UNIVERSIDADE FEDERAL DE MINAS GERAIS
DEPARTAMENTO DE ESTATÍSTICA

Avaliação de Risco de Crédito utilizando grupo de classificadores

Renato Hudson de Andrade

Belo Horizonte
Out/2008

Renato Hudson de Andrade

Avaliação de Risco de Crédito utilizando grupo de classificadores

Dissertação apresentada ao Curso de Mestrado do Departamento de Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Luis Henrique Duczmal

Belo Horizonte

Out/2008

Agradecimentos

Ao Professor Doutor Luis Henrique Duczmal pela ajuda, motivação, incentivo, paciência e sabedoria que foi capaz de transmitir durante todos os momentos em que este trabalho foi feito.

Aos Professores Enrico Colossimo, Frederico Cruz, Roberto Quinino e Suely Aparecida pelas contribuições durante todo o curso.

À todos os professores e fucionários do Departamento de Estatística que de algum modo colaboraram com meus estudos realizados neste instituto.

À minha família pelo apoio incondicional, e especialmente meus pais pela dedicação e incentivo desde o início.

À minha esposa Clarice, pelas constantes e ternas palavras de incentivo e apoio, mesmo diante dos obstáculos aparentemente intransponíveis.

Obrigado a todos. Esta é certamente uma conquista NOSSA!

Resumo

Uma das maiores aplicações da Estatística no ambiente financeiro são os modelos de análise de risco de crédito. Esta dissertação avalia o desempenho dos seguinte modelos de classificação no contexto de risco de crédito: regressão logística, rede bayesiana ingênua, máquina de vetor suporte, aprendizagem baseada em racioncínio e árvore de classificação. Ao final caracterizou-se as principais vantagens e desvantagens de cada uma das abordagens.

A seguir, avaliou-se as alternativas existentes para a combinação dos classificadores. O grau de acerto das predições quando o grupo de classificadores é utilizado no lugar de um único modelo de classificação é superior ao melhor classificador individual.

Abstract

Models of analysis of credit risk are among the major applications of Statistics in the financial environment. This dissertation assesses the performance of the following models ofclassification in the context of credit risk:logistic regression, naive Bayesian networks, support vector machines, k next-neighbourhoods and classification trees. The main advantages and disadvantages of each approach are discussed.

Their combinations are evaluated as possible alternatives to the use of a single model of classification. The degree of success of predictions using the group of binders is verified through numerical simulations to be higher than the best individual classifier.

Sumário

1 Introdução **1**

2 Classificação **3**

2.1 Teoria Estatística de Decisão 4

2.2 Avaliação do Desempenho 7

2.3 Análise Discriminante 13

2.4 Árvores de Classificação 15

2.4.1 Processo de Particionamento 17

2.4.2 Definição do rótulo da classe 20

2.4.3 Critério de Parada ou poda de uma árvore 21

2.4.4 Boosting 22

2.5 Regressão Logística 25

2.5.1 Estimativa do Modelo 26

2.5.2 Seleção de Modelos 26

2.6 Máquina de Vetor Suporte 28

2.7 Aprendizagem baseada em instâncias 35

2.8 Aprendizagem Bayesiana 37

3 Descrição da Base de dados **40**

3.1 Análise descritiva univariada 41

3.2 Análise descritiva bivariada e multivariada 46

3.3 Dados Omissos 49

4 Resultados dos Classificadores **52**

4.1 Árvore de classificação e Boosting 52

4.2 Regressão 53

4.3 Máquina de vetor suporte 54

4.4 Aprendizagem Baseada em Instâncias 55

4.5 Aprendizagem Bayesiana 56

5 Agregação dos Classificadores **57**

6 Conclusão **62**

Referências Bibliográficas **62**

Capítulo 1

Introdução

Nos últimos anos, esforços contínuos têm sido feitos na busca de eficiência na gestão de crédito [3]. A questão central é o controle de risco implementado no momento da concessão do crédito. O agente financeiro define a taxa de juros do empréstimo levando-se em consideração a possibilidade do cliente não honrar com as suas obrigações. Se a inadimplência de um credor for maior que o valor previsto, a rentabilidade do negócio provavelmente será comprometida.

Os estudos neste campo de aplicação desenvolveram-se tanto, que receberam um nome específico: *credit score*; sendo mais conhecido no Brasil como análise de crédito. Esta é uma área de intenso uso dos métodos estatísticos multivariados. Dentre as aplicações da estatística, em destaque estão os métodos de classificação e previsão.

Como o mercado de crédito tem assumido proporções cada vez maiores na economia, pequenas melhorias na gestão de crédito resultam em ganhos expressivos. É importante destacar o papel econômico do aumento de crédito como fator de crescimento econômico de um país.

O desafio imposto pela gestão de crédito é a construção de modelos de previsão de inadimplência que sejam capazes de trabalhar em um ambiente com variáveis qualitativas e quantitativas resultando em um modelo com alto poder de discriminação entre os clientes inadimplentes e adimplentes.

Para alcançar um desempenho superior, propomos a utilização de várias técnicas de

classificação, minimizando o risco de concessão de crédito a um cliente inadimplente. Outro ponto de interesse é a maximização do valor da carteira de empréstimo do banco.

No segundo capítulo é feito uma abordagem estatística dos métodos de classificação avaliando as vantagens e desvantagens de cada um dos métodos no contexto da análise de crédito. Outra importante questão avaliada, é a medição do desempenho dos métodos de classificação.

O terceiro capítulo traz uma descrição da base de dados e avaliação da qualidade da informação disponível para a construção do modelo. Pontos importantes como valores ausentes, valores extremos, desbalanceamento de classes e associação entre as variáveis são discutidas e analisadas neste capítulo.

No quarto capítulo vemos as diferentes formas de realizar o agrupamento das respostas dos classificadores de forma que resultem em uma única classificação. Neste capítulo também se pode medir o ganho financeiro vindo do uso de vários classificadores.

No quinto capítulo temos a conclusão do trabalho apresenta as principais vantagens e desvantagens em se utilizar vários classificadores na classificação do risco de crédito.

Capítulo 2

Classificação

A classificação é uma técnica, em que um determinado objeto ou elemento é rotulado como pertencente a uma classe sem conhecer a princípio a classe a que este objeto pertence. Outras características do elemento são utilizadas para inferir a classe a que ele pertence. Quanto maior for o relacionamento entre as características e a classe, mais preciso é o processo de classificação.

Este processo tem iníncio com uma amostra onde os elementos estão classificados. Chama-se de variável respostas a classe de interesse. Neste contexto estamos interessados em classificar o cliente entre inadimplente ou adimplente. O conjunto onde já se sabe quais são os clientes inadimplentes e quais os adimplentes, este conjunto recebe o nome de amostra de treinamento.

Os métodos de classificação procuram por padrões que caracterizem as classes. O classificador aprende como as características dos elementos se relacionam com a classe a ser predita, por exemplo, na classe inadimplente é observado um nível superior de pessoas com pouco tempo de contratação.

Quando um novo elemento é obtido e não se conhece a que classe pertence, utiliza-se os parâmetros aprendidos no algoritmo de classificação e obtém-se a estimativa de qual é a classe mais provável deste objeto. A técnica de classificação é, em essência, uma previsão feita para os elementos através de inferências realizadas na amostra de treinamento. A diferença mais significativa entre a previsão e a classificação, é que a variável resposta é

qualitativa nos métodos de classificação e nos métodos de previsão a variável resposta é quantitativa.

Normalmente esta forma de aprendizagem em que se possui um conjunto de rótulos já definidos, é chamado de aprendizado supervisionado ou com professor, isto porque existe uma amostra de treinamento que possui a classificação correta de cada um dos elementos.

Os classificadores devem possuir certas propriedades para que sejam reconhecidos como úteis. A primeira característica é a precisão do classificador. Outra característica é a escalabilidade que é o desempenho do classificador quando a sua amostra aumenta de tamanho. A velocidade com que se consegue aprender os padrões e fazer previsão para um novo objeto é importante. A capacidade de interpretar as regras do classificador permite que o aprendizado da máquina seja absorvido pelas pessoas que se relacionam com problema.

Neste capítulo iremos tratar dos aspectos estatísticos relacionados à classificação e à teoria de decisão. Existem inúmeros tipos de classificadores. Neste trabalho adotamos alguns dos classificadores que estão entre os principais métodos [11],[6] e [10] .

2.1 Teoria Estatística de Decisão

A classificação é um fenômeno estocástico descrito por um vetor aleatório X consistindo de um conjunto de p variáveis independentes (qualitativas e quantitativas), e uma variável resposta categórica Y com k classes. As variáveis pertencentes ao vetor X são chamadas de variáveis explicativas, e variável Y é chamada de variável resposta.

Suponha também que tenhamos N realizações do vetor aleatório X representadas por $\{x_i\}_{i=1}^N$, e um conjunto correspondente de realizações da variável dependente Y, representada por $\{Y_i\}_{i=1}^N$. Estas realizações constituem a amostra de treinamento representadas por

$$\Im = \{(X_i, Y_i)\}_{i=1}^N \tag{2.1}$$

A princípio não se conhece a relação funcional exata entre X e Y, assim o seguinte

modelo é proposto:

$$Y = f(X) + \epsilon \tag{2.2}$$

Onde $f(\cdot)$ é uma função determinística, e ϵ é um erro aleatório que representa a incerteza sobre a relação entre Y e X. Na construção do modelo, o erro ϵ é em geral uma variável aleatória com média nula e com a menor variância menor possível. Quando o erro possui média zero para qualquer realização de X, o modelo é chamado de não viciado. Assim temos a seguinte resposta de predição

$$f(X) = E\left[Y \mid x\right] = E\left[Y\right] \tag{2.3}$$

Outro ponto importante na construção no modelo, é a independência entre os erros e a variável resposta. Isso nos garante que todas as informações sobre as classes são disponibilizadas através do vetor X, e está codificada na função de $f(X)$, ou seja, os erros não estão correlacionados com o vetor X.

O modelo é uma descrição matemática de um ambiente estocástico. O nosso propósito é utilizar o vetor X para explicar, e prever a variável dependente. Assim codificamos o conhecimento empírico, representado pela amostra de treinamento $\Im$ em um vetor correspondente de parâmetros do modelo.

$$\Im \rightarrow \beta \tag{2.4}$$

O modelo fornece uma aproximação para a função real $f(\cdot)$ que não é conhecida. Podemos representar a resposta do modelo de previsão produzido em resposta ao vetor de entrada x a predição y.

$$y = F(X, \beta) \tag{2.5}$$

A função $F(\cdot, \beta)$ é uma função de entrada-saída. Conhecidos os dados de treinamento $\Im$, o vetor β é obtido pela minimização da função perda.

$$L(\beta) = E(Y_i - F(x_i, \beta))^2 \tag{2.6}$$

A função perda quadrática é a mais utilizada quando a variável resposta é contínua. O erro quadrático médio (EQM) é a esperança da função perda quadrática. O EQM pode ser decomposto entre duas partes, a variância e o vício. Para se ter um bom desempenho global, tanto a variância quanto o vício da função aproximativa devem ser pequenas.

O vício é a tendência que o modelo possui de subestimar ou superestimar os resultados. A variância representa a adequação da informação contida na amostra de treinamento $\Im$ acerca da função de regressão $f(x)$, sendo o fator mais importante na avaliação do erro de estimativa, já que existem abordagens que diminuem o vício nos estimadores.

Quando a variável resposta é dicôtonica continuamos com o mesmo paradigma, havendo mudança apenas na forma da função perda. O estimador $\hat{Y}$ pode assumir todas as k classes possíveis para Y. A função perda passa a ser representada por uma matriz L de tamanho $k \times k$ onde k = cardinalidade(Y). L pode ser zero na diagonal e é não negativa, onde L(m,n) é a probabilidade por classificar uma observação da classe Y_m como classe Y_n. Quanto maiores os valores na diagonal maior é qualidade do classificador. Esta matriz é conhecida como matriz de classificação.

Em algumas situações um erro é mais custoso do que os outros, por isso podemos acrescentar o custo dos erros de classificação. Vamos chamar o custo de classificação errada como sendo $c(n|m)$, que é o custo de classificar um elemento como sendo da classe n quando ele é na verdade da classe m. Na aplicação de análise de crédito tem sentido utilizar custo classificação errada, já que o prejuízo de um cliente inadimplente é maior em módulo que o lucro de um cliente adimplente. Entretanto não encontramos um processo formalizado para estimar os custos de erro da classificação. Então uma sugestão é utilizar a razão da rentabilidade média dos clientes inadimplentes sobre a rentabilidade média dos clientes adimplentes.

2.2 Avaliação do Desempenho

Após a construção da função de classificação, é necessário avaliar a sua qualidade. Existem várias medidas diferentes que avaliam o desempenho estatístico de um classificador. Logo a seguir veremos as principais medidas e as suas características. Entretanto existem outros aspectos a serem avaliados. O valor financeiro ganho pelo modelo é uma métrica que avalia o desempenho econômico do modelo. Mesmo que o modelo seja o estado da arte, se não traz ganhos financeiros do ponto de vista do negócio, ele não possui utilidade.

Desempenho estatístico

A limitação da função de classificação esta condicionada às informações que o modelo dispõe para sua construção. A precisão de um modelo é avaliada comparando suas predições com os resultados observados. A tabela abaixo mostra os possíveis resultados associados à comparação do resultado com a predição. A forma mais usada na avaliação da precisão do modelo é a matriz de classificação.

Tabela 2.1: Construção de uma matriz de classificação

Observado	Predito		
	Adimplente	Inadimplente	Total
Adimplente	a	b	$a+b$
Inadimplente	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Os valores das caselas a e d são os elementos que foram corretamente classificados. Pode-se então determinar a proporção de acertos, ou porcentagem de casos classificados corretamente, somando-se os elementos $a+d$ e dividindo-se o resultado pelo número total de casos.

$$\hat{p}(acerto) = \frac{a+d}{a+b+c+d} \tag{2.7}$$

Quanto maior a proporção de acertos melhor. Também existe o complementar desta medida, o erro de classificação, que é o número de elementos que foram classificados de forma errada sob o número total de classificações realizadas. Esta métrica sintetiza a qualidade preditiva do modelo. Quanto menor o erro de classificação melhor.

$$\hat{p}(erro) = 1 - \hat{p}(acerto) \tag{2.8}$$

Outras características importantes são a especificidade e sensibilidade que um modelo possui. Sensibilidade é definida como a proporção entre o número de indivíduos cujo o resultado do modelo foi predito como sendo adimplente sobre o número total de clientes adimplentes, evidentemente que quanto mais próximo de 1 estiver a sensibilidade melhor. Um modelo muito sensível aponta os clientes adimplentes com alto grau de acerto.

$$Sensibilidade = \frac{a}{a+b} \tag{2.9}$$

A especificidade é a proporção entre o número de clientes que o modelo previu como inadimplente sobre o número total de inadimplentes. Quanto mais próximo de 1 estiver a especificidade do modelo, melhor será este teste. Um modelo muito específico aponta os clientes inadimplentes com alto grau de acerto. O ideal é que tanto a sensibilidade quanto a especificidade do modelo seja alta [2].

$$Especificidade = \frac{d}{c+d} \tag{2.10}$$

Quando os custos relacionados aos erros de classificação são diferentes esta informação entra na avaliação do resultado da seguinte forma:

$c(a \mid i)$: é o custo do cliente inadimplente ser classificado como adimplente.

$c(i \mid a)$: é o custo do cliente adimplente ser classificado como inadimplente.

Neste caso, o custo esperado de classificação incorreta (CI) é:

$$CI = c(a \mid i) \cdot \frac{c}{c+d} + c(i \mid a) \cdot \frac{b}{a+b} \tag{2.11}$$

A fórmula acima considera que as duas classes possuem probabilidades iguais. Caso exista um desbalanceamento, incluimos as proporções respectivas a cada classe.

$$CI = c(a \mid i) \cdot \frac{c}{c+d} \cdot p(inadimplente) + c(i \mid a) \cdot \frac{b}{a+b} \cdot p(adimplente) \tag{2.12}$$

onde:

p(adimplente): é a proporção de clientes adimplentes

p(inadimplente): é a proporção de clietnes inadimplentes.

Todas as métricas avaliadas até agora são pontuais. Entretanto, pode-se extrair mais informações a respeito da predição. A maioria do modelos associa predição à probabilidade de ocorrência. Saber que um cliente possui probabilidade de 80% de ser adimplente, produz mais confiança na classificação do que saber apenas a classificação. Quanto maior a probabilidade do cliente pertencer a uma classe maior é a confiança que pode ser dada a classificação, mantendo os outros fatores constantes.

O gráfico ROC (*Reciever Operating Characteristic*) [7] permite uma visualização melhor da sensibilide e especificidade do modelo, e mostra o *tradeoff* entre a sensibilidade e a especificidade. Na figura **??** são exibidas a sensibilidade (True Positive) e (1 - especificidade) (False Positive).

Normalmente considera-se o ponto de corte de 0,5. Se a chance de um cliente ser adimplente é maior do que 0,5, então é predita a classe de adimplente para ele. Entretanto

nada impede que o ponto de corte seja qualquer outro valor, como por exemplo 0,3. Se a probabilidade do cliente ser adimplente for maior que 0,3, então ele é classificado como adimplente. Para cada ponto de corte pode-se construir uma matriz de classificação e através desta matriz obtem-se os valores de sensibilidade e (1-especificidade) da regra de classificação e plota-se os valores no gráfico ROC.

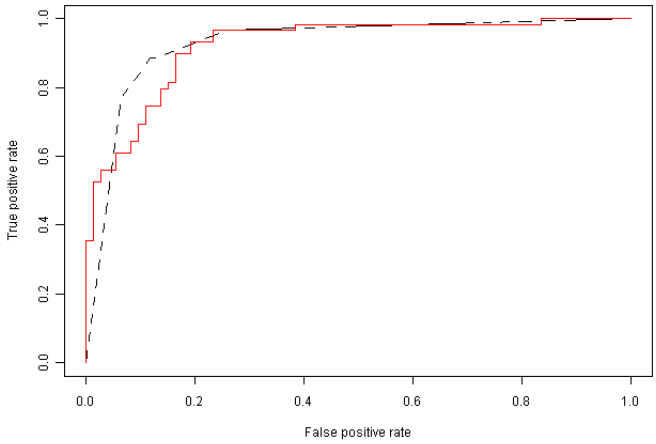


Figura 2.1: ROC de dois modelos de classificação

A curva pontilhada em vermelho é de um modelo de regressão logística, a curva preta é de uma árvore de classificação. Quanto mais alto à esquerda a curva passa melhor é o ajuste do modelo. Se o elementos fossem classificados de forma aleatória o gráfico mostraria uma reta diagonal crescente. Se no gráfico existir uma curva superior a outra, então o modelo superior domina o modelo inferior. No gráfico exibido isso não ocorre, no início da curva percebe-se que a regressão logística é superior já na parte central; a árvore de classificação é superior e no fim, a regressão logística passa a ser superior novamente.

Até agora todos os estimadores foram estimativas pontuais, entretanto avaliar a dispersão dos resultados é importante para saber se os testes apresentam diferenças significativas. Provavelmente o método mais utilizado e um dos mais simples para estimar o erro de predição é a validação cruzada (*cross validation, CV*). Se existem dados suficientes,

uma parte dos dados é reservada. Esta amostra é chamada de amostra de validação (CV_i), enquanto o restante é utilizado na estimação do modelo. Os dados que foram reservados são então utilizados na avaliação do ajuste do modelo. Os dados são divididos em k partes iguais, então, utilizando-se cada um dos k sub-conjuntos obtem-se um conjunto de estimativas do erro de predição do modelo[7].

Tipicamente o valor de k varia de 5 a 10, se o valor de k for igual ao tamanho da amostra então o método de validação é chamado de *leave-one-out-cross*. Os elementos são divididos entre os grupos de maneira aleatória de forma que o número de elementos em cada conjunto seja igual. Cada sub-conjunto é utilizado para estimar o erro apenas uma vez. O valor estimado para o erro de classificação é igual a média dos erros de classificação. Se dois modelos tiverem erro de classificação iguais, então o modelo com a menor variabilidade nas classificações deve ser utilizado.

Desempenho econômico

Para avaliar o desempenho econômico do modelo, é necessário utilizar as probabilidades a priori das classes adimplentes e inadimplentes. Conhecendo estas probabilidades, pode-se calcular o resultado econômico do modelo. Inicialmente analisa-se qual o impacto de não utilizar nenhuma modelagem na concessão de crédito.

Considerando que o percentual de clientes adimplentes seja p(a) e p(i) seja o percentual de clientes inadimplentes, tal que p(a+i)=1; sendo os custos de conceder empréstimo a um clinte inadimplente c e o lucro de conceder empréstimo a um cliente adimplente l; então a rentabilidade pode ser calculada como:

$$Rentabilidade = p(a) \cdot (l+c) - c \tag{2.13}$$

Para esta estratégia ser rentável a proporção dos cliente adimplentes, p(a) deve ser:

$$p(a) \geq \frac{c}{l+c} \tag{2.14}$$

No outro extremo temos a concessão de crédito apenas aos clientes adimplentes. Mesmo sabendo que não existe forma de atingir tal propósito, este cálculo permite estabelecer o limite que um classificador pode alcançar. Assim a rentabilidade é igual à:

$$Rentabilidade = p(a) \cdot l \tag{2.15}$$

Entre estes dois extremos estão os modelos de classificação. Assumindo que quando um cliente adimplente não recebe crédito a perda é igual a l, pode-se construir a seguinte equação de rentabilidade do modelo.

$$Rentabilidade = p(a) \cdot l \cdot (2 \cdot sensibilidade - 1) - p(i) \cdot c \cdot (1 - especificidade) \tag{2.16}$$

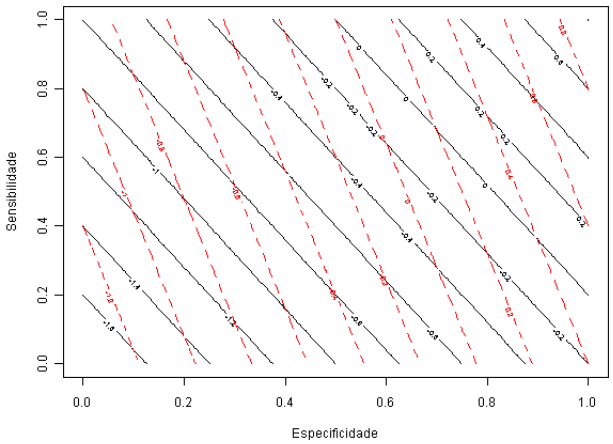


Figura 2.2: Sensibilidade da rentabilidade

A figura 2.2 mostra o resultado financeiro, o c=5, l=1 e p(a)=0.9 e 0.8. As retas pontilhadas em vermelho mostram o impacto quando a p(a)=0.9 e as retas pretas quando p(a)=0.8. A sensibilidade é mais importante quando a proporção de clientes inadimplentes é maior ou o custo da inadimplencia é elevado. A especificidade é importante se o lucro é alto e a proporção de clientes adimplentes é alta.

2.3 Análise Discriminante

Análise Discriminante é uma técnica de análise de dados onde a variável resposta é categórica e as variáveis explicativas são contínuas [16]. Utilizando a amostra de treinamento estima-se as funções discriminantes que melhor discriminem entre as categorias da variável resposta. Durante este processo as variáveis explicativas são avaliadas para descobrir quais são as mais influentes.

Neste trabalho dispomos apenas de duas populações: adimplentes (1) e os inadimplentes(0). Seja $f_1(x)$ e $f_2(x)$ as funções de probabilidade da variavél explicativa X para o grupo adimplente e inadimplente. Então existe R_1 que é uma região que classifica os elementos como pertencente a classe 1 (adimplente). Por consequencia $R_0 = \omega - R_1$, é a região que classifica os elementos como vindo da classe 0 (inadimplente).

A probabilidade de classificarmos o elemento como sendo 0 (inadimplente), quando na verdade é 1 (adimplente) pode ser traduzida pela equação 2.17. A equação 2.18 é a probabilidade de classificar o elemento como 1 dado que é 0.

$$P(0 \mid 1) = P(X \in R_0 \mid 1) = \int_{R_0=\omega-R_1} f_1(X) \tag{2.17}$$

$$P(1 \mid 0) = P(X \in R_1 \mid 0) = \int_{R_1} f_0(X) \tag{2.18}$$

Utilizando a amostra de treinamento também podemos estimar as probabilidades apriori de pertencer aos grupos 1 e 0. Utilizando todas estas informações e adicionando o custo de classificação errada estimado na seção anterior podemos definir o custo esperado de erro de classificação (ECM). Na equação do ECM, vemos que a diferença de tamanho entre os grupos tem grande impacto no valor final. Na análise de crédito $p(a) >> p(i)$, existe mais elementos adimplentes do que inadimplentes na amostra. A regra de decisão que minimiza o custo de erro de classificação é dado pela seguinte forma:

$$\left[R_1 : \frac{f_1(X)}{f_0(X)} \geq \frac{c(1 \mid 0)}{c(0 \mid 1)} \frac{p(i)}{p(a)}\right] \tag{2.19}$$

Se R_1 for falso, então R_0 é verdadeiro e o cliente foi classificado como sendo inadimplente. Note que como existe uma razão maior de clientes adimplentes, e o custo de

classificar um cliente inadimplente como adimplente é maior do que perder um cliente adimplente que foi erroneamente classificado como inadimplente.

Normalmente, quando trabalha-se com uma distribuição multivariada normal pode-se utilizar algumas propriedades que torna a análise discriminante mais fácil de ser utilizada. Entretanto é necessário verificar, se as matrizes de variância-covariância são estatisticamente diferentes ou não. Se assumirmos a hipótese de igualdade das matrizes var-cov, então pode-se escrever a regra de decisão como sendo:

$$R_1 : exp\{-\frac{1}{2}(x-\mu_1)^{'}\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{'}\Sigma^{-1}(x-\mu_0)\} \geq \frac{c(1|0)}{c(0|1)}\frac{p_0}{p_1} \tag{2.20}$$

Se R_1 for falso então o elemento é classificado como sendo inadimplente. Entretanto, quando algumas das variáveis não possuem distribuição normal, contornamos este problema aplicando a transformação de Box-Cox para aproximar as variáveis à normalidade [16]. Caso as matrizes de var-cov sejam diferentes pode-se utilizar a seguinte regra de decisão:

$$R_1 : \Big(-\frac{1}{2}x^{'}(S_1^{-1} - S_0^{-1})x + (\overline{x}_1^{'}S_1^{-1} - \overline{x}_0^{'}S_0^{-1})x - k\Big) \geq ln\Big(\frac{c(1 \mid 0)}{c(0 \mid 1)} \cdot \frac{p_0}{p_1}\Big) \tag{2.21}$$

onde k é:

$$k = \frac{1}{2}ln(\frac{|\Sigma_1|}{|\Sigma_0|}) + \frac{1}{2}(\mu_1^{'}\Sigma_1^{-1}\mu_1 - \mu_0^{'}\Sigma_0^{-1}\mu_0) \tag{2.22}$$

Algumas das desvantagens deste método são:

- A forma da distribuição das variáveis deve ser convexa(elipse), caso não seja os métodos não-lineares são mais indicados.
- A pouca flexibilidade ao trabalhar com variáveis não-normais surgindo a necessidade de transformações nas variáveis para reduzir a assimetria.
- Quando existem variáveis explicativas categóricas é indicado o uso de outras técnicas como a regressão logística [16].

Após considerar os pontos acima, não abordaremos a técnica de análise discriminante.

2.4 Árvores de Classificação

A árvore de classificação é um mecanismo para classificar dados, sendo representada graficamente através dos elementos: nó raiz, nós de decisão e folhas. O nó de decisão contém um teste sobre o valor de um atributo. Cada ramo para sub-árvore ou folha tem um conjunto de valores associados. Cada sub-árvore contém a mesma estrutura de uma árvore. Este processo cria um único caminho da raiz até a folha. O CART (*Classification and Regression Trees - CART*) é um dos métodos mais utilizados.

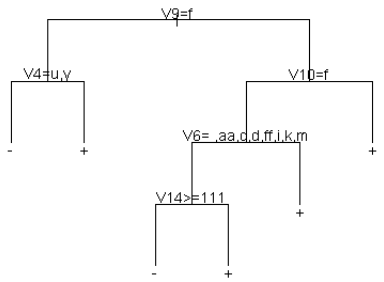


Figura 2.3: Árvore de classificação de crédito

Os elementos são particionados até que o grau de homogeneidade dos elementos dentro de cada nodo seja alto, ou o nodo contenha um número muito pequeno de elementos para ser novamente particionado. A figura 2.3 exemplifica uma árvore de classificação aplicada ao contexto de análise de crédito. Se a variável V9 do cliente for igual a F, então o cliente vai para o nodo esquerdo, e se a variável V4 do cliente é L então vai para o nodo direito. Como o cliente alcançou uma folha, o perfil de crédito é predito como sendo adimplente.

Os nodos finais são normalmente chamados de folhas. As folhas recebem os rótulos que caracterizam os elementos nelas contidos. Já nos nodos, são realizados testes sobre

os elementos de forma a produzirem partições com elementos mais homogêneos do que o nodo superior. Uma árvore de classificação distribui elementos de uma base de dados em um número finito de classes. Em uma árvore de decisão, os nodos da árvore representam os nomes de atributos, as ligações representam possíveis valores para o atributo e as folhas representam as diferentes classes.

Resumindo, os passos na criação de uma árvore de classificação são:

- Definir uma forma de particionar os dados de forma que os nodos sejam mais homogêneos que o conjunto inicial;
- Definir uma regra de parada no processo de particionamento;
- Definir o rótulo de classificação.

Um indivíduo é classificado seguindo o caminho da raiz da árvore até uma folha, enquanto as suas características satisfizerem os nodos e suas ligações. Ao atingir a folha o rótulo da folha é aplicado ao elemento.

A árvore de classificação pode ser aplicada a grandes conjuntos de dados e são adequadas para qualquer tipo de dados (qualitativos ou quantitativos), [7]. Outra vantagem é a facilidade de serem entendidas, e o resultado do algoritmo pode ser usado diretamente pelo usuário. Algumas das principais vantagens são:

- baixa influência dos outliers, já que os pontos que cortam são definidos com base na ordem dos elementos;
- baixa influência do número de variavéis utilizadas no modelo, já que apenas as variáveis que discriminam são utilizadas na construção da árvore;
- a árvore é invariante a toda transformação monotônica nas variáveis;
- facilidade de aplicar e armazenar o modelo.

2.4.1 Processo de Particionamento

A idéia fundamental é realizar a partição que produz os nodos mais puros ou homogêneos que o conjunto inicial. Entre os critérios disponíveis para a escolha do atributo estão o método da entropia, o critério de Gini, o método da paridade, a técnica de Laplace e também escolha aleatória dentre outros.

Critério de Entropia

A entropia é um conceito geral e está associada à tendência natural dos processos macroscópicos para a desordem. Na teoria da informação, a entropia foi associada à medição do conhecimento sobre o estado do sistema. Este critério mede a quantidade de bits necessários para a codificação da informação [9]. Se a frequência de cada classe for igual, a entropia é máxima, e se todos os elementos pertencerem a uma única classe a entropia é zero.

$$E(y) = -\sum p_i \cdot log(p_i) \tag{2.23}$$

A entropia pode então ser usada para calcular a estimativa do grau de impuridade de um nodo.

Critério de Gini

Uma maneira de avaliar o grau de concentração de uma variável é utilizando o índice de Gini. O índice de Gini mede o quanto uma dada distribuição difere de uma distribuição eqüitativa ou em quanto duas distribuições diferem uma da outra.

Aplicando o índice de Gini como critério de particionamento, o valor máximo é alcançado quando todas as classes tem igual probabilidade de ocorrer, e será igual a zero quando os elementos pertencerem a uma única classe.

$$G(y) = 1 - \sum_{i=1}^{k} p_i^2 \tag{2.24}$$

Os critérios avaliados acima apenas medem o impacto do particionamento. Eles não indicam quais variáveis devem ser particionadas nem onde a variável deve ser particionada. O ganho de informação obtido com a partição pode ser calculado da seguinte forma:

$$\Delta i = G(y) - p_d G(y_d) - p_e G(y_e) \tag{2.25}$$

ou

$$\Delta i = E(y) - p_d E(y_d) - p_e E(y_e) \tag{2.26}$$

onde o índice (d) e (e) significa nodo da direita e da esquerda. O ganho de informação pode ser visto como a diferença entre a impureza inicial e a média da impureza dos novos nodos. O índice de Gini e a entropia medem o ganho de informação que é inversamente proporcional ao nível de impureza de cada nodo.

Para exemplificar o processo vamos comparar o ganho de informação das variáveis V9 e e V10 (figura 2.3) para o nodo raiz. Ambas as variáveis são dicotômicas.

Figura 2.4: Avaliando o ganho de informação

Utilizando a entropia para avaliar a partição através da variável V9

$$E(y) = \frac{-383}{690} \cdot log\left(\frac{383}{690}\right) - \frac{307}{690} \cdot log\left(\frac{307}{690}\right) = 0.6870$$
$$E_e(y) = \frac{-77}{361} \cdot log\left(\frac{77}{361}\right) - \frac{284}{361} \cdot log\left(\frac{284}{361}\right) = 0.5182$$
$$E_d(y) = \frac{-306}{338} \cdot log\left(\frac{306}{338}\right) - \frac{23}{338} \cdot log\left(\frac{-23}{338}\right) = 0.2534$$

Ganho de informação para a variável V9 é:

$$\Delta E = 0.6870 - 0.4768 \cdot 0.2534 - 0.5231 \cdot 0.51829 = 0.30729 \tag{2.27}$$

Utilizando o índice de Gini para avaliar a partição variável V9

$$G(y) = 1 - \left(\frac{383}{690}\right)^2 - \left(\frac{307}{690}\right)^2 = 0.4939$$

$$Ge(y) = 1 - \left(\frac{77}{361}\right)^2 - \left(\frac{284}{361}\right)^2 = 0.3356021$$

$$Gd(y) = 1 - \left(\frac{306}{338}\right)^2 - \left(\frac{23}{338}\right)^2 = 0.1300431$$

Ganho de informação para a variável V9 é

$$\Delta G = 0.4939 - 0.5231 \cdot 0.3356021 - 0.4768 \cdot 0.1300431 = 0.256342 \tag{2.28}$$

Utilizando a entropia para avaliar a partição através da variável V10

$$E(y) = -\frac{383}{690} \cdot log\left(\frac{383}{690}\right) - \frac{307}{690} \cdot log\left(\frac{307}{690}\right) = 0.6870$$

$$E_e(y) = -\frac{86}{295} \cdot log\left(\frac{86}{295}\right) - \frac{209}{295} \cdot log\left(\frac{209}{295}\right) = 0.6035$$

$$E_d(y) = -\frac{297}{395} \cdot log\left(\frac{297}{395}\right) - \frac{98}{395} \cdot log\left(\frac{98}{395}\right) = 0.5602$$

Ganho de informação para a variável V10 é:

$$\Delta i = 0.6870 - 0.4275 \cdot 0.6035 - 0.5724 \cdot 0.5602 = 0.1082 \tag{2.29}$$

Utilizando o índice de Gini para avaliar a partição da variável V10

$$G(y) = 1 - \left(\frac{383}{690}\right)^2 - \left(\frac{307}{690}\right)^2 = 0.4939$$

$$Ge(y) = 1 - \left(\frac{86}{295}\right)^2 - \left(\frac{209}{295}\right)^2 = 0.4130$$

$$Gd(y) = 1 - \left(\frac{297}{395}\right)^2 - \left(\frac{98}{395}\right)^2 = 0.3730$$

Ganho de informação para a variável V10 é

$$\Delta G = 0.4939 - 0.4275 \cdot 0.4130 - 0.5724 \cdot 0.3730 = 0.1037 \tag{2.30}$$

A variável V9 possui a capacidade de diminuir o grau de impureza dos nodos de forma mais acentuada do que a variável V10, ou seja, o ganho de informação é maior se a variável V9 for utilizada para particionar os elementos.

Definição da Partição

Ao estabelecer o atributo a ser utilizado, é necessário definir o particionamento das ligações. O tipo da variável altera a forma de realizar o particionamento. As variáveis qualitativas são trabalhadas de forma diferente das variáveis quantitativas.

- Variável Dicotômica

  Este é o caso mais simples; a variável só pode ser dividida em dois grupos. Após a divisão é avaliado o ganho de informação alcançado e o valor é comparado com os das demais variáveis explicativas utilizadas.

- Variável Multinomial

  Neste caso existem $(2^{n-1} - 1)$ formas de se agrupar os elementos. Caso o número de classes na variável seja elevado o tempo computacional necessário para processar todas as combinações possíveis pode se tornar significativo.

- Variável Ordinal

  Para um atributo com k classes, após a ordenação das classes avalia-se todas as k-1 partições possíveis para escolher a melhor partição.

- Variável Contínua

  Escolhe para a partição o ponto considerado com o melhor valor pelo critério de avaliação adotado (entropia ou Gini), testando e avaliando todas as partições possíveis com base em cada uma das partições possíveis. O ponto de particionamento consiste em um teste binário com resultado $x \leq A$ e $x \geq A$ onde A é o valor de particionamento. Para encontrar o melhor ponto de particionamento ordena-se a variável de forma crescente e avalia-se o ponto médio entre dois valores consecutivos. Avalia-se todos os valores possíveis.

2.4.2 Definição do rótulo da classe

Durante o crescimento da árvore de classificação é necessário avaliar se existe um próximo nodo ou se o processo vai se encerrar. Se todos os elementos de um nodo per-

tencem a uma mesma classe, o nodo pode ser considerado como sendo uma folha, ou seja, os elementos que alcançarem a folha receberão o rótulo da classe mais freqüente na folha. Caso os elementos tenham classes diferentes mas todas as características nas variáveis respostas sejam iguais, então a folha recebe o rótulo da classe mais provável.

Se o número de elementos dentro de uma classe for menor do que um limite pré-estabelecido (5 elementos) ou o resultado do melhor atributo não apresenta diferença do nodo anterior, então a folha recebe o rótulo da classe mais provável.

2.4.3 Critério de Parada ou poda de uma árvore

As árvores de classificação possuem a tendência de decorar a amostra de treinamento. Este comportamento é conhecido como *overfitting* e reduz a capacidade de generalização do classificador. Existem algumas estratégias que reduzem o impacto deste tipo de problema.

A primeira e mais simples é estabelecer um critério de parada, de forma que durante o processo de construção da árvore se avalie se o ganho de informação é maior que um determinado limite. Caso o ganho de informação não seja maior que o valor definido o nodo se transforma em uma folha. Uma das críticas ao método de pré-poda é a definição do valor limite, caso seja muito alto a árvore será maior que o necessário e o seu poder de generalização será menor, caso seja muito baixo o poder de predição da árvore será menor.

Outra forma de contornar o overfitting, é deixando que a árvore seja construída sem nenhuma restrição e após a sua estimação uma poda seja feita de forma a corrigir os nodos onde a capacidade de generalização ficou comprometida. Uma das formas mais simples de avaliar o grau de ajuste da árvore é através da validação cruzada. Para cada aumento do árvore o seu erro é também avaliado na amostra de validação. A partir de um determinado tamanho da árvore o erro de predição na amostra de treinamento cresce com o aumento da dimensão da árvore. Este seria o ponto em que a árvore de classificação está perdendo a capacidade de generalização e começando a decorar as características da amostra de treinamento. Portanto a árvore final seria esta árvore podada.

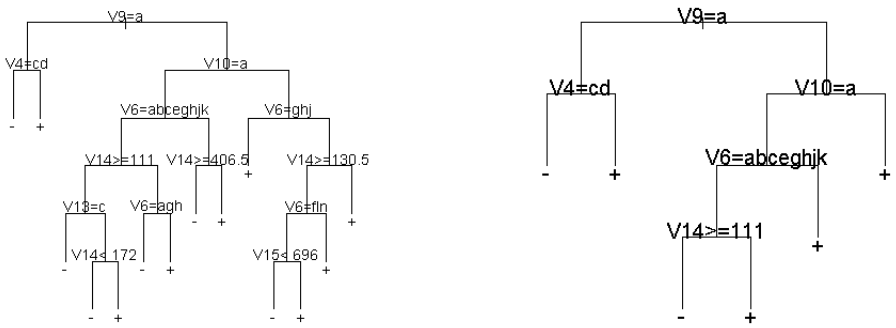


Figura 2.5: Diferenças de uma árvore completa para uma podada

Outra forma é avaliar o custo de complexidade. A árvore é avaliada através da seguinte fórmula:

$$C(t) = E(t) + \alpha \cdot \mid \text{número de folhas} \mid \tag{2.31}$$

Onde:

c(t) é o custo de complexidade;

α é o fator de complexidade.

Se o valor de α for igual a zero, então o tamanho da árvore não possui nenhum impacto no custo de complexidade. Com aumentos no valor de α a dimensão da árvore já começa a impactar no custo de complexidade, se o valor de α for muito alto é melhor ter uma árvore bem menor e aceitar um nível superior de erro. Para cada valor de α existe uma poda que minimiza o custo de complexidade

2.4.4 Boosting

As árvores de classificação sofreram um forte impacto com a criação do método de boosting. De acordo como o princípio de dividir para conquistar, um tarefa complexa é resolvida dividindo-a em um número de tarefas mais simples. Esta mesma abordagem pode ser aplicada ao problema de classificação. Através de uma série de classificadores é feita a classificação do elemento.

No método de boosting ao invés de fazer um único modelo, é construido uma série de modelos, e uma votação ponderada entre os classificadores é feita de forma a classificar o elemento com a classe mais frequente. A importância deste método está no fato de alterar a importância dos exemplos de classificação de acordo com a dificuldade. Os elementos de classificação mais fáceis são menos importantes do que os elementos mais difíceis, por isso a ponderação dada aos exemplos difíceis ou próximos da borda de classificação são maiores do que os exemplos periféricos. O método de boosting altera as bases de exemplos utilizadas pelos classificadores criando conjuntos de dados com distribuições inteiramentes diferentes.

Uma das abordagens de maior sucesso entre os métodos de boosting é o AdaBoost [6]. O AdaBoost difere dos algoritmos de reforço em dois aspectos. Primeiro, o ajuste é alcançado de forma adaptativa em relação aos erros da hipótese retornada pelo modelo de aprendizagem; segundo o desempenho depende apenas do modelo de aprendizagem sobre aquelas distribuições que são realmente geradas durante o processo de aprendizagem. A classificação final é calculada por uma votação ponderada das hipóteses construídas.

Algoritmo de Boosting

Entrada - Amostra de treinamento $(x_i, d_i)_{i=1}^{N}$, modelo de aprendizagem e número de iterações do modelo. Inicialização: Faça $D_1(i) = 1/N$ para todo i.

Computação:

1. Chame o modelo de aprendizagem, fornecendo a distribuição D_n;
2. Retorne o modelo $M_n : X \longrightarrow Y$;
3. Calule o erro do modelo M_n

$$\epsilon_n = \sum_{i:M_n(x_i)\neq d_i} D_i$$

4. Faça

$$\beta_n = \frac{\epsilon_n}{1-\epsilon_n}$$

5. Atualizar a distribuição D_n:

$$D_{n+1}(i) = \frac{D_n(i)}{Z_n} \times \begin{cases} \beta_n \text{ se } M_n(x_i) = d_i \\ 1 \text{ caso contrario} \end{cases}$$

onde Z_n é uma constante de normalização (escolhida de modo que $D_{n+1}(i)$ seja uma distribuição de probabilidade.

Saída: O modelo final é

$$M(x) = argmax_{d\in D} \sum_{n:M_n(x)=d} log\frac{1}{\beta_n}$$

2.5 Regressão Logística

Quando a variável resposta é dicotômica (resposta 0 ou 1), um modelo muito utilizado é a regressão logística. A vantagem da regressão logística sobre análise discriminante é a facilidade de incluir variáveis qualitativas e quantitativas.

A regressão logística deve ser utilizada neste contexto porque a regressão linear apresenta as seguintes desvantagens quando a variável resposta é dicotômica são [16]:

- não-normalidade dos resíduos.
- heteroscedasticidade dos resíduos.
- possibilidade do valor previsto ficar fora do limite 0-1.
- relacionamento linear entre a probabilidade da variável resposta e as variáveis explicativas.

O modelo de regressão logística mantém a variável resposta dentro do limite 0-1 e a relação entre probabilidade de classsifação e as variáveis resposta não é linear. A curva em formato de um "S" é a forma característica de uma distribuição logística.

A equação abaixo mostra como é a forma do relacionamento entre as variáveis explicativas e a variável resposta:

$$P_i = E(Y = 1 \mid X_i) = \frac{1}{1 + e^{-\beta_1 + \beta_2 X_i}} \tag{2.32}$$

Enquanto os β e o X podem assumir qualquer valor no conjunto dos reais indo do $-\infty$ a ∞, a P_i varia entre 0 e 1. Pode-se relacionar $P_i/(1 - P_i)$ com a simples razão de probabilidades a favor e a probabilidade contra. Assim, se $P_i = 0.8$, significa que as chances são de 4 para 1 em favor a adimplência. Se aplicarmos o logaritmo na razão de chance, obteremos um resultado que lineariza a equação.

$$ln\Big(\frac{P_i}{1 - P_i}\Big) = \beta_1 + \beta_2 X_i \tag{2.33}$$

A interpretação do modelo logit para o β_2 é a variação da chance em log para cada mudança unitaria em X. O intercepto é o valor da chance em log se a variável explicativa for zero.

2.5.1 Estimativa do Modelo

Se utilizarmos os dados em nível micro, não podemos estimar os parâmetros pela rotina de Mínimos Quadrados. Por isso, utilizaremos o método de máxima verossimilhança para estimar o parâmetro.

O processo iterativo para obter $\widehat{\beta}$ pode ser expresso como um processo iterativo de mínimos quadrados ponderados

$$\beta^{m+1} = (X^T V^{(m)} X)^{-1} X^T V^{(m)} z^{(m)} \tag{2.34}$$

onde $V = diag\{\pi_1(1-\pi_1), \ldots, \pi_n(1-\pi_n)\}$, z é a variável dependente modificada, $z_i = \eta_i + (y_i - n_i\pi_i)/\{n_i\pi_i(1-\pi_i)\}$. Assintoticamente,

$$\widehat{\beta} - \beta \sim N_p(0, (X^T V X)^{-1}). \tag{2.35}$$

2.5.2 Seleção de Modelos

Uma maneira de encontrar um modelo reduzido que inclua apenas as covariáveis e interações mais importantes para explicar a probabilidade de adimplência $\pi(X)$ é o método Stepwise. O método baseia-se num algoritmo misto de inclusão e eliminação de variáveis segundo a importância das mesmas de acordo com algum critério estatísticio (distribuição χ^2, F). Esse grau de importância pode ser avaliado pelo nível de significância do teste de razão de verossimilhança entre os modelos que excluem ou incluem as variáveis em questão.

Descrevemos uma variante desse algoritmo usada por Hosmer e Lemeshow. A etapa inicial começa com o ajustamento do modelo apenas com o intercepto e é completada pelos passos seguintes:

1. constroem-se testes de razão de verossimilhança entre o modelo inicial e os modelos lógisticos simples formados como cada uma das covariáveis do estudo. O menor dos níveis descritivos associados a cada teste será comparado com limite superior (PE). Se PE for maior incluímos a covariável referente àquele nível e passamos ao passo seguinte; caso contrário, paramos a seleção e adotamos o último modelo;

2. partindo do modelo, incluindo a covariável selecionada no passo anterior, introduzimos individualmente as demais covariáveis. Cada um desses novos modelos é testado contra o modelo inicial desse passo. Novamente, o menor valor dos níveis descritivos é comparado com PE. Se for menor do que PE, implica na inclusão no modelo da covariável correspondente e a passagem ao passo seguinte. Caso contrário, paramos a seleção;

3. comparamos o desvio do modelo logístico contendo as covariáveis selecionadas nos passos anteriores com o desvio dos modelos que dele resultam por exclusão individual de cada uma das covariáveis. Se o maior nível dos testes da razão de verossimilhança for menor do que PS, a covariável associada a esse nível descritivo permanece no modelo. Caso contrario, ela é removida. Em qualquer circunstância, o algoritmo segue para o passo seguinte;

4. o modelo resultante do passo anterior será ajustado, no entanto, antes de tornar-se o modelo inicial da etapa 2 (seleção de interações de primeira ordem entre as covariáveis incluídas), avaliamos a significância de cada um dos coeficientes das covariáveis selecionadas, por exemplo através de um teste de Wald. Se alguma covariável ou fator não for significativo podemos excluí-los do modelo.

5. uma vez selecionadas as covariáveis mais importantes, ou efeitos principais, damos entrada na etapa 2 com o passo 1 que agora envolve apenas iterações de primeira ordem entre as covariáveis selecionadas, e assim por diante.

Uma desvantagem do procedimento descrito é exigir as estimativas de máxima verossimilhança em cada passo, o que encarece o trabalho computacional, particularmente

quando há muitas covariáveis ou fatores.

Avaliando o Erro de Classificação

Como estamos aplicando a regressão logística no contexto de análise de crédito, sabemos que o custo de rejeitar um cliente adimplente é menor que o custo de aceitar um cliente inadimplente. Essa diferença existente sobre o tipo de erro deve ser considerada na classificação do cliente.

Como exemplo, se o custo de um cliente inadimplente é $ 100 e o lucro de um cliente adimplente é $ 20 existe uma relação 1:5 para o cliente inadimplente. Então, se a probabilidade de um determinado cliente ser adimplente for de 80%, é mais seguro não aceita-lo, mas se for de 84% o cliente já deve ter o seu crédito concedido.

Este procedimento permite que a diferença nos erros de classificação sejam consideradas na análise do cliente. Se os custos dos erros de classificação forem iguais, a fronteira de decisão passa a ser os 50%, Se um cliente tem uma chançe de ser adimplente maior que 50%, então ele recebe crédito. Como não se sabe qual o custo do erro de classificação assume-se que seja igual.

A regressão logística tem sido considerada a técnica padrão em análise de crédito[15]. Um dos maiores problemas da regressão logística é a quantidade de detalhes que podem reduzir a capacidade preditiva do modelo, como outliers e dados omissos.

2.6 Máquina de Vetor Suporte

A idéia principal de uma máquina de vetor suporte é construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre os exemplos seja máxima.

A máquina de vetor suporte (Support Vector Machines - SVM) é uma implementação do método de minimização estrutural de risco. Este princípio indutivo é baseado no fato de que a taxa de erro de uma máquina de aprendizagem sobre os dados de teste é limitada pela soma da taxa de erro de treinamento e por um termo que depende da

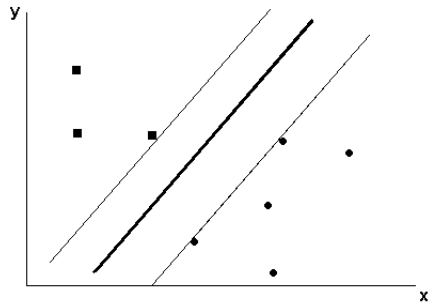
dimensão Vapnik-Chernovenckis(V-C).

Consequentemente, a máquina de vetor suporte pode fornecer um bom desempenho de generalização em problemas de classifição de padrões, apesar do fato de que ela não incorporar conhecimento do domínio do problema. Este atributo é único das máquinas de vetor de suporte[14].

Em termos conceituais, os vetores de suporte são aqueles elementos cujas coordenadas se encontram mais próximos da superfície de decisão e são portanto, os mais difíceis de classificar. Dessa forma, tem uma influência direta na localização ótima da superfície de decisão. O hiperplano ótimo é único no sentido de que o vetor w fornece a máxima separação possível entre exemplos positivos e negativos [6]. A classificação feita por este método maximiza a capacidade de generalização. A SVM é uma técnica que tem alcançado bons resultados em diferentes contextos de classificação.

A SVM incorpora o princípio de minimização do risco estrutural e Vapanik mostra que este princípio é superior ao de minimização de risco empírico. O princípio de minimização do risco estrutural leva à minimização do limite superior do erro de generalização, enquanto a minimização do risco empírico implica na minimização dos erros de classificação da amostra de treinamento.

Figura 2.6: separação máxima entre os dois grupos



A figura 2.6 apresenta um vetor que separa as duas classes (quadrado e círculo). Este plano é o resultado do SVM e é um plano ótimo de separação entre as classes, já que a distância dos elementos até a margem de separação é máxima. Um hiperplano que maximiza a margem de separação entre as classes, garante uma capacidade de generalização

superior aos métodos já apresentados.

Para facilitar a compreensão abordaremos inicialmente o caso mais simples de separação de classes lineares, depois iremos identificar as mudanças necessárias para incluir a não-linearidade no modelo.

A equação de uma superfície de decisão na forma de um hiperplano que realize a função de separação entre as classes é

$$w^T x + b = 0 \tag{2.36}$$

Onde x é o vetor de entrada, w é um vetor de peso ajustável e b é um bias. Pode-se escrever assim o problema

$$w^T x_i + b \geq 0 \ \forall \ y = +1$$

$$w^T x_i + b \leq 0 \ \forall \ y = -1$$

Para um dado vetor w e bias b, a separação entre o hiperplano e o ponto de dado mais próximo é denominado margem de separação, representado por ρ. O objetivo de uma máquina de vetor suporte é encontrar o hiperplano para o qual a margem de separação ρ é máxima. Sob esta condição, a superfície de decisão é ótima. O hiperplano ótimo, representando uma superfície de decisão linear multidimensional no espaço de entrada é definido por:

$$w_o^T x + b_o = 0 \tag{2.37}$$

A função discriminante $g(x) = w_o^T x + b_o$ fornece uma medida algébrica da distância de x até o hiperplano.

$$r = \frac{g(x)}{\|w_o\|} \tag{2.38}$$

Onde r é a distância do vetor de suporte aos elementos mais próximos.

Para encontrar os parâmetros w_o e b_o para o hiperplano ótimo, os vetores devem satisfazer a seguinte restrição:

$$w_o^T x_i o + b \geq +1 \ \forall \ y = +1$$

$$w_o^T x_i o + b \leq -1 \ \forall \ y = -1$$

Se os padrões forem linearmente separáveis, os pontos para os quais a primeira e segunda linha são satisfeitos com o sinal de igualdade são chamados de vetores de suporte. Os vetores de suporte são os pontos que se encontram mais próximos da superfície de decisão e são, portanto, os mais difíceis de classificar.

Considere que ρ represente o valor ótimo da margem de separação entre as duas classes que constituem o conjunto de treinamento.

$$\rho = 2r \tag{2.39}$$

Maximizar a margem de separação entre classes é equivalente a minimizar a norma euclidiana do vetor de coeficiente w. O problema de otimização pode ser formulado como:

Dado a amostra de treinamento encontre os valores ótimos do vetor de coeficientes e bias b de modo que satisfaçam a seguinte restrição

$$w^T x_i + b \geq 1 \ \forall \ i = 1, \cdots, N \tag{2.40}$$

e o vetor peso w minimize a função custo

$$\Phi(w) = \frac{1}{2} w^T w \tag{2.41}$$

O problema possui o seguinte dual que pode ser resolvido da seguinte forma.

Dado a amostra de treinamento, encontre os multiplicadores de lagrange $\{\alpha_i\}$ que maximizem a função objetivo

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - 1 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j \tag{2.42}$$

sujeito as seguintes restrições

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \ \alpha_i \geq 0 \ \forall \ i \tag{2.43}$$

logo

$$w_o = \sum_{i=1}^{N} \alpha_{o,i} y_i x_i b_o = 1 - w_o^t x \tag{2.44}$$

Todas as equações vistas até agora foram para um separador linear, a generalização para o caso não-linear corresponde ao uso de uma superfície multidiemensional para interpolar os dados de teste. A função de base radial gaussiana fornece um conjunto de funções que constituem uma base arbitrária para os vetores de entrada quando eles são expandidos sobre o novo espaço.

A construção de uma função de base radial envolve três camadas. A camada de entrada que é constituida por valores associados a cada elemento. A segunda camada, aplica uma transformação não-linear do espaço de entrada para o espaço oculto. O espaço oculto é de alta dimensionalidade. A saída é linear fornecendo a resposta aos padrões das variáveis de entrada.

O teorema de Cover sobre a separabilidade de padrões diz:

Um problema complexo de classificação de padrões dispostos em um espaço de alta dimensão tem maior probabilidade de ser linearmente separável do que em um espaço de baixa dimensionalidade.

Considere que H represente um conjunto de N padrões (vetores) $X_1, ..., X_N$ cada um dos quais é atribuido a uma de duas classes H_1 e H_2. Para cada padrão $x \in H$, defina um vetor constituído de um conjunto de funções de valor real $\{\varphi(x) \mid i = 1, .., m\}$,como mostrado por

$$\varphi(x) = [\varphi_1(x), \varphi_2(x), \cdots, \varphi_m(x)]^T \tag{2.45}$$

O vetor $\varphi(x)$, então mapeia pontos no espaço de entrada de dimensão p para pontos em um novo espaço de dimensão m.

Uma dicotomia $\{H_1, H_2\}$ de H é dita ser separavel por φ, se existir um vetor w de dimensão m para o qual podemos escrever

$$w^T\varphi(x) \geq 0 \ \forall \ x \in H_1 \tag{2.46}$$

$\mathrm{w}^T\varphi(x) \leq 0 \ \forall \ x \in H_2(2.46)$

O hiperplano definido pela equação

$$w^T\varphi(x) = 0 \tag{2.47}$$

descreve a superfície de separação no espaço φ.

O problema de classificação para um dado conjunto de N pontos diferentes $\{x_i \in \Re^m \mid i = 1, ..., N\}$ e um conjunto correspondente de N número reais $\{y_i \in \Re \mid i = 1, ..., N\}$, encontre uma função $F : \Re^m \rightarrow \Re^1$ que satisfaça a condição $F(x_i) = y_i$.

A técnica de função de base radial gaussiana consiste em escolher uma função F que tem a seguinte forma:

$$F(x) = \sum_{i=1}^{N} w_i\varphi(\|x - x_i\|) \tag{2.48}$$

onde $\{\varphi(\|x-x_i\|) \mid i = 1, ..., N\}$ é um conjunto de N funções gaussianas e $\|.\|$ representa uma norma euclidiana. Os pontos de dados conhecidos $x_i \in \Re^m$, são tomados como centro das funções de base radial.

$$\varphi(r) = exp\Big(\frac{-r^2}{2\sigma^2}\Big) \tag{2.49}$$

Inserindo as condições de interpolação, obtem-se o seguinte conjunto de equações lineares para os coeficientes desconhecidos w_i

$$\begin{bmatrix} \varphi_{11} & \cdots & \varphi_{1N} \\ \vdots & \ddots & \vdots \\ \varphi_{N1} & \cdots & \varphi_{NN} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \tag{2.50}$$

onde

$$\varphi_{ji} = \varphi(\|x_j - x_i\|) \tag{2.51}$$

Os vetores N por 1 y e w representam o vetor resposta e o vetor de coeficientes. Considere que Φ represente uma matriz N por N com elementos φ_{ji}. Esta matriz é chamada de matriz de interpolação. Pode-se reescrever na forma compacta

$$\Phi w = y \tag{2.52}$$

Assumindo que a função radial gaussiana seja usada, a matriz Φ possui inversa, e o vetor w pode ser calculado da seguinte forma

$$w = \Phi^{-1} y \tag{2.53}$$

Retornando a função objetivo para o problema linear, podemos substituir os valores de $x_i^T x_j$ pelo função gaussiana e teremos um classificador não-linear.

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - 1 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.54}$$

onde

$$k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \tag{2.55}$$

sujeita as restrições

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{2.56}$$

$0 \leq \alpha_i \leq C \; forall \; ilogo$

$$w_o = \sum_{i=1}^{N} \alpha_{o,i} y_i \varphi(x_i) \tag{2.57}$$

O número de funções de base radial e seus centros são determinados automaticamente pelo número de vetores de suporte e seus valores. Os parâmetros de σ e C devem ser escolhidos pelo usuário e podem ser avaliados através do método de validação cruzada.

2.7 Aprendizagem baseada em instâncias

Os seres humanos possuem uma grande capacidade de tomar decisões baseadas na sua experiência passada. Decisões rápidas e precisas são feitas seguindo uma simples lógica. O primeiro passo é identificar entre os elementos já observados, aqueles que são os mais similares com o elemento a ser classificado, o próximo passo é inferir qual dentre os resultados dos elementos já observados é o valor mais provável do elemento a ser classificado. Este tipo de resolução de problema assume que o passado possui grande chance de repetir.

A aprendizagem baseada em instâncias, também é conhecida como Raciocínio Baseado em Memória ou K-next-neighbour (knn) é uma metodologia não-paramêtrica. A idéia principal é que no banco de dados encontre-se exemplos já rotulados que sejam próximos do elemento desconhecido e as características dos exemplos sejam inferidas ao elemento. Quanto mais próximos os exemplos do elemento, maiores são as semelhanças sendo que a vizinhança do elemento passa a descreve-lo. A figura 2.7 exemplifica o processo de predição.

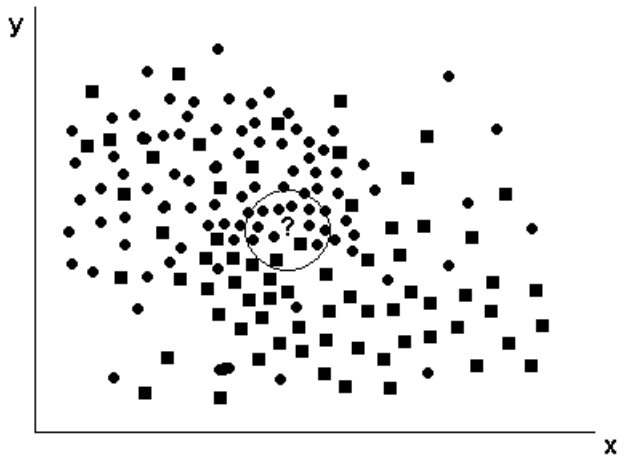


Figura 2.7: Classificação através dos elementos mais próximos

Para que possamos identificar os vizinhos mais próximos de um ponto de consulta, precisamos de uma métrica de distância. O tipo das variáveis utilizadas definem quais são as métricas de distância mais apropriadas.

A distância de Malahanobis é utilizada quando as variáveis são numéricas e leva em consideração a covariância e variância entre as variáveis. Quando os dados são categóricos, as métricas como Jaccard ou Hamming podem ser utilizadas e quando ocorre uma mistura de variáveis categóricas e numéricas a métrica de Gower pode ser utilizada.

Distância de Malahanobis

$$d_{i,j} = [(X_i - X_j)'\Sigma^{-1}(X_i - X_j)]^{\frac{1}{2}} \tag{2.58}$$

onde:

Σ é a matriz de variância-covariância

Distância de Gower

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_{ijk} \cdot s_{ijk}}{\sum_{k=1}^{p} w_{ijk}} \tag{2.59}$$

onde:

s_{ijk} é a similaridade entre o i-ésimo e j-ésimo indivíduo na k-ésima variável; w_{ijk} é 0 ou 1 caso as variáveis sejam comparáveis.

Para variável quantitativa

$$s_{ijk} = 1 - \frac{\mid x_{ik} - x_{jk} \mid}{\text{Amplitude de k}} \tag{2.60}$$

O principal problema deste método é a dependência do número de elementos com o aumento da dimensão dos dados. Para exemplificar: em um conjunto de treinamento com 1.000.000 de casos e 100 variáveis teriamos que selecionar 89% do espaço para encontrar 10 pontos vizinhos. A melhor forma de evitar o problema da dimensionalidade é selecionar as variáveis que possuem o maior poder de predição na variável resposta.

As curvas da figura 2.8 mostram o impacto da dimensionalidade na distância entre os elementos. Suponha um hipercubo de dimensão 10. Para ter 10% dos dados é necessario abranger 80% do volume o hipercubo. Na dimensão 4 é necessário cobrir 56% do volume e para dimensão 3 é necessário cobrir 46%. A dimensionalidade pode rapidamente deteriorar

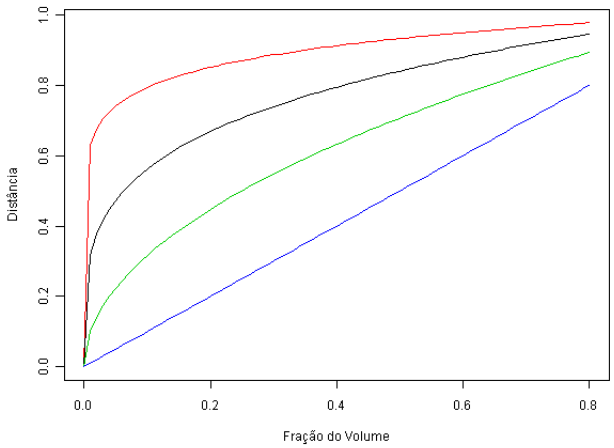


Figura 2.8: Impacto da dimensionalidade

a capacidade de predição deste método, já que os elementos mais próximos não estão assim tão próximos.

O número de vizinhos a serem considerados é uma variável que deve ser avaliada na construção do modelo. Na prática, o uso de um valor de k entre 3 e 10 fornece bons resultados para a maioria dos conjuntos de dados. Normalmente o valor do k é avaliado através da validação cruzada.

O problema da dimensionalidade é evitado selecionando-se as variáveis que diminuem o erro de classificação, percebe-se uma certa consistência na seleção das variáveis.

2.8 Aprendizagem Bayesiana

O raciocínio bayesiano provê um acesso à inferência probabilística. É fundamentado na suposição que as quantidades de interesse são governadas por uma probabilidade de distribuição e que decisões podem ser tomadas pelo raciocínio sobre essas probabilidades em conjuntos com os dados observados.

Durante esta dissertação iremos utilizar o classificador bayesiano ingênuo (naive bayes), que esta entre os mais simples e possui uma grande efetividade. As prinicipais características do método incluem:

- cada exemplo de treinamento observado pode incrementar ou decrementar a probabilidade estimada de que a hipótese esteja correta;
- o conhecimento a priori pode ser combinado com os dados observados e determinar a probabilidade final da hipótese;
- novos itens podem ser classificados pela combinação de predição de múltiplas hipóteses ponderadas pelas suas respectivas probabilidades;
- capacidade de incorporar os dados categôricos com missing na modelagem.

Para definir o teorema de Bayes precisamente, primeiro precisamos introduzir a seguinte notação: pode-se definir $P(h)$ como sendo a probabilidade inicial da hipótese antes dos dados serem observados. Frequentemente $P(h)$ é chamada de priori de h e reflete o conhecimento inicial. Pode-se definir $P(D)$ como sendo a probabilidade de observar o resultado D na base de dados de treinamento. O próximo passo é relacionar estas quantidades com o $P(D \mid h)$ que é a probabilidade de observar os dados de treinamento D dado à ocorrência da hipótese h. De fato o interesse é em $P(h \mid D)$ que é a probabilidade de observarmos a hipótese h dado que os dados D foram obtidos. Esta quantidade é chamada de probabilidade a posteriori de h, porque reflete a confiança na hipótese h, após observarmos os dados D.

O teorema de Bayes é a pedra fundamental porque provê uma forma de calcular a probabilidade a posteriori $P(h \mid D)$, dado a priori $P(h)$ junto com a $P(D \mid h)$ que foi obtida nos dados de treinamento.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{2.61}$$

O método de aprendizagem bayesiano mais simples é o método bayesiano ingênuo. Em alguns domínios sua performance é equiparada a métodos bem mais complexos, como redes neurais e a árvore de classificação.

O classificador bayesiano aplica-se à aprendizagem de tarefas onde cada elemento é descrito como sendo a combinação de variáveis e relaciona estas variáveis a uma classe.

Observe que $P(x_1, x_2, \ldots, x_n)$ pode ser desconsiderado já que é uma constante de normalização independente de $\hat{c}$. Como estamos interessados apenas em avaliar qual é a hipótese mais provável podemos simplificar a equação acima da seguinte forma.

$$h_{max} = P(D \mid h)P(h) \tag{2.62}$$

O classificador bayesiano ingênuo possui a capacidade de trabalhar com mais de uma variável explicativa. Neste caso vamos considerar x_n como sendo um vetor de n variáveis associados a um determinado elemento ou caso. Para simplificar os cálculos o classificador bayesiano ingênuo assume a hipótese que as variáveis explicativas são independentes.

Um fato que merece atenção é o fato de que a equação 2.59 é totalmente baseada nos dados de treinamento. E fácil estimar o $P(Y_i)$ pela simples contagem da frequência que cada classe possui. Note entretanto, que $P(x_1, x_2, \ldots, x_n \mid Y_i)$ é um pouco mais problemático já que o número de elementos deve ser muito maior que o número de variáveis uma vez que a chance de se observar elementos em todas as combinações possíveis é praticamente impossível.

O classificador bayesiano ingênuo assume a suposição simplificadora de que as variáveis explicativas são condicionamente independentes as classes. Assim sendo, dado que observamos um elemento a probabilidade de observarmos a intercessão dos $x_1, x_2, \ldots, x_n$ é o produto das probabilidades de cada variável individual.

$$P(x_1, x_2, \ldots, x_n \mid Y_i) = \prod_{j=1}^{n} P(x_j \mid Y_i) \tag{2.63}$$

substituindo a equação acima pela equação 2.59 temos que

$$\hat{C} = argmax P(Y_i) \prod_{j=1}^{n} P(x_j \mid Y_i) \tag{2.64}$$

onde $\hat{Y}$ é a classe que possui a maior probabilidade de ocorrer conhecendo os valores das variáveis explicativas $x_1, x_2, \ldots, x_n$.

Capítulo 3

Descrição da Base de dados

Antes de passar para aplicação dos métodos de classificação é importante ter uma visão crítica das características dos dados. Este passo permite uma compreensão básica dos dados e das relações entre as variáveis.

A base de dados utilizadas nesta dissertação é pública, podendo ser encontrada no repositório da UCI, http://archive.ics.uci.edu/ml/datasets/Credit+ Approval. A base de dados contém ao todo a classificação de crédito de 690 clientes. Associado a cada cliente existem 15 variáveis. Destas nove variáveis são qualitativas e seis são quantitativas. Os clientes adiemplentes recebem o rótulo de "+" na décima sexta variável, e os inadimplentes "-" . Todas as variáveis estão codificadas de forma a impossibilitar a identificação dos clientes.

Um ponto interessante nesta base de dados é o fato das classes adimplentes e inadimplentes serem equilibradas, o que nem sempre ocorre na análise de crédito. Normalmente o percentual de cliente inadimplentes é pequeno se considerado em relação ao percentual de adimplentes. Grandes diferenças entre as relações de adimplentes/inadimplentes impactam no desempenho dos classificadores. Entre as abordagens mais utilizadas estão a sub-amostragem dos casos mais frequentes ou a sobre-amostragem dos casos menos frequentes [4]. Um exemplo é a base de dados Nursey do repositório UCI, a relação entre as classes é de 1:34. Este grau de desequilíbrio é comum em alguns contextos como: propensão de cancelamento, risco de crédito e diagnóstico de doenças, entre outros. Entre

as abordagens mais utilizadas para lidar com o desequilíbrio de classes estão *tomek link* e SMOTE.

Outro problema comum em classificação de risco de crédito são os diferentes valores dos custos de classificação errada. Rejeitar um cliente adimplente possui um custo menor do que aceitar um cliente inadimplente. Esta diferença de custo pode passar da relação de 5:1, por isso os erros de classificação devem ser avaliados de forma a considerar estas diferenças.

3.1 Análise descritiva univariada

Avaliar as características das variáveis aumenta a capacidade de percepção das interações entre os modelos e os diferentes tipos de variáveis. Portanto, uma análise gráfica auxília não só na construção e validação dos modelos, como também é um importante aspecto na aprendizagem das propriedades e diferenças dos modelos que serão abordados [8].

A tabela 3.1 apresenta as principais estatísticas univariadas das variáveis. A terceira coluna apresenta a amplitude caso a variável seja quantitativa, e se a variável for qualitativa exibe o número de classes da variável.

As diferenças de número de classes entre as variáveis é algo que chama atenção. As variáveis qualitativas V6 e V7 possuem um grande número classes (14, 9). O número excessivo de classes nas variáveis qualitativas também é um obstáculo a mais na estimação dos modelos não só pelo acréscimo de variáveis a serem estimadas, mas também pela associação de classes infrequentes.

A quarta e sétima colunas da tabela mostram os valores de mínimo e máximo das variáveis, caso a variável seja qualitativa é exibido a frequência da menor/maior classe. Entre as variáveis qualitativas, cinco delas apresentaram classes infrequentes ($\simeq$ 1%). Este fato já chama atenção para a possibilidade de agrupar as classes menos frequentes para a simplificação da variável.

Tabela 3.1: Resumo estatístico das variáveis

Nome	Tipo	N° de classes/R	mín.	$\bar{x}$	$\tilde{x}$	máx.	s	missing
V1	Quali	2	31%	–	–	69%	–	12
V4	Quali	3	0,3%	–	–	75,2%	–	6
V5	Quali	3	0,3%	–	–	75,2%	–	6
V6	Quali	14	0,4%	–	–	19,8%	–	9
V7	Quali	9	0,3%	–	–	57,8%	–	9
V9	Quali	2	47,7%	–	–	52,3%	–	–
V10	Quali	2	42,7%	–	–	57,2%	–	–
V12	Quali	2	45,8%	–	–	54,2%	–	12
V13	Quali	3	1,1%	–	–	90,6%	–	–
V16	Quali	2	44,5%	–	–	55,5%	–	–
V2	Quanti	66,5	13,7	31,5	28,4	80,2	12	–
V3	Quanti	28	0	4,7	2,7	28	5	–
V8	Quanti	28,5	0	2,2	1	28	3,3	–
V11	Quanti	67	0	2,4	0	67	4,8	–
V14	Quanti	2000	0	184	160	2000	173	13
V15	Quanti	100000	0	1074	5	100000	5210,1	–

Avaliando a quinta e sexta colunas pode-se perceber pela diferença da média para a mediana que as variáveis quantitativas tem uma tendência assimétrica a direita. As variáveis V3 e V11 possuem coeficientes assimétricos de Pearson acima de 1, ou seja, ela são significativamentes assimétricas. Apesar de não ser visível através da tabela, as variáveis V11, V14 e V15 apresentaram uma excessiva frequência de zeros (57%, 19% e 42%).

A última coluna da tabela 3.1 mostra o número de observações com valores ausentes ou missing. A maioria dos valores ausentes estão nas variáveis qualitativas. A última observação a ser feita é que as variáveis V4 e V5 possuem distribuições idênticas mudando apenas os rótulos das classes.

Para facilitar na compreensão dos dados, serão apresentados através de histogramas as variáveis para que a forma da distribuição dos dados possa ser avaliada de uma forma melhor.

A variável V7 apresenta 10 categorias sendo que a soma das seis classes menos frequentes somam 5%. A análise gráfica das variávies qualitativas mostrou que existem problemas de valores omissos, número excessivo de classes e concentração de casos em algumas classes.

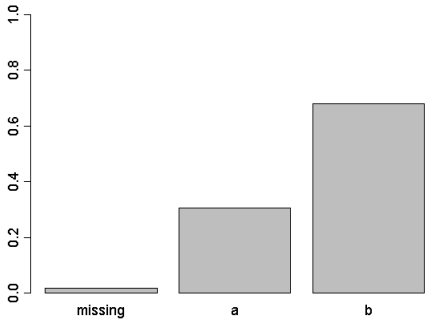


Figura 3.1: Distribuição de V1

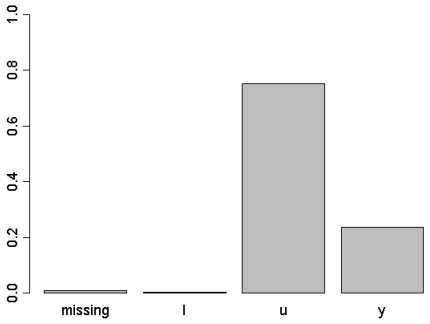


Figura 3.2: Distribuição de V4

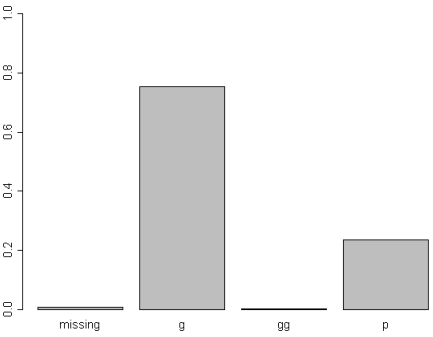


Figura 3.3: Distribuição de V5

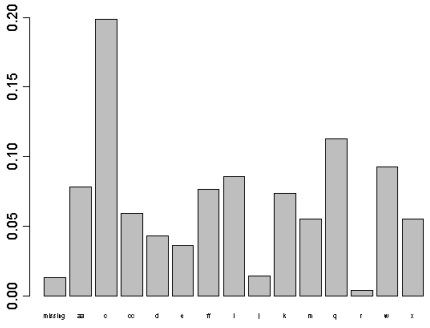


Figura 3.4: Distribuição de V6

Algumas das característica comuns entre as variáveis quantitativas é assimetria a direita e ausência de números negativos. Alguns modelos estatísticos fazem a suposição de que as variáveis tenham distribuição normal. Quando se depara com distribuições assimétricas e deseja-se utilizar modelos que possuem a suposição de normalidade os dados

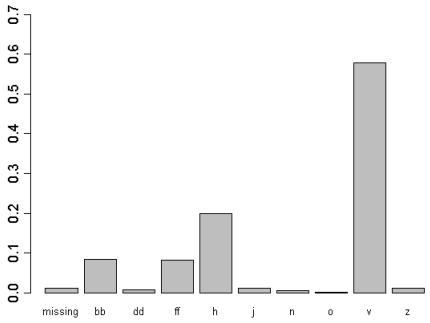


Figura 3.5: Distribuição de V7

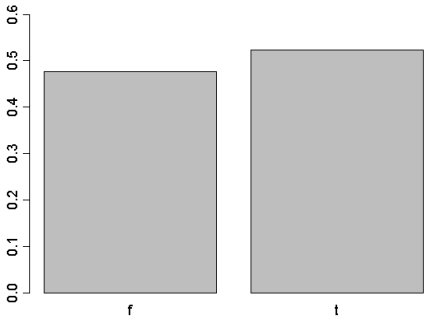


Figura 3.6: Distribuição de V9

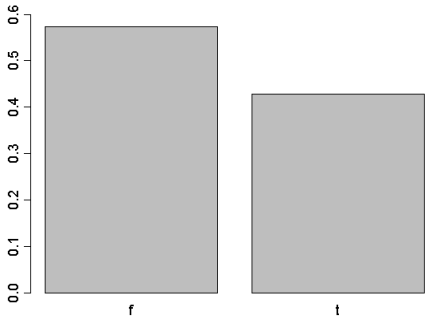


Figura 3.7: Distribuição de V10

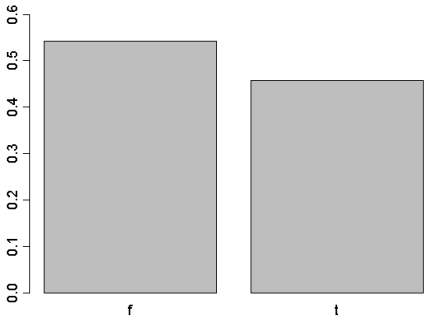


Figura 3.8: Distribuição de V12

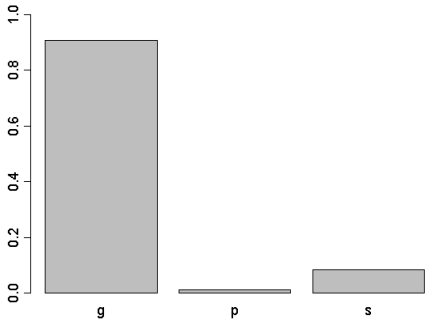


Figura 3.9: Distribuição de V13

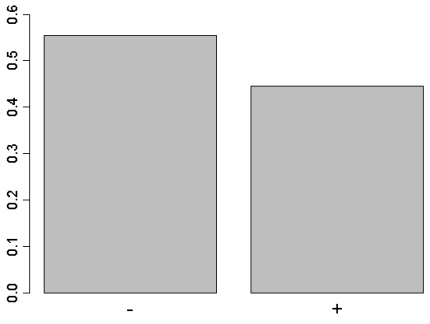


Figura 3.10: Distribuição de V16

devem ser transformados para que a distribuição fique mais simétrica. As transformações mais comuns são ln, raiz quadrada, inverso e transformação de Box-Cox [16].

Analisando as variáveis quantitativas, percebe-se que pelo menos um dos problemas: assimetria, excesso de zeros e/ou dados omissos estão presentes em todas as variáveis.

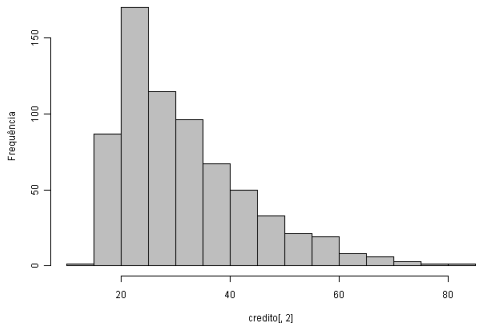


Figura 3.11: Distribuição de V2

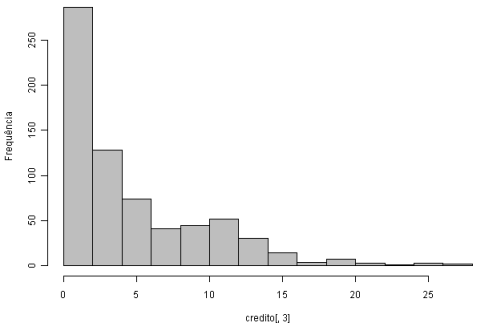


Figura 3.12: Distribuição de V3

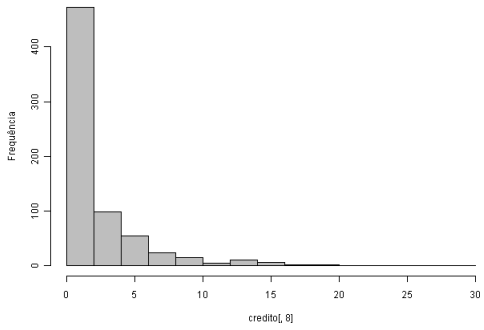


Figura 3.13: Distribuição de V8

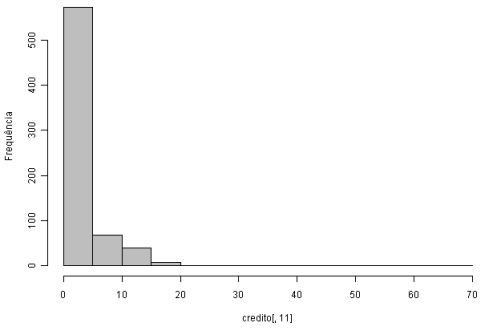


Figura 3.14: Distribuição de V11

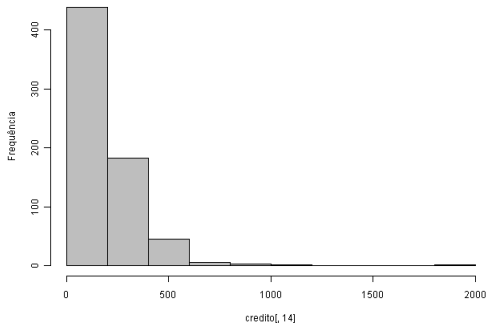


Figura 3.15: Distribuição de V14

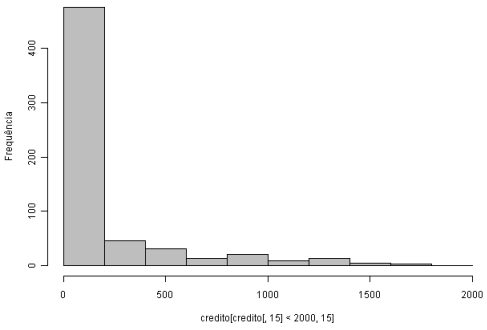


Figura 3.16: Distribuição de V15

Estas características exigem modelos estatísticos mais flexíveis. Através do conhecimento das características das variáveis é possível avaliar qual o grau de afastamento do dados às hipóteses que os modelos possuem, permitindo uma melhor escolha dos métodos a serem utilizados.

3.2 Análise descritiva bivariada e multivariada

Para comparar as diferenças entre as variáveis qualitativas entre os grupos adimplentes e inadimplentes vamos utilizar o gráfico de barras para visualizar as diferenças de proporção entre as classes.

A variável V9, figura 3.21, possui diferenças nas classes entre os grupos de clientes adimplentes e inadimplentes. A ocorrência da classe f é uma evidência a favor do grupo inadimplente enquanto a classe t é o contrário. A variável V10, figura 3.22, também possui diferenças nas classes entre os clientes.

Para avaliar as diferenças entre as variáveis quantitativas pode-se utilizar o gráfico de Boxplot. Os riscos no centro da caixa representam a mediana, os limites inferiores e superiores representam o 1° e 3° quartil respectivamente.

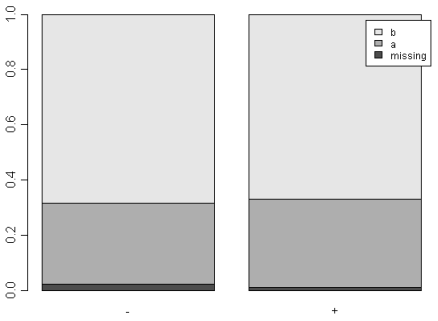


Figura 3.17: Variável V1

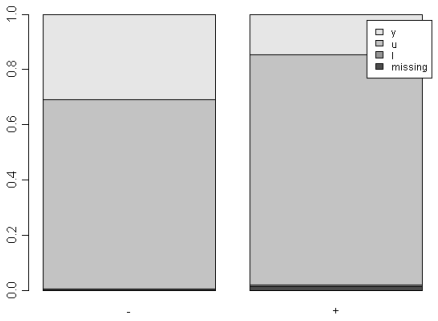


Figura 3.18: Variável V4

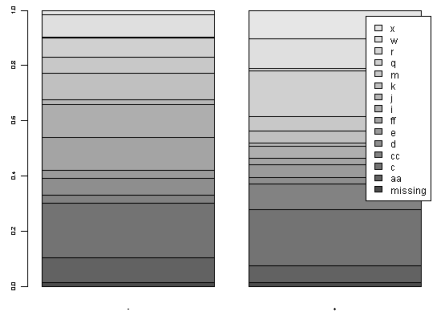


Figura 3.19: Variável V6

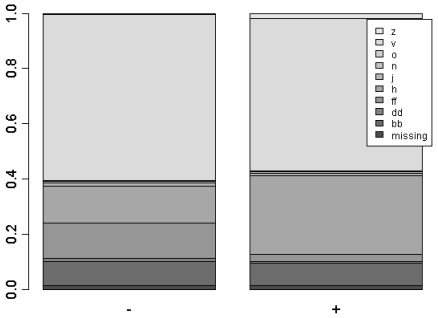


Figura 3.20: Variável V7

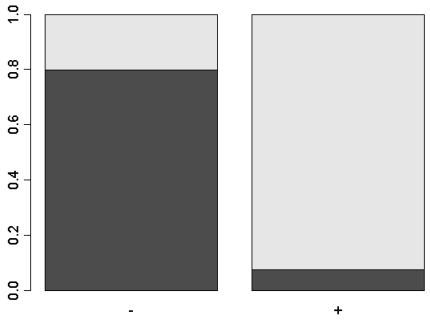


Figura 3.21: Variável V9

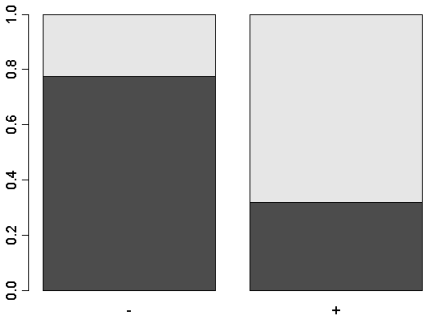


Figura 3.22: Variável V10

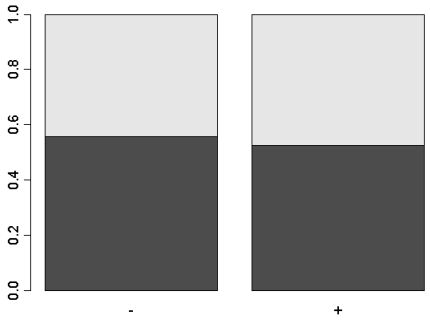


Figura 3.23: Variável V12

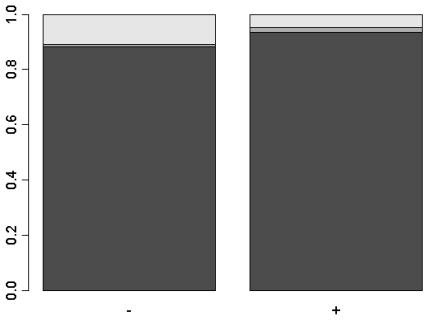


Figura 3.24: Variável V13

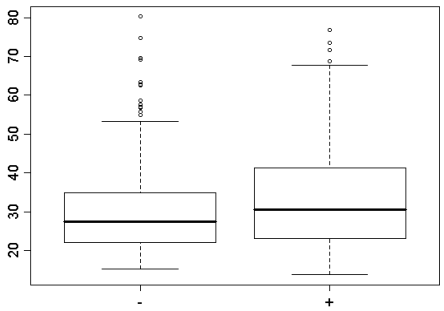


Figura 3.25: Variável V2

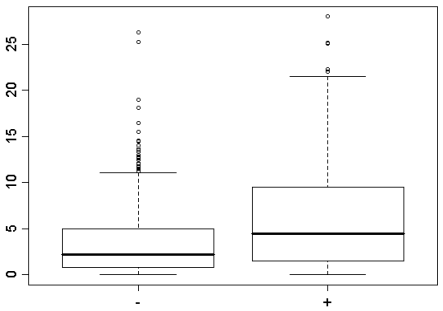


Figura 3.26: Variável V3

Ambas as variáveis V8 e V11 são as que visualmente mais discriminam o risco de crédito. A variável V8 possui uma assimetria que esta presente em ambos os grupos. Observa-se que o 3° quartil do grupo inadimplente está próximo do valor do 1° quartil do grupo adimplente. Em ambos os grupos ocorreram valores extremos. A variável V11

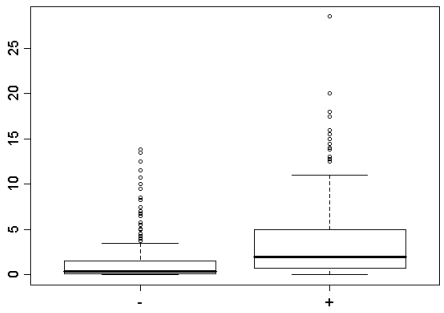


Figura 3.27: Variável V8

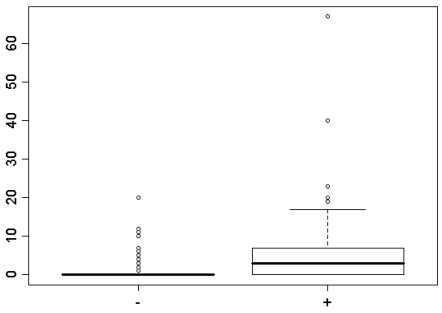


Figura 3.28: Variável V11

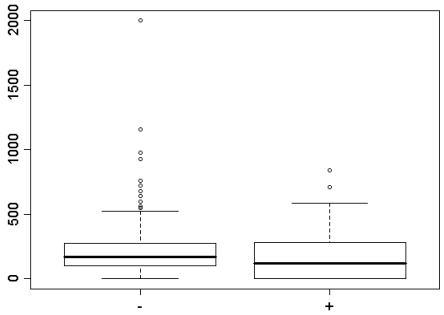


Figura 3.29: Variável V14

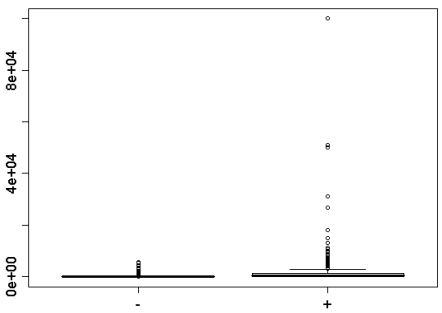


Figura 3.30: Variável V15

parece que é a que possui maior poder discriminatório do risco de crédito. Os casos que tiverem valor acima de zero provavelmente são adimplentes. Os modelos deverão dar maior importância à variável V11 já que pela análise gráfica é a variável que individualmente possui uma alta associação entre os grupos de risco de crédito.

A variável V14 possui valores próximos em ambos os perfis de risco de crédito, entretanto a dispersão entre os clientes adimplentes é maior. A variável V15 possui um intervalo de variação muito grande para o nível de dispersão que possui, por isso vemos apenas pontos extremos. Uma transformação nesta variável como $ln(x)$ pode ajudar na visualização de padrões.

Avaliando as variáveis qualitativas as variáveis V9 e V10 são as que mais discriminam o risco de crédito já nas variáveis quantitativas percebe-se que a variável V8 e V11 são as que aparentemente mais discriminam os grupos. Os clientes adimplentes possuem

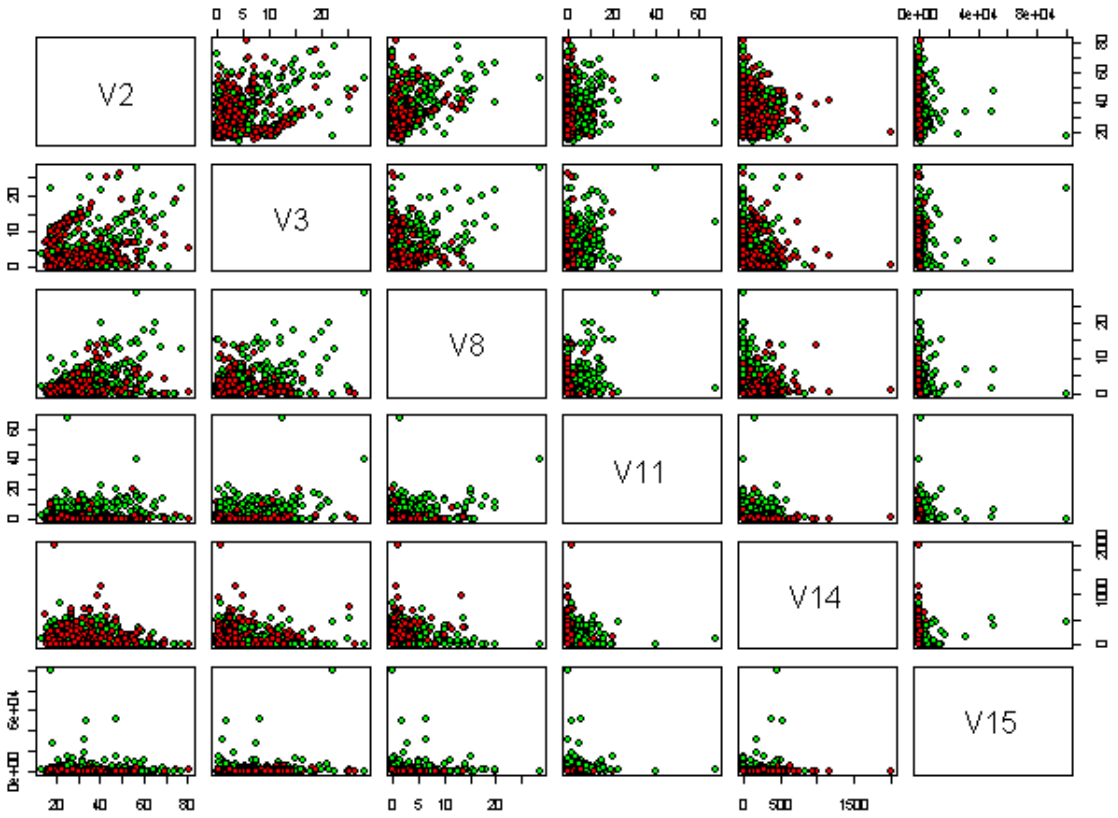


Figura 3.31: Relacionamento entre as variáveis quantitativas

valores bem acima dos clientes inadimplentes. Já as variáves V2,V3 e V14 parecem não diferenciar tanto os grupos, e a variável V15 possui tanta assimetria que a visualização da diferença entre os grupos fica difícil de ser feita. Outra constatação importante é que em praticamente todas as variáveis ocorreram valores discrepantes.

Avaliando a figura 3.31, a dispersão entre as variáveis quantitativas mostra que as associações são fracas e a forte assimetria é uma caracteristica em todas as variáveis.

3.3 Dados Omissos

Na base de dados, aproximadamente 5% dos clientes possui pelo menos um valor omisso. Esta omissão de valores é normalmente chamada de missing e é um fato comum na maioria das aplicações estatísticas multivariadas. Avaliar a forma como estes valores omissos

ocorreram é importante para considerar a validade da amostra.

Um processo de dados perdidos é um evento externo ao respondente. Alguns processos de dados perdidos são conhecidos e podem ser acomodados na modelagem (exemplo censura em análise de sobrevivência). Como não se sabe qual foi o processo de perda, uma avaliação dos elementos com missing foi realizada buscando identificar o padrão dos dados perdidos.

Para amostras univariadas, quando os elementos contendo valores ausentes são um subconjunto aleatório dos elementos da amostra original, então este é um mecanismo ignorável de valores ausentes. Já quando a probabilidade de uma variável Y estar ausente depende do próprio valor de Y, diz-se que este é um mecanismo não ignorável, ou seja, o simples desprezo dos objetos com valores ausentes pode gerar resultados tendenciosos.

Em [5] foi introduzida uma nomeclatura que até hoje é aceita em amostras multivariadas, nas quais apenas uma variável (Y) possui valores ausentes. Considerando-se duas variáveis X e Y da amostra, esta nomeclatura define três padrões de mecanismos de valores ausentes:

- a probabilidade do valor de Y ser ausente independe do valor de X e de Y também. Neste caso o mecanismo é chamado de não-ignorável (NI);

- a probabilidade do valor de Y ser ausente independe do valor de X, mas não do valor de Y. Neste caso, diz-se que o mecanismo é *Missing at random*, ou seja, perda aleatória.

- a probabilidade de Y ser ausente independe do valor de X e do valor de Y. Neste caso os valores ausentes são completamente aleatórios *MCAR - Missing completly at random*

Para se buscar um parâmetro na definição se um mecanismo de valores ausentes é MCAR ou não, pode-se realizar um estudo comparativo com os casos completos e os casos incompletos. Neste estudo pode-se comparar as distribuições das duas amostras através de um teste t ou um teste não-paramêtrico.

A primeira verificação é identificar se a distribuição dos valores omissos é ao acaso entre as observações ou se existe padrões distintos. As variáveis V1, V2, V4, V5, V6, V7 e V14 possuiram dados perdidos, ao todo são 49 valores omissos ou seja 0,5% do total de valores. Estes valores estão distribuidos em 37 casos, o que corresponde a 5% da amostra.

Se os valores não observados da variável são verdadeiramente uma amostra aleatória de todos os valores, sem qualquer processo inerente que conduza a tendências para os dados observados, tem-se um processo de dados perdidos completamente ao acaso (MCAR missing completely at random).

Se essa é a forma do processo de dados perdidos, quaisquer atitudes corretivas podem ser aplicadas sem levar em consideração o impacto de qualquer outra variável ou do processo de dados perdidos. Realizou-se um teste geral de aleatoriedade para determinar se os dados perdidos são MCAR. As variáveis V1, V12 e V14 tiveram valores não significativos. Já nas variáveis V4, V5, V6 e V7 os dados não são perdidos de forma aleatória.

O tratamento mais simples e direto para lidar com dados perdidos é incluir somente as observações com dados completos. A abordagem de caso completo deve ser usada somente se os dados perdidos são MCAR, caso contrário os resultados não são generalizavéis à população.

As conclusões sobre dados omissos é que os elementos que tiveram valores omissos na variável V2 e V4 são descartados, e os dados omissos nas variáveis V4, V5, V6 e V7 são marcados como missing. Os elementos com valores omissos nas variáveis V1, V12 e V14 são MCAR.

Capítulo 4

Resultados dos Classificadores

Neste capítulo os diferentes classificadores são comparados e os pontos negativos e positivos são avaliados. As principais métricas selecionadas para avaliação do desempenho são: erro de classificação, sensibilidade, especificidade e tempo gasto na estimação e predição dos modelos.

4.1 Árvore de classificação e Boosting

Comparando os métodos CART e Boosting o resultado obtido pelo método Boosting foi superior, o erro foi menor que o erro do método CART em 3 classificações cruzadas, e o erro máximo de classifacação utilizando boosting foi 7% menor do que o erro do método CART. O tempo gasto na predição é aproximadamente 5 vezes menor do que o tempo de construção do algoritmo. Já o método de boosting exige muito mais tempo para predizer e construir, na amostra o tempo de predição é menor que o dobro do tempo de construção. Para estimar a variabilidade das predições, utilizou-se o método de validação cruzada com 5 amostras de validação (CV).

Tabela 4.1: Comparação entre os erros de classificação do Adaboostin e CART

	CV1	CV2	CV3	CV4	CV5	Média
Erro de classificação CART	11.6%	18.8%	12.3%	15.2%	11.6%	13.9%
Sensibilidade CART	79.4%	87.7%	90.5%	92.3%	85.3%	87.4%
Especificidade CART	89.8%	80.3%	90.2%	85.2%	83.1%	85.7%
Tempo de construção	0.28s	0.27s	0.28s	0.29s	0.95s	0.41s
Tempo de predição	0.05s	0.06s	0.07s	0.07s	0.05s	0.07s
Erro de classificação Boos.	15.9%	17.4%	10.1%	11.6%	15.2%	14.0%
Sensibilidade Boos.	79.2%	88.3%	89.5%	91.6%	85.7%	86.9%
Especificidade Boos.	90.2%	78.2%	90.4%	83.6%	83.6%	87.2%
Tempo de construção	3.64s	3.33s	3.12s	3.25s	4.27s	3.52s
Tempo de predição	2.69s	2.51s	2.50s	2.54s	4.25s	2.90s

4.2 Regressão

As tabelas 4.2 e 4.3 mostram os resultados obtidos pelos métodos de regressão linear e regressão logística. No capítulo 2 apresentamos as vantagens do método de regressão logística e os violações de hipóteses que o modelo linear de regressão apresenta. Apenas para exemplificar mostraremos os resultados do modelo linear para que possa ser comparado com a regressão logística.

Tabela 4.2: Avaliação do modelo de regressão linear

	CV1	CV2	CV3	CV4	CV5	Média
Erro de Classificação	15.9%	15.0%	8.9%	12.1%	16.4%	13.7%
Sensibilidade	74.0%	77.2%	89.3%	83.3%	76.0%	80.0%
Especificidade	96.6%	90.8%	94.1%	94.4%	93.2%	93.8%
Tempo de construção	0.04s	0.07s	0.04s	0.04s	0.02s	4.2s
Tempo de predição	0.01s	0.01s	0.01s	0.01s	0.01s	0.01s

Tabela 4.3: Avaliação do modelo de regressão logística

	CV1	CV2	CV3	CV4	CV5	Média
N° de váriaveis	9	8	11	10	11	9.8
Erro de Classificação	15.1%	10.0%	12.6%	17.4%	13.4%	13.7%
Sensibilidade	80.8%	91.98%	85.7%	88.5%	84.0%	86.2%
Especificidade	89.8%	87.8%	90.2%	74.1%	89.8%	86.4%
Tempo de construção	0.67s	0.19s	0.46s	0.23s	0.19s	0.35s
Tempo de predição	0.11s	0.12s	0.13s	0.13s	0.07s	0.11s

O erro médio de classificação das regressões linear e logística foram equivalentes. Entretanto, avaliando a dispersão do erro médio, observamos que o desvios padrões da regressão logística é menor (2.76) do que o desvio padrão do erro médio da regressão linear (3.13). Utilizando a metodologia de *stepwise* o tempo gasto na construção da função de regressão foi pequeno e o tempo gasto na predição é de aproximadamente um terço do tempo gasto na construção no caso da regressão logística. O método de regressão logística nesta amostra foi superior à regressão linear.

4.3 Máquina de vetor suporte

Dentre os classificadores que foram avaliados, a máquina de vetor suporte foi o método que apresentou o menor valor máximo de erro de classificação 15.9%: os tempo gastos na construção e predição do modelo foram pequenos e o tempo de predição é cerca da metade do tempo gasto na construção, os dados podem ser observados na tabela 4.4. O erro médio de classificação do kernel gaussiano foi de 13.96%.

Tabela 4.4: Avaliação do modelo máquina de vetor de suporte kernel radial

	CV1	CV2	CV3	CV4	CV5	Média
N° de vetores suporte	236	231	256	234	243	240
Erro de Classificação	15.9%	15.7%	10.4%	12.9%	14.9%	13.9%
Sensibilidade	80.8%	84.2%	86.9%	85.9%	80.0%	83.6%
Especificidade	88.1%	84.2%	94.1%	88.8%	91.5%	89.3%
Tempo de construção	0.09s	0.11s	0.13s	0.10s	0.11s	0.11s
Tempo de predição	0.08s	0.08s	0.06s	0.07s	0.07s	0.07s

4.4 Aprendizagem Baseada em Instâncias

Um dos pontos que pode-se observar neste método, é que nas amostras de validação que utilizaram apenas o vizinho mais próximo para estimar os elementos tiveram o melhor e pior valores de erro médio de classificação.

Tabela 4.5: Comparação entre os erros de classificação do KNN

	CV1	CV2	CV3	CV4	CV5	Média
N° de vizinhos	9	3	1	1	9	4.6
N° das variáveis	(3,9,11,15)	(9,11)	(3,8,9,11,15)	(3,9,11,15)	(3,9,11,15)	-
Erro de clas.	12.87%	18.78%	8.88%	20.45%	12.68%	14.7%
Erro de treinamento	11.42%	6.9%	8.85%	0.18%	12.33%	7.9%
Sensibilidade	93.2%	73.6%	90.2%	77.7%	89.8%	84.9%
Especificidade	82.1%	91.2%	91.6%	80.7%	85.3%	86.2%
Tempo de construção	0.12s	0.12s	0.12s	0.12s	0.10s	0.11s
Tempo de predição	0.13s	0.09s	0.09s	0.10s	0.11s	0.10s

O erro de classificação médio foi igual a 14.75% ou seja o maior valor se comparado aos outros métodos utilizados. Os números de vizinhos considerados apresentou grande

variação. Apesar do tempo gasto na construção ser pequeno, quase não houve diferença entre o tempo gasto na predição e classificação.

4.5 Aprendizagem Bayesiana

Tabela 4.6: Comparação entre os erros de classificação do Naive Bayes

	CV1	CV2	CV3	CV4	CV5	Média
Erro de Classificação	10.86%	19.56%	9.42%	13.04%	12.31%	13.0%
Erro de Treinamento	13.04%	11.7%	14.31%	13.4%	13.58%	13.2%
Sensibilidade	89.6%	88.3%	94.2%	90.4%	90.9%	90.6%
Especificidade	88.5%	74.3%	84.6%	81.8%	83.6%	82.6%
Tempo de construção	0.17s	0.13s	0.12s	0.20s	0.10s	0.14s
Tempo de predição	0.07s	0.06s	0.07s	0.05s	0.05s	0.06s

O erro médio de classificação do método foi de 13.03%. Avaliando o desempenho com as complexidades e hipóteses feitas pelo modelo pode-se dizer que é um modelo interessante para aplicações onde não se tem muito tempo para validação de hipóteses. Os tempos gastos na construção e predição foram pequenos e não apresentaram grandes variações.

Comparando os resultados do método bayesiano com o método de regressão em árvore vemos que os resultados ficaram muito próximos, entretanto é interessante notar que os valores da sensibilidade foram maiores do que o da especificidade.

Capítulo 5

Agregação dos Classificadores

Uma estratégia comum em problemas computacionalmente complexos, é dividir o problema em problemas mais simples e resolver os pequenos problemas que formam o problema inicial. Thomas [15] propõem utilizar vários classificadores diferentes na tentativa de minimização do erro. A fusão das opiniões permite obter uma resposta superior a cada classificador individualmente. Cada classificador utilizado possui um conjunto de hipóteses e têm uma forma diferente de construir as fronteiras de classificação. Quanto menor for a correlação entre os classificadores escolhidos maiores serão os ganhos obtidos.

A idéia de reunir um comitê remonta a Nilson[6]; após os classificadores obterem as suas previsões uma votação era feita definindo a previsão do elemento. Desde então esta abordagem também é conhecida como máquina de comitê, e elas podem ser classificadas em duas grandes categorias:

- Estruturas estáticas. As respostas de vários classificadores são combinados por meio de um mecanismo que não envolve o sinal de entrada. Ou seja uma votação é feita, e com base no resultado temos a previsão para o elemento em questão. A votação pode ser feita de forma simples ou ponderada. Um refinamento na votação simples pode ser alcançada se a estimativa da probabilidade de um elemento pertencer a uma classe for utilizada. Pode-se combinar a probabilidade da classe de todas os classificadores, resultando na probabilidade média do elemento pertencer à determinada classe.

- Estruturas dinâmicas. O sinal de entrada está envolvido no mecanismo que integra a saída dos classificadores individuais. Esta categoria é mais conhecida como mistura de especialistas.

Apesar da combinação dos resultados dos classificadores através de uma votação simples parecer uma metodologia simples, os resultados alcançados na votação são robustos.

Normalmente um modelo com alta sensibilidade não possui alta especificidade, as duas características são importantes na seleção do modelo. Vendo as figuras 5.1 a 5.5 vemos que os modelos não tiveram um resultado consistente em todas as amostras de validação. Graficamente pode-se perceber a inexistência de um classificador dominante, ou seja, um método de classificação com valores elevados na sensibilidade e especificidade em todas as amostras de validação.

Após combinarmos os resultados dos classificadores através de uma votação, obtivemos os seguintes resultados: o erro de classificação médio foi de 12.55% com uma sensibilidade de 87.5% e especificidade de 91.7%. Isto significa uma melhora de 4.2% sobre o melhor modelo individualmente (Regressão Logística). Na figura 5.6 podemos observar o erro médio de classificação de cada uma das amostra de validação. A média dos erros é representado pelo sinal de "+".

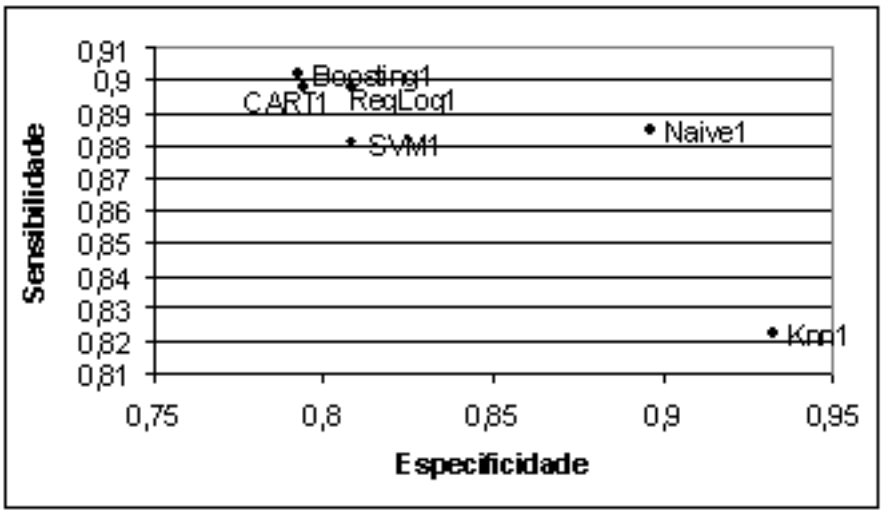


Figura 5.1: Sensi. x Especi. CV1

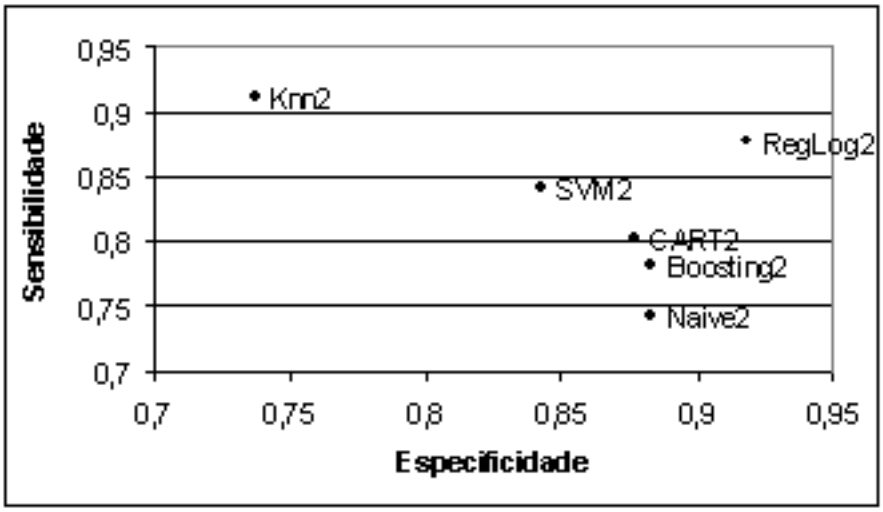


Figura 5.2: Sensi. x Especi. CV2

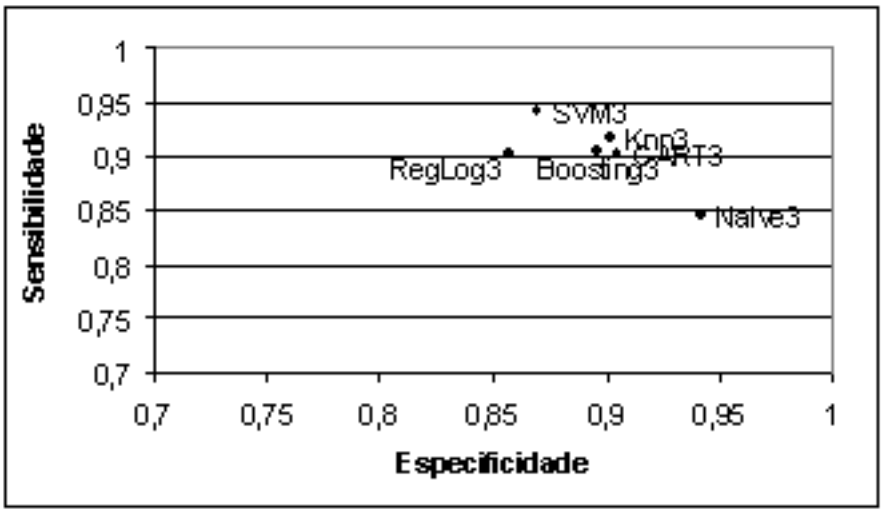


Figura 5.3: Sensi. x Especi. CV3

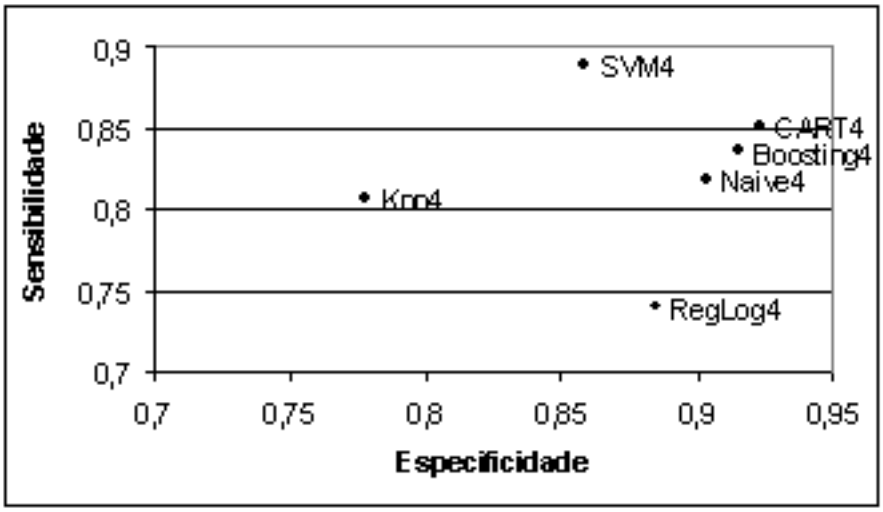


Figura 5.4: Sensi. x Especi. CV4

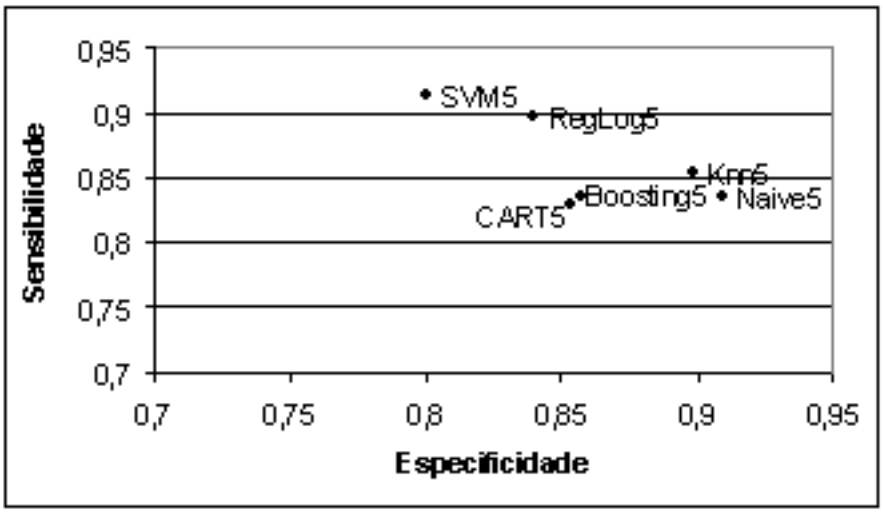


Figura 5.5: Sensibilidade vs Especificidade CV5

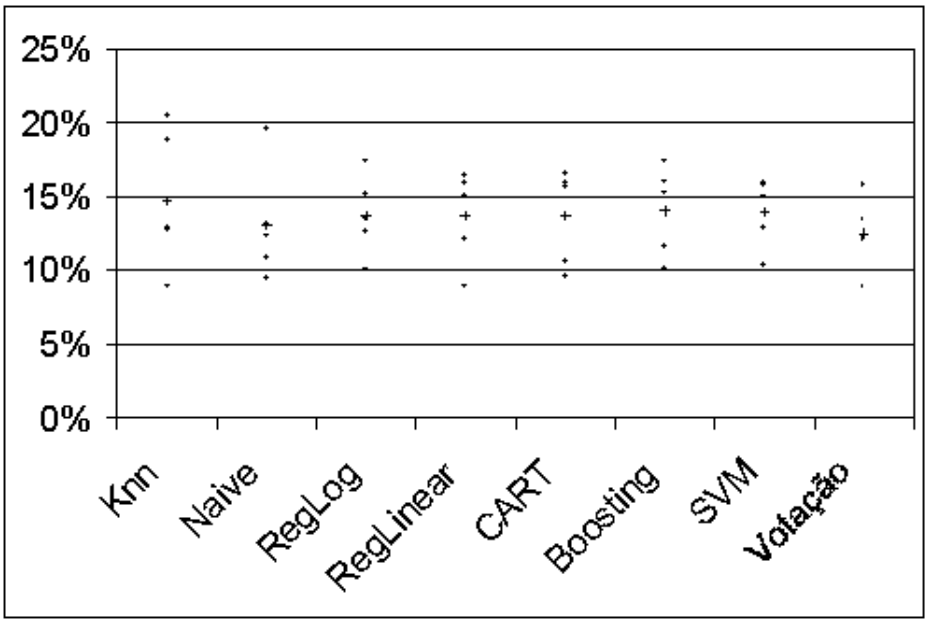


Figura 5.6: Comparando o Erro entre os modelos

Capítulo 6

Conclusão

Os modelos não apresentaram uma característica dominante em relação ao erro de classificação. Diferente do esperado, novas abordagens como boosting e máquina de vetor suporte não tiveram melhora nos resultaos tão expressivas apesar do grande acréscimo de complexidade.

A regressão lógistica e o classificado bayesiano poderiam ser considerados como os que tiveram melhor desempenho geral se considerar a complexidade e a precisão. Um ponto que chama atenção é que apesar da simplicidade assumida na hipótese de que as variáveis são independentes não obteve grande impacto nos resultados do classificador bayesiano.

Os métodos de Boosting e máquina de vetor de suporte utilizam a mesma estratégia de considerar alguns pontos mais importantes do que outros. Outra estratégia encontrada nas máquinas de vetor de suporte é a expanção do vetor de variáveis para um espaço de maior dimensionalidade.

A votação do rótulo do cliente através da combinação dos classificadores foi a que resultou nos melhores resultados. Os resultados melhoraram quase 5% ao utilizar um grupo de classificadores no lugar de um único classificador. Mesmo que 5% não seja uma melhora estatisticamente significativa com base na amostra, no aspecto financeiro já passa a ser um resultado significativo.

Referências Bibliográficas

[1] Jack Dongarra Al Geist, Adam Beguelin. *Parallel Virtual Machine; A User's Guide and Tutorial for Networked Parallel Computing*. MIT Press, 3 ed. edition, 1994.

[2] Héctor Gustavo Arango. *Bioestatística teórica e computacional*. Ed. Guanabara Koogan, 2 ed. edition, 2005.

[3] Alexandre Jorge Chaia. *Modelos de Gestão do Risco de Crédito e sua Aplicabilidade ao Mercado Brasileiro*. USP, 1 ed. edition, 2003.

[4] Nitesh V. Chawla. *SMOTE: Syntetic Minority Over-sampling Techinique*. Journal Of Artificial Intelligence Research, 16 edition, 2002.

[5] D.B.Rubim. *Multiple Imputation for Nonresponse in Surveys*. Ed. Wiley Series, 1 ed. edition, 1987.

[6] Simon HAYKIN. *Rede neurais: princípios e prática*. Bookman, 2 ed. edition, 2001.

[7] Robet Tibshirani Jerome Friedman and Trevor Hastie. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 1 ed. edition, 2001.

[8] R.L.Tatham W.C.Black J.F.Hair Jr., R.E.Anderson. *Análise Multivariada de Dados*. Bookman, 5 ed. edition, 2005.

[9] Gordon Linoff Michael J.A.Berry. *Data Mining Techiniques for marketing, sale and customer support*. John Wiley, 1 ed. edition, 1997.

[10] Sueli Aparecida Mingoti. *Análise de dados através de métodos de estatística multivariada*. Ed. UFMG, 1 ed. edition, 2007.

[11] Tom M.Mitchell. *Machine Learning*. McGraw HIll, 1 ed. edition, 1997.

[12] Peter S. Pacheco. *Parallel Progamming with MPI*. Morgan Kaufmann, 1 ed. edition, 1997.

[13] Marcos Pitanga. *Computação em Cluster*. Brasport, 1 ed. edition, 2003.

[14] Peter Norving Stuart Russel. *Artificial Intelligence; A Modern Approach*. Prentice Hall, 1 ed. edition, 1995.

[15] David B. Edelman Thomas C. LYN and Jonathan N. Crook. *Credit Scoring and Its Applications*. SIAM, 1 ed. edition, 2002.

[16] Dean W. Wichern and Richard A. Johnson. *Applied Multivariate Statistical Analysis*. Prentice Hall, 4 ed. edition, 1998.

[17] Ewing Lusk William Gropp and Thomas Sterling. *Beowulf Cluster Computing with Linux*. MIT Press, 2 ed. edition, 2003.