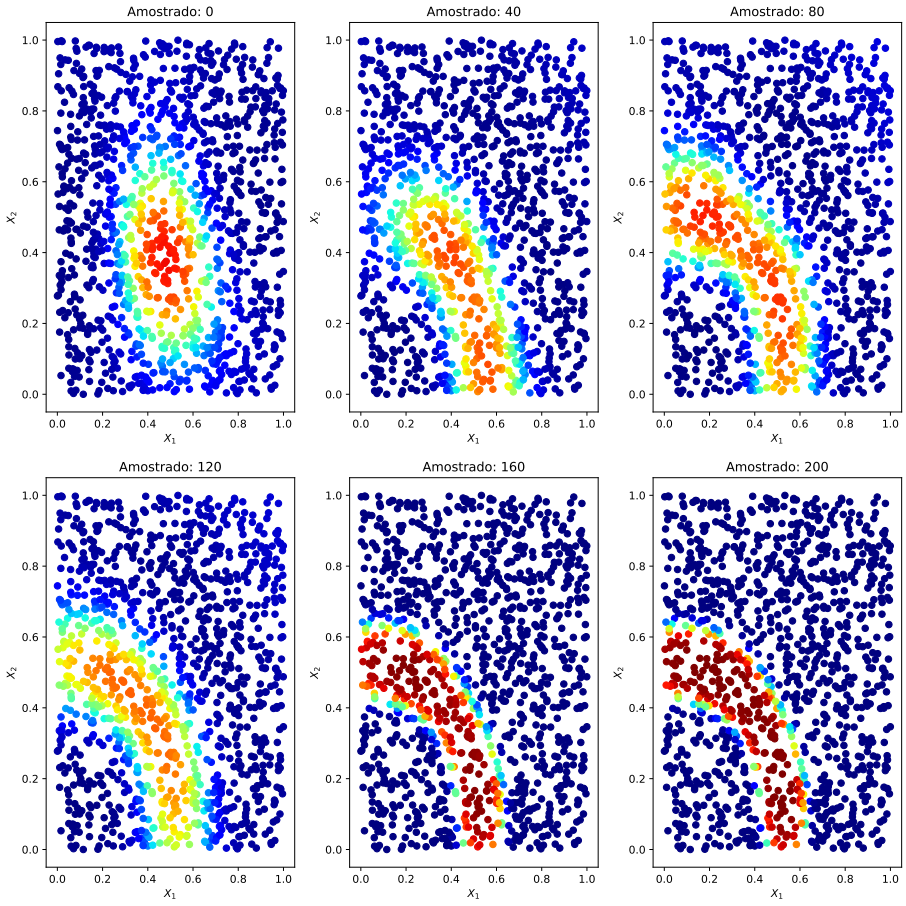


Figura 4.2: Dados de treinamento inicial

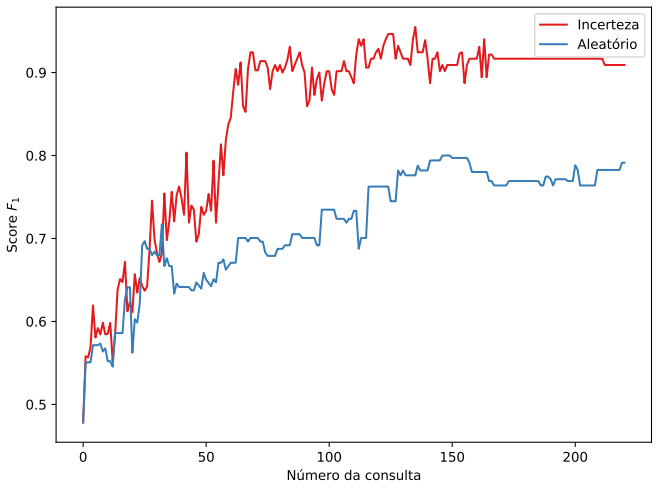
Voltando à experimentação, tendo um conjunto de teste $\mathcal{T}$ de tamanho $n_{\mathcal{T}} = 240$ e um conjunto não rotulado $\mathcal{U}$ de tamanho $n_{\mathcal{U}} = 940$, fazemos 220 consultas sobre o banco de dados, rotulando as instâncias $\boldsymbol{x}$ de acordo com a estratégia da amostragem por incerteza sob o modelo *SVM* descrito. Paralelamente, utilizando-se os mesmos conjuntos $\mathcal{T}$ e $\mathcal{U}$, escolhe-se cada instância $\boldsymbol{x}$ para rotulação de forma aleatória. A Figura 4.3 mostra os resultados dessa experimentação. Percebe-se primeiramente na Figura 4.3a que gradativamente, a medida que se avança as consultas, a probabilidade estimada pelo modelo *SVM* vai convergindo para a distribuição vista na Figura 4.1, mostrando um resultado satisfatório do aprendizado ativo. Além disso, notamos pela Figura 4.3b uma convergência do *score* F_1 em valores mais satisfatórios na amostragem por incerteza do que na amostragem aleatória.

Vemos então que ao escolher um modelo adequado para a modelagem dos dados simulados, o aprendizado ativo tem de fato excelente retornos, com uma convergência relativamente rápida para o resultado esperado na Figura 4.1 em comparação a amostragem aleatória que estabiliza em menores valores de F_1. Ainda podemos averiguar quais observações cada metodologia amostrou em cada etapa na Figura 4.4. Visualiza-se na amostragem por incerteza que todas as observações consultadas estão perto da fronteira entre as classes 0 e 1 vista na Figura 4.1, principalmente as 50 primeiras observações. Já a amostragem aleatória não apresenta um padrão específico, o que é esperado, visto que simplesmente estamos sorteando a observação a ser rotulada em cada iteração. Portanto,

com o modelo certo, o aprendizado ativo é de fato mais inteligente e eficiente na busca de observações mais informativas em comparação a amostra aleatória.



(a) Probabilidades $\hat{P}(Y = 1|\boldsymbol{x})$ estimadas para cada ponto simulado em diferentes números de consultas no aprendizado ativo



(b) *Score* F_1 por consulta para a amostragem por incerteza (vermelho) e amostragem aleatória (azul)

Figura 4.3: Resultados do aprendizado ativo usando o SVM com Kernel gaussiano.

Além do SVM, podemos utilizar nesse exemplo também a floresta aleatória e o KNN que são modelos também adequados e flexíveis para a predição de classes com padrões não lineares. Compara-se novamente o desempenho dos aprendizados ativo e passivo para ambos modelos na Figura 4.5.

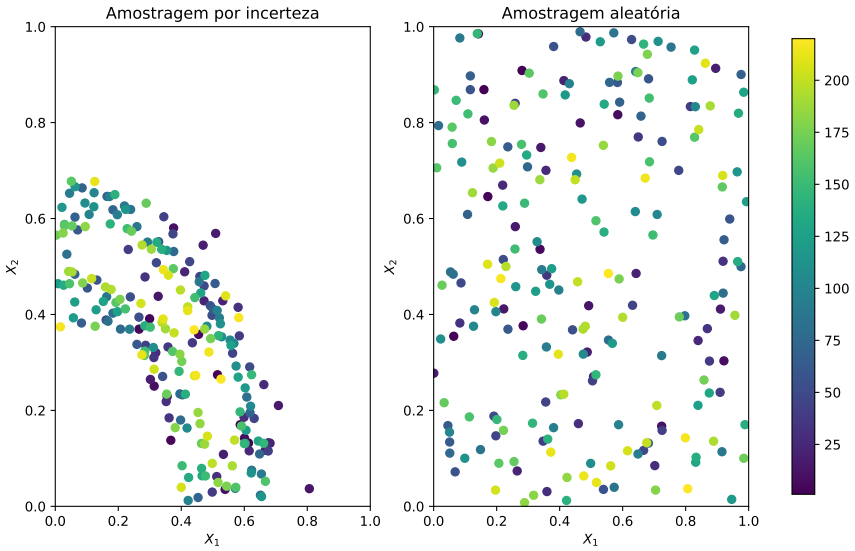


Figura 4.4: Observações amostradas em cada consulta para o aprendizado ativo e passivo no modelo SVM com Kernel gaussiano e hiper-parâmetro de regularização $C = 1$

Visualiza-se novamente uma convergência do aprendizado ativo em valores satisfatórios de F_1 para ambos modelos, tendo um melhor desempenho que a amostragem aleatória.

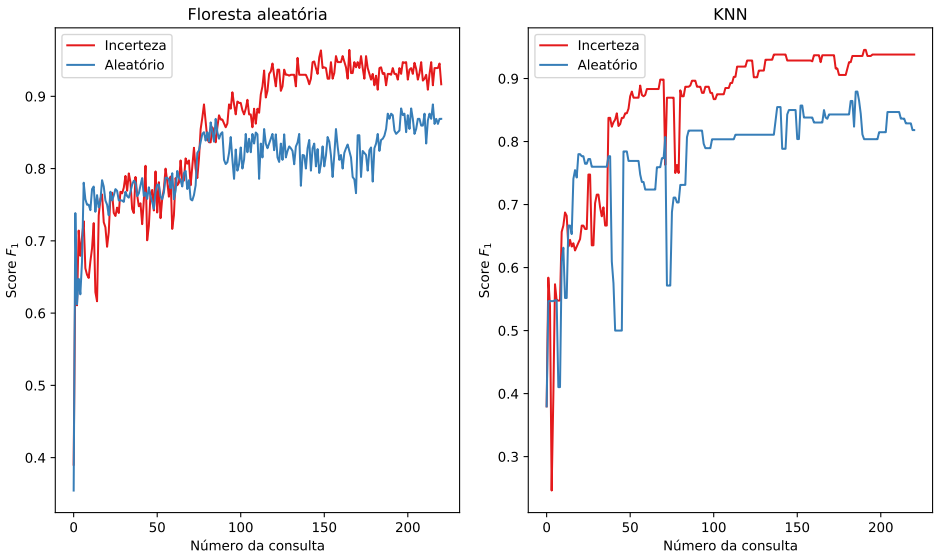


Figura 4.5: F_1 por consulta para o aprendizado ativo e passivo nos modelos de Floresta Aleatória e KNN

Para todos esses experimentos conduzidos, obtivemos resultados satisfatórios para o aprendizado ativo em termos de melhora na performance dos modelos. No entanto, o

aprendizado ativo ainda tem grandes limitações e pode não render bons resultados dependendo do modelo utilizado ou amostra inicial. Isso pode ser ilustrado tomando novamente o modelo *SVM* com kernel gaussiano, agora com hiper-parâmetro de regularização $C = 0.01$, aplicando aprendizado ativo e passivo neste classificador. A Figura 4.6 mostra um pior desempenho do aprendizado ativo em comparação ao passivo, tendo uma grande oscilação do *score* para a amostragem por incerteza, enquanto a amostragem aleatória apresenta uma boa convergência deste. No final,o aprendizado passivo é preferível ao aprendizado ativo neste caso.

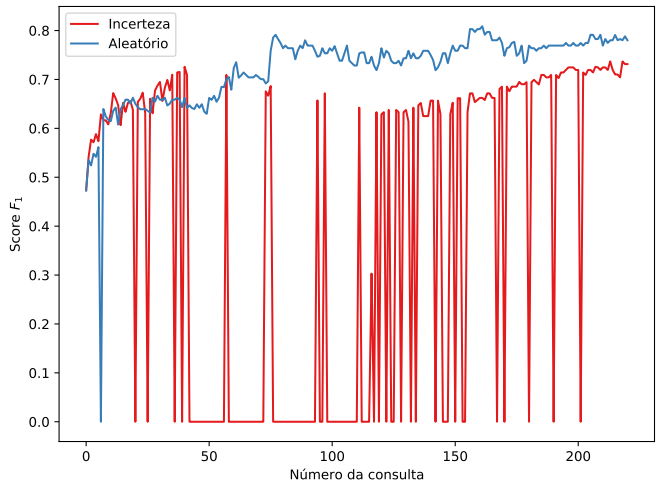


Figura 4.6: F_1 por consulta para o aprendizado ativo e passivo no modelo SVM com hiper-parâmetro de regularização $C = 0.01$. Dessa vez, a amostragem aleatória mostrou-se mais satisfatória.

Para entender o porquê de isso acontecer, a Figura 4.7 mostra quais observações foram amostradas em cada iteração do aprendizado ativo. Vemos dessa vez uma maior dispersão dos pontos, com a região de consulta tendo um formato circunférico ao invés de um formato de disco. De fato, vemos que os pontos localizados mais à lateral esquerda do gráfico, entre 0.6 e 0.4, não foram amostrados, havendo um maior foco no centro do disco. Ou seja, o classificador escolhido teve maiores incertezas sobre os valores do centro do disco, modelando a região de fronteira como um círculo ao invés de um disco. Isso pode se dever ao fato de que quando diminuímos C, aumentamos mais o *overfitting* sobre o conjunto de treinamento, fazendo com que o modelo enxergue os dados de forma estritamente circular e amostre as próximas observações com base nisso.

Portanto, vemos nesse caso um pior desempenho do aprendizado ativo pelo fato de a amostragem estar viesada às limitações do modelo ao tomarmos $C = 0.01$. Como a

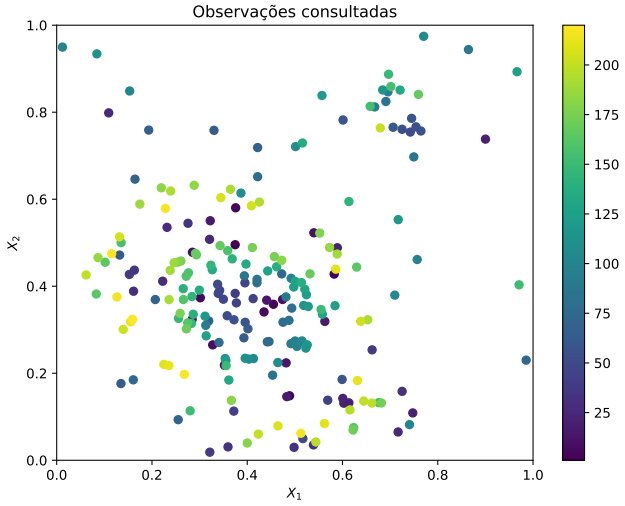


Figura 4.7: Instâncias amostradas em cada iteração do aprendizado ativo

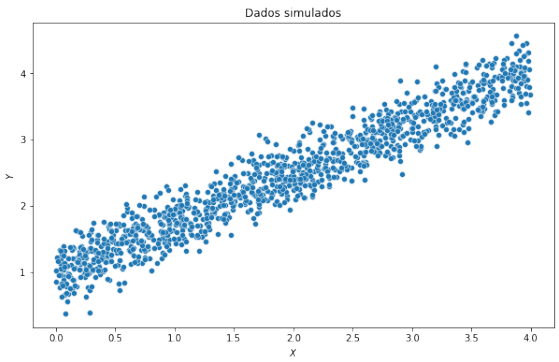
amostragem aleatória é totalmente agnóstica às limitações e especificações do modelo, ela não é afetada pelo viés de amostragem e portanto, rende melhores performances para o modelo. Por fim, tal experimento ilustra de forma simples os cuidados que devemos ter ao utilizar aprendizado ativo discutidos no Capítulo 2, elucidando-se a necessidade de se ter certa confiança sobre o modelo utilizado.

4.2 Novo método de regressão ativa

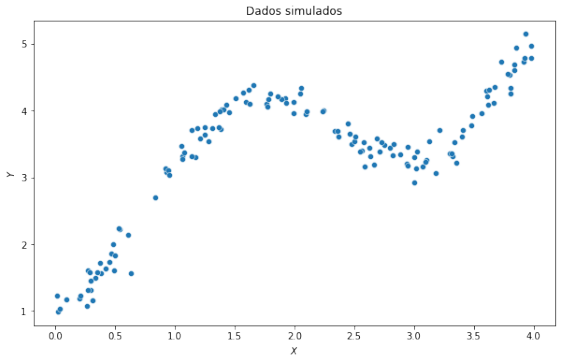
A seguir iremos realizar diversas experimentações buscando comparar nosso método a amostragem aleatória e a outros dois métodos de regressão ativa presentes na literatura, estudando a qualidade e veracidade de nosso método e analisando se este de fato apresenta alguma utilidade no contexto de regressão ativa. Tais experimentações serão conduzidas em dados simulados e reais, tendo tais dados melhores descritos a seguir:

- **Dados simulados**: Foram feitas 3 simulações univariadas distintas, cada uma nomeada de D1, D2, D3 respectivamente, tendo cada uma um total de $N = 1200, 200$ e 200 observações simuladas respectivamente. Seus formatos e fórmulas são mostrados na Figura 4.8.

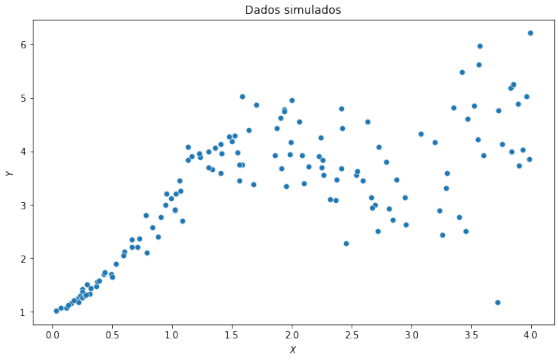
  Assim, D1 são dados lineares simples, D2 são dados não lineares também simples e D3 são dados não lineares heterocedasticos. Além disso, são simulados mais 3 conjuntos de alta dimensão ao adicionar 30 covariáveis $X_2, \ldots, X_{30}$ sem relação alguma com Y, adicionando-se vários ruídos nos dados. Tais conjuntos serão denominados



(a) Dados 1: $Y_i = 1 + 0.75X_i + \varepsilon_i$



(b) Dados 2: $Y_i = 1 + 0.75X_i + 2 \cdot \text{sen}(X_i)^2 + \varepsilon_i$



(c) Dados 3:$Y_i = 1 + 0.75X_i + 2 \cdot \text{sen}(X_i)^2 + \varepsilon_i$ com $\varepsilon_i \sim \text{Normal}(0, \frac{X_i}{4})$

Figura 4.8: Dados simulados para comparação dos métodos de regressão ativa

D1-dim, D2-dim e D3-dim. Portanto, consideraremos ao total 6 conjuntos simulados no estudo, tendo um melhor entendimento do que está acontecendo em cada situação, visto que conhecemos a natureza de nossos dados. A Tabela 4.1 sumariza os dados simulados que serão utilizados.

Tabela 4.1: Descrição dos dados simulados

Dados simulados	**Característica**
D1	Linear com $N = 1200$
D2	Não linear com $N = 200$
D3	Não linear heterocedastico com $N = 200$
D1 - dim	Linear com ruído (30 covariaveis a mais)
D2 - dim	Não linear com ruído (30 covariaveis a mais)
D3 - dim	Não linear heterocedastico com ruído (30 covariaveis a mais)

- **Dados reais**: Para os dados reais, foram considerados 4 bancos de dados distintos disponíveis no repositório UCI [1], sendo eles: a base de dados de aerofólio, de vinho branco, vinho vermelho e concreto. A Tabela 4.2 sumariza as informações dos dados reais a serem utilizados.

[1] https://archive.ics.uci.edu/ml/datasets.php

Tabela 4.2: Sumário dos bancos de dados utilizados para os experimentos

Banco de dados	**Número de observações**	**Número de variáveis**
Aerofólio	1503	6
Vinho branco	4898	12
Vinho vermelho	1599	12
Concreto	1030	9

Para os dados simulados, consideramos uma amostra inicial de tamanho $|\mathcal{L}| = 10$ e número de consultas $n = 30$, enquanto para os dados reais consideramos a amostra inicial como $|\mathcal{L}| = 16$ e um número de consultas $n = 60$. Salienta-se que cada conjunto de dados foi dividido em *pool* ($\mathcal{U}$) e teste ($\mathcal{T}$), com proporção de 0.3 de teste para os dados simulados e 0.2 de teste para os dados reais. A amostra inicial de treinamento é então extraída da *pool*, compondo inicialmente o conjunto de treinamento $\mathcal{L}$. Para garantir resultados consistentes, amostrou-se 30 amostras iniciais diferentes para os dados simulados e 35 para os dados reais, calculando-se ao final a curva de risco média para cada método. Dessa maneira diminuímos possíveis flutuações nas curvas dos riscos e potenciais vieses causados por certas amostras iniciais.

Quanto a nosso método, ressalta-se que utilizaremos o número de *folds* $k = 2$, garantindo que os conjuntos de treino e validação tenham tamanho parecido e o algoritmo seja mais rápido de se executar. Para comparar nosso método com a amostragem por incerteza e consulta por comitê utilizaremos como modelo de regressão ativa processos gaussianos para amostragem por incerteza, e um *ensemble* composto por uma floresta aleatória, knn e *bagging* de árvores de regressão para a consulta por comitê. Já para comparar nosso método a amostragem aleatória, consideraremos como modelos de regressão ativa $\hat{g}(\boldsymbol{x})$ os modelos de regressão linear, floresta aleatória e knn. Em ambas comparações, utilizaremos como modelos de validação para nosso método os modelos de regressão linear, floresta aleatória e knn. Por fim, salienta-se que será tomado como risco o erro quadrático médio dado pela fórmula:

$$\hat{R}_{pred}(\hat{g}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{g}(\boldsymbol{x}_i))^2 \ ,$$

sendo este calculado em cada consulta no conjunto de teste.

4.2.1 Experimentos simulados: redução de viés versus amostragem aleatória

A seguir, compararemos as curvas de risco para cada método em cada conjunto de dados simulado.

- **D1**: Para os dados lineares, obteve-se os gráficos apresentados na Figura 4.9.

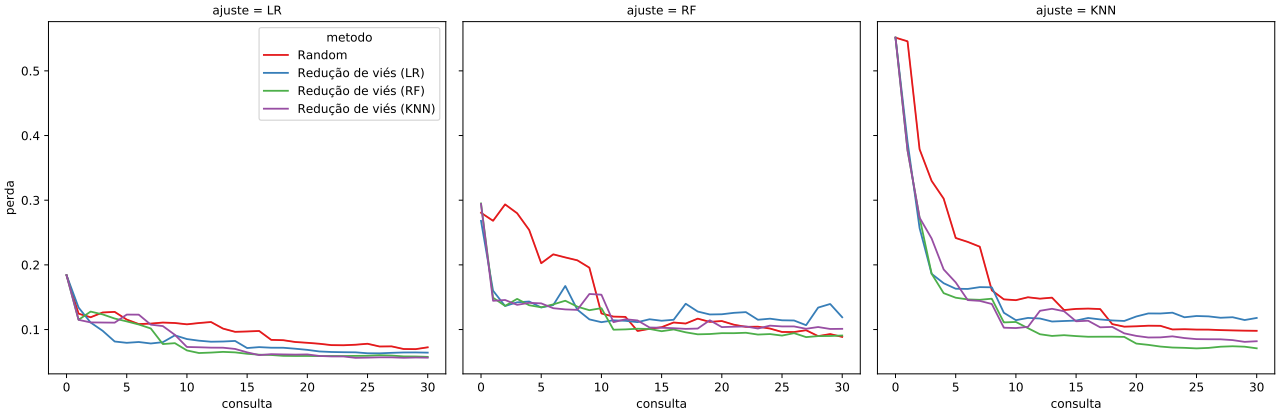


Figura 4.9: Curvas de risco do aprendizado ativo para dados simulados lineares. Colunas: Modelos de aprendizado ativo utilizados para o ajuste.

Vemos pela Figura 4.9 que para o ajuste da regressão linear, nosso método possui curvas menores que a amostragem aleatória para todos os métodos de validação, enquanto para os dois modelos não paramétricos, apenas os modelos de validação de floresta aleatória e knn possuem menores curvas que a amostragem aleatória. Em particular, notamos que para um número de consultas menor/igual a 10, o método de redução de viés já possui curvas razoavelmente menores que a amostragem aleatória para todos os ajustes, não tendo grandes mudanças na curva após as 10 consultas. Ou seja, o método nestes casos parece convergir rapidamente para valores pequenos de risco. Como esse conjunto de dados simulado é relativamente trivial, esse resultado já nos mostra que nosso método é ao menos minimamente vantajoso.

- **D2**: A Figura 4.10 nos dá as curvas de risco para os dados não lineares.

  Para os dados não lineares, notamos que a redução de viés com um modelo de validação de regressão linear possui um desempenho ruim em comparação a amostragem aleatória para todos os modelos de ajuste, tendo um aumento da perda a medida que se aumenta as consultas para o modelo de ajuste de regressão linear.

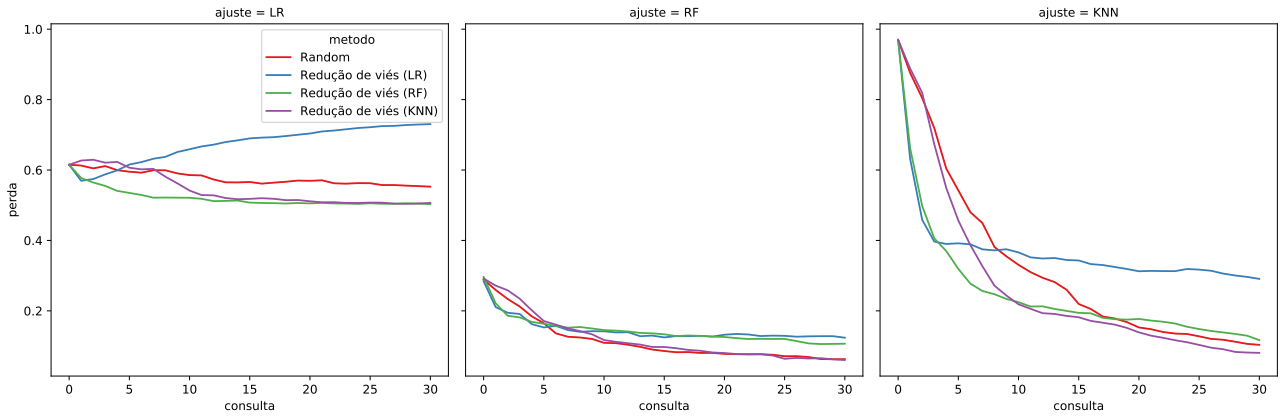


Figura 4.10: Curvas de risco do aprendizado ativo para dados simulados não lineares homocedásticos. Colunas: Modelos de aprendizado ativo utilizados para o ajuste.

Isso é justificado pelo fato de que a regressão linear para validação não capta bem o comportamento não linear dos erros para cada modelo.

Já se tomarmos os modelos não paramétricos como modelos de validação, notamos curvas de risco satisfatórias para os ajustes de regressão linear e knn (menores que a amostragem aleatória) e similares a amostragem aleatória para o ajuste da floresta aleatória (apenas usando knn como modelo de validação). Em particular, notamos que para os ajustes de knn e regressão linear, nosso método já possui curvas razoavelmente menores que a amostragem aleatória para um número de consultas menor/igual a 10. Em geral, a curva do modelo de validação do knn está consideravelmente menor ou igual a amostragem aleatória para todos os ajustes. Assim, para este caso, o modelo de validação do knn produz um melhor desempenho nos modelos, com os resultados ainda nos mostrando um bom potencial para nosso método.

- **D3**: A Figura 4.11 mostra as curvas de risco para os dados não lineares heterocedásticos. Nota-se que, apesar da heterocedasticidade, nosso método parece ter um bom desempenho nos ajustes da regressão linear e floresta aleatória, tendo praticamente todos os modelos de validação um melhor desempenho que a amostragem aleatória para a última consulta, tendo ao menos 2/3 curvas abaixo da curva da amostragem aleatória para todas as consultas.

  Em particular destaca-se a regressão linear para os dois ajustes, tendo um melhor desempenho entre os demais modelos, algo que pode ser explicado pelo fato dos dados se encaixarem bem em uma regressão linear dado que variância dos dados é crescente, atenuando o comportamento não linear visualizado nos Dados 2. Ressalta-

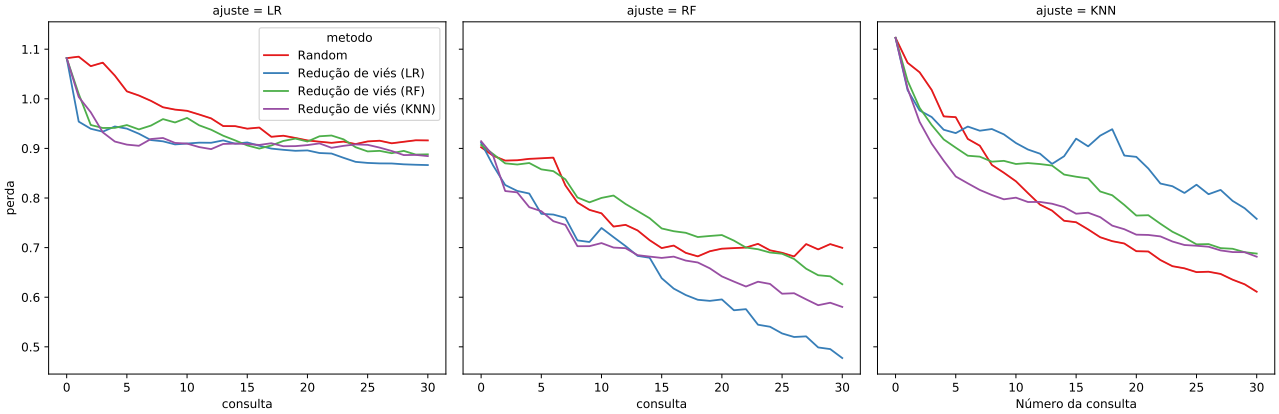


Figura 4.11: Curvas de risco do aprendizado ativo para dados simulados não lineares heterocedásticos. Colunas: Modelos de aprendizado ativo utilizados para o ajuste.

se principalmente a curva de risco da floresta aleatória com validação pela regressão linear, que tem o melhor desempenho final dentre todos os modelos da figura, e tem uma curva acentuadamente decrescente.

Porém, para o knn vemos que, em geral, a amostragem aleatória tem curva inferior a todas as demais curvas do método de redução de viés, tendo a curva com validação do knn a mais próxima da curva da amostragem aleatória. Assim, vemos que para dados não lineares heterocedásticos nosso método tem um desempenho relativamente bom sendo que ainda não adicionamos ao algoritmo uma estimativa da variância intrínseca $\mathbb{V}[Y|\boldsymbol{X}]$, um elemento que pode melhorar consideravelmente nosso método para esses tipos de dados.

- **D1-dim**: Plotando agora as curvas de risco para dados simulados lineares com ruído, obtemos a Figura 4.12. Percebe-se nesta que o modelo de validação de floresta aleatória em geral possui as menores curvas, destacando-se principalmente nos ajustes com Lasso e knn, em que se percebe uma grande diferença do método de redução de viés para a amostragem aleatória. Para o ajuste com florestas aleatórias, verifica-se poucas diferenças entre as curvas, com todas as curvas tendo um comportamento similar, apesar de na consulta 30 o método de redução de viés com modelo de validação de floresta aleatória resultar em um modelo com desempenho um pouco melhor que os demais modelos produzidos.

  A pequena diferença entre os métodos no modelo de regressão ativa de floresta aleatória pode ser explicado pelo fato da floresta aleatória já ser um algoritmo que realiza seleção de variáveis, estabelecendo um grau de importância para as variáveis.

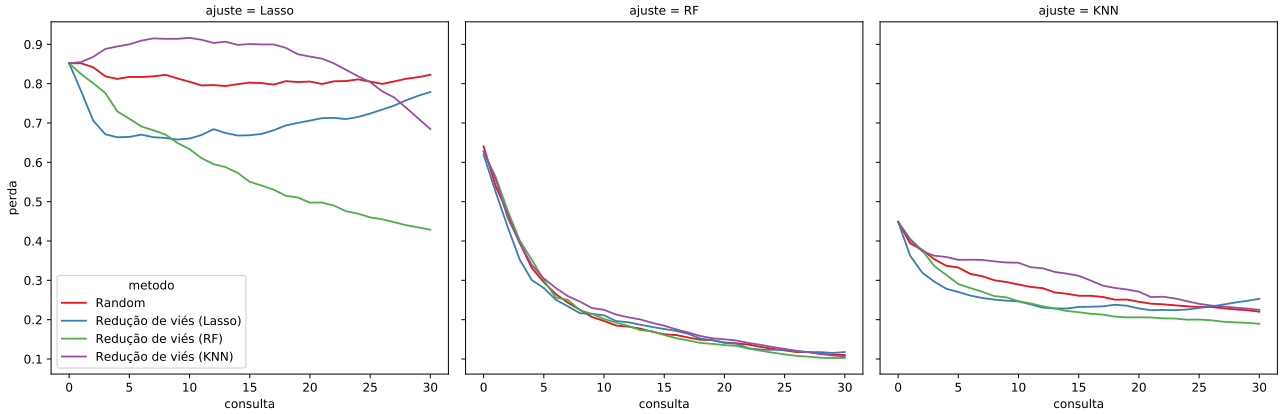


Figura 4.12: Curvas de risco do aprendizado ativo para dados simulados lineares com ruído. Colunas: Modelos de aprendizado ativo utilizados para o ajuste.

Dessa forma, apenas X_1 acaba sendo considerado de fato importante, visto que as demais variáveis são ruídos, sendo portanto um modelo facilmente melhorável, apenas dependendo de novas consultas para isso.

- **D2-dim**: Para os dados simulados não lineares com ruído obtemos as curvas de risco da Figura 4.13.

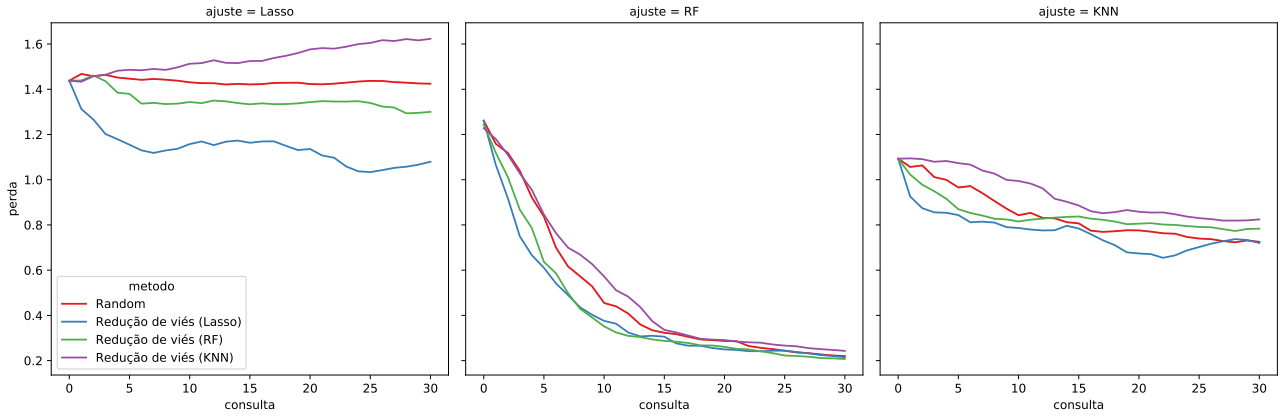


Figura 4.13: Curvas de risco do aprendizado ativo para dados simulados não lineares com ruído. Colunas: Modelos de aprendizado ativo utilizados para o ajuste.

Para os dados não lineares com ruído, percebe-se uma boa performance de nosso método nos ajustes do lasso e floresta aleatória, considerando os mesmos também como modelo de validação, tendo curvas abaixo da amostragem aleatória para os dois modelos de validação nos dois ajustes considerados. Já para o ajuste com knn, percebemos que o método de redução de viés utilizando o lasso como modelo de validação produz bons resultados, porém possui um comportamento indesejável de crescimento da curva após a consulta 20, tendo depois um pequeno decrescimento

na consulta 30. Isso também pode se dever ao fato do knn naturalmente ser um método pior para dados ruidosos, tendo um alto distúrbio no ajuste ao adicionar-se novas observações que são potencialmente informativas de acordo com o lasso.

- **D3-dim**: Por fim, para os dados não lineares heterocedásticos com ruídos, obtemos as curvas de risco mostradas na Figura 4.14.

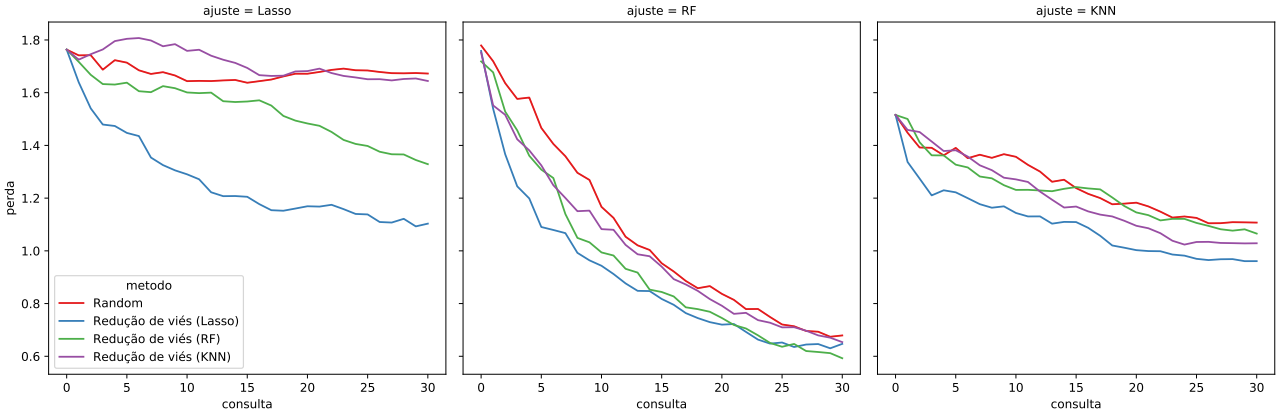


Figura 4.14: Curvas de risco do aprendizado ativo para dados simulados não lineares heterocedásticos com ruído. Colunas: Modelos de aprendizado ativo utilizados para o ajuste.

Destaca-se na Figura 4.14 uma relevante melhora de desempenho em todos os ajustes ao tomar o lasso como modelo de validação, tendo para todos ajustes uma curva muito abaixo da amostragem aleatória. O modelo de validação de floresta aleatória também possui um bom desempenho em todos os modelos, tendo também uma curva de risco relativamente menor que a da amostragem aleatória.

Em resumo, notamos através desses diferentes exemplos simulados um desempenho promissor de nosso método para a regressão ativa, tendo em geral, melhores resultados que o aprendizado passivo, principalmente em situações de mais ruído. De fato, percebe-se um excelente desempenho do método de redução de viés em dados com relativo ruído, tendo o uso de modelos de validação que realizem seleção de variáveis para garantir certa robustez a ruídos.

4.2.2 Experimentos simulados: redução de viés versus amostragem por incerteza e consulta por comitê

A seguir, compararemos nosso método com a amostragem por incerteza e consulta por comitê. Iremos partir diretamente dos dados não lineares, ignorando os dados lineares e

lineares com ruído por esses serem conjuntos muito triviais para essas comparações que faremos. Para realizar as comparações entre os métodos, utilizaremos processos gaussianos para comportar a amostragem por incerteza e um modelo de ensemble para comportar a consulta por comitê, utilizando esses mesmos modelos como modelos de ajuste para nosso método. Também utilizaremos a amostragem aleatória para cada exemplo, tendo uma linha de base para os métodos de aprendizado ativo comparados em cada banco de dados.

- **D2**: Para os dados não lineares, obtemos:
  - **RV versus Amostragem por incerteza**: As curvas de risco dos métodos de amostragem por incerteza, redução de viés e amostragem aleatória são mostradas na Figura 4.15. Percebe-se que a redução de viés com modelos de va-

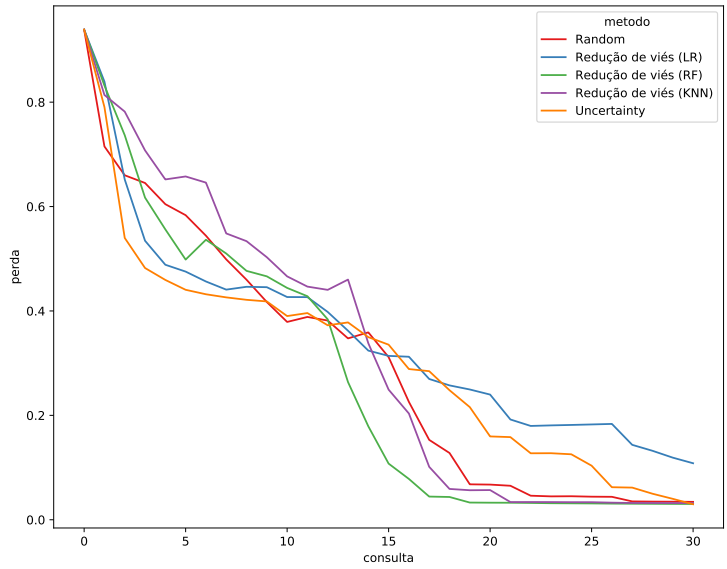


Figura 4.15: Curvas de risco do aprendizado ativo para cada método nos dados simulados não lineares, comparando nosso método de aprendizado ativo à amostragem por incerteza e amostragem aleatória

lidação da floresta aleatória e knn decaem rapidamente para um risco próximo de zero, tendo curvas abaixo de todos os demais métodos a partir da consulta 15. Destaca-se principalmente a floresta aleatória pela sua curva relativamente acentuada, com rápido decrescimento, tendo uma curva praticamente constante no menor risco possível após a consulta 15. Já, tomando a amostragem por incerteza, apesar da curva ser menor que as demais até a décima consulta, observa-se um declínio razoavelmente lento após esse número de consultas. De

fato, observa-se sua curva acima da própria amostragem aleatória, convergindo ao mesmo risco da dessa e da redução de viés apenas na última consulta. Assim, vemos um cenário razoavelmente favorável a nosso método em comparação a amostragem por incerteza.

– **RV versus Consulta por comitê**: As curvas de risco dos métodos de consulta por comitê, redução de viés e amostragem aleatória são mostradas na Figura 4.15.

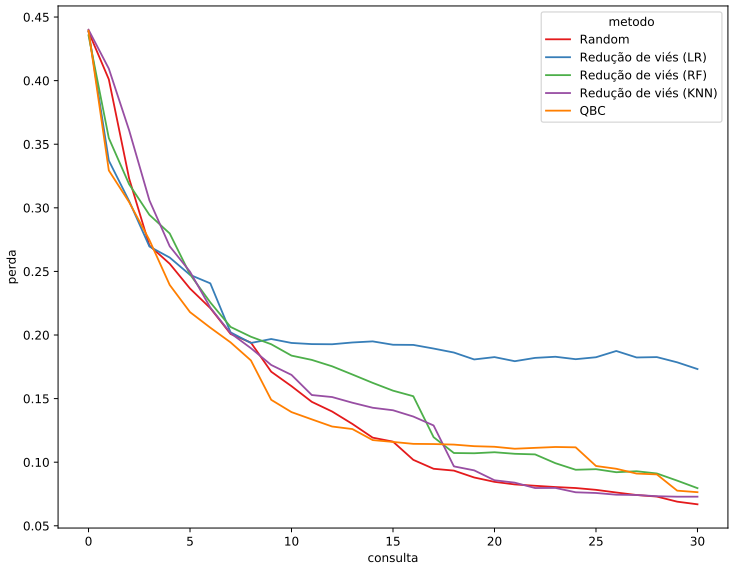


Figura 4.16: Curvas de risco do aprendizado ativo para cada método nos dados simulados não lineares, comparando nosso método de aprendizado ativo à consulta por comitê e amostragem aleatória

Vemos que em geral, até a consulta 15, a consulta por comitê tem uma curva menor que os demais métodos, e após isso, a amostragem aleatória e o knn parecem possui melhores desempenhos, com a amostragem aleatória em geral possuindo a curva mais satisfatória dentre os demais métodos. Apesar disso, vemos que o knn em geral está próximo a amostragem aleatória e passa a ter menor curva que a consulta por comitê a partir de uma consulta entre 15 e 20, com uma grande proximidade da amostragem aleatória. Assim, apesar da amostragem aleatória ser ao final o melhor método nesse exemplo, podemos dizer que nosso método possui um melhor desempenho que a consulta por comitê no modelo de validação do knn.

- **D3**: Considerando os dados não lineares heterocedásticos, obtemos:
  - **RV versus Amostragem por incerteza**: As curvas de risco dos métodos de amostragem por incerteza, redução de viés e amostragem aleatória para os dados heterocedásticos são mostradas na Figura 4.17.

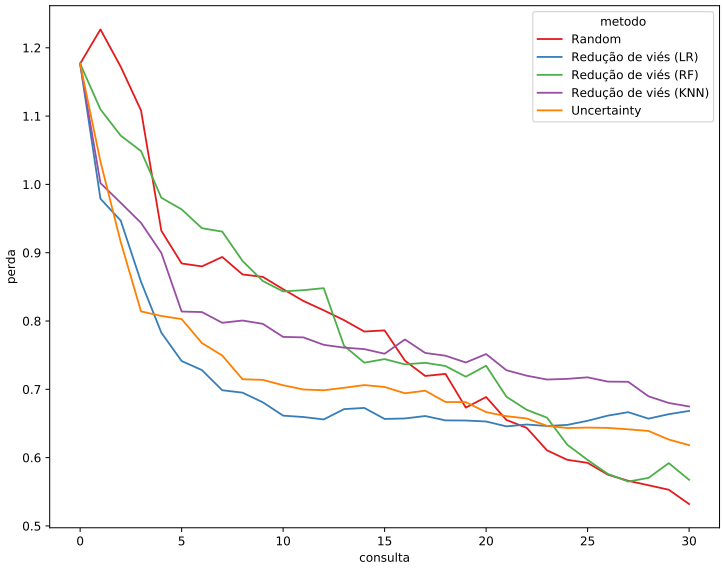


Figura 4.17: Curvas de risco do aprendizado ativo para cada método nos dados simulados não lineares heterocedásticos, comparando nosso método de aprendizado ativo à amostragem por incerteza e amostragem aleatória

Visualiza-se que até a consulta 20, a redução de viés com a validação por regressão linear tem sua curva abaixo das demais, e após isso, a amostragem aleatória e redução de viés com a validação por floresta aleatória possuem menores curvas, tendo a amostragem aleatória com a menor curva que os demais métodos. Ressalta-se também que a amostragem por incerteza tem um desempenho pior que a redução de viés com modelo de validação por florestas aleatórias após a consulta 20, e pior que a regressão linear antes da consulta 20. Assim, apesar de nosso método estar um pouco pior que a amostragem aleatória, ainda tem um desempenho melhor que a amostragem por incerteza.

  - **RV versus Consulta por comitê**: As curvas de risco dos métodos de consulta por comitê, redução de viés e amostragem aleatória para os dados heterocedásticos são mostradas na Figura 4.18.

    Neste caso, a amostragem aleatória esta disparadamente mais abaixo das de-

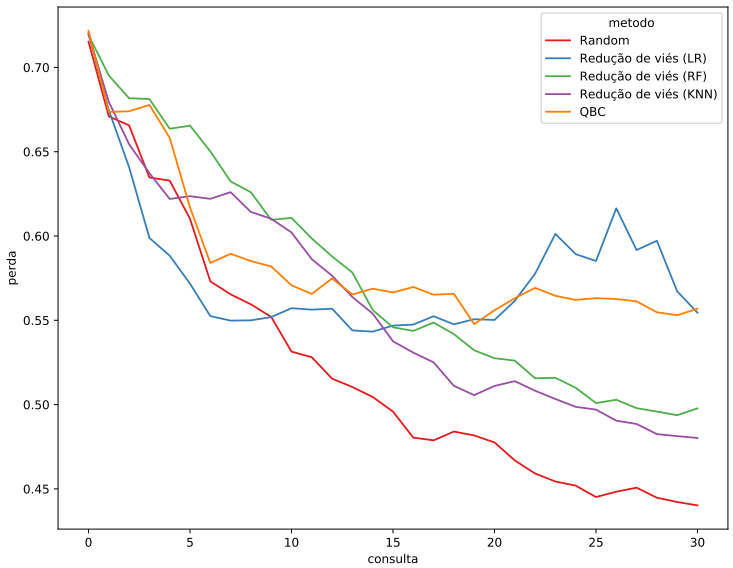


Figura 4.18: Curvas de risco do aprendizado ativo para cada método nos dados simulados não lineares heterocedásticos, comparando nosso método de aprendizado ativo à consulta por comitê e amostragem aleatória

mais curvas, deixando claro que para esses dados, com o ensemble tomado anteriormente, o aprendizado ativo não é muito efetivo, sendo preferível o aprendizado passivo. Apesar disso, vemos que a curva de risco da redução de viés com validação de floresta aleatória e knn são estritamente decrescentes e possuem um melhor comportamento que a consulta por comitê, que tem uma curva relativamente estável na perda 0.55. Destaca-se também que esse tipo de problema pode-se dever ao modelo de *ensemble* escolhido, visto que a arbitrariedade da escolha dos modelos que compõe o comitê tem naturalmente certo impacto sobre os resultados.

Destaca-se em geral que para os dados não lineares heterocedásticos, a amostragem aleatória apresentou melhor comportamento que os demais modelos de regressão ativa, algo que pode se dever tanto ao comportamento dos dados, quanto ao como os modelos de processo gaussiano e o ensemble escolhido lidam com tais dados. Além disso, podemos ainda melhorar nosso método com a estimativa da variância intrinseca, o que da mais crédito para nosso método em sua atual versão.

- **D2-dim**: Para os dados não lineares com ruído obtemos:

– **RV versus Amostragem por incerteza**: As curvas de risco dos métodos de amostragem por incerteza, redução de viés e amostragem aleatória para os dados não lineares com ruídos são mostradas na Figura 4.19.

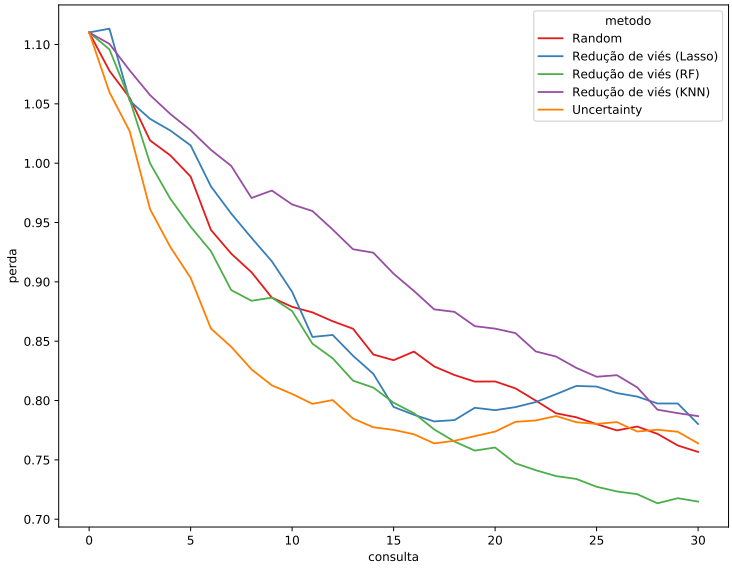


Figura 4.19: Curvas de risco do aprendizado ativo para cada método nos dados simulados não lineares com ruídos, comparando nosso método de aprendizado ativo à amostragem por incerteza e amostragem aleatória

Vemos que em geral, a redução de viés pela floresta aleatória tem uma curva estritamente decrescente e bem abaixo das demais após a consulta 20 apesar da amostragem por incerteza possui menor curva antes desse número de consulta. Assim vemos que mesmo aumentamos os ruídos dos dados, nosso método ainda possui um melhor desempenho que os demais para os dados não lineares.

– **RV versus consulta por comitê**: As curvas de risco dos métodos de consulta por comitê, redução de viés e amostragem aleatória para os dados não lineares homocedásticos são mostradas na Figura 4.18. Ao contrário do cenário observado em **D2**, vemos uma curva menor tanto da validação por floresta aleatória quanto por lasso em todas as consultas em comparação a consulta por comitê e amostragem aleatória. Tal comportamento se observou também anteriormente na Subseção 4.2.1, confirmando a robustez de nosso método frente a dados muito ruidosos, algo que a consulta por comitê em geral tende a apresentar problemas, tendo um risco maior que a própria amostragem aleatória ao final

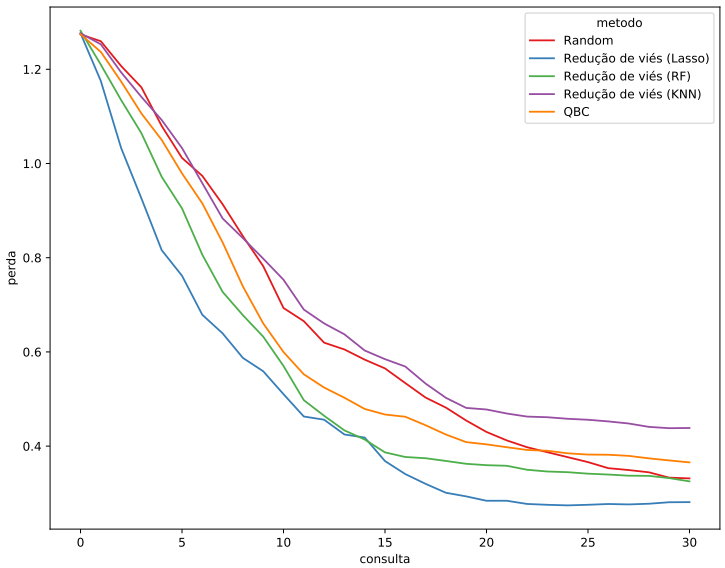


Figura 4.20: Curvas de risco do aprendizado ativo para cada método nos dados simulados não lineares com ruídos, comparando nosso método de aprendizado ativo à consulta por comitê e amostragem aleatória

das 30 consultas.

Assim, vemos que o comportamento observado anteriormente na Subseção 4.2.1 se mantém também para o processo gaussiano e o *ensemble* fixado, mostrando melhor desempenho que os outros dois métodos também investigados.

- **D3-dim** Para os dados não lineares heterocedásticos com ruído obtemos:

  - **RV versus Amostragem por incerteza**: A seguir, a Figura 4.21 mostra as curvas de risco dos métodos de amostragem por incerteza, redução de viés e amostragem aleatória para os dados não lineares heterocedásticos com ruídos. Percebe-se imediatamente um comportamento relativamente errático da amostragem por incerteza, tendo uma curva de risco estacionada em uma perda alta enquanto as demais curvas tem um comportamento decrescente.

    Além disso, notamos que as curvas são relativamente não suaves mesmo tomando-se um número de tentativas razoável como 30, algo que pode se dever por algumas divergências do processo gaussiano durante os ajustes. Quanto as curvas do método de redução de viés com todos os modelos de validação, vemos que praticamente todos estão um pouco abaixo ou empatados com a curva

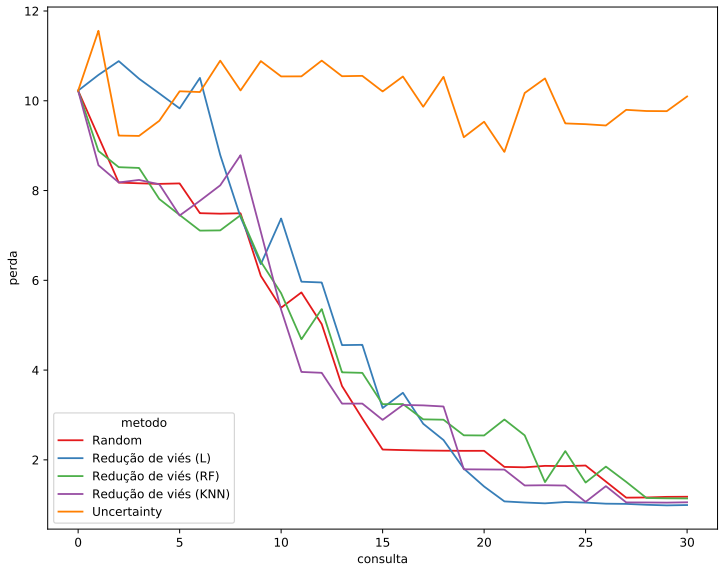


Figura 4.21: Curvas de risco do aprendizado ativo para cada método nos dados simulados não lineares heterocedásticos com ruídos, comparando nosso método de aprendizado ativo à amostragem por incerteza e amostragem aleatória

da amostragem aleatória, tendo ao final, na 30ª consulta, um menor risco que este. Entre os modelos de validação, destaca-se principalmente o lasso, que converge para um risco pequeno já na 20ª consulta.

– **RV versus consulta por comitê**: A Figura 4.22 mostra as curvas de risco dos métodos de consulta por comitê, redução de risco e amostragem aleatória para os dados não lineares heterocedásticos com ruído.

  Nota-se que em geral, as curvas estão relativamente próximas, havendo maior diferenciação entre elas a partir da consulta 10, tendo a redução de viés com validação por florestas aleatória um maior destaque dos demais métodos. De fato, vemos que a curva da floresta aleatória passa a ter a menor perda a partir da consulta 10, sendo menor que a amostragem aleatória e a consulta por comitê. Além disso, destaca-se essa última (consulta por comitê) como a maior curva de risco entre as demais, demonstrando novamente uma performance não muito boa frente ao método de redução de viés

Portanto, novamente, reforça-se a qualidade de nosso método frente a dados ruidosos, notando-se os resultados positivos nos mais diferentes casos simulados e modelos usados, seja usando os modelos desta Subseção ou da Subseção 4.2.1.

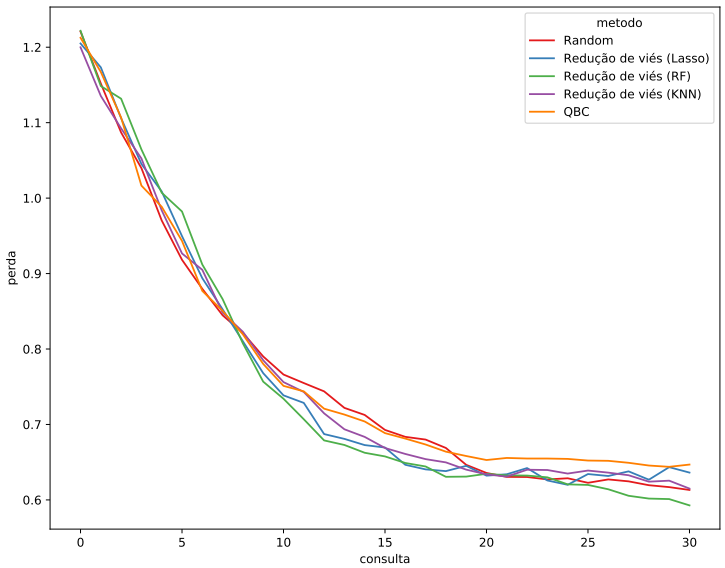


Figura 4.22: Curvas de risco do aprendizado ativo para cada método nos dados simulados não lineares heterocedásticos com ruídos, comparando nosso método de aprendizado ativo à consulta por comitê e amostragem aleatória

Em geral, observamos nessas comparações com outros modelos de regressão ativa, que de fato, nosso método tem certas vantagens em comparação a estes e possui potencial para ser um método útil para regressão ativa, possuindo desempenho parecido ou melhor que os demais métodos nesses casos simulados, com a vantagem de ser flexível para outros diferentes tipos de modelos além de processos gaussianos e modelos de *ensemble*. A seguir, iremos comparar os diferentes métodos em dados reais, usando tanto os modelos de ajuste da Subseção 4.2.1, quanto os modelos da Subseção atual.

4.2.3 Experimentos em dados reais

Na mesma dinâmica da Subseção 4.2.1, compararemos as curvas de risco de cada método de interesse em cada conjunto de dados real.

- **Aerofólio**: Para os dados de aerofólio obtivemos os seguintes resultados:
  - **RV versus Amostragem aleatória**: As curvas de risco de cada modelo e método são dadas na Figura 4.23. Percebemos em geral que a validação pelo knn apresentou melhores resultados dentre os demais métodos de validação considerados, tendo uma curva abaixo da amostragem aleatória para as 30

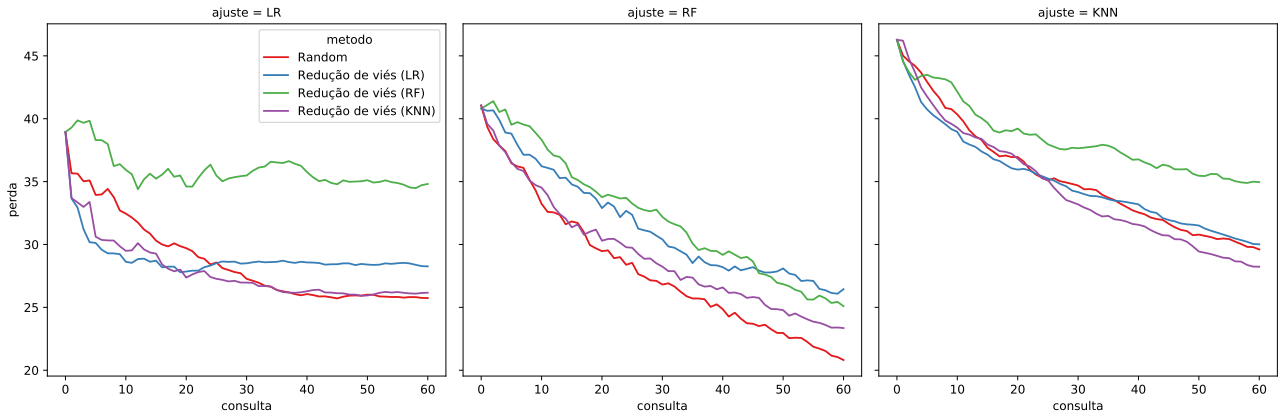


Figura 4.23: Curvas de risco do aprendizado ativo para cada método nos dados de aerofólio, comparando nosso método de aprendizado ativo à amostragem aleatória, utilizando processos gaussianos como modelo de ajuste

primeiras consultas no ajuste por regressão linear, convergindo depois para o mesmo valor de perda, enquanto para o ajuste por knn, apresentou uma curva menor que a amostragem aleatória para praticamente todas as consultas. Já para o ajuste por florestas aleatórias, nota-se claramente uma menor curva para amostragem aleatória, tendo nosso método um pior desempenho para esse modelo que o aprendizado passivo.

– **RV versus Amostragem por incerteza**: Para a comparação das curvas de risco entre a amostragem por incerteza e a redução de viés, obtemos a Figura 4.24. Observa-se nessa que os modelos de validação por floresta aleatória e regressão linear possuem um comportamento muito errático, sendo modelos de validação de fato não adequados, enquanto a validação por knn apresenta um comportamento razoável para o aprendizado ativo. Apesar disso, vemos que a curva da amostragem por incerteza e amostragem aleatória são menores que a curva do knn, tendo a amostragem aleatória a menor curva dentre todas as demais. Observa-se portanto, resultados relativamente desfavoráveis para nosso método na comparação com a amostragem por incerteza, no modelo de processos gaussianos com *kernel* polinomial.

– **RV versus Consulta por comitê**: Para a comparação entre a redução de viés e consulta por comitê, obtemos a Figura 4.25.

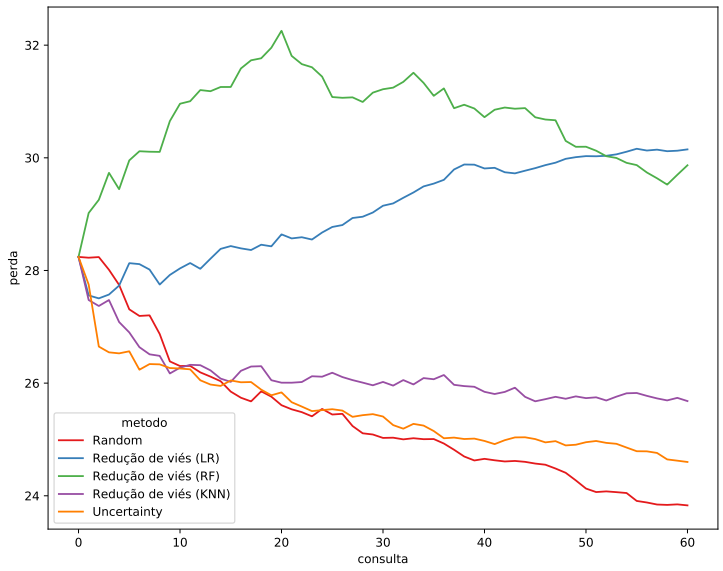


Figura 4.24: Curvas de risco do aprendizado ativo com processos gaussianos para cada método nos dados de aerofólio, comparando nosso método de aprendizado ativo à amostragem por incerteza e amostragem aleatória

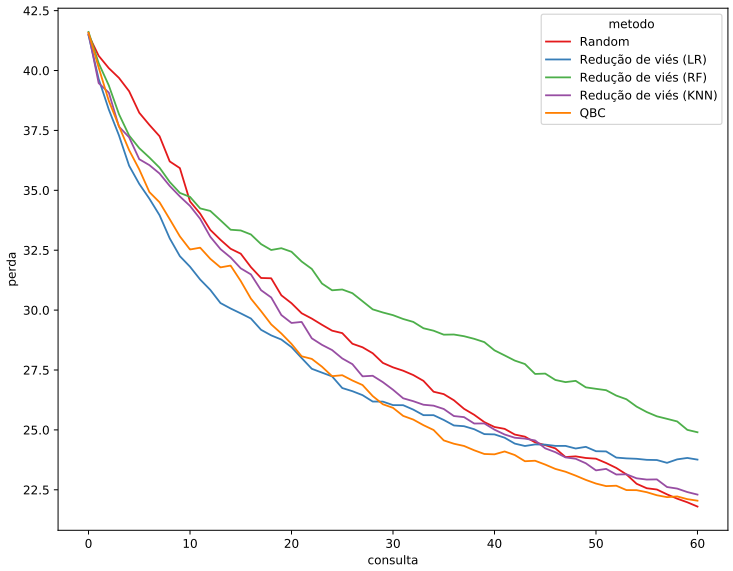


Figura 4.25: Curvas de risco do aprendizado ativo para cada método nos dados de aerofólio, comparando nosso método de aprendizado ativo à consulta por comitê e amostragem aleatória, utilizando um modelo de *ensemble* como modelo de ajuste

Percebe-se por esta que para as 30 primeiras consultas, a validação por regressão linear resulta em menores perdas para nosso método, e após tal número

de consultas, a consulta por comitê passa a ter menores perdas. Ao final das 60 consultas, percebe-se certo empate entre a amostragem aleatória, a consulta por comitê e nosso método com validação por knn, tendo a amostragem aleatória a menor perda entre os três. Assim, temos resultados parcialmente bons para nosso método, visto que o modelo de validação knn acompanha bem a curva da amostragem aleatória e consulta por comitê, enquanto a própria validação por regressão linear possui uma menor curva para as 30 primeiras consultas.

- **Vinho branco**: Para os dados de vinho branco, os seguintes resultados foram obtidos:
  - **RV versus Amostragem aleatória**: As curvas de risco de cada modelo e método são dadas na Figura 4.26. Em geral, nota-se que tanto para o ajuste

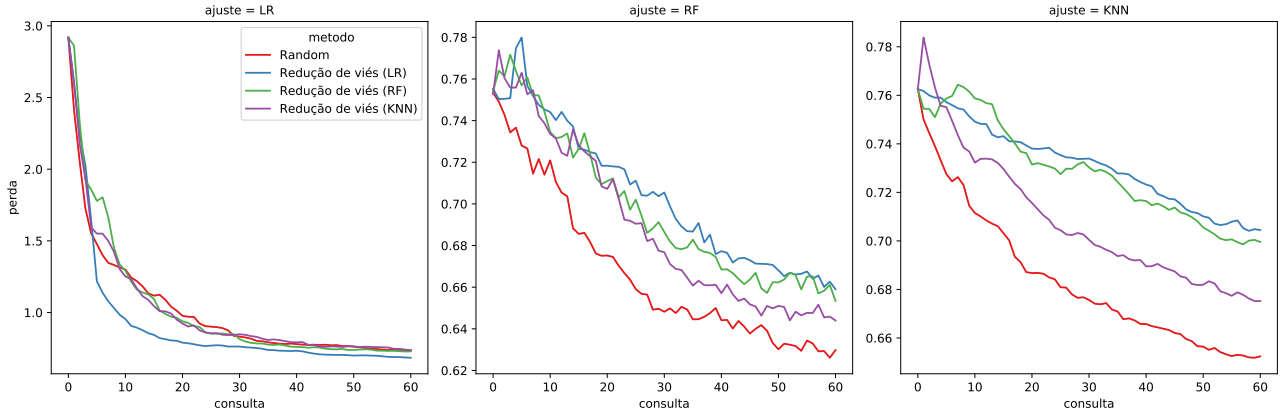


Figura 4.26: Curvas de risco do aprendizado ativo para cada método nos dados de vinho branco, comparando nosso modelo de aprendizado ativo à amostragem aleatória

por florestas aleatórias quanto para o knn a curva da amostragem aleatória é menor que todas as demais, tendo a perda uma escala relativamente pequena. Em contraste, no ajuste de regressão linear, o modelo de validação de regressão linear possui a menor curva dentre todos os demais métodos desde a primeira consulta, com tal ajuste possuindo uma escala relativamente grande. Assim, apesar da amostragem aleatória possuir melhor desempenho nos dois ajustes não paramétricos, temos um grande destaque de nosso método na regressão linear.

  - **RV versus Amostragem por incerteza**: A Figura 4.27 nos da as curvas que comparam nosso método à amostragem por incerteza e amostragem aleatória.

Percebe-se que, após a décima consulta, nosso método com validação por re-

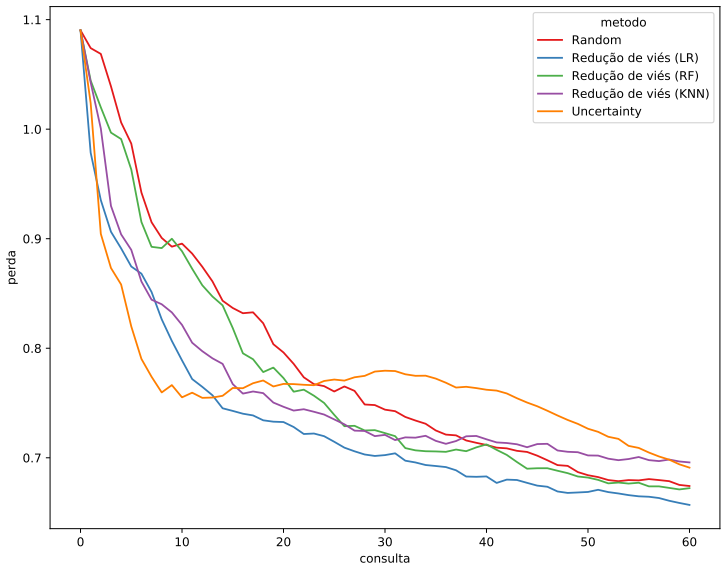


Figura 4.27: Curvas de risco do aprendizado ativo para cada método nos dados de vinho branco, comparando nosso método de aprendizado ativo à amostragem por incerteza e amostragem aleatória, utilizando processos gaussianos como modelo de ajuste

gressão linear possui uma curva menor que todos demais métodos, mostrando um desempenho muito melhor que a amostragem por incerteza. Essa por sua vez, apesar de possuir menor curva até a décima consulta, tem um aumento e decrescimento de perda que configura uma curva maior que quase todos os demais métodos após a décima consulta.

– **RV versus Consulta por comitê**: Por fim, para a comparação entre a redução de viés e consulta por comitê, obtemos a Figura 4.28. Observa-se nesta que em geral, os métodos de redução de viés com modelo de validação knn, consulta por comitê e a amostragem aleatória tem certo empate na maior parte das consultas, com a amostragem aleatória se destacando mais que os outros dois na últimas 10 consultas. Além disso, destaca-se a escala da perda que é consideravelmente baixa, com máximo de 0.68 e mínimo de 0.60. Assim, apesar da amostragem aleatória se destacar nas últimas consultas, consideraremos um empate na performance dos três métodos devido a escala muito baixa e a grande proximidade das curvas para quase todas as consultas.

- **Vinho vermelho**: Para os dados de vinho vermelho, os seguintes resultados foram

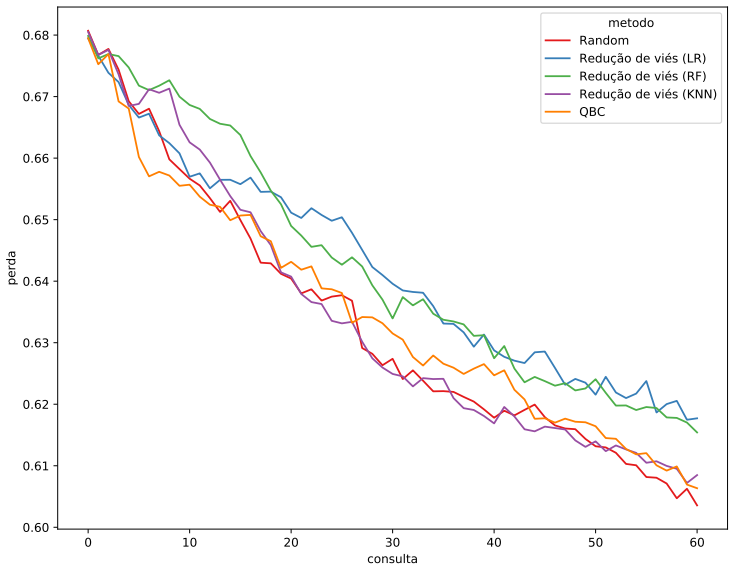


Figura 4.28: Curvas de risco do aprendizado ativo para cada método nos dados de vinho branco, comparando nosso método de aprendizado ativo à consulta por comitê e amostragem aleatória, utilizando um modelo de *ensemble* como modelo de ajuste

obtidos:

– **RV versus Amostragem aleatória**: As curvas de risco de cada modelo e método são dadas na Figura 4.29. . Nesse caso, percebe-se que para os mode-

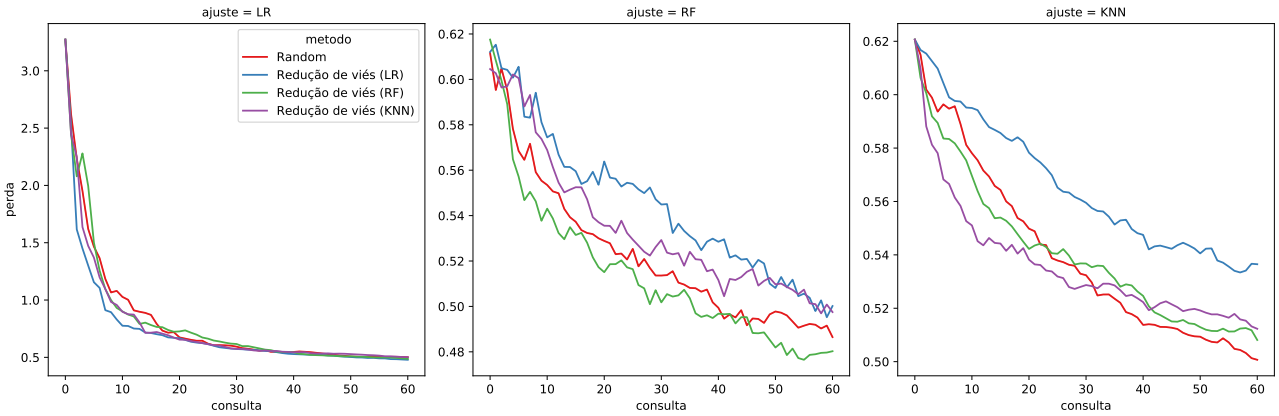


Figura 4.29: Curvas de risco do aprendizado ativo para cada método nos dados de vinho vermelho, comparando nosso modelo de aprendizado ativo à amostragem aleatória

los de ajuste de regressão linear e floresta aleatória, os modelos de validação de regressão linear e floresta aleatória respectivamente possuem melhores desempenhos que os demais métodos. Para o do ajuste e validação por regressão linear, nota-se uma curva abaixo dos demais métodos até a consulta 20, tendo

todos os métodos um desempenho muito parecido, praticamente estacionado em 0.5, após essa consulta.

Para o ajuste por floresta aleatória, percebe-se que a curva da validação por floresta aleatória esta abaixo de todos os demais métodos para todas as consultas, configurando-se assim como melhor método que os demais. Já para o knn, apesar da validação por knn apresentar melhores resultados até a consulta 30 que os demais métodos, a amostragem aleatória passa a possuir um melhor desempenho que os demais métodos a partir da consulta 30. Destaca-se também as escalas dos ajuste por floresta aleatória e por knn que estão relativamente baixas em comparação a escala da regressão linear que vai de 3 a 0.5, indicando que mesmo com um conjunto de treinamento inicial pequeno estes modelos já possuem um desempenho muito bom.

– **RV versus Amostragem por incerteza**: A seguir, comparamos a redução de viés à amostragem aleatória e amostragem por incerteza pela Figura 4.30.

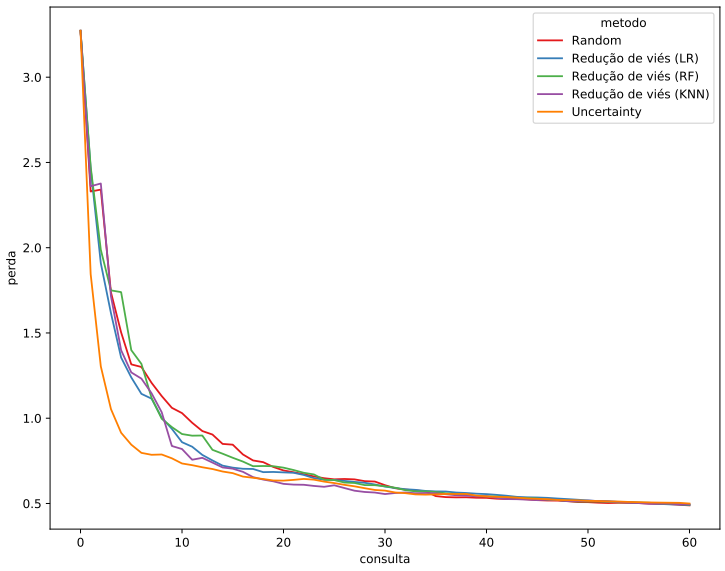


Figura 4.30: Curvas de risco do aprendizado ativo para cada método nos dados de vinho vermelho, comparando nosso método de aprendizado ativo à amostragem por incerteza e amostragem aleatória, utilizando processos gaussianos como modelo de ajuste

Nota-se neste caso que a amostragem por incerteza tem um melhor desempenho que os demais métodos até a consulta 30, tendo uma curva menor que todas as demais. Após esse número de consultas, todos os métodos se juntam em uma

perda próxima de 0.5, com uma curva já estacionada. Um comportamento semelhante é observado na Figura 4.29 no painel do ajuste por regressão linear. Salienta-se também que, até a consulta 30, o método de redução de viés com os mais variados modelos de validação tem uma curva abaixo da amostragem aleatória, mostrando um desempenho razoável, apesar de pior que a amostragem por incerteza. Assim, neste caso, a amostragem por incerteza tem maior destaque, apesar de mais nas consultas finais, todos os métodos convergirem para a mesma perda.

– **RV versus Consulta por comitê**: Por fim, comparando-se os métodos de redução de viés e consulta por comitê, obtemos a Figura 4.29 com as curvas de cada método. Nota-se novamente certo empate entre a consulta por comitê,

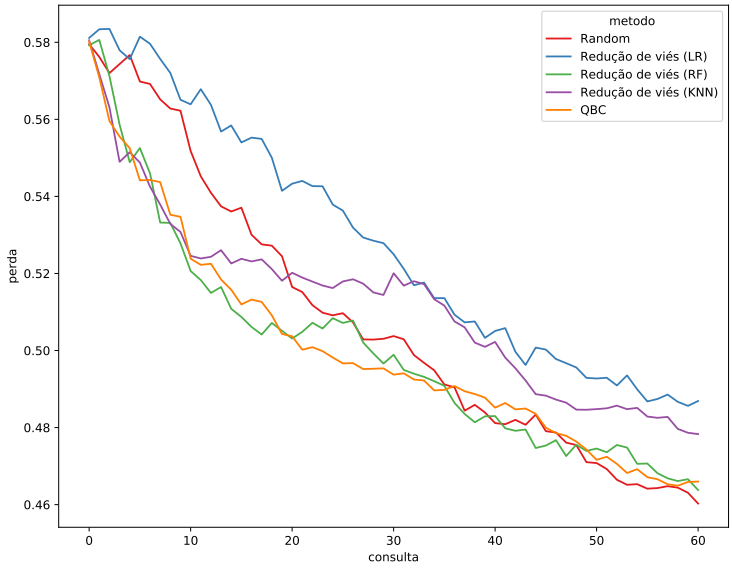


Figura 4.31: Curvas de risco do aprendizado ativo para cada método nos dados de vinho vermelho, comparando nosso método de aprendizado ativo à consulta por comitê e amostragem aleatória, utilizando um modelo de *ensemble* como modelo de ajuste

redução de viés e amostragem aleatória, tendo até a consulta 30, uma disputa apenas entre consulta por comitê e redução de viés com validação por floresta aleatória, e após essa, um empate entre os três métodos. Nas consultas finais, a amostragem aleatória passa a ter uma menor perda que os outros dois métodos, mas ainda assim, vemos uma grande proximidade entre os três quanto a perda, visto também que a escala da perda é relativamente baixa.

- **Concreto** Por fim, para os dados de concreto obtemos os seguintes resultados:

  – **RV versus Amostragem aleatória**: As curvas de risco de cada modelo e método são dadas na Figura 4.32. Para esse caso, percebe-se que no ajuste de

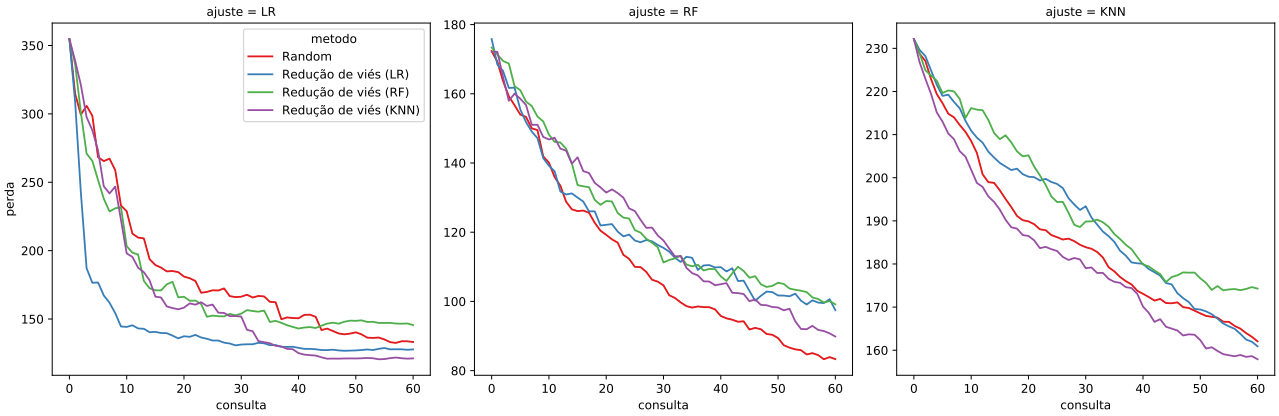


Figura 4.32: Curvas de risco do aprendizado ativo para cada método nos dados de concreto, comparando nosso modelo de aprendizado ativo à amostragem aleatória

regressão linear, o método de redução de viés com validação por regressão linear e knn possuem de fato os melhores desempenhos dentre os demais métodos, tendo curvas abaixo das demais. Também se observa um melhor desempenho da redução de viés para o ajuste por knn, mais especificamente para a validação por knn, tendo uma curva abaixo das demais. Já para a floresta aleatória, vemos que praticamente para todas as consultas, a amostragem aleatória possui menor curva que a redução de viés com todos os possíveis modelos de validação possíveis, tendo assim o melhor desempenho para esse modelo de ajuste.

  – **RV versus Amostragem por incerteza**: Comparamos agora a redução de viés à amostragem por incerteza e amostragem aleatória pela Figura 4.33. Em geral, nota-se que para as 30 primeiras consultas, a amostragem por incerteza tem uma curva menor que todas as demais, chegando a alcançar a perda 120. Porém, após a consulta 30, o método diverge com uma curva crescente, tendo, nesse trecho final, um desempenho pior que a redução de viés com validação por regressão linear e knn e a amostragem aleatória.

  Já a redução de viés mantém um comportamento decrescente para todas as consultas, possuindo o melhor desempenho geral ao tomar como validação o modelo de regressão linear, dados que esse tem uma curva muito afastada da amostragem aleatória e decresce para a perda 120 muito rapidamente, estabi-

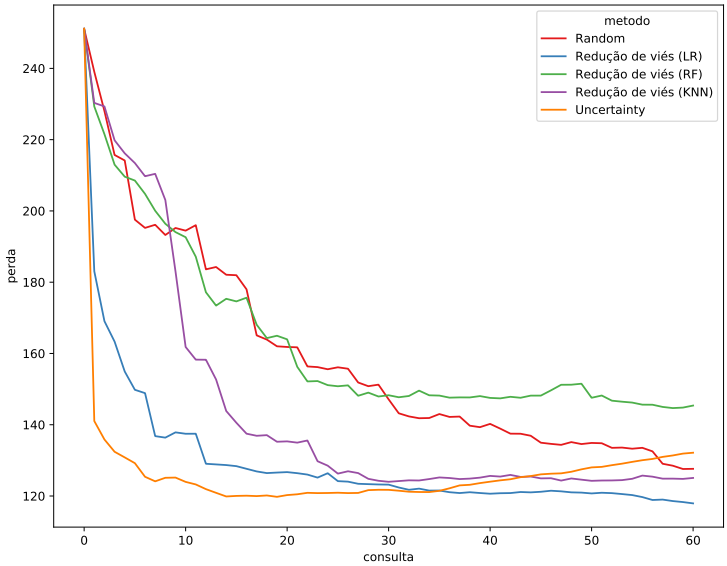


Figura 4.33: Curvas de risco do aprendizado ativo para cada método nos dados de concreto, comparando nosso método de aprendizado ativo à amostragem por incerteza e amostragem aleatória, utilizando processos gaussianos como modelo de ajuste

lizando em tal perda. Destaca-se também a validação por knn, que também possui uma curva satisfatória, menor que a amostragem aleatória e com rápido decrescimento.

– **RV versus Consulta por comitê**: Por fim, compara-se os métodos de redução de viés, consulta por comitê e amostragem aleatória através da Figura 4.34.

  Em geral, para todas as consultas, vemos que a amostragem aleatória possui uma menor curva que os demais modelos, tendo de fato, um melhor desempenho. Já, a consulta por comitê, possui pior desempenho que praticamente todas as validações da redução de viés, observando-se que nas consultas finais, as curvas da redução de viés estão consideravelmente abaixo da consulta por comitê, principalmente a validação por regressão linear.

  Assim, observamos dessa vez um cenário que a amostragem é melhor que os demais métodos, porém a consulta por comitê não é melhor que a redução de viés.

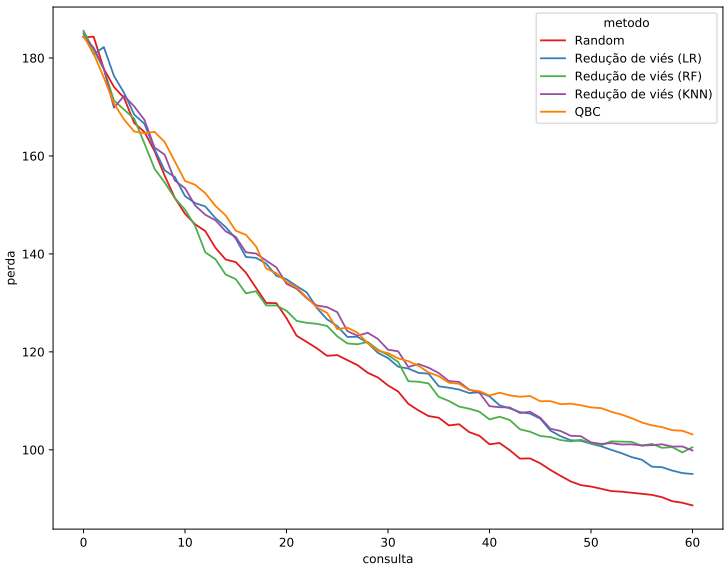


Figura 4.34: Curvas de risco do aprendizado ativo para cada método nos dados de concreto, comparando nosso método de aprendizado ativo à consulta por comitê e amostragem aleatória

Capítulo 5

Conclusão dos experimentos e considerações finais

A área de aprendizado ativo é caracterizada pela consulta e rotulação de observações mais informativas ao conjunto de treinamento associado a um modelo de predição em um contexto de dados semi-rotulados. Existem muitos métodos que podem ser utilizados nos mais diferentes cenários, cada um com sua vantagem e desvantagem. Nesse aspecto, existem cenários que a amostragem aleatória pode ser melhor que o aprendizado ativo. Observamos um exemplo desse tipo de situação na Subseção 4.1.1.

Destaca-se também nesse contexto o fato de não existir na literatura de aprendizado ativo muitos métodos para modelos de regressão ativa, tendo um maior foco para problemas de classificação. Dessa maneira, propomos uma nova metodologia para regressão ativa, ainda em uma versão mais simplificada, e comparamos esta à outras metodologias existentes para regressão ativa.

Apesar de nosso método consistir em uma versão inicial, observamos bons resultados para esse em comparação aos métodos já existentes e à amostragem aleatória tanto para os dados simulados quanto os dados reais. Para os dados simulados, os resultados são expressivos, com nosso método tendo um melhor desempenho em comparação aos métodos de amostragem aleatória, amostragem por incerteza e consulta por comitê.

Já para os dados reais, os resultados foram mais discretos, com muito espaço para melhoramentos. Como visto pela Tabela 5.1, nas comparações com a consulta por comitê enxergou-se três empates e um pior desempenho, enquanto nas comparações com a amostragem por incerteza, observam-se duas comparações com pior desempenho de nosso método e duas com melhor desempenho. No caso da amostragem aleatória, nota-se que

em geral para cada banco de dados, nosso modelo possui melhor desempenho que a amostragem aleatória para 2 ajustes distintos. Em particular, o ajuste por floresta aleatória possui os piores desempenhos, com 3 desempenhos piores de nosso método em comparação a amostragem aleatória, enquanto para a regressão linear observa-se um cenário oposto, tendo melhores desempenhos para todos os bancos de dados. Assim, apesar de resultados um pouco piores que os observados dos dados simulados, observa-se em geral, bons resultados de nosso método.

Tabela 5.1: Tabela comparando as metodologias de aprendizado ativo existentes com a metodologia de redução de viés. ✓ indica a redução de viés teve resultados melhores que os demais métodos, ✗ indica um pior desempenho da redução de viés, e ∼ indica empate entre nosso método e o método concorrente

Banco de dados	**Métodos de comparação**				
	Amostragem aleatória			**Amostragem por incerteza**	**Consulta por comitê**
	R.L.	F.A	KNN	P.G	Ensemble
Aerofólio	✓	✗	✓	✗	∼
Vinho branco	✓	✗	✗	✓	∼
Vinho vermelho	✓	✓	✗	✗	∼
Concreto	✓	✗	✓	✓	✗

Por fim, como o método de redução de viés ainda pode ser muito melhorado ao incorporar a estimativa da variância $\mathbb{V}[Y|\boldsymbol{X}]$ ao nosso método, além de outras possíveis modificações para o modelo de validação, pode-se dizer que os resultados observados são positivos, tendo assim o nosso método um grande potencial de se tornar um novo método útil para regressão ativa. Em trabalhos futuros, exploraremos as possíveis melhorias ressaltadas, para tornar nosso método ainda mais competitivo.

Referências Bibliográficas

Abe, N. (1998). Query learning strategies using boosting and bagging. *Proc. of 15ˆ¡ th¿ Int. Cmf. on Machine Learning (ICML98)*, páginas 1–9.

Balcan, M.-F., Broder, A. e Zhang, T. (2007). Margin based active learning. Em *International Conference on Computational Learning Theory*, páginas 35–50. Springer.

Berger, A., Della Pietra, S. A. e Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, **22**(1), 39–71.

Beygelzimer, A., Dasgupta, S. e Langford, J. (2009). Importance weighted active learning. Em *Proceedings of the 26th annual international conference on machine learning*, páginas 49–56.

Blum, A. e Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. Em *Proceedings of the eleventh annual conference on Computational learning theory*, páginas 92–100.

Breiman, L. (1996). Bagging predictors. *Machine learning*, **24**(2), 123–140.

Burbidge, R., Rowland, J. J. e King, R. D. (2007). Active learning for regression based on query by committee. Em *International conference on intelligent data engineering and automated learning*, páginas 209–218. Springer.

Cardoso, T. N., Silva, R. M., Canuto, S., Moro, M. M. e Gonçalves, M. A. (2017). Ranked batch-mode active learning. *Information Sciences*, **379**, 313–337.

Chaloner, K. e Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, páginas 273–304.

Cohn, D., Atlas, L. e Ladner, R. (1994). Improving generalization with active learning. *Machine learning*, **15**(2), 201–221.

Cohn, D. A., Ghahramani, Z. e Jordan, M. I. (1996). Active learning with statistical models. *Journal of artificial intelligence research*, **4**, 129–145.

Cover, T. M. e Thomas, J. A. (2006). Elements of information theory second edition solutions to problems. *Internet Access*, páginas 19–20.

Culotta, A. e McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. Em $AAAI$, volume 5, páginas 746–751.

Dagan, I. e Engelson, S. P. (1995). Committee-based sampling for training probabilistic classifiers. Em *Machine Learning Proceedings 1995*, páginas 150–157. Elsevier.

Danka, T. e Horvath, P. (2018). modal: A modular active learning framework for python. *arXiv preprint arXiv:1805.00979*.

Dasgupta, S. e Hsu, D. (2008). Hierarchical sampling for active learning. Em *Proceedings of the 25th international conference on Machine learning*, páginas 208–215.

Fedorov, V. V. (2013). *Theory of optimal experiments*. Elsevier.

Geman, S., Bienenstock, E. e Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, **4**(1), 1–58.

Grandvalet, Y., Bengio, Y. *et al.* (2005). Semi-supervised learning by entropy minimization. *CAP*, **367**, 281–296.

Hanneke, S. (2007). A bound on the label complexity of agnostic active learning. Em *Proceedings of the 24th international conference on Machine learning*, páginas 353–360.

Hastie, T., Tibshirani, R. e Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York. ISBN 9780387848587.

Hwa, R. (2001). *Learning probabilistic lexicalized grammars for natural language processing*. Harvard University.

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki.

James, G., Witten, D., Hastie, T. e Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kullback, S. e Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, **22**(1), 79–86.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise.

Lewis, D. D. e Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. Em *Machine learning proceedings 1994*, páginas 148–156. Elsevier.

Lewis, D. D. e Gale, W. A. (1994). A sequential algorithm for training text classifiers. Em *SIGIR'94*, páginas 3–12. Springer.

McCallum, A., Nigam, K. *et al.* (1998). A comparison of event models for naive bayes text classification. Em *AAAI-98 workshop on learning for text categorization*, volume 752, páginas 41–48. Citeseer.

Mckay, D. (1992). Information-based objective functions for active data selections. *Neural Computation*, **4**(4), 590–604.

Melville, P., Yang, S. M., Saar-Tsechansky, M. e Mooney, R. (2005). Active learning for probability estimation using jensen-shannon divergence. Em *European conference on machine learning*, páginas 268–279. Springer.

Nemhauser, G. L., Wolsey, L. A. e Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, **14**(1), 265–294.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. *et al.* (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, **12**, 2825–2830.

Roy, N. e McCallum, A. (2001). Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, **2**, 441–448.

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media.

Schütze, H., Velipasaoglu, E. e Pedersen, J. O. (2006). Performance thresholding in practical text classification. Em *Proceedings of the 15th ACM international conference on Information and knowledge management*, páginas 662–671.

Settles, B. (2009). Active learning literature survey.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **6**(1), 1–114.

Settles, B. e Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. Em *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, páginas 1070–1079.

Settles, B., Craven, M. e Friedland, L. (2008). Active learning with real annotation costs. Em *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:.

Seung, H. S., Opper, M. e Sompolinsky, H. (1992). Query by committee. Em *Proceedings of the fifth annual workshop on Computational learning theory*, páginas 287–294.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, **27**(3), 379–423.

Snoek, J., Larochelle, H. e Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, **25**.

Srinivas, N., Krause, A., Kakade, S. M. e Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Tomanek, K. e Morik, K. (2011). Inspecting sample reusability for active learning. Em *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, páginas 169–181. JMLR Workshop and Conference Proceedings.

Tomanek, K., Wermter, J. e Hahn, U. (2007). An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. Em *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, páginas 486–495.

Tomanek, K., Laws, F., Hahn, U. e Schütze, H. (2009). On proper unit selection in active learning: co-selection effects for named entity recognition. Em *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, páginas 9–17.

Wallace, B. C., Small, K., Brodley, C. E. e Trikalinos, T. A. (2010). Active learning for biomedical citation screening. Em *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 173–182.

Wu, D. (2018). Pool-based sequential active learning for regression. *IEEE transactions on neural networks and learning systems*, **30**(5), 1348–1359.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. Em *33rd annual meeting of the association for computational linguistics*, páginas 189–196.

Zhu, X. e Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, **3**(1), 1–130.

Zhu, X. J. (2005). Semi-supervised learning literature survey.

Apêndice A

Um Problema relacionado

Até o momento, foram apresentados métodos que atacam especificamente o problema ligado a bancos de dados semi-rotulados, cuja ausência de rótulos se deve pelo custo de obtenção destes. As abordagens até agora tratadas têm como principal objetivo consultar e adicionar observações para o conjunto de treinamento de dado modelo de acordo com diferentes medidas de informatividade. Um método que tem um espírito semelhante mas um contexto diferente é o método de **otimização bayesiana**.

A **otimização bayesiana** tem como principal objetivo encontrar a observação $\boldsymbol{x}$ que minimize ou maximize funções caras de se avaliar ou caixas pretas. Esse tipo de otimização é utilizado em casos de *tunagem* de hiper-parâmetros de modelos complexos, geralmente redes neurais, cujo risco estimado para cada variação de hiper-parâmetro é de dificil obtenção pelo custo computacional de tais modelos. A ideia básica de otimização Bayesiana é de se avaliar os valores da função de caixa preta $f(\cdot)$ em pontos (valores de hiper parâmetros) $\boldsymbol{x}$ onde a promessa de se achar melhores valores (menores para caso de minimização e maiores para maximização) é maior, tentando-se dessa forma encontrar pontos de máximo ou mínimo global (Snoek *et al.*, 2012).

Comumente, assimila-se um Processo Gaussiano como uma priori de tal tipo de função $f(\boldsymbol{x})$, de forma que para nossas observações de treinamento (hiper-parâmetros fixos) $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ ($\boldsymbol{x}_i$ sendo os hiper-parâmetros do modelo e y_i o risco estimado para esse i-ésima coordenada) temos $y_i \sim \text{Normal}(f(\boldsymbol{x}_i), \nu)$, com ν uma dada variância proveniente de um ruído ε (Snoek *et al.*, 2012). Fixada tal priori e utilizando-se os dados de treinamento, induze-se uma posteriori sobre tal função a partir da regressão por processos gaussianos, na mesma família (utilizando-se da conjugação normal) de distribuições que a priori, com atualizações na estimativa da média e variância associada a cada ponto de teste no espaço

de hiper parâmetros. Obtendo-se uma distribuição a posteriori de tais dados, tem-se interesse em consultar uma nova observação que maximize ou minimize mais a função $f(\cdot)$. Para fazer isso, utiliza-se de **funções de aquisições** $a : \mathcal{X} \to \mathbb{R}^+$ que determinam qual o próximo ponto pertencente ao espaço de hiper parâmetros $\mathcal{X}$ que deve ser avaliado de forma que $\boldsymbol{x}_A^* = \arg\max_{\boldsymbol{x}} a(\boldsymbol{x})$ tal que $a(\cdot)$ apenas dependa das observações passadas e dos hiper parâmetros do processo gaussiano. Snoek *et al.* (2012) denotou tais dependências como $a(\boldsymbol{x}, \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta)$. Sob a definição de uma priori por Processos Gaussianos, a classe de funções $a(\cdot)$ apenas dependerá da predição da média $\mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta)$ e da variância $\sigma^2(\boldsymbol{x}, \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta)$ para cada instância $\boldsymbol{x}$. Denota-se para os métodos a seguir o melhor valor $\boldsymbol{x}_M$ anterior a cada iteração da otimização bayesiana como $\boldsymbol{x}_M = \arg\min_{\boldsymbol{x}_i\ i\in\mathcal{L}} f(\boldsymbol{x}_i)$ ou $\boldsymbol{x}_M = \arg\max_{\boldsymbol{x}_i\ i\in\mathcal{L}} f(\boldsymbol{x}_i)$ dependendo do objetivo do problema (maximizar ou minimizar a função) e as funções de distribuição acumulada e densidade probabilidade de uma normal padrão como $\Phi(\cdot)$ e $\phi(\cdot)$ respectivamente. Assim, define-se três principais funções de aquisição muito utilizadas na área de otimização bayesiana (Snoek *et al.*, 2012):

- **Probabilidade de Melhora**: Criada por Kushner (1964), segue a intuição de consultar uma nova observação que maximize a probabilidade de melhora sobre o atual melhor valor $\boldsymbol{x}_M^*$. Isso é computado de forma analítica utilizando-se as propriedades de um PG. Para maximização, tal função de aquisição é definida como:

$$a_{PM}\left(\boldsymbol{x}; \{x_i, y_i\}_{i=1}^n, \theta\right) = \Phi\left(\frac{\mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}, \theta) - f(\boldsymbol{x}_M^*)}{\sigma(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}, \theta)}\right) \tag{A.1}$$

  No caso de minimização basta inverter a subtração, com o denominador dado por $f(x_M^*) - \mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}, \theta)$. Para simplificação, tomemos também $\gamma(\boldsymbol{x}) = \frac{\mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}, \theta) - f(\boldsymbol{x}_M^*)}{\sigma(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}, \theta)}$

- **Melhora Esperada**: Alternativamente à probabilidade de melhora, podemos maximizar a melhora esperada (ME) sobre o atual melhor valor $\boldsymbol{x}_M^*$ (Snoek *et al.*, 2012) de forma também analítica como dado abaixo para o caso de maximização:

$$\begin{aligned} a_{ME}\left(\boldsymbol{x}; \{x_i, y_i\}_{i=1}^n, \theta\right) = \left(\mu\left(\boldsymbol{x}; \{x_i, y_i\}_{i=1}^n, \theta\right) - f(\boldsymbol{x}_M^*)\right)\Phi(\gamma(\boldsymbol{x})) \\ + \sigma(\boldsymbol{x}; \{x_i, y_i\}_{i=1}^n, \theta)\phi(\gamma(\boldsymbol{x})) \end{aligned} \tag{A.2}$$

  Para minimização, basta novamente tomar $f(x_M^*) - \mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}, \theta)$, redefinindo $\gamma(\boldsymbol{x})$ com essa nova subtração.

- **Limite Superior de Confiança do PG**: Uma ideia mais recente para funções de aquisições é a de se explorar os limites de intervalos de confiança (superior para maximização e inferior para minimização) para construir funções de aquisição que minimizem ainda mais possíveis consultas indesejáveis durante a otimização (Srinivas *et al.*, 2009; Snoek *et al.*, 2012). Tais funções de aquisição possuem a fórmula:

$$a_{LSC}\left(\boldsymbol{x}; \{x_i, y_i\}_{i=1}^{n}, \theta\right) = \mu\left(\boldsymbol{x}; \{x_i, y_i\}_{i=1}^{n}, \theta\right) + \kappa\sigma\left(\boldsymbol{x}; \{x_i, y_i\}_{i=1}^{n}, \theta\right) \tag{A.3}$$

  Substituindo-se mais por menos caso queiramos minimizar $f(\cdot)$ (ou seja, caso estejamos computando $a_{LIC}\left(\boldsymbol{x}; \{x_i, y_i\}_{i=1}^{n}, \theta\right)$). O hiper-parâmetro κ também é *tunável*, balanceando o grau de exploração sobre limites superiores ou inferiores do PG.

Um pequeno e breve exemplo *toy* que mostra brevemente o uso de otimização bayesiana é o de se otimizar a função apresentada pela Figura A.1. Note que essa função é não convexa, existindo diversos máximos e mínimos locais. Com apenas um exemplo

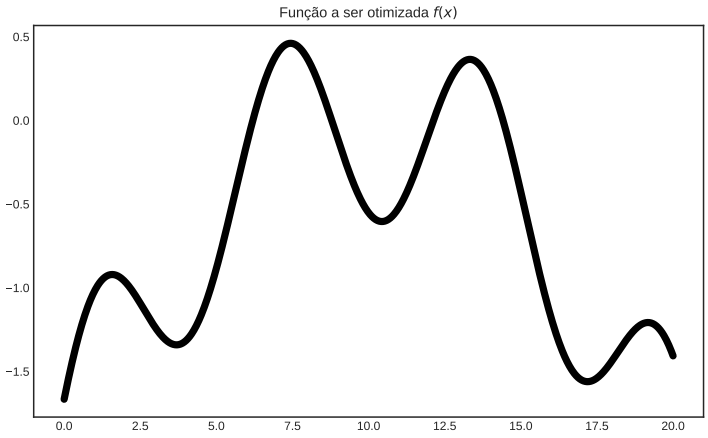


Figura A.1: Função a ser otimizada dada por $f(x) = \frac{sen(x)}{2} - \frac{(10-x)^2}{50} + \frac{cos(x)}{3}$ com $x \in (0, 20)$

de treinamento $(3, f(3) \approx -1.24)$, visualiza-se pela Figura A.2 as etapas da otimização bayesiana para $f(x)$, fixando como função de aquisição a **melhora esperada**. Percebe-se também pela Figura A.2 que o máximo global da função gerada foi alcançado já na 3ª consulta ao espaço de hiper parâmetros $\mathcal{X} = (0, 20)$. Assim, exemplifica-se por tais gráficos a flexibilidade da otimização bayesiana com priori dada por PG's, sendo um procedimento que consegue achar mínimos e máximos globais em funções complicadas não convexas em troca de um custo computacional de consultar e testar diferentes pontos $\boldsymbol{x} \in \mathcal{X}$ (Snoek *et al.*, 2012).

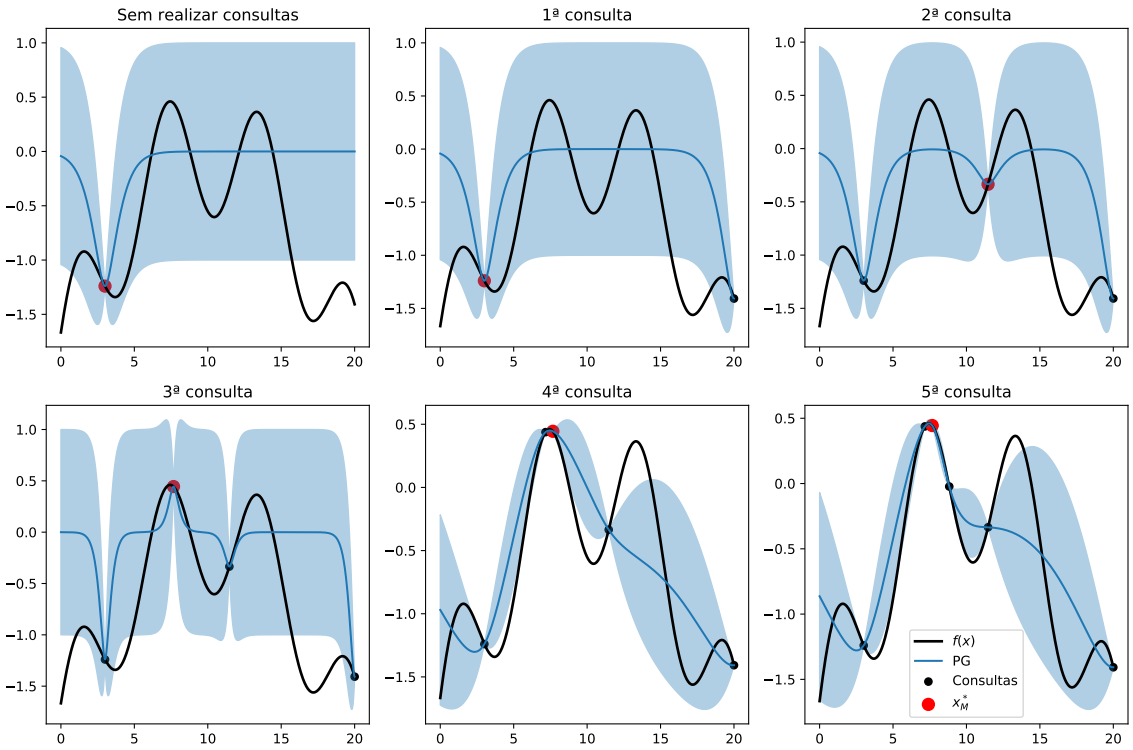


Figura A.2: Cada etapa de uma otimização bayesiana usando a função de aquisição de **melhora esperada**. Os pontos em vermelhos são os melhores valores de cada iteração x_M^* e área sombreada de azul nos da o desvio padrão associado a cada ponto x.

Apêndice B

Códigos utilizados

Todos os códigos utilizados para gerar figuras e os resultados de aprendizado ativo podem ser encontrados no github pelo link https://github.com/Monoxido45/Active_learning_tests