MÉTODO DE POTENCIAL PARA A CLASSIFICAÇÃO SUPERVISIONADA

VALMIR ROGERIO DA SILVA

Orientador: Prof. Dr. Borko Stosics

Área de Concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau
de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, fevereiro de 2008

Silva,Valmir Rogerio da
Método de potencial para a classificação supervisionada / Valmir Rogerio da Silva. – Recife: O Autor, 2008.
xiv, 66 folhas : il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CCEN. Departamento de Estatística, 2008.

Inclui bibliografia e apêndices.

1. Análise discriminante. 2. Método de potencial. 3. Método de núcleo. 4. Reconhecimento de padrão. I. Título.

519.535 CDD (22.ed.) MEI2008-027

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

27 de fevereiro de 2008
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Valmir Rogerio da Silva

intitulada

"Método de potencial para classificação supervisionada"

seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.

Klaus Leite Pinto Vasconcellos
Coordenador da Pós-Graduação em Estatística

Banca Examinadora:

Bob D. Stosic
Borko Darko Stosic orientador

Barroso
Lúcia Pereira Barroso (USP)

Manoel Raimundo Sena Junior

Este documento será anexado à versão final da dissertação.

Este trabalho é carinhosamente dedicado

a Valdeci (*in memoriam*) e Marcos (*in memoriam*).

Agradecimentos

À Natureza.

Aos meus pais, Severina e Valdeci (*in memoriam*), pelo amor incondicional, carinho e significado da minha existência, e ao companheiro de mamãe, Antônio, pelo carinho.

Aos meus irmãos, Almir, Antônio Filho, Rita de Cássia, Alessandra, pelo amor incondicional, carinho, confiança e amizade, e à minha sobrinha, Antonelly, pela alegria e momentos de criança.

Ao meu primo Marcos (*in memoriam*), pelos momentos inesquecíveis que compartilhamos nesta vida. E também aos meus primos, Lucas, Emanuel, Eliaquim, Rafael, Rodrigo, Macioneide, Márcia, Amanda, Luana, Roberto.

Aos meus avós, Auxiliadora e Eduardo (*in memoriam*), Anísia (*in memoriam*), pelo amor, carinho e incentivo.

Às minhas tias, Maria, Ana, Marlene, pelo carinho e apoio, e ao meu tio Erivan.

Ao professor Borko Stosics, pela orientação, atenção, paciência, confiança e amizade.

Aos professores da UFPE, especialmente a Maria Cristina e Claúdia Regina, pelo carinho, atenção e orientações para a vida; Andrei Toom, pela atenção e amizade; Getúlio José Amorim do Amaral, pelas sugestões e atenção e Manuel Sena, pela atenção e incentivo.

Aos meus amigos da UFPE, Abraão, Andréa, Andrea Prudente, Alessandra, Ângela, Edwin, Emília, Ênio, Esdras, Hemílio, Jane, Josemir, Juliana, Lídia, Líliam, Lilian, Luis Henrique, Manoel, Marcelo, Olga, Raphael, Rejane, Silvia, Wilton, pela amizade, carinho e pelos momentos de alegria compartilhados.

Aos meus amigos, Damião (Man), Josivan, Gleidson, Leonardo, Charles, Elizandro, Valdemir, Elson, Ewerton (Java), pela amizade, carinho e companheirismo.

Aos companheiros do calabouço (CEU-M/UFPE), Alexandre e Bruno, pela harmonia e amizade.

A Cândido, Maurício, Cícero, Antônio, Volnylson, pela amizade e atenção.

Às bibliotecárias, Raquel, Jane, Mercês, Joana, e ao bolsista Maurício, pela atenção e excelente serviço prestado.

À Valéria Bittencourt, pelo enorme carinho, paciência e amizade com que sempre tratou a mim e aos demais alunos do mestrado.

À banca examinadora, pelas valiosas críticas e sugestões.

À CAPES, pelo apoio financeiro.

Resumo

A classificação supervisionada representa uma parte das técnicas empregadas no contexto de Data Mining, com crescente impacto nos estudos em várias áreas do conhecimento, possibilitado pelo crescimento exponencial da capacidade de processamento e disponibilidade de recursos computacionais. Além do fato que já existem várias técnicas bem conhecidas e estabelecidas na literatura, a importância desta área exige esforços contínuos no sentido da comparação de performance de diferentes métodos, e da sua aplicabilidade aos diferentes tipos de dados, bem como propostas de novas metodologias que poderão contribuir para o estado da arte atual.

Esta dissertação apresenta um novo método de classificação, o método de potencial. Este método é construído com base em conceitos da física, através do mapeamento de observações no espaço p-dimensional dos dados para o sistema virtual de partículas interagentes no espaço Euclidiano p-dimensional. O método é formalizado com todos os detalhes necessários para a definição da regra de classificação com base na teoria de decisão de Bayes. As características mais relevantes do método também são apresentadas.

O método de núcleo é utilizado para comparação com o método de potencial por apresentar boas propriedades e ser bastante difundido e estudado no meio acadêmico. Os dois métodos se diferenciam basicamente pela forma funcional com que estimam as densidades que são utilizadas para se construir a regra de classificação. Os classificadores propostos pelos métodos são então avaliados com respeito a discriminabilidade da regra de decisão para dados reais e simulados, respectivamente, através das técnicas bootstrap e holdout. O estudo atual, além de ser longe de ser completo, mostra que o desempenho dos métodos é semelhante na maioria dos casos, mas que por outro lado existem situações quando cada um deles tem vantagens em relação ao outro.

Palavras-chave: Classificação; Análise Discriminante de Potencial; Análise Discriminante de Núcleo; Bootstrap.

Abstract

Supervised classification represents a part of the techniques employed for Data Mining, with an increasing impact on studies in a variety of areas of knowledge, made possible by exponential growth of processing capacity and accessibility of computational resources. Besides the fact that various techniques already exist that are well known and established in the literature, importance of this area requires continuous efforts in the direction of comparison of performance of different methods, their applicability do different types of data, as well as search for new methodologies that may contribute to the state of the art.

This dissertation presents a new method of classification, termed the method of potential. This method is constructed on the basis of concepts of physics, through mapping of observations in p-dimensional space to a virtual system of interacting particles in a p-dimensional Euclidean space. The formalism of the method is presented, with details necessary for the definition of classification rule on the basis of Bayes decision theory.

The method of potential is compared with the kernel method, as the latter presents good properties, and is widely used by the research community. The basic difference between the two methods is the functional form used to estimate the probability density, which is used to construct the classification rule. The classifies offered by the methods are then evaluated with respect to the discriminability of the decision rules for real and simulated data, respectively, through the techniques, bootstrap and holdout. Besides being far from complete, the current study shows that the performance of the two methods is similar in most cases, while there exist situations when each of them demonstrates advantage in relation to the other.

Keywords: Classification; Potential Discriminant Analysis; Kernel Discriminant Analysis; Bootstrap;

Índice

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

1.1 Introdução

A área de reconhecimento de padrões é bastante diversificada e dispõe de um grande número de técnicas [Duda et al. (2001)]. Um subconjunto muito importante e intensamente estudado destas técnicas é composto por métodos estatísticos [Jain et al. (2000), Friedman et al. (2000), Webb (2002)]. Os métodos estatísticos apresentam bons resultados quando as condições necessárias para a aplicação destes são satisfeitas. Além dos métodos estatísticos, outras técnicas estão avançando consideravelmente e decorrem da aplicação de conceitos da física a problemas de naturezas diversas, mediante a semelhança desses problemas com sistemas físicos. Quando consideramos a parte de aprendizagem superviosionada, mais especificamente o problema da classificação, dependendo da estrutura dos dados disponíveis para a aprendizagem, métodos estatísticos clássicos, tais como discriminante linear de Fisher, regressão logística e discriminante de núcleo são técnicas muito poderosas e estão entre as mais utilizadas. Por outro lado, também é possível abordar a análise discriminante no contexto de um processo físico, e essa e a principal contribuição desta dissertação. Mais especificamente, propomos um novo método de classificação, o discriminante de potencial, onde o espaço das observações é considerado como espaço físico Euclidiano, e daí extraímos as propriedades do método. O espaço físico do qual tratamos refere-se ao espaço das partículas que interagem entre si, espaço esse que é isomorfo ao espaço

real dos dados. Durante o texto estaremos lidando com métodos não paramétricos, o método de potencial, com a definição, propriedades, avaliação do desempenho; e o método de núcleo, que é utilizado para efeito de comparação por ser "parecido" com o método de potencial e apresentar boas propriedades.

No problema da classificação estamos interessados na proposição de um classificador, e para formalizarmos um procedimento de classificação temos que considerar as seguintes definições: padrão, vetor de dados p-dimensional $\boldsymbol{x} = (x_1, \ldots, x_p)'$, onde as componentes x_i são medidas de características de um objeto (o espaço de todos os valores possíveis de $\mathbf{x}$ é designado por $\mathcal{X}$); populações, classes ou grupos, são os conjuntos $\Pi_1, \ldots, \Pi_g$, $g \geq 2$, de onde provém os padrões, sendo que para cada padrão $\boldsymbol{x}$ existe uma variável categorizada y para identificar a população a qual o padrão pertence; amostra de treinamento, é uma amostra de padrões das g classes $T = \{\boldsymbol{x}_j, y_j\}_{j=1}^{n}$, onde n é o tamanho da amostra, utilizada para construir o classificador; amostra de teste, parecida com a amostra treinamento, porém é utilizada para avaliar o desempenho do classificador; regra de decisão, é uma função $F(\mathbf{x}) : \mathbf{R}^p \to \mathcal{C}$, $\mathcal{C}= \{1, \cdots, g\}$, utilizada para a classificação de uma observação p-dimensional em uma das g classes. Assim, um procedimento de classificação consiste na proposição da regra F a partir da amostra T, de tal forma que para cada $\boldsymbol{x} \in \mathcal{X}$, $F(\boldsymbol{x})$ designa uma classe em $\mathcal{C}$. Em essência, o desempenho do método de classificação está diretamente relacionado com a estrutura da regra de classificação.

1.2 Estrutura da Dissertação

Esta dissertação é composta por seis capítulos. No capítulo 2 consideramos o método de núcleo, que será utilizado para comparação com o método de potencial. O método de núcleo tem boas propriedades, e está completamente formalizado. É um método que vem sendo estudado desde os anos de 1950 [Fix & Hodges (1951)], e nos dias atuais recebe atenção dos pesquisadores [Scott (1992), Simonoff (1996)]. A aplicação do método de núcleo na análise discriminante é apenas uma das aplicações possíveis. Na realidade o método consiste em estimar densidades, e a partir da estimação muitas aplicações são possíveis. Assim, para o método de núcleo iremos considerar a parte de estimação das densidades com os detalhes necessários (funções núcleo,

escolha do parâmetro de suavização, critérios de erro, etc), e a análise discriminante de núcleo, construída com base nas densidades estimadas. No capítulo 3 abordamos o método de potencial. Neste caso vamos tratar da formalização do método. A construção é feita com base em conceitos da física, e todos os conceitos necessários são descritos (lei de Coulomb, campo elétrico, potencial elétrico, etc). Dentro da análise discriminante de potencial consideramos as duas formalizações possíveis, a definição da regra de classificação a partir das densidades estimadas pelo método de potencial, ou a partir da comparação do potencial resultante para cada classe. No capítulo 4 é abordada a parte referente às técnicas de comparação de classificadores. O foco é para a discriminabilidade da regra de classificação, que é avaliada através da taxa de erro do classificador. Os procedimentos considerados são holdout, cross-validation e bootstrap. Especial atenção é dada ao método de bootstrap devido o estimador .632+ da taxa de erro, por apresentar uma correção de viés para a estimativa. No capítulo 5 são apresentados estudos de simulação, e a análise de um conjunto de dados reais, para avaliar a performance dos métodos de potencial e de núcleo. Desta forma, para a simulação são considerados seis cenários com estruturas diferentes, com o tamanho do conjunto de treinamento variando, e para cada classificador são apresentados as taxas de erro estimadas com os respectivos desvios-padrões – estimados através do método holdout. Para o conjunto de dados reais será utilizado o procedimento bootstrap na estimação da taxa de erro. O capítulo 6 traz as considerações finais com conclusões e direcionamento para trabalhos futuros.

1.3 Plataforma Computacional

Nesta dissertação utilizamos o R em sua versão 2.6.1 [R Development Core Team (2006)] para fazer as simulações, análise de dados e construção dos gráficos com as regiões de classificação. O R é um software gratuito e encontra-se disponível através do site `http://www.R-project.org`. O Maple 11 foi utilizado na exploração das propriedades numéricas, estimação de densidades e gráficos 3D para o método de potencial. O Maple não é gratuito, e para mais informações visite `http://www.maplesoft.com/`. Este texto foi tipografado em LaTeX [Lamport (1994)] e mais detalhes podem ser encontrados em `http://www.tex.ac.uk/CTAN/latex`.

CAPÍTULO 2

Análise Discriminante de Núcleo

2.1 Introdução

No problema de análise discriminante, utilizamos uma regra de decisão $F(\mathbf{x}) : \mathbf{R}^p \rightarrow \{1, \cdots, g\}$ para a classificação de uma observação p-dimensional em uma das g classes. Utilizando o enfoque da teria de decisão de Bayes [Duda et al. (2001)], associamos uma observação para a classe com maior probabilidade a posteriori, o que pode ser descrito como

$$d(\mathbf{x}) = \underset{j \in \{1,\ldots,g\}}{\operatorname{argmax}} \quad p(j|\mathbf{x}) = \underset{j \in \{1,\ldots,g\}}{\operatorname{argmax}} \quad \pi_j f_j(\mathbf{x}) \tag{2.1}$$

onde $\pi_j's$ são as probabilidades a priori e $f_j(\mathbf{x})'s$ são as funções densidade de probabilidade das respectivas classes $(\Pi_1, \cdots, \Pi_g, g \geq 2)$.

Na prática as funções densidade $f_j(\mathbf{x})'s$ são usualmente não conhecidas, e podem ser estimadas a partir do conjunto de treinamento de forma paramétrica ou não paramétrica. Nos casos onde se dispõe de conhecimentos sobre a forma das distribuições envolvidas, ou porque se conhecem bem os mecanismos que geram os padrões ou porque se realizam testes estatísticos não paramétricos (χ^2, Kolmogorov-Smirnov, ou outros), fica faltando apenas estimar os parâmetros da distribuição, e isto pode ser feito utilizando técnicas clássicas da teoria de estimação [Mood et al. (1974);Hoel et al. (1971)]. Conseqüentemente, a performance da regra de discriminação paramétrica depende fortemente da validade dos modelos paramétricos [Kohn (1998)].

Nos casos em que não há informações sobre as distribuições, podem ser feitas estimações diretamente a partir dos dados. Este tipo de estimação, denominada de estimação não paramétrica, é mais flexível e não depende da suposição de um modelo paramétrico para cada população.

O método do núcleo estimador é um método bem conhecido de estimação não paramétrica de funções densidade [Silverman (1986)], que para a sua implementação depende da definição de alguns critérios, como a função núcleo (kernel) e a janela ótima (h) para suavização dos dados.

O uso do método de núcleo estimador para estimar densidades em problemas de análise discriminante é bastante popular na literatura existente [Silverman (1986); Scott (1992); Simonoff (1996); Jain et al. (2000); Friedman et al. (2000); Duda et al. (2001)] e em muitos softwares estatísticos (p.ex., R, SAS).

Na seção 2.2 abordamos a estimação de densidades através do método do núcleo para o caso unidimensional e multivariado, com destaque para a escolha do parâmetro de suavização. Em seguida, na seção 2.3, consideramos a técnica de classificação pelo método do núcleo, acompanhada com um algoritmo.

2.2 Método de Núcleo para a Estimação de Densidades

Conforme Scott (1992), é notável que o histograma permaneceu como o único estimador não paramétrico de densidade até os anos de 1950, quando simultaneamente progressos substanciais foram feitos na estimação de densidade e na estimação espectral de densidade. Partiu de Fix & Hodges (1951) a proposição do algoritmo básico de estimação não-paramétrica da função de densidade. Eles abordaram o problema da discriminação estatística quando a forma paramétrica da densidade amostral não é conhecida. Ainda segundo Scott (1992), durante as décadas seguintes, diversos algoritmos gerais e modos teóricos alternativos de análise foram introduzidos [p.ex., Rosenblatt (1956); Parzen (1962); Epanechnikov (1969)]. Nos anos de 1970 veio os primeiros trabalhos focando a aplicação prática destes modelos (referências para estes trabalhos, bem como os anteriores, podem ser encontradas em Scott (1992) e Silverman (1986)).

Vamos inicialmente considerar o método do histograma, por sua importância, e a partir dele construir o método de núcleo.

2.2.1 Histogramas

O mais antigo estimador da função de densidade é o histograma. A sua construção é feita a partir de um ponto de origem x_0 e um comprimento das classes h. As classes do histograma são os intervalos $[x_0 + mh, x_0 + (m+1)h]$, para valores positivos e negativos de m. Da maneira formal, define-se o histograma como:

$$\hat{f}(x) = \frac{1}{nh}(\text{números de } X_i \text{ na classe de } x). \tag{2.2}$$

Este estimador não é contínuo e depende fortemente da escolha de h e x_0. O parâmetro h é conhecido como parâmetro de suavização, e variando o valor de h obtemos diferentes formas de $\hat{f}(x)$. Por exemplo, quando aumentamos o valor de h temos intervalos maiores e conseqüentemente o histograma será uma representação mais suave. Nos extremos, digamos, quando $h \to 0$, o histograma torna-se uma representação muito ruidosa dos dados. Ao passo que para $h \to \infty$, o histograma torna-se uma representação muito suave dos dados.

Uma tentativa de eliminar a dependência do ponto de origem x_0 na construção do histograma, é o uso do estimador conhecido na literatura como "ingênuo" (ver Silverman (1986), pp. 11-13). Assim, se a variável aleatória X tem função densidade $f(x)$,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h),$$

então, um estimador natural da função densidade, escolhendo h pequeno, é dada por:

$$\hat{f}(x) = \frac{1}{nh}(\text{números de } X_i \text{ em } (x-h, x+h)) \quad \text{(Estimador Ingênuo)}. \tag{2.3}$$

Para expressar mais transparentemente o estimador em (2.3), definimos a função peso w por

$$w(x) = \begin{cases} \frac{1}{2} & \text{se } |x| < 1 \\ 0 & \text{caso contrário.} \end{cases} \tag{2.4}$$

Então é fácil ver que o estimador ingênuo pode ser escrito como

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right) \tag{2.5}$$

Segue de (2.4) que o estimador é construído colocando-se uma "caixa" de largura $2h$ e altura $(2nh)^{-1}$ em cada observação e então somando para obter a estimativa $\hat{f}$. Não é difícil notar que

a escolha de h ainda controla a quantidade de suavização a ser feita e que $\hat{f}$ não é uma função contínua, tendo derivada nula em todos os pontos exceto nos pontos de salto $X_i \pm h$ (onde a derivada geralmente diverge).

O estimador de núcleo aparece naturalmente quando generalizamos o estimador (2.5) para superar algumas dificuldades relacionadas a este tipo de estimador. De fato, todos algoritmos não paramétricos são assintoticamente métodos de núcleo, chamado de classe geral das "seqüências delta" [Scott (1992)]. Para tanto, fazendo-se a substituição da função $w(x)$ por uma função núcleo não-negativa K que satisfaça a condição de continuidade na reta e,

$$\int_{-\infty}^{+\infty} K(x)dx = 1.$$

Usualmente, mas não sempre, utiliza-se uma função de probabilidade simétrica, como por exemplo, a função de probabilidade normal para K. Neste caso particular, $\hat{f}$ será uma curva suave com derivadas de todas as ordens.

2.2.2 Estimação da Função Densidade

Suponha que $X_1, \cdots, X_n$ são observações de uma função densidade de probabilidade f. O estimador de núcleo $\hat{f}$ de f é definido por

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \tag{2.6}$$

onde K é a função núcleo e h é o parâmetro de suavização, positivo e não-aleatório, também chamado por alguns autores de largura da janela ou largura da banda.

Conforme Epanechnikov (1969) e Silverman (1982), a escolha da função núcleo não é crucial para a performance estatística do método, e é mais razoável escolher um núcleo que auxilie na eficiência computacional (além do fato de que noções sobre eficiência computacional tem sido radicalmente modificadas desde as décadas de setenta e oitenta). A tabela (2.1) apresenta exemplos de funções núcleo univariadas. Dentre as funções núcleo apresentadas, para o critério de otimalidade em termos do "erro quadrático integrado médio" (2.11), o núcleo que minimiza esse erro é Epanechnikov.

Tabela 2.1: Funções núcleo comumente utilizadas com dados univariados.

Função Núcleo	Forma analítica, $K(x)$
Retangular	$\frac{1}{2}$ para $\|x\| < 1$, 0 caso contrário
Triangular	$1 - \|x\|$ para $\|x\| < 1$, 0 caso contrário
Biweight	$\frac{15}{16}(1-x^2)^2$ para $\|x\| < 1$, 0 caso contrário
Normal	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
Epanechnikov	$\frac{3}{4}\left(1 - x^2/5\right)/\sqrt{5}$ para $\|x\| < \sqrt{5}$, 0 caso contrário

De Silverman (1986), a definição do núcleo estimador como a soma de "lombadas" centradas nas observações é facilmente generalizada para o caso multivariado. Assim, o estimador de núcleo multivariado com núcleo K e parâmetro de suavização h é definido por

$$\hat{f}(\boldsymbol{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} K\left(\frac{1}{h}(\boldsymbol{x} - \boldsymbol{X}_i)\right), \tag{2.7}$$

onde $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$ e $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$. A função núcleo $K(\boldsymbol{x})$ é agora uma função definida no espaço p-dimensional, satisfazendo

$$\int_{\mathbb{R}^p} K(\boldsymbol{x}) d\boldsymbol{x} = 1. \tag{2.8}$$

Usualmente, tendo em vista que a escolha da função núcleo não é crucial, K será uma função densidade de probabilidade unimodal gradualmente simétrica, por exemplo a distribuição normal padrão multivariada

$$K(\boldsymbol{x}) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}\right).$$

Uma outra possibilidade é a função núcleo Epanechnikov multivariada

$$K(\boldsymbol{x}) = \begin{cases} \frac{(1-\boldsymbol{x}'\boldsymbol{x})(p+2)}{2c_p} & \text{para } |\boldsymbol{x}| < 1 \\ 0 & \text{caso contrário,} \end{cases}$$

onde $c_p = \frac{\pi^{p/2}}{\Gamma((p/2)+1)}$ é o volume de uma esfera unitária p-dimensional.

Uma forma de estimação da função densidade multivariada é a soma do produto de funções núcleo, sem a implicação de independência entre as variáveis [Ferreira (2007)], o que resulta em

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n}\frac{1}{h_1 \cdots h_p}\sum_{i=1}^{n}\prod_{j=1}^{p} K_j\left(\frac{[\boldsymbol{x} - \boldsymbol{X}_i]_j}{h_j}\right), \tag{2.9}$$

onde existem diferentes parâmetros de suavização associados com cada variável. De forma mais geral, podemos assumir alguma das formas univariadas apresentadas na tabela 2.1 para os K_j, $j = 1, \ldots, p$. Por outro lado, usualmente a mesma forma é assumida para todos os K_j.

Uma metodologia utilizada para reduzir o problema da estimação de uma densidade multivariada é apresentada por Fukunaga (1972) e utilizada por Cavalcante (2004). O procedimento é simples e consiste em se utilizar uma transformação inversa na densidade multivariada com o objetivo de tornar as variáveis não correlacionadas. A partir daí se utiliza a abordagem já existente na literatura relacionada a estimação do parâmetro de suavização para o caso univariado. Ou seja, a janela h pode ser estimada individualmente para cada variável transformada, uma vez que estas não são correlacionadas.

Por fim, podemos ter uma representação mais geral para a estimação de uma densidade pelo método do núcleo

$$\hat{f}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}; \mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X_i}), \tag{2.10}$$

onde $K(\mathbf{x})$ é o núcleo, que é uma função densidade de probabilidade simétrica; $\mathbf{H}$ é a matriz denominada de largura de banda, que é simétrica e positiva-definida; e $K_{\mathbf{H}} = |\mathbf{H}|^{-1/2}K(\mathbf{H}^{-1/2}\boldsymbol{x})$.

De acordo com Wand & Jones (1993) a parametrização mais útil da matriz largura de banda $\mathbf{H}$ é a diagonal $\mathbf{H} = \text{diag}(h_1^2, h_2^2, \cdots, h_p^2)$ e a geral ou não restrita que não tem restrição sobre $\mathbf{H}$, desde que $\mathbf{H}$ permaneça positiva-definida e simétrica.

2.2.3 Escolha do Parâmetro de Suavização

O critério de erro utilizado para medir a performance de $\hat{f}$ (amplamente utilizado pelos pesquisadores nesta área) é o *erro quadrático integrado médio* (EQIM)[Rosenblatt (1956)], definido como

$$\text{EQIM}(\mathbf{H}) = E\left\{\int_{\mathbb{R}^p} [\hat{f}(\boldsymbol{x}; \mathbf{H}) - f(\boldsymbol{x})]^2 d\boldsymbol{x}\right\}. \tag{2.11}$$

Nosso objetivo na seleção do parâmetro de suavização é para estimar

$$\mathbf{H}_{\text{EQIM}} = \underset{\mathbf{H}\in\mathcal{H}}{\text{argmin}}\, \text{EQIM}(\mathbf{H}), \tag{2.12}$$

onde $\mathcal{H}$ é o espaço das matrizes simétricas, positivas-definidas de dimensão $(p \times p)$. Contudo, o EQIM apresenta forma fechada apenas se f é uma mistura de distribuições normais e K é a função núcleo normal [Wand & Jones (1995)]. Neste caso, para contornar esse problema emprega-se uma aproximação assintótica, conhecida como *erro quadrático integrado médio assintótico* (EQIMA)

$$\text{EQIMA}(\mathbf{H}) = \frac{1}{n}R(K)|\mathbf{H}|^{-1/2} + \frac{1}{4}\mu_2(K)^2 \int_{\mathbb{R}^p} \text{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x}))d\boldsymbol{x}, \tag{2.13}$$

onde $R(v) = \int_{\mathbb{R}^p} v(\boldsymbol{x})^2 d\boldsymbol{x}$ para alguma função integrável quadrada v; $\mu_2(K)\mathbf{I}_p = \int_{\mathbb{R}^p} \boldsymbol{x}\boldsymbol{x}' K(\boldsymbol{x})d\boldsymbol{x}$, com $\mu_2(K) < \infty$ e $\mathbf{I}_p$ é a matriz identidade de dimensão $(p \times p)$; e $D^2 f(\boldsymbol{x})$ é a matrix Hessiana de f.

Nós fazemos uso da tratabilidade do EQIMA procurando

$$\mathbf{H}_{\text{EQIMA}} = \underset{\mathbf{H}\in\mathcal{H}}{\text{argmin}}\, \text{EQIMA}(\mathbf{H}). \tag{2.14}$$

O próximo passo é estimar o EQMI ou o EQMIA . Por Duong (2007), o seletor do parâmetro de suavização guiado pelos dados é

$$\hat{\mathbf{H}}_{\text{EQIM}} = \underset{\mathbf{H}\in\mathcal{H}}{\text{argmin}}\, \widehat{\text{EQIM}}(\mathbf{H}) \quad \text{ou} \quad \hat{\mathbf{H}}_{\text{EQIMA}} = \underset{\mathbf{H}\in\mathcal{H}}{\text{argmin}}\, \widehat{\text{EQIMA}}(\mathbf{H}). \tag{2.15}$$

Existem diferentes métodos para a estimação em (2.15), que podem ser encontrados em Duong (2004). Por exemplo, o método *plug-in* estima o EQIMA, enquanto o método de mínimos quadrados por validação cruzada (LSCV, sigla em inglês) estima o EQIM diretamente. Para o nosso propósito, classificação, utilizaremos o LSCV.

Para ver que o LSCV estima o EQIM diretamente, considere a versão multivariada do LSCV

$$\text{LSCV}(\mathbf{H}) = \int_{\mathbf{R}^p} \hat{f}(\mathbf{x};\mathbf{H})^2 d\mathbf{x} - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{-i}(\mathbf{X}_i;\mathbf{H}),$$

onde o estimador *leave-one-out* [seção 4.3] é

$$\hat{f}_{-i}(\mathbf{x};\mathbf{H}) = (n-1)^{-1}\sum_{\substack{j=1\\ j\neq i}} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_j).$$

O LSCV seletor $\hat{\mathbf{H}}_{\text{LSCV}}$ é o argumento mínimo de LSCV($\mathbf{H}$). Conforme Duong (2007), pode-se mostrar que $\mathbf{E}[\text{LSCV}(\mathbf{H})] = \text{EQIM}(\mathbf{H}) - \int_{\mathbf{R}^p} f(\mathbf{x})^2 d\mathbf{x}$, indicando que o LSCV estima o EQIM diretamente.

2.3 Análise Discriminante de Núcleo

A regra discriminante de núcleo (RDN) é obtida da regra discriminante de Bayes (2.1) substituindo f_j por sua estimativa através do método de núcleo

$$\hat{f}_j(\mathbf{x}; \mathbf{H}) = n_j^{-1} \sum_{i=1}^{n_j} K_{\mathbf{H}_j}(\mathbf{x} - \mathbf{X}_{ji}),$$

e π_j é usualmente substituído pela proporção amostral n_j/n, onde $n = \sum_{j=1}^{g} n_j$; desta forma

$$\text{RDN: Alocar } \mathbf{x} \text{ para a população } \Pi_0 \text{ onde } \Pi_0 = \underset{j \in \{1,\ldots,g\}}{\operatorname{argmax}} \quad \hat{\pi}_j \hat{f}_j(\mathbf{x}; \mathbf{H}_j). \tag{2.16}$$

Agora levanta-se a questão sobre qual é a melhor escolha do parâmetro de suavização para esta estimação de densidades. Embora a escolha de $\mathbf{H}_j$ seja ótima em termos do EQIM, o mesmo não se verifica quando consideramos o erro de classificação. Por exemplo, em um estudo de simulação Taylor (1997) gerou amostras de tamanho 100 das distribuições $\text{N}(0; 1)$ e $\text{N}(0; 0.5^2)$; para estas distribuições, $(h_1; h_2) = (0,422; 0,211)$ minimiza o EQIM assintótico (EQIMA), visto que a atual taxa de erro é minimizada para $(h_1; h_2) = (0,720; 0,215)$. Uma alternativa é selecionar o parâmetro de suavização satisfazendo uma medida de erro adaptada para a análise discriminante [Hall & Kang (2005)], porém pelo fato de ser bastante computacionalmente intensivo, não foi incluso na biblioteca ks do R [Duong (2007)]. Felizmente, Hall & Kang (2005) mostraram que seus seletores ótimos para a análise discriminante é da mesma ordem assintótica dos seletores ótimos para a estimação das densidades. Outro trabalho nessa linha é o de Ghosh & Chaudhuri (2004), que apresentam um método para a escolha apropriada do parâmetro de suavização, quando a densidade é estimada para o conjunto de treinamento em um problema de classificação.

2.3.1 Algoritmo para Análise Discriminante de Núcleo

Um algoritmo para análise discriminante é descrito em Ferreira (2007). A descrição do algoritmo, com uma leve adaptação, é apresentada a seguir

1. Para cada amostra de treinamento $\boldsymbol{X}_j = \{\boldsymbol{X}_{j1}, \boldsymbol{X}_{j2}, \ldots, \boldsymbol{X}_{jn_j}\}$, $j = 1, \ldots, g$, calcule a estimativa da densidade

$$\hat{f}(\boldsymbol{x}; \mathbf{H}_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} K_{\mathbf{H}_j}(\boldsymbol{x} - \boldsymbol{X}_{ji}),$$

onde $\mathbf{H}_j$ deve ser escolhida através de algum método sensível.

2. Se as probabilidades a priori das classes estão disponíveis, utilize-as. Caso contrário, estime-as através das proporções amostrais no conjunto de treinamento, $\hat{\pi}_j = n_j/n$, $j = 1, \ldots, g$.

3. (a) Classifique as observações do conjunto de teste de acordo com a regra (2.16).

   (b) Classifique as observações do conjunto de treinamento de acordo com a regra (2.16).

4. Estime a taxa de má-classificação para os conjuntos de teste e de treinamento.

CAPÍTULO 3

Método Discriminante de Potencial

3.1 Introdução

A análise discriminante de potencial é uma nova técnica não paramétrica de classificação supervisionada, cuja formulação representa a contribuição fundamental deste trabalho. A construção da regra de classificação também utiliza o enfoque da teoria de decisão de Bayes [Duda et al. (2001)], porém esta pode ser feita de duas formas equivalentes, ou estimando as densidades diretamente com o auxílio da função de potencial, conceito que será explicado na seção 3.6, para as classes, ou a partir da comparação do potencial resultante em relação às classes. As propriedades resultantes do método de potencial são bastante interessantes e dentre elas podemos citar: a regra de decisão é superajustada, ou seja, a regra não apresenta erro de decisão para classificar o conjunto de treinamento e toda observação cria uma fronteira de classificação; a estimação da densidade não depende de parâmetros de suavização, e a normalização pode ser feita numericamente; o método pode ser aplicado em problemas onde os conjuntos das classes não são linearmente separáveis, como também em problemas onde o vetor de observações pertence a uma dimensão muito alta.

As razões que podem fazer com que o método de potencial seja utilizado estão na natureza do problema em si. Por exemplo, consideremos um método de diagnóstico onde a partir dele seja tomada a decisão se determinado paciente deve ou não se submeter a um tratamento específico

para uma patologia. Pode-se cometer dois erros, o paciente não está doente e é submetido ao tratamento, gerando assim uma série de problemas (efeitos colaterais do tratamento, custo com o tratamento, etc); ou o paciente que está doente não recebe o tratamento, prolongando mais o seu sofrer, e demais conseqüências. Neste caso, ambos os erros devem ser evitados, porém como não se consegue, é importante um método que apresente a menor chance de cometer esses erros. A partir das evidências de que para um conjunto de pacientes utilizado como treinamento, todos foram corretamente classificados entre os que devem ser tratados e os que não devem, o método de potencial apresenta-se como uma interessante alternativa a ser considerada pela a sua capacidade de classificação de novos pacientes. Em outro caso, por exemplo, pessoal que buscam ter crédito de uma instituição financeira, neste caso não dá crédito a um bom pagador não gera prejuízo para a instituição, porém fornecer crédito a um mau pagador sim. Assim, conseguir identificar os maus pagadores é um esforço constante da instituição, e novamente o método de potencial se torna uma opção por conseguir identificar corretamente todos os maus pagadores do conjunto de treinamento e fazer com que cada mau pagador seja um protótipo para comparar com um novo cliente que solicita crédito.

Antes de abordarmos a análise discriminante de potencial, vamos inicialmente apresentar os conceitos físicos necessários para a formalização do método. Assim, os seguinte temas da física serão abordados: carga elétrica, lei de Coulomb, campo elétrico, potencial elétrico (energia potencial elétrica, superfícies eqüipotenciais, potencial devido a uma carga, potencial devido a um grupo de cargas). A elaboração dos tópicos de física foi feita com base em Halliday (2006). E por fim, apresentamos a metodologia de classificação pelo método de potencial.

3.2 Carga Elétrica

O fenômeno de eletricidade já era conhecido pelos primeiros filósofos gregos. Eles sabiam que ao se esfregar um pedaço de âmbar, ele atraia pequenos pedaços de palhas. Por isso, não é de surpreender que o termo *elétron* vem da palavra grega que significa âmbar. Diversos outros fenômenos são observados que envolvem propriedades eletromagnéticas, e se devem à presença e ao movimento de imensa quantidade de *carga elétrica* que está armazenada nos objetos materiais.

Em particular, a **carga elétrica** é uma característica intrínseca das partículas fundamentais que formam esses objetos.

Não se observa normalmente a imensa quantidade de carga de um objeto, pois existem dois tipos de cargas, *cargas positivas* e *cargas negativas*, e geralmente um objeto contém quantidades iguais dos dois tipos de cargas. Com equilíbrio de cargas, um objeto está *eletricamente neutro*; ou seja, sua carga resultante é nula. Dizemos que um objeto está *carregado* ou *eletrizado* quando a carga resultante é não nula (excesso de um de dois tipos de carga). O desequilíbrio é geralmente muito pequeno quando comparado com as quantidades totais de carga positiva e negativa contidas no objeto.

Uma característica fundamental dos objetos carregados é que eles interagem exercendo forças uns sobre os outros e se constata a seguinte regra: cargas com o mesmo sinal elétrico se repelem, e cargas com sinais elétricos contrários se atraem. A expressão desta regra de forma quantitativa é dada na seção seguinte pela lei de Coulomb para a força eletrostática.

3.3 Lei de Coulomb

Deve-se ao francês Charles Augustin de Coulomb a formulação, em 1785, da lei que rege as interações entre as partículas eletrizadas.

A interação eletrostática entre partículas eletrizadas caracteriza-se por meio de forças de atração ou repulsão, dependendo dos sinais das cargas. Considere duas partículas eletrizadas (também chamadas de *cargas pontuais*) com intensidades de cargas q_1 e q_2 separadas por uma distância r. A força eletrostática, de atração ou repulsão, atuando sobre a partícula 2 devido à presença da partícula 1 é dada por

$$\vec{F}_{12} = kq_1q_2\frac{\vec{r}_{12}}{|\vec{r}_{12}|^3} \qquad \text{(Lei de Coulomb)}, \tag{3.1}$$

onde k é a constante eletrostática. Também, a partícula 2 exerce força $\vec{F}_{21} = -\vec{F}_{12}$, com mesma intensidade, sobre a outra partícula; as duas forças formam um par de forças da terceira lei de Newton (forças de ação e reação).

A figura 3.1 apresenta as direções e os sentidos das forças que agem em cada uma das partícu-

las dependendo do sinal da carga que cada uma contém. Assim, na figura 3.1.a as partículas se repelem (repulsão) e na figura 3.1.b as partículas se atraem (atração).

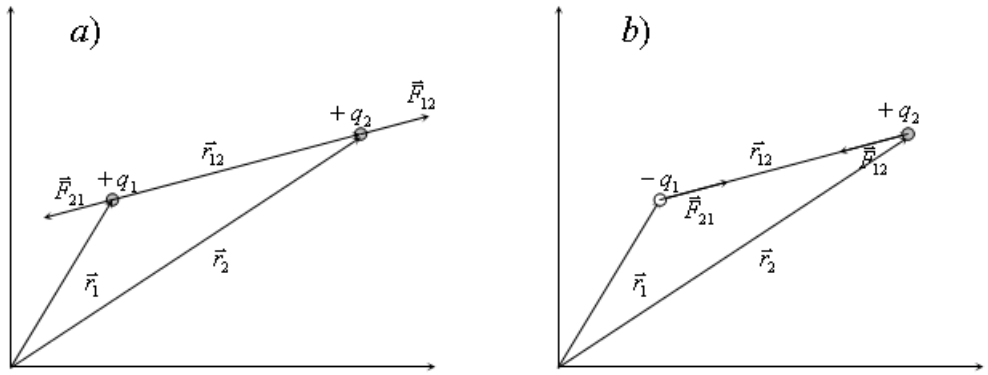

Figura 3.1: Direção e sentido da força eletrostática conforme o sinal das cargas.

No caso onde existem n partículas carregadas, elas interagem independentemente aos pares, e a força sobre qualquer uma delas, digamos sobre a partícula i, é dada pela soma vetorial (força resultante)

$$\vec{F}_{i,res} = \vec{F}_{i,1} + \vec{F}_{i,2} + \cdots + \vec{F}_{i,(i-1)} + \vec{F}_{i,(i+1)} + \cdots + \vec{F}_{i,n}, \tag{3.2}$$

onde, $\vec{F}_{i,j}$ é a força atuando sobre a partícula i devido à presença da partícula $j \neq i$.

3.4 Campo Elétrico

Uma carga pontual positiva q estabelece um campo elétrico no espaço que a cerca. Em um ponto P qualquer neste espaço o campo possui intensidade, direção e sentido. Ou seja, o campo elétrico é um campo vetorial; ele é formado por uma distribuição de vetores, um para cada ponto ao redor de um objeto carregado.

Para definirmos o campo elétrico necessitamos de uma carga positiva q_0, chamada de carga de teste, que é colocada num ponto P desse campo. Medimos então a força eletrostática $\vec{F}$ que age sobre a carga de teste. O campo elétrico $\vec{E}$ no ponto P devido ao objeto carregado é então

definido como

$$\vec{E} = \frac{\vec{F}}{q_0} \qquad \text{(campo elétrico)}. \tag{3.3}$$

Assim, a intensidade do campo elétrico $\vec{E}$ no ponto P é $E = \frac{F}{q_0}$, e a direção e o sentido de $\vec{E}$ são os da força $\vec{F}$ que age sobre a carga de teste positiva.

É simples determinar o campo elétrico devido a uma carga pontual q (ou partícula carregada) em qualquer ponto a uma distância r da carga pontual. O passo inicial é colocar uma carga de teste positiva q_0 nesse ponto. Através da lei de Coulomb (Equação 3.1) determina-se a intensidade da força eletrostática que age sobre q_0, o que nos dá

$$F = k\frac{|q||q_0|}{r^2}. \tag{3.4}$$

Dependendo se q é positiva, a força $\vec{F}$ aponta para longe da carga pontual ao longo da direção que une as cargas, e se q for negativa, aponta em direção à carga pontual ao longo da direção que une as cargas. Pela Equação 3.3, a intensidade do vetor campo elétrico é

$$E = \frac{F}{q_0} = k\frac{|q|}{r^2} \qquad \text{(carga pontual)}. \tag{3.5}$$

Novamente temos que a direção e o sentido de $\vec{E}$ são os mesmos que os da força sobre a carga de teste.

No caso em que temos n cargas pontuais e uma carga de teste positiva q_0, podemos determinar o campo elétrico resultante. Pela Equação 3.2, a força resultante $\vec{F}_0$ das n cargas pontuais que agem sobre a carga de teste é

$$\vec{F}_0 = \vec{F}_{01} + \vec{F}_{02} + \cdots + \vec{F}_{0n}.$$

Portanto, pela Equação 3.3, o campo elétrico resultante na posição da carga de teste é

$$\begin{aligned} \vec{E} &= \frac{\vec{F}_0}{q_0} = \frac{\vec{F}_{01}}{q_0} + \frac{\vec{F}_{02}}{q_0} \cdots + \frac{\vec{F}_{0n}}{q_0} \\ &= \vec{E}_1 + \vec{E}_2 + \cdots + \vec{E}_n. \end{aligned} \tag{3.6}$$

Nesta equação, $\vec{E}_i$ é o campo elétrico que seria estabelecido pela carga pontual i atuando sozinha. Podemos então perceber da Equação 3.6 e da Equação 3.2 que o princípio da superposição se aplica tanto a campos elétricos quanto a forças eletrostáticas.

3.5 Potencial Elétrico

3.5.1 Energia Potencial Elétrica

A força eletrostática é uma força conservativa, e quando ela atua entre duas ou mais partículas carregadas dentro de um sistema de partículas, a esse sistema podemos atribuir um **potencial elétrico**. Se há mudança no sistema do estado inicial i para um estado final diferente f, a força eletrostática realiza um trabalho W sobre as partículas. A variação resultante ΔU da energia potencial do sistema é

$$\Delta U = U_f - U_i = -W. \tag{3.7}$$

Devido a característica conservativa da força eletrostática, o trabalho realizado pela força eletrostática independe da trajetória. Assim, quando uma partícula carregada dentro do sistema se move do ponto i até o ponto f, enquanto a força eletrostática entre ela e o resto do sistema atua sobre ela, e o resto do sistema não varia, o trabalho W realizado pela força é o mesmo para todas as trajetórias entre os pontos i e f.

É comum utilizarmos como *configuração de referência* de um sistema de partículas carregadas a situação na qual todas as partículas estão infinitamente separadas uma das outras, para a qual a energia potencial de referência correspondente é definida como nula. Consideremos um estado inicial i com várias partículas cuja separação é infinita (configuração de referência), e outro estado f no qual essas partículas ficam agrupadas de maneira a se tornarem próximas. Como a energia potencial inicial U_i por definição tem valor zero, representemos por W_∞ o trabalho realizado pelas forças eletrostáticas entre as partículas durante o deslocamento do infinito até as posições próximas. Então da Equação 3.7 a energia potencial final U do sistema é

$$U = -W_\infty. \tag{3.8}$$

3.5.2 Potencial Elétrico

Em um campo elétrico a energia potencial de uma partícula carregada depende da intensidade da sua carga. Entretanto, a energia potencial *por unidade de carga* possui um valor único em qualquer ponto em um campo elétrico. Assim, a energia potencial por unidade de carga é uma

característica apenas do campo elétrico, e independe da carga q da partícula que é observada. A energia potencial por unidade de carga em um ponto de um campo elétrico é chamada de **potencial elétrico** V (ou simplesmente potencial) nesse ponto, grandeza escalar dada por

$$V = \frac{U}{q}. \tag{3.9}$$

A diferença de potencial elétrico ΔU entre dois pontos quaisquer i e f em um campo elétrico é igual a diferença de energia potencial por unidade de carga entre os dois pontos. Isto é

$$\Delta V = V_f - V_i = \frac{U_f}{q} - \frac{U_i}{q} = \frac{\Delta U}{q}. \tag{3.10}$$

Da Equação 3.7 temos que $\Delta U = -W$ e substituindo na Equação 3.10 temos

$$\Delta V = V_f - V_i = \frac{-W}{q} \qquad \text{(definição de diferença de potencial)}. \tag{3.11}$$

Portanto, a diferença de potencial entre dois pontos é igual ao valor negativo do trabalho realizado pela força eletrostática para mover uma carga unitária de um ponto a outro. Uma diferença de potencial pode ser positiva, negativa, ou nula, dependendo dos sinais e das intensidades de q e W.

Se adotamos $U_i = 0$ no infinito como a energia potencial de referência, então, pela Equação 3.9, o potencial elétrico no infinito também deve ser nulo, e a partir de 3.8 e 3.9 segue que o potencial elétrico em qualquer ponto em um campo elétrico e dado por

$$V = \frac{-W_\infty}{q}, \tag{3.12}$$

onde W_∞ é o trabalho realizado pelo campo elétrico sobre uma partícula carregada quando esta partícula se move do infinito até o ponto f. Um potencial V pode ser positivo, negativo ou nulo, dependendo dos sinais e das intensidades de q e W_∞.

3.5.3 Superfícies Eqüipotenciais

A **superfície eqüipotencial** é formada por pontos adjacentes que possuem o mesmo potencial elétrico. Quando uma partícula carregada se move entre dois pontos i e f sobre uma mesma superfície eqüipotencial nenhum trabalho resultante W é realizado pelo campo elétrico sobre a

partícula. Isto decorre da Equação 3.12, pois W deve ser nulo se $V_f = V_i$. Como o trabalho independe da trajetória (portanto da energia potencial e do potencial), $W = 0$ para qualquer trajetória ligando os pontos i e f, não importando se essa trajetória está ou não inteiramente sobre a superfície eqüipotencial.

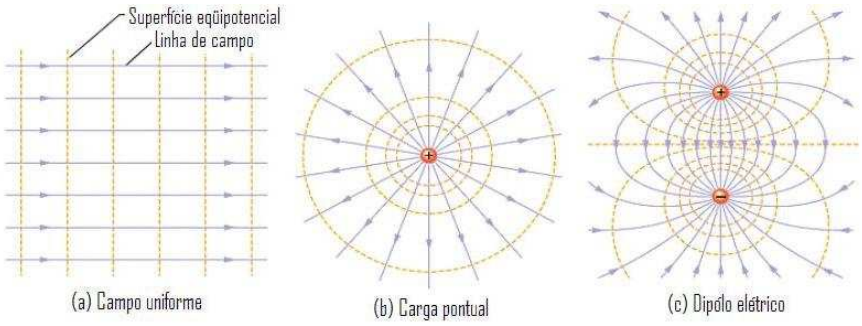

Figura 3.2: Linhas de campo elétrico (linhas cheias) e cortes transversais de superfícies equipotenciais (linhas tracejadas).

A figura 3.2 mostra linhas de campo elétrico (linhas cheias) e cortes transversais de superfícies eqüipotenciais (linhas tracejadas) para um campo elétrico uniforme e para um campo associado a uma carga pontual e a um dipolo elétrico.

3.5.4 Calculando o Potencial a partir do Campo

Conhecendo o vetor campo elétrico $\vec{E}$ ao longo de qualquer trajetória que ligue dois pontos i e f em um campo elétrico podemos calcular a diferença de potencial entre esses pontos. O cálculo é feito encontrando o trabalho realizado pelo campo sobre uma carga de teste positiva quando ela se move de i para f, e depois usando a Equação 3.12. O que nos dá

$$V_f - V_i = -\int_i^f \vec{E} \cdot d\vec{s}. \tag{3.13}$$

Assim, a diferença de potencial $V_f - V_i$ entre dois pontos i e f quaisquer em um campo elétrico é igual a menos a integral de linha (significando a integral ao longo de uma trajetória particular) de $\vec{E} \cdot d\vec{s}$ de i até f. Entretanto, todas as trajetórias (sejam elas fáceis ou difíceis de serem usadas) fornecem o mesmo resultado, pois a força eletrostática é conservativa.

3.5.5 Potencial devido a uma Carga Pontual

Considere um ponto P a uma distância R de uma partícula fixa com carga positiva q conforme figura 3.3. Para mover uma carga de prova q_0 desde o ponto P até o infinito, como a trajetória adotada não é importante, escolhemos a mais simples de todas – uma linha radial a partir da partícula fixa, passando por P, até o infinito. Como o campo $\vec{E}$ também é radial o deslocamento diferencial $d\vec{s}$ tem a mesma direção de $\vec{E}$. Assim,

$$\vec{E} \cdot d\vec{s} = E \cos\theta ds = Edr.$$

Então, pela Equação 3.13,

$$V_f - V_i = -\int_i^f \vec{E} \cdot d\vec{r} = \int_R^\infty Edr.$$

Fazendo $V_f = 0$ (no infinito) e $V_i = V$ (em R) e usando definição de campo elétrico dada por Equação 3.5 temos

$$0 - V = -k\int_R^\infty \frac{1}{r^2}dr = k\left[\frac{1}{r}\right]_R^\infty = -k\frac{q}{R}.$$

Resolvendo para V e trocando R por r, temos então

$$V = k\frac{q}{r} \tag{3.14}$$

como o potencial elétrico devido a uma partícula com carga q, em qualquer distância radial r da partícula.

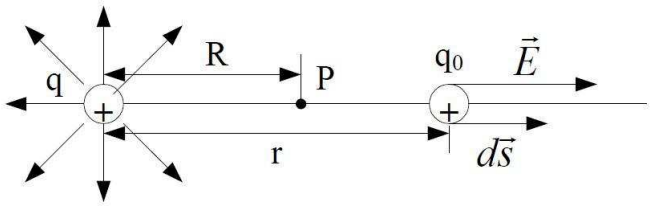

Figura 3.3: A carga pontual positiva q produz um campo elétrico $\vec{E}$ e um potencial elétrico V no ponto P. Determinamos o potencial movendo uma carga de teste q_0 de P até o infinito.

A Equação 3.14 foi deduzida para uma partícula carregada positivamente, mas a dedução também se aplicada a uma partícula carregada negativamente e, neste caso, q é uma grandeza

negativa. Observe que o sinal de V e de q são os mesmos. Desta forma: uma partícula carregada positivamente produz um potencial elétrico positivo; uma partícula carregada negativamente produz um potencial elétrico negativo.

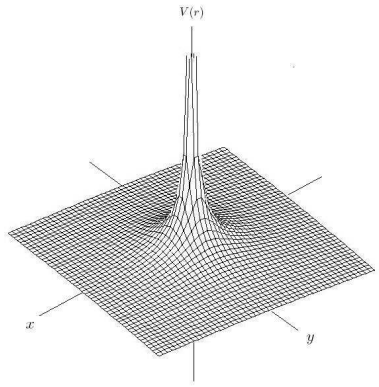

Figura 3.4: Gráfico do potencial elétrico $V(r)$ devido a uma carga carga pontual positiva localizada na origem de um plano xy.

A figura 3.4 mostra um gráfico do potencial elétrico para uma partícula carregada positivamente localizada na origem de um plano xy; a intensidade de V está plotada na vertical. Observe que a intensidade aumenta quando $r \to 0$.

3.5.6 Potencial devido a um Grupo de Cargas Pontuais

O potencial resultante em um ponto devido a um grupo de cargas pontuais pode ser determinado com a ajuda do princípio da superposição. O procedimento é calcular separadamente o potencial resultante no ponto dado, devido à presença de cada carga, usando a Equação 3.14 com o sinal de carga incluído. Depois os potenciais são somados. Assim, para n cargas, o potencial resultante é

$$V = \sum_{i=1}^{n} V_i = k \sum_{i=1}^{n} \frac{q_i}{r_i} \qquad (n \text{ cargas pontuais}), \tag{3.15}$$

onde q_i é o valor da i-ésima carga e r_i é a distância radial do ponto dado à i-ésima carga.

A figura 3.5 apresenta o potencial elétrico resultante devido à presença de quatro cargas positivas e duas cargas negativas no plano, onde a terceira dimensão é reservada para representar o potencial. As cargas positivas q_1, q_2, q_3 e q_4, estão localizadas, respectivamente, nos pontos

$p_1 = (2, 2)$, $p_2 = (2, 8)$, $p_3 = (7, 7)$ e $p_4 = (9, 8)$ do plano xy. As cargas negativas q_5 e q_6 estão localizadas, respectivamente, nos pontos $p_5 = (5, 5)$, $p_6 = (7, 3)$ do plano xy.

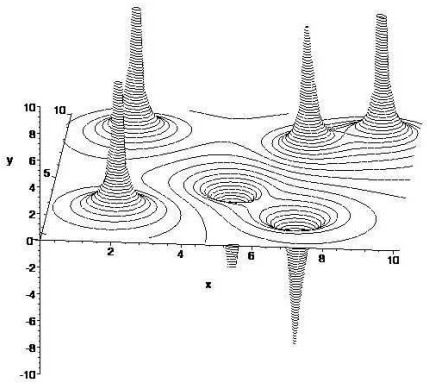

Figura 3.5: Gráfico do potencial elétrico resultante V devido a presença de cargas pontuais positivas e cargas pontuais negativas localizadas no plano xy.

A interpretação intuitiva da figura 3.5 pode ser obtida a partir da noção de que todos os sistemas físicos "tentam" minimizar a energia, do ponto de vista que para se colocar um sistema em um estado (ou configuração) com energia elevada, deve-se fazer o trabalho correspondente (pelas forças externas de qualquer natureza). Mais precisamente, fica mais "fácil" (exige menos trabalho) colocar uma carga positiva q num lugar com potencial mais baixo da figura 3.5 do que num lugar com potencial mais alto, em analogia com a força e o potencial gravitacionais. Para cargas negativas (que não têm analogia com a massa e forças gravitacionais) a situação é inversa: fica mais "fácil" colocá-las em lugares "altos" do que em lugares "baixos". Para agilizar a visualização intuitiva para este caso, ao invés do potencial V poderia ser observado o seu valor negativo (reflexão da figura 3.5 em relação ao plano xy), onde de novo os lugares "altos" serão mais "difíceis" de realizar do que os lugares "baixos". Em síntese, regiões com potencial negativo representam áreas de atração para cargas positivas (aumento de potencial corresponde ao aumento de atração), enquanto regiões com potencial positivo representam áreas de atração para cargas negativas.

3.6 Análise Discriminante de Potencial

De forma geral, os métodos de classificação supervisionada baseiam-se em conceitos de proximidade às observações individuais do conjunto de treinamento pertencentes às g populações. Deste ponto de vista, para cada observação individual poderia ser associado um campo, exercendo forças atrativas para as observações da mesma população, e forças repulsivas para observações de outras populações, decaindo com distância da origem. A superposição de tais campos para todas observações do conjunto de treinamento pode ser utilizada para estimar as densidades de probabilidade das classes em todo o espaço.

Adotando forças virtuais que decaem com o quadrado da distância, para o caso de duas populações podemos diretamente aplicar o formalismo da eletrostática, descrito em seções anteriores. A única modificação necessária é a troca de sinal do sentido da força, porque cargas elétricas de mesmo tipo se repelem, e no caso atual se atraem. Para o conjunto de treinamento com seis observações, $p_1 = (2, 2)$, $p_2 = (2, 8)$, $p_3 = (7, 7)$ e $p_4 = (9, 8)$ da população A, e $p_5 = (5, 5)$ e $p_6 = (7, 3)$ da população B, o potencial resultante mostrado na figura 3.5 apresenta duas regiões: região com potencial positivo, que representa área de atração para população A; e região com potencial negativo, área de atração para a população B.

Para um número de populações $g > 2$ não existe analogia no mundo real ($g = 1$ é análogo com forças gravitacionais, e $g = 2$ tem analogia com forças eletrostáticas com inversão de sinal), e para a aplicação do método atual de Discriminante de Potencial é necessário generalizar o formalismo físico. Para isto, vamos considerar um caso simples de três partículas virtuais dos tipos A, B e C, situadas no plano bidimensional nos pontos $(0, 1)$, $(0, 0)$ e $(0, -1)$, respectivamente, onde partículas do mesmo tipo exercem forças atrativas entre si, e partículas de tipos diferentes se repelem, com forças decaindo com o quadrado da distância (por simplicidade, vamos adotar unidades com constante de proporcionalidade unitária, e "cargas" unitárias). A força resultante atuando em uma partícula de teste vai depender da sua posição e do seu tipo (A, B ou C). A figura 3.6 mostra os três casos para partículas de teste de todos os três tipos, no ponto $(1, \frac{1}{2})$. Deste exemplo segue que precisamos definir três potenciais distintos, um para cada tipo de

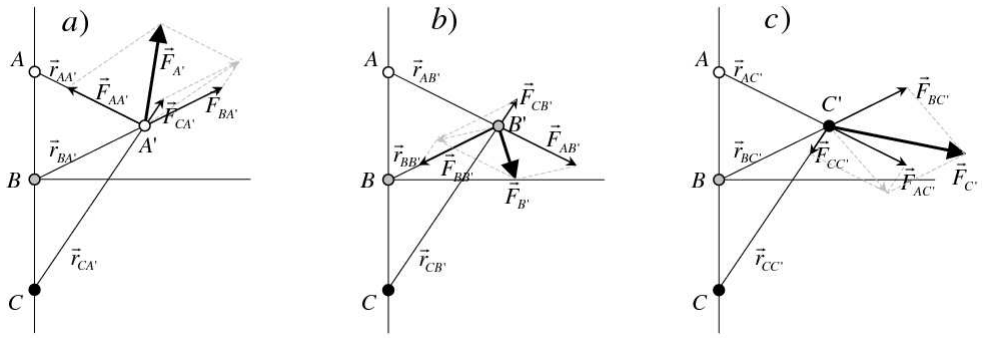

Figura 3.6: Gráfico da força resultante exercida pelas partículas A, B e C situadas, respectivamente, nos pontos $(0, 1)$, $(0, 0)$ e $(0, -1)$, nas partículas de teste para tipos diferentes, posicionadas em $(1, \frac{1}{2})$.

partícula, porque o trabalho necessário para trazer uma partícula do infinito para um certo ponto depende do seu tipo. De fato, no caso de forças eletrostáticas também poderíamos ter definido dois potenciais $V_+ = -V_-$ para cargas de teste positivas e negativas, respectivamente, como já comentado em relação a figura 3.5. Neste caso, a energia potencial seria calculada usando só valores absolutas da carga, enquanto seu sinal entraria na definição dos potenciais V_+ e V_-.

Suponha agora que temos $\Pi_1, \Pi_2, \cdots, \Pi_g$ $(g \geq 2)$ populações, e de cada população dispomos de uma amostra $\boldsymbol{X}_j = \{\boldsymbol{X}_{j1}, \boldsymbol{X}_{j2}, \ldots, \boldsymbol{X}_{jn_j}\}$, $j = 1, \ldots, g$ e $n = \sum_{j=1}^{g} n_j$. Então, a força resultante exercida pelo conjunto de treinamento na observação de teste pertencente à população Π_i, situada no ponto $\mathbf{x} = (x_1, x_2, \cdots, x_p)'$, pode ser definida por

$$\mathbf{F}_{\Pi_i}(\mathbf{x}) = \sum_{j=1}^{g} (1 - 2\delta_{ij}) \sum_{k=1}^{n_j} \frac{\mathbf{x} - \mathbf{X_{jk}}}{|\mathbf{x} - \mathbf{X_{jk}}|^3}, \tag{3.16}$$

onde δ_{ij} é o delta de Kronecker. Por simplicidade adotamos "cargas" e constante de proporcionalidade unitárias. Como a força resultante na partícula de teste representa soma vetorial de forças individuais de todas as partículas gerando o campo, o potencial resultante no ponto $\mathbf{x} = (x_1, x_2, \cdots, x_p)'$ com relação a população Π_i é a soma simples (escalar) de potenciais individuais. Estes potenciais individuais não dependem da forma como partícula de teste chega ao ponto observado, e a integração de cada termo no lado direito da equação 3.16 do infinito até $\mathbf{x}$ pode ser calculada (de forma mais fácil) ao longo do raio $\mathbf{x} - \mathbf{X_{jk}}$, onde o deslocamento é colinear

com a força, em analogia com o caso de forças eletrostáticas. Assim, para o potencial resultante temos

$$V_{\Pi_i}(\mathbf{x}) = \sum_{j=1}^{g} (1 - 2\delta_{ij}) \sum_{k=1}^{n_j} \frac{1}{|\mathbf{x} - \mathbf{X_{jk}}|}. \tag{3.17}$$

Pela definição atual, para cada população se tem uma função escalar negativa de potencial, de p-variáveis. Para transformar estas funções em densidades de probabilidade correspondentes, é necessário fazer uma transformação não-linear para torná-las finitas e integrar por todo espaço para encontrar a constante de normalização, e dividir a função transformada por esta constante (para garantir o valor unitário da integral da densidade). Cada ponto do espaço p-dimensional agora pode ser classificado como pertencente aquela população cuja função densidade tem valor máximo neste ponto. Por outro lado, se para a transformação não-linear entre o potencial e a função densidade de probabilidade escolhemos uma função não-crescente (e.g, $f(V) = V^2/(1 + V^2)$ para $V < 0$ e $f(V) = 0$ para $V \geq 0$), a população com o menor valor de potencial terá o maior valor da densidade de probabilidade. Podemos também estimar as densidades independentemente, considerando o potencial V_{ind} apenas em relação aos elementos de uma população e aplicarmos a transformação $V_{ind}^2/(1+V_{ind}^2)$. Assim, para a divisão do espaço p-dimensional em regiões correspondentes às populações individuais não é necessário calcular as densidades, é suficiente observar o valor mais baixo do potencial em cada ponto.

Para ilustrarmos o que foi apresentado, vamos considerar o caso simples onde temos três populações, Π_1, com observações $X_{11} = (2, 2)$, $X_{12} = (2, 8)$, Π_2 com observações $X_{21} = (7, 7)$ e $X_{22} = (9, 8)$ e Π_3 com observações $X_{31} = (5, 5)$, $X_{32} = (7, 3)$. Os potenciais resultantes para as populações Π_1, Π_2 e Π_3 estão representados na figura 3.7.a. Assim, todas as observações (pontos no $\mathbf{R}^2$) cujo potencial resultante em relação a população Π_1 é menor do que Π_2 e Π_3, são classificadas como sendo da população Π_1, todas as observações com o menor potencial resultante em relação a população Π_2 são classificadas como sendo da população Π_2, e semelhante observações com o menor potencial resultante em relação a população Π_3 são classificadas como sendo da população Π_3.

A regra discriminante de potencial (RDP) é definida calculando-se o potencial resultante

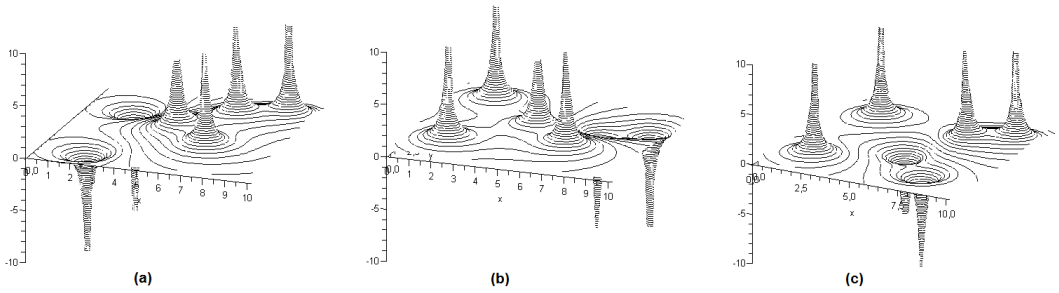

Figura 3.7: Gráfico do potencial resultante em relação a população Π_1 (a), gráfico do potencial resultante em relação a população Π_2 (b), e gráfico do potencial resultante em relação a população Π_3 (c).

para a observação $\mathbf{x}$ em relação a cada população, equação 3.17, e classificando $\mathbf{x}$ pertencente a população do menor potencial resultante. Desta forma temos

$$\text{RDP: Alocar } \mathbf{x} \text{ para a população } \Pi_0 \text{ onde } \Pi_0 = \min(V_{\Pi_1}, V_{\Pi_2}, \ldots, V_{\Pi_g}) \tag{3.18}$$

Analisando a regra discriminante de potencial vemos que ela difere da regra discriminante de núcleo na estimação das densidades. Enquanto no método de núcleo as densidades são estimadas com base em uma medida de erro, no método de potencial não se faz uso deste recurso, e neste caso as densidades estimadas podem não refletir precisamente as densidades reais. Assim, no método de potencial a estimação das densidades é voltada para o problema de classificação, pois em caso contrário, o método seria semelhante ao método de núcleo.

3.6.1 Algoritmo para Análise Discriminante de Potencial

Como no método de potencial para a classificação de um novo padrão são necessárias n computações para determinar o potencial resultante para cada classe, a complexidade computacional do algoritmo do método é $O(n)$, e o procedimento tem os seguintes passos.

1. Para cada observação do conjunto de teste calcule o potencial resultante em relação a todas as classes Π_j, $j = 1, \cdots, g$, utilizando a expressão (3.17);

2. Utilize os potenciais encontrados em (1) para fazer a classificação segundo a expressão (3.18);

3. Após classificar todas as observações estime a taxa de erro.

Como uma ilustração do método de potencial consideremos o cenário demonstrativo da tabela 3.1, cujas respectivas curvas de níveis das densidades são apresentas na figura 3.8. A partir de uma amostra de treinamento simulada com $n = 40$, com 20 observações para cada classe, são estimadas as densidades de forma independente pelo método de potencial, figura 3.9, e as regiões de classificação, figura 3.10. Vemos então que as densidades estimadas estão sobrepostas, e nas regiões onde uma função densidade estimada é mais elevada que outra, são as regiões onde a regra de classificação atribui observações como pertencentes a esta classe. Podemos então interpretar as regiões de classificação da figura 3.10 como as projeções das densidades estimadas no plano xy, segundo o critério que para cada ponto do plano apenas uma densidade é projetada, a que tem maior valor neste ponto.

Tabela 3.1: Distribuições consideradas no cenário demonstrativo.

Cenário	Distribuições
Demostrativo	$\Pi_1 : f_1 \sim \frac{1}{2}N\left(\begin{bmatrix}-\frac{3}{2}\\-\frac{3}{2}\end{bmatrix};\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\-\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right)+\frac{1}{2}N\left(\begin{bmatrix}\frac{1}{2}\\\frac{1}{2}\end{bmatrix};\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\-\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right);$ $\Pi_2 : f_2 \sim \frac{1}{2}N\left(\begin{bmatrix}\frac{3}{2}\\\frac{3}{2}\end{bmatrix};\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\-\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right)+\frac{1}{2}N\left(\begin{bmatrix}-\frac{1}{2}\\-\frac{1}{2}\end{bmatrix};\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\-\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right)$

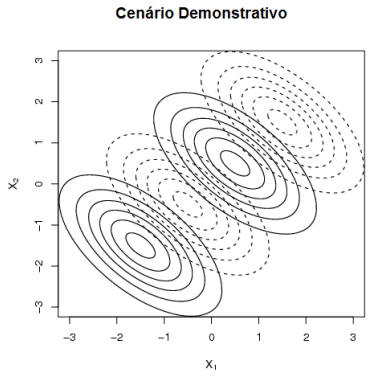

Figura 3.8: Curvas de nível das densidades considerada no cenário demostrativo: linhas sólidas – Π_1, linhas tracejadas – Π_2.

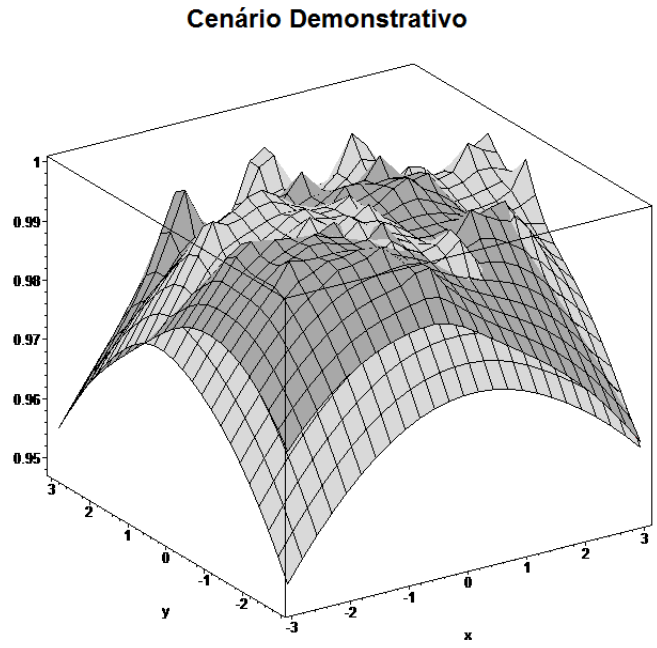

Figura 3.9: Densidades estimadas a partir do método de potencial para as populações do cenário demonstrativo : cor escura – Π_1, cor clara – Π_2.

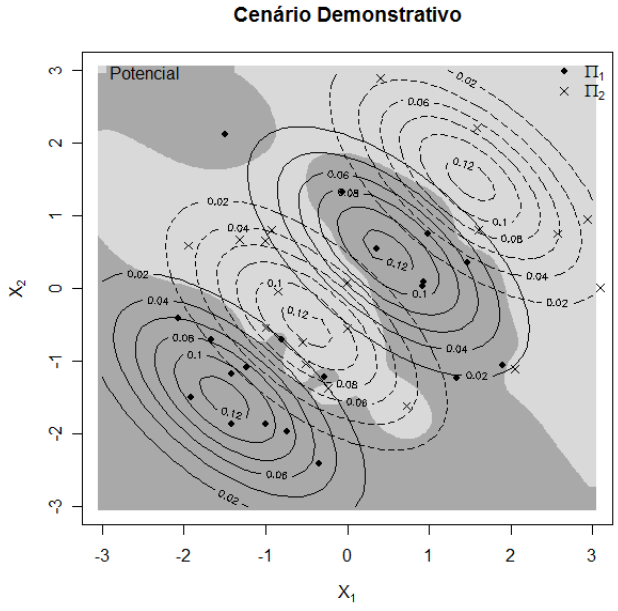

Figura 3.10: Regiões de classificação devida ao método de potencial para as populações do cenário demonstrativo a partir da amostra de treinamento simulada.

CAPÍTULO 4

Avaliação do Desempenho de um Classificador

4.1 Introdução

Dentre os vários critérios utilizados para se avaliar o desempenho de um classificador, a discriminabilidade da regra de classificação (como ele classifica dados fora do conjunto de treinamento) é uma forma usual. Como existem muitas medidas de discriminabilidade, a mais comum é a *taxa de erro* ou *taxa de má-classificação* [Webb (2002)], que representa a proporção de padrões que foram incorretamente classificados. Geralmente, é muito difícil de se obter uma expressão analítica para a taxa de erro, e portanto esta deve ser estimada a partir dos dados disponíveis. Existe uma vasta literatura sobre a taxa de erro, porém esta tem uma desvantagem por ser uma simples medida de desempenho, atribuindo pesos iguais tanto para os padrões que foram corretamente classificados, quanto para os incorretamente classificados (correspondendo a uma função perda do tipo zero-um). Em adição ao cálculo da taxa de erro, podemos também construir uma matriz de *confusão* ou matriz de *má-classificação*. O (i, j)-ésimo elemento desta matriz é o número de padrões da classe i que foram classificados como pertencentes a classe j.

Seja $F(\mathbf{x})$ a regra de decisão construída utilizando o conjunto de treinamento T, definido na seção 1.1, para o classificador F, e seja $L(y, F(\boldsymbol{x}))$ a *função perda 0-1* definida por

$$L(y, F(\boldsymbol{x})) = \begin{cases} 0 & \text{se } y = F(\boldsymbol{x}) \\ 1 & \text{caso contrário.} \end{cases} \tag{4.1}$$

A taxa de erro aparente, e_A, ou a taxa de erro de treinamento é obtida utilizando o conjunto de treinamento para estimar a taxa de erro, ou seja,

$$e_A = n^{-1} \sum_{i=1}^{n} L(y_i, F(\boldsymbol{x}_i)). \tag{4.2}$$

Este procedimento de estimação da taxa de erro aparente é consistente, mas viciado [Johnson & Wichern (1998)] e tende a subestimar a verdadeira taxa de erro Err $= E[L(Y, F(\boldsymbol{X}))]$, que é a probabilidade esperada de classificar incorretamente um padrão selecionado aleatoriamente. Essa probabilidade é a taxa de erro em um conjunto de teste independente, infinitamente grande, retirado de uma população com a mesma distribuição do conjunto de treinamento. O classificador tende a se adaptar aos dados de treinamento, fazendo com que o erro aparente ou erro de treinamento seja uma estimativa otimista da taxa de erro verdadeira [Ferreira (2007)]. Assim, em problemas de super-ajustamento, a taxa de erro aparente será nula.

Vamos considerar três métodos de avaliação do desempenho de um classificador, tendo em vista as vantagens e desvantagens do uso de cada um. O primeiro será o procedimento holdout discutido na seção (4.2). Em seguida o cross-validation, discutido na seção (4.3). E por último, o bootstrap na seção (4.4).

4.2 Holdout

O método holdout divide os dados em dois subconjuntos mutuamente excludentes, um que será utilizado para construir a regra de classificação (amostra de treinamento), e o outro para se avaliar o desempenho da regra (amostra de teste ou validação). A desvantagem desse método está em fazer uso ineficiente dos dados (usando apenas parte para treinar o classificador), e fornecer estimativas viesadas pessimistas [Jain et al. (2000),Webb (2002)]. Para que este método seja eficiente, é recomendável que se deixe, "a parte", de 25 a 50% dos elementos da amostra original [Mingoti (2005)], desta forma, o método não pode ser utilizado em amostras pequenas. Outra característica do método holdout é que ele é melhor que o erro aparente para amostras grandes.

4.3 Cross-validation

Este procedimento, que vem sendo bastante utilizado por pesquisadores na estimação de erros de predição [Efron (1983); Efron & Tibshirani (1993); Stone (1974); Stone (1977)], também conhecido como método de *validação cruzada*, ou *leave-one-out*, calcula o erro utilizando $n-1$ elementos amostrais no conjunto de treinamento, e utiliza o elemento que ficou fora para o conjunto de teste. Isto é repetido para todos n subconjuntos de tamanho $n-1$. Este método estima diretamente a taxa de erro verdadeira, que é o erro de generalização $\text{Err} = E[L(Y, F(\boldsymbol{X}))]$ quando aplicamos o classificador F a uma amostra de teste independente com mesma distribuição que o conjunto de treinamento. Para n grande, o método é computacionalmente caro, requerendo a construção de n classificadores. Portanto, é aproximadamente não viciado, embora tenha uma grande variância [Jain et al. (2000),Webb (2002)]. Denotando por $F^{-(i)}$ o classificador construído sem a observação $\mathbf{x}_i$, então o estimador cross-validation da taxa de erro é

$$e_{CV} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, F^{-(i)}(\boldsymbol{x}_i)). \tag{4.3}$$

O *método rotacional* ou *k-fold cross validation* particiona o conjunto de treinamento em k subconjuntos, utilizando $k-1$ para o treinamento e testando o subconjunto que ficou fora. Este procedimento é repetido para cada um dos k subconjuntos e a taxa de erro final será uma combinação das taxas de erro calculadas em cada uma das k partes. Se $k = n$ nós temos o método cross-validation padrão. Este método é intermediário entre o holdout e o cross-validation, tendo um vício menor que o procedimento holdout, e menos computacionalmente intensivo quando comparado ao cross-validation.

4.4 Bootstrap

A técnica bootstrap é composta por uma classe de procedimentos de amostragem da distribuição empírica, com reposição, para gerar conjuntos de observações que podem ser utilizados para a correção de viés. Introduzida por Efron (1979), esta técnica recebeu considerável atenção na literatura durante as últimas décadas. Ela fornece uma estimação não paramétrica do viés e da variância de um estimador e, como um método de estimação da taxa de erro, ela tem se

mostrado superior a muitas outras técnicas. Embora seja computacionalmente intensiva, é uma técnica muito atrativa, e houve muitos desenvolvimentos desde a abordagem básica [Efron & Tibshirani (1993); Davison & Hinkley (1997)].

Vamos considerar o procedimento bootstrap para estimar a correção de viés da taxa de erro aparente. Suponha que dispomos de uma amostra de treinamento $T = \{\boldsymbol{x}_j, y_j\}_{j=1}^n$ e seja $\hat{G}$ a distribuição empírica. Sob a amostra conjunta ou misturada, $\hat{G}$ é a distribuição com massa $1/n$ em cada observação $\mathbf{x}_i$, $i = 1, \cdots, n$. Sob a amostra separada, $\hat{G}_i$ é a distribuição com massa $1/n_i$ na observação $\mathbf{x}_i$ da classe C_i (n_i padrões na classe C_i). O procedimento é como segue

1. Gere um novo conjunto de dados (amostra bootstrap) $T^{*(B)}$ de acordo com a distribuição empírica;

2. Construa o classificador utilizando $T^{*(B)}$;

3. Calcule a taxa de erro aparente para esta amostra e denote isto por $\tilde{e}_A$;

4. Calcule a verdadeira taxa de erro para este classificador (tendo o conjunto T como a população inteira) e denote isto por $\tilde{e}_c$;

5. Calcule $w_b = \tilde{e}_A - \tilde{e}_c$;

6. Repita os passos 1-5 B vezes;

7. O viés bootstrap da taxa de erro aparente é

$$W_{boot} = E[\tilde{e}_A - \tilde{e}_c]$$

   onde a esperança é com respeito ao mecanismo de amostragem que gerou os conjuntos $T^{*(B)}$, isto é

$$W_{boot} = \frac{1}{B}\sum_{b=1}^{B} w_b$$

8. A versão corrigida pelo viés da taxa de erro aparente é dada por

$$e_A^{(B)} = e_A - W_{boot} \tag{4.4}$$

No passo 1, sob a amostra misturada, n observações independentes são geradas da distribuição $\hat{F}$; algumas destas podem ser repetidas na formação do conjunto $T^{*(B)}$ e pode acontecer que uma ou mais classes não sejam representadas na amostra bootstrap. Sob a amostra separada, n_i observações são geradas usando $\hat{F}_i$, $i = 1, \cdots, C$. Assim, todas as classes são representadas na amostra bootstrap na mesma proporção dos dados originais.

Conforme Webb (2002), o número de amostras bootstrap, B, usado para estimar W_{boot} pode ser limitado por considerações computacionais, porém para a estimação da taxa de erro ele pode ser considerado como sendo da ordem de 25-100.

O estimador bootstrap $e_A^{(B)}$, equação (4.4), não fornece em geral boas estimativas. A razão é que as amostras bootstrap estão atuando como conjuntos de treinamento, enquanto que o conjunto de treinamento original está atuando como conjunto de teste e estes dois conjuntos terão observações em comum. Como conseqüência, os classificadores ficam super-ajustados e fornecem estimativas da taxa de erro irrealisticamente boas [Friedman et al. (2000)].

Para obter uma melhor estimativa bootstrap, utiliza-se um procedimento que imita o método cross-validation, com as observações que não aparecem nas amostras bootstrap. O procedimento para cada observação faz com que apenas consideremos as predições feitas a partir de amostras bootstrap que não contém aquela observação. Assim, define-se a estimativa *bootstrap leave-one-out* da taxa de erro por

$$e_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, F^{*(b)}(\boldsymbol{x}_i)), \tag{4.5}$$

onde C^{-i} é o conjunto de índices das amostras bootstrap b que não contém a observação i, e $\left|C^{-i}\right|$ é o número de tais amostras. No cálculo de e_0 nós temos que escolher um número de amostras bootstrap suficientemente grande para assegurar que todos os $\left|C^{-i}\right|$ sejam maiores que zero, ou então nós podemos apenas deixar de fora os termos em (4.5) correspondentes aos $\left|C^{-i}\right|$ que são zero. Este estimador soluciona o problema do super-ajustamento sofrido por $e_A^{(B)}$, contudo será ainda uma estimativa viesada (para cima) da taxa de erro verdadeira.

O estimador ".632" proposto por Efron (1983) é uma combinação linear do erro aparente e do estimador bootstrap leave-one-out da taxa de erro. Este estimador tenta corrigir o viés do

estimador $e_A^{(B)}$, sendo definido por

$$e_{.632} = 0,368 \times e_a + 0,632 \times e_0. \tag{4.6}$$

A derivação do estimador .632 é complexa; intuitivamente ele tenta trazer o valor da estimativa bootstrap leave-one-out para próximo do erro aparente, reduzindo assim seu viés. O uso da constante .632 está relacionado com o valor aproximado da probabilidade da i-ésima observação pertencer à b-ésima amostra bootstrap de tamanho n, ou seja,

$$\begin{aligned} P\{i\text{-ésima observação} \in b\text{-ésima amostra bootstrap}\} &= 1 - \left(1 - \frac{1}{n}\right)^n \\ &\approx 1 - e^{-1} \\ &= 0,632. \end{aligned}$$

O estimador .632 tem bom desempenho em situações de subajustamento, porém pode ter problemas nos casos de super-ajustamento dos classificadores. Um caso similar ao que acontece com o classificador construído pelo método de potencial, que é super-ajustado, é discutido por Friedman et al. (2000) quando se considera o estimador .632. Suponha que nós temos duas classes de iguais tamanhos, com "alvos" independentes dos rótulos das classes, e nós aplicamos o regra do vizinho mais próximo. Temos que $e_A = 0$, $e_0 = 0,5$ e $e_{.632} = 0,632 \times 0,5 = 0,316$. Contudo a verdadeira taxa de erro é 0,5.

Pode-se melhorar o estimador .632 tendo em conta a quantidade de super-ajustamento. Primeiro nós definimos γ para ser a *taxa de erro de não-informação*: esta é a taxa de erro de nossa regra de predição se as entradas e os rótulos das classes são independentes. Uma estimativa de γ é obtida avaliando a regra de predição sobre todas as possíveis combinações dos rótulos y_i e preditores $\mathbf{x}_{i'}$

$$\hat{\gamma} = \sum_{i=1}^{N} \sum_{i'=1}^{N} L(y_i, f(\mathbf{x}_{i'})). \tag{4.7}$$

Por exemplo, considere um problema de classificação dicotômico: seja $\hat{p}_1$ a proporção observada de respostas y_i iguais a 1, e seja $\hat{q}_1$ a proporção observada de predições $f(\mathbf{x}_{i'})$ iguais a 1. Então

$$\hat{\gamma} = \hat{p}_1(1 - \hat{q}_1) + (1 - \hat{p}_1)\hat{q}_1. \tag{4.8}$$

Com uma regra como o primeiro vizinho mais próximo para qual $\hat{q}_1 = \hat{p}_1$ o valor de $\hat{\gamma}$ é $2\hat{p}_1(1-\hat{p}_1)$. A generalização para o caso de múltiplas categorias de (4.8) é $\hat{\gamma} = \sum_l \hat{p}_l(1-\hat{q}_l)$.

Usando isto, a *taxa relativa de super-ajustamento* é definida para ser

$$\hat{R} = \frac{e_0 - e_A}{\hat{\gamma} - e_A}, \tag{4.9}$$

Uma quantidade que varia de 0, se não há super-ajustamento ($e_0 = e_A$) a 1, se o super-ajustamento é igual ao valor de não-informação $\hat{\gamma} - e_A$. Finalmente, nós definimos o estimador ".632+" [Efron & Tibshirani (1997)] por

$$\begin{aligned} e_{.632+} &= (1-\hat{w}) \times e_A + \hat{w} \times e_0 \\ \text{com } \hat{w} &= \frac{0,632}{1-0,632\hat{R}}. \end{aligned} \tag{4.10}$$

O peso w varia de 0.632 se $\hat{R} = 0$ a 1 se $\hat{R} = 1$, e $e_{.632+}$ varia de $e_{.632}$ a e_0. Para o problema do discriminante de potencial com os rótulos das classes independentes das entradas, $\hat{w} = \hat{R} = 1$, então $e_{.632+} = 0$, que tem o correto valor esperado de 0,5. Em outros problemas com menor grau de super-ajustamento, $e_{.632+}$ ficará entre e_A e e_0.

CAPÍTULO 5

Comparação entre Análise Discriminante de Núcleo e de Potencial

5.1 Introdução

Com o objetivo de comparar o desempenho dos métodos de análise discriminante e de núcleo, avaliamos a discriminabilidade da regra de classificação, de ambos os métodos, para dados simulados e reais. Assim, na seção 5.2, utilizamos simulação de Monte Carlo, sobre alguns cenários de interesse, para estimar a taxa de erro de classificação pelo método Holdout (seção 4.2). Na seção 5.3, consideramos um conjunto de dados reais, onde a estimação da taxa de erro é obtida utilizando-se o estimador bootstrap .632+ (seção 4.4).

5.2 Comparação para Dados Simulados

O estudo de simulação considerou seis cenários de interesse, que foram extraídos de Ferreira (2007). Os cenários são apresentados na tabela 5.2, e as curvas de níveis para cada cenário são mostradas na figura 5.1. Conforme podemos observar, cada cenário apresenta certa particularidade, e esta pode se refletir em um problema de classificação simples, cenários A, B e C e difícil, cenários D, E e F, para um classificador ótimo para conjuntos linearmente separáveis. Foram utilizadas 1000 réplicas de Monte Carlo, e amostras de treinamento de tamanhos 20, 50, 100, 500 e 1000, tendo cada grupo tamanhos iguais. Para cada cenário foi construído o classificador utilizando a amostra de treinamento gerada em cada réplica, com a avaliação do desempenho do

Tabela 5.1: Distribuições consideradas em cada cenário de simulação.

Cenário	Distribuição
A	$\Pi_1 : f_1 \sim N\left(\begin{bmatrix}1\\-1\end{bmatrix}; \begin{bmatrix}\frac{4}{9} & \frac{14}{45}\\ \frac{14}{45} & \frac{4}{9}\end{bmatrix}\right); \Pi_2 : f_2 \sim N\left(\begin{bmatrix}-1\\1\end{bmatrix}; \begin{bmatrix}\frac{4}{9} & 0\\ 0 & \frac{4}{9}\end{bmatrix}\right)$
B	$\Pi_1 : f_1 \sim N\left(\begin{bmatrix}-1\\1\end{bmatrix}; \begin{bmatrix}\frac{2}{3} & \frac{1}{5}\\ \frac{1}{5} & \frac{1}{3}\end{bmatrix}\right); \Pi_2 : f_2 \sim N\left(\begin{bmatrix}1\\-\frac{1}{2}\end{bmatrix}; \begin{bmatrix}\frac{2}{3} & \frac{1}{5}\\ \frac{1}{5} & \frac{1}{3}\end{bmatrix}\right)$
C	$\Pi_1 : f_1 \sim N\left(\begin{bmatrix}-1\\\frac{1}{2}\end{bmatrix}; \begin{bmatrix}\frac{2}{3} & \frac{1}{5}\\ \frac{1}{5} & \frac{4}{9}\end{bmatrix}\right); \Pi_2 : f_2 \sim N\left(\begin{bmatrix}1\\-\frac{1}{2}\end{bmatrix}; \begin{bmatrix}\frac{2}{3} & \frac{1}{5}\\ \frac{1}{5} & \frac{4}{9}\end{bmatrix}\right)$
D	$\Pi_1 : f_1 \sim \frac{1}{2}N\left(\begin{bmatrix}-\frac{3}{2}\\-\frac{3}{2}\end{bmatrix}; \begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right) + \frac{1}{2}N\left(\begin{bmatrix}\frac{1}{2}\\\frac{1}{2}\end{bmatrix}; \begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right);$ $\Pi_2 : f_2 \sim \frac{1}{2}N\left(\begin{bmatrix}\frac{3}{2}\\\frac{3}{2}\end{bmatrix}; \begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right) + \frac{1}{2}N\left(\begin{bmatrix}-\frac{1}{2}\\-\frac{1}{2}\end{bmatrix}; \begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right)$
E	$\Pi_1 : f_1 \sim \frac{1}{2}N\left(\begin{bmatrix}-\frac{3}{2}\\0\end{bmatrix}; \begin{bmatrix}\frac{1}{12} & \frac{1}{4}\\ \frac{1}{4} & 1\end{bmatrix}\right) + \frac{1}{2}N\left(\begin{bmatrix}\frac{3}{2}\\0\end{bmatrix}; \begin{bmatrix}\frac{1}{12} & \frac{1}{4}\\ \frac{1}{4} & 1\end{bmatrix}\right);$ $\Pi_2 : f_2 \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}; \begin{bmatrix}\frac{4}{9} & \frac{1}{5}\\ \frac{1}{5} & \frac{4}{9}\end{bmatrix}\right)$
F	$\Pi_1 : f_1 \sim \frac{1}{2}N\left(\begin{bmatrix}-\frac{3}{2}\\0\end{bmatrix}; \begin{bmatrix}\frac{2}{10} & \frac{1}{4}\\ \frac{1}{4} & \frac{3}{10}\end{bmatrix}\right) + \frac{1}{2}N\left(\begin{bmatrix}\frac{3}{2}\\0\end{bmatrix}; \begin{bmatrix}\frac{3}{10} & \frac{1}{4}\\ \frac{1}{4} & \frac{3}{10}\end{bmatrix}\right);$ $\Pi_2 : f_2 \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}; \begin{bmatrix}\frac{4}{5} & \frac{2}{5}\\ \frac{2}{5} & 1\end{bmatrix}\right)$

classificador feita utilizando-se uma amostra de teste, independente, de tamanho 1000, com 500 observações para cada grupo, também gerada em cada réplica.

A tabela 5.2 apresenta os resultados da simulação para o cenário A. Por se tratar de um problema de classificação relativamente simples, dado que a sobreposição das densidades é muito pequena, os métodos tendem a apresentar valores baixos para a taxa de erro estimada. Percebe-se no entanto, o melhor desempenho do método de potencial, especialmente para o tamanho de amostra de treinamento muito pequeno, $n = 20$, onde a taxa de erro estimada foi de 0,4479%, contra 7,6025% para o discriminante de núcleo. Nota-se também o baixo desvio-padrão apresentado pelo método de potencial, não variando muito quando o conjunto de treinamento cresce.

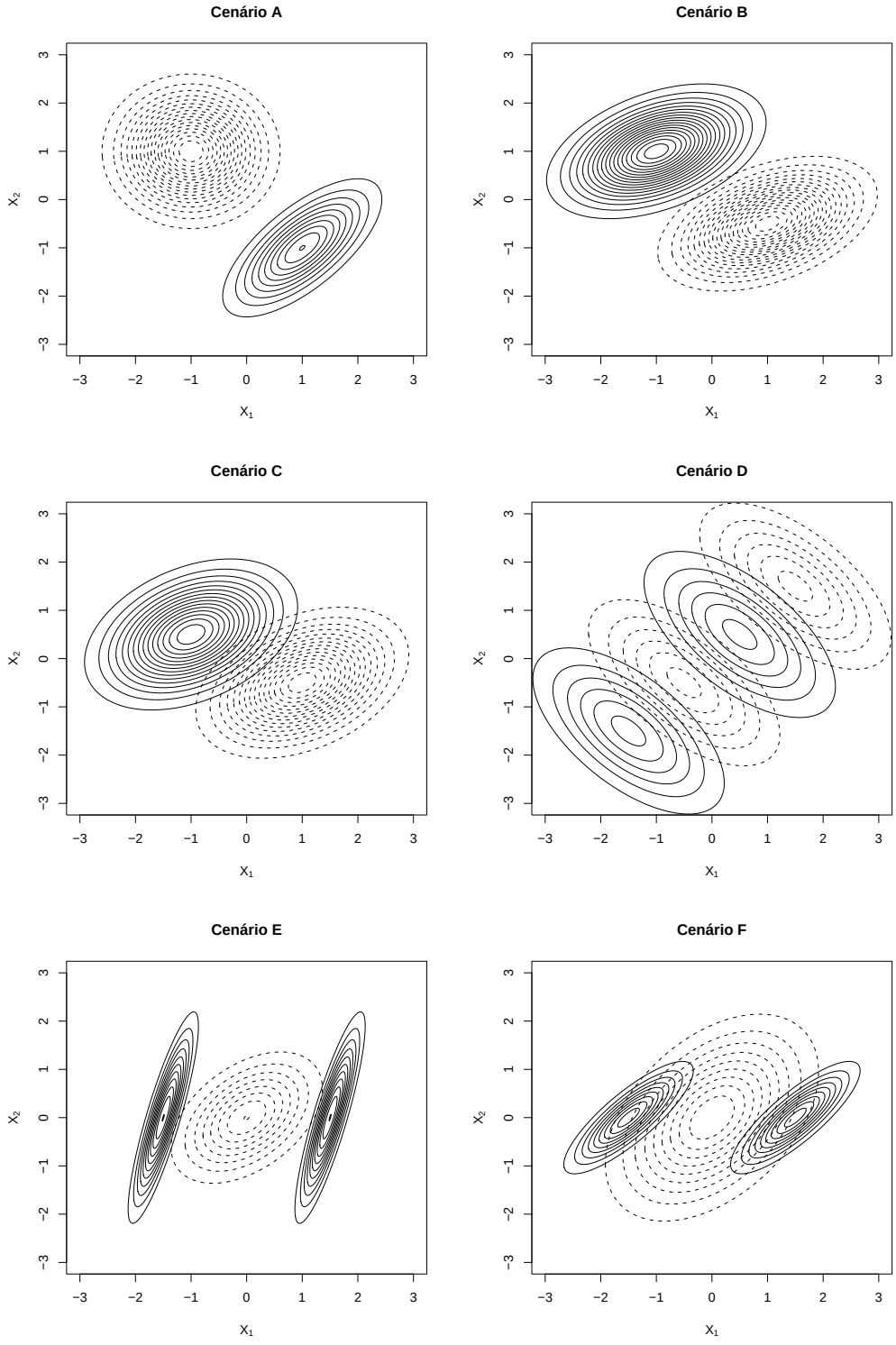

Figura 5.1: Curvas de nível das densidades consideradas em cada cenário de simulação: linhas sólidas – Π_1, linhas tracejadas – Π_2.

Tabela 5.2: Média e Desvio-padrão (D. P.) das taxas de erro (%) de classificação sobre as 1000 réplicas de Monte Carlo para as funções discriminantes de Núcleo e de Potencial considerando o Cenário A.

		$n = 20$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Discriminante de Núcleo	Média	7,6025	1,7939	0,7236	0,3707	0,3345
	D. P.	10,3176	3,4506	1,1373	0,2034	0,1852
Discriminante de Potencial	Média	0,7594	0,6327	0,5909	0,5515	0,5353
	D. P.	0,4479	0,3182	0,2782	0,2495	0,2418

Na tabela 5.3 temos o resultado da simulação para o cenário B. Também por se tratar de um problema de classificação relativamente simples, as estimativas da taxa de erro tendem a ser pequenas. Com os métodos apresentando desempenhos similares. Porém, mais uma vez observa-se que para o tamanho de amostra pequeno, o método de potencial tem um melhor desempenho, apresentando taxa de erro estimada de 1,6703% contra 10,1413% do método de núcleo, para $n = 20$, e pelo baixo desvio-padrão.

Tabela 5.3: Média e Desvio-padrão (D. P.) das taxas de erro (%) de classificação sobre as 1000 réplicas de Monte Carlo para as funções discriminantes de Núcleo e de Potencial considerando o Cenário B.

		$n = 20$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Discriminante de Núcleo	Média	10,1413	3,3496	1,6951	1,0619	1,0084
	D. P.	11,4162	4,3990	1,1379	0,3384	0,3279
Discriminante de Potencial	Média	1,6703	1,4024	1,2905	1,1519	1,1405
	D. P.	0,7347	0,4995	0,4014	0,3482	0,3499

Analisando a tabela 5.4, vemos que para o cenário C, a diferença de desempenho entre os dois métodos se dá principalmente quando $n = 20$ e $n = 50$, onde o método de potencial se sai consideravelmente melhor que o método de núcleo. Por exemplo, para $n = 20$, a taxa de erro estimada para o método de potencial é de 5,3480% e de 15,0591% para o método de núcleo, No entanto, em termos gerais, neste problema de classificação de complexidade fácil, ambos os métodos, como esperado, apresentam bons desempenhos.

O cenário D é o mais difícil, e o resultado da simulação para este cenário é apresentado na

Tabela 5.4: Média e Desvio-padrão (D. P.) das taxas de erro (%) de classificação sobre as 1000 réplicas de Monte Carlo para as funções discriminantes de Núcleo e de Potencial considerando o Cenário C.

		$n = 20$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Discriminante de Núcleo	Média	15,0591	7,3173	5,1660	4,0051	3,8934
	D. P.	11,5520	4,9048	1,9443	0,6370	0,6121
Discriminante de Potencial	Média	5,3480	4,7266	4,4152	4,1096	4,0703
	D. P.	1,4042	0,9336	0,7263	0,6473	0,6302

tabela 5.5. A alta superposição das densidades faz com que os métodos tenham dificuldade para alocar com eficiência as observações do conjunto de teste. Isso se reflete na taxa de erro estimada, sendo alta para os dois métodos e decrescendo com o tamanho do conjunto de treinamento. Em termos gerais, ambos os métodos apresentam desempenhos similares, com taxas de erro estimadas relativamente próximas, diferindo por no máximo 3%, para $n \geq 50$.

Tabela 5.5: Média e Desvio-padrão (D. P.) das taxas de erro (%) de classificação sobre as 1000 réplicas de Monte Carlo para as funções discriminantes de Núcleo e de Potencial considerando o Cenário D.

		$n = 20$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Discriminante de Núcleo	Média	33,9649	23,9493	19,2534	15,6282	15,2869
	D. P.	9,1111	6,4039	3,3594	1,2694	1,1568
Discriminante de Potencial	Média	29,8165	25,0999	22,3974	18,0783	17,2562
	D. P.	5,1491	4,0261	3,2016	1,7072	1,3482

Para o cenário E, os métodos de potencial e de núcleo obtiveram bons desempenhos nas simulações [tabela 5.6]. Embora se trate de um problema de classificação relativamente difícil, as taxas de erro estimadas para ambos os métodos, por exemplo em $n = 20$, foram de 16,8942% para o método de núcleo, e de 16,0193% para o método de potencial, decrescendo para valores próximos de 3% para $n = 1000$, mostrando que basicamente os métodos não diferem.

Por fim, a tabela 5.7 apresenta os resultados da simulação para o cenário F, que se caracteriza por ser um difícil problema de classificação. O desempenho dos métodos é relativamente bom, principalmente quando o tamanho da amostra de treinamento é grande. Analisando a diferença

Tabela 5.6: Média e Desvio-padrão (D. P.) das taxas de erro (%) de classificação sobre as 1000 réplicas de Monte Carlo para as funções discriminantes de Núcleo e de Potencial considerando o Cenário E.

		$n = 20$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Discriminante de Núcleo	Média	16,8942	7,8224	4,9369	3,4115	3,3147
	D. P.	12,0020	5,4887	2,4012	0,6022	0,5641
Discriminante de Potencial	Média	16,0193	9,4987	6,3359	3,9765	3,7493
	D. P.	5,4177	3,5729	1,9472	0,6866	0,6308

de desempenho entre os métodos, percebe-se que para $n \geq 50$, as diferenças entre as taxas de erro estimadas não excedem 2%, ou seja, os métodos têm desempenhos similares.

Tabela 5.7: Média e Desvio-padrão (D. P.) das taxas de erro (%) de classificação sobre as 1000 réplicas de Monte Carlo para as funções discriminantes de Núcleo e de Potencial considerando o Cenário F.

		$n = 20$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
Discriminante de Núcleo	Média	27,7767	18,9139	15,1291	12,4111	12,1559
	D. P.	9,9746	5,7672	2,804127	1,0677	1,0619
Discriminante de Potencial	Média	22,6427	17,8150	15,5821	13,6194	13,3951
	D. P.	5,8435	3,0535	1,8241	1,1566	1,1232

Como forma de ilustrar os métodos de núcleo e de potencial, para cada cenário considerado na tabela 5.2, foi gerada uma amostra de treinamento de tamanho 100 (50 observações para cada classe), com a finalidade de desenhar a região de classificação para os dois métodos a partir dessa amostra. Assim, nas figuras 5.2 a 5.13, estão as regiões de classificação devidas aos métodos de potencial e de núcleo. Para cada cenário temos duas figuras, sendo que a primeira é a região de classificação para o conjunto $\{(x_1, x_2) \in \mathbf{R}^2 | -3 \leq x_1 \leq 3, -3 \leq x_2 \leq 3\}$ e a segunda é para o conjunto $\{(x_1, x_2) \in \mathbf{R}^2 | -5 \leq x_1 \leq 5, -5 \leq x_2 \leq 5\}$. A utilidade dessas duas figuras para cada cenário é para mostrar aspectos locais e globais, e observar regiões atípicas de classificação. Por exemplo, para o cenário A vemos que o método de potencial tende a dividir a região do plano em duas partes, conforme a visão próxima da figura 5.2(a) e da visão afastada da figura 5.3(a). Já o método de núcleo resulta em uma região de classificação atípica, figuras 5.2(b) e 5.3(b).

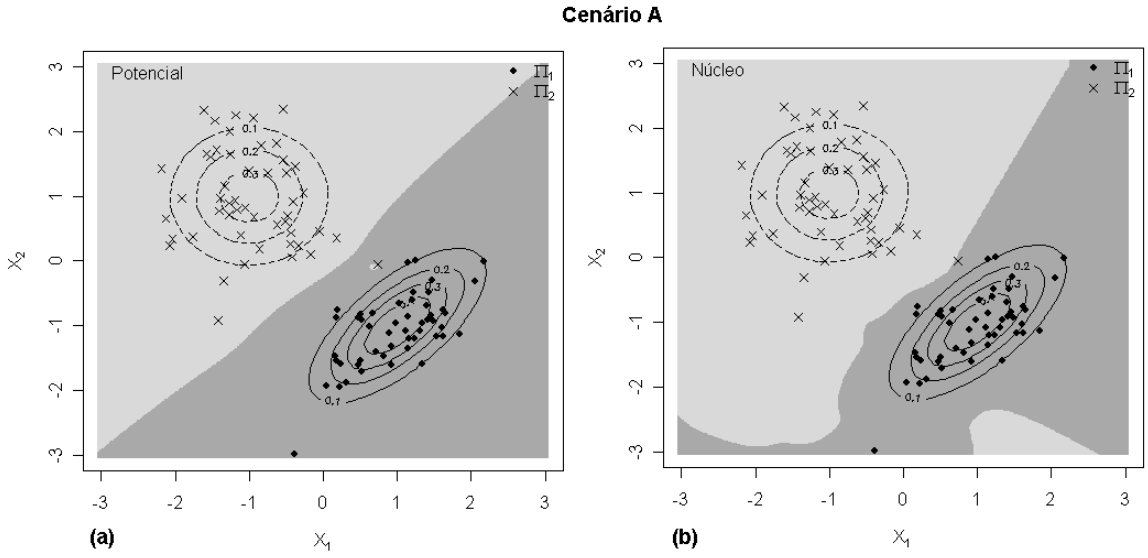

Figura 5.2: Visão próxima das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário A.

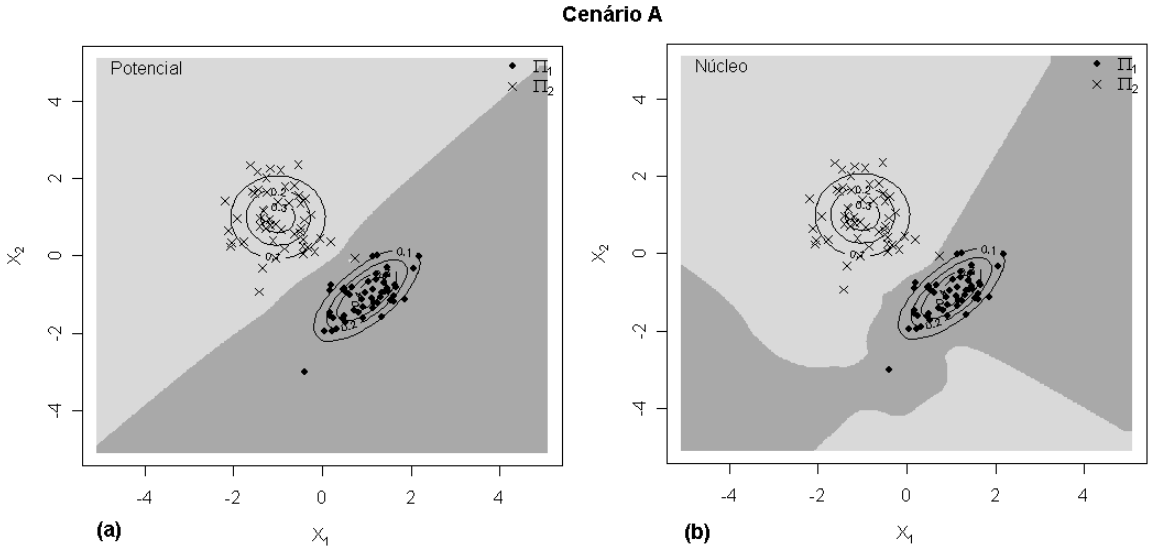

Figura 5.3: Visão afastada das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário A.

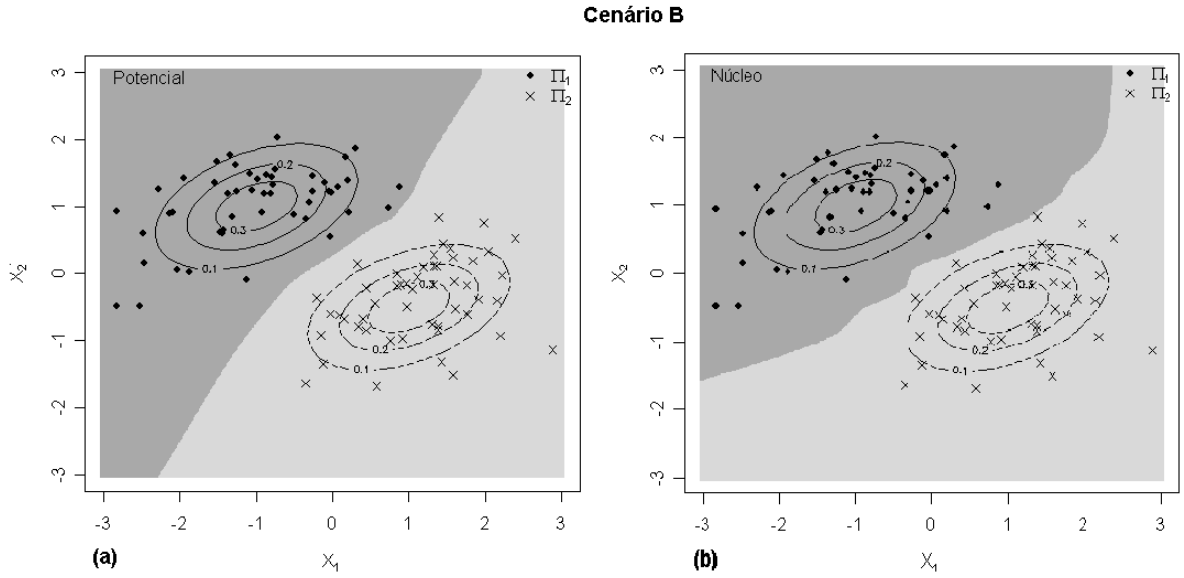

Figura 5.4: Visão próxima das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário B.

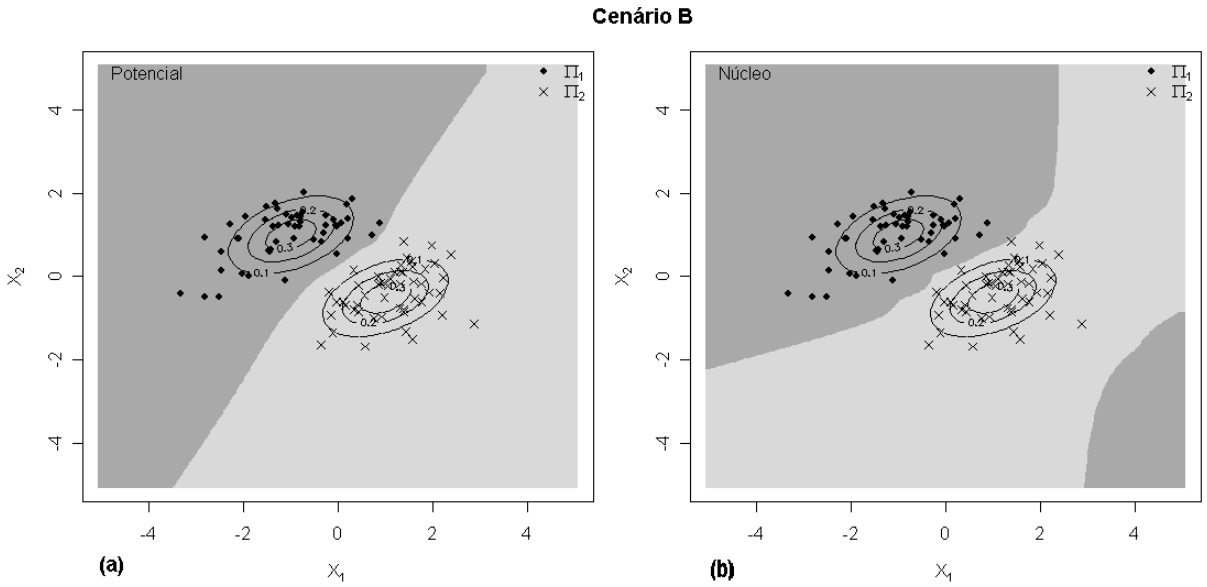

Figura 5.5: Visão afastada das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário B.

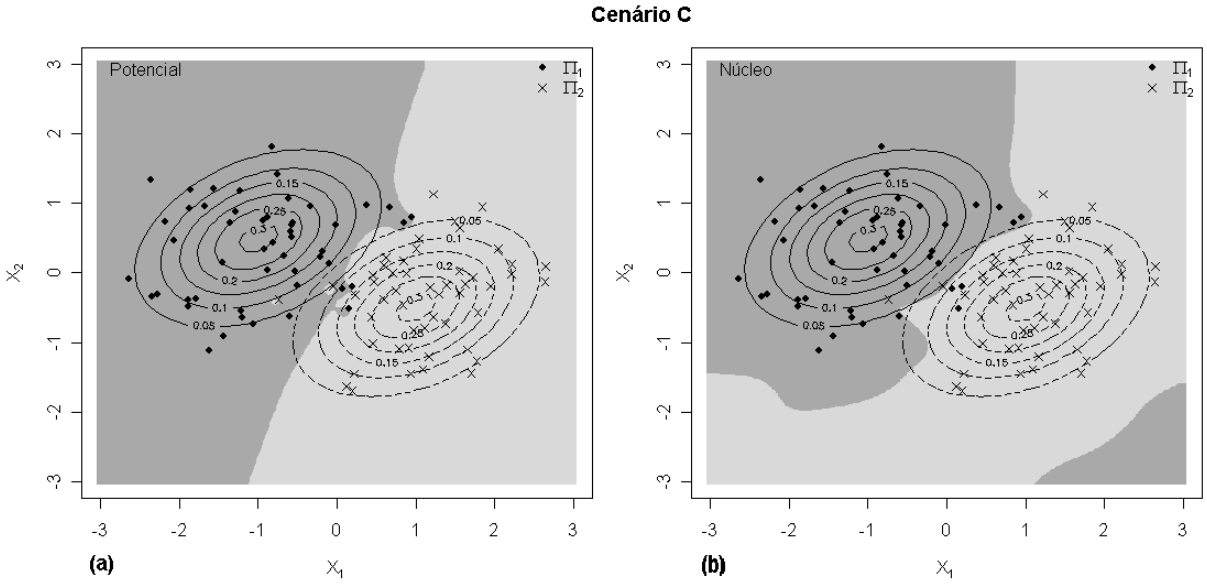

Figura 5.6: Visão próxima das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário C.

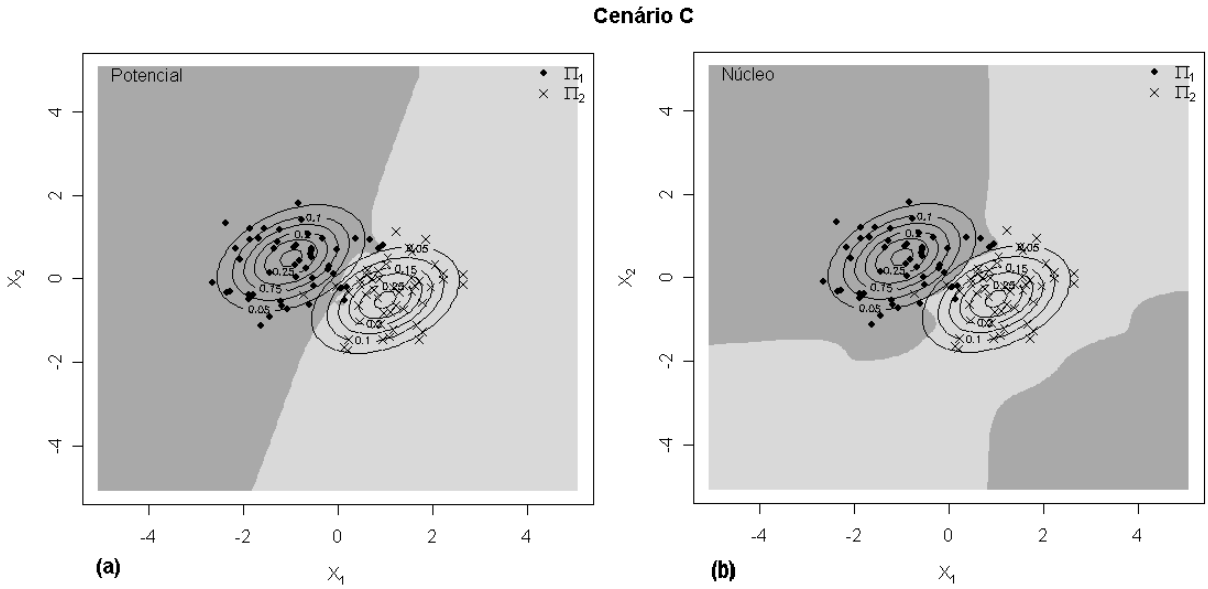

Figura 5.7: Visão afastada das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário C.

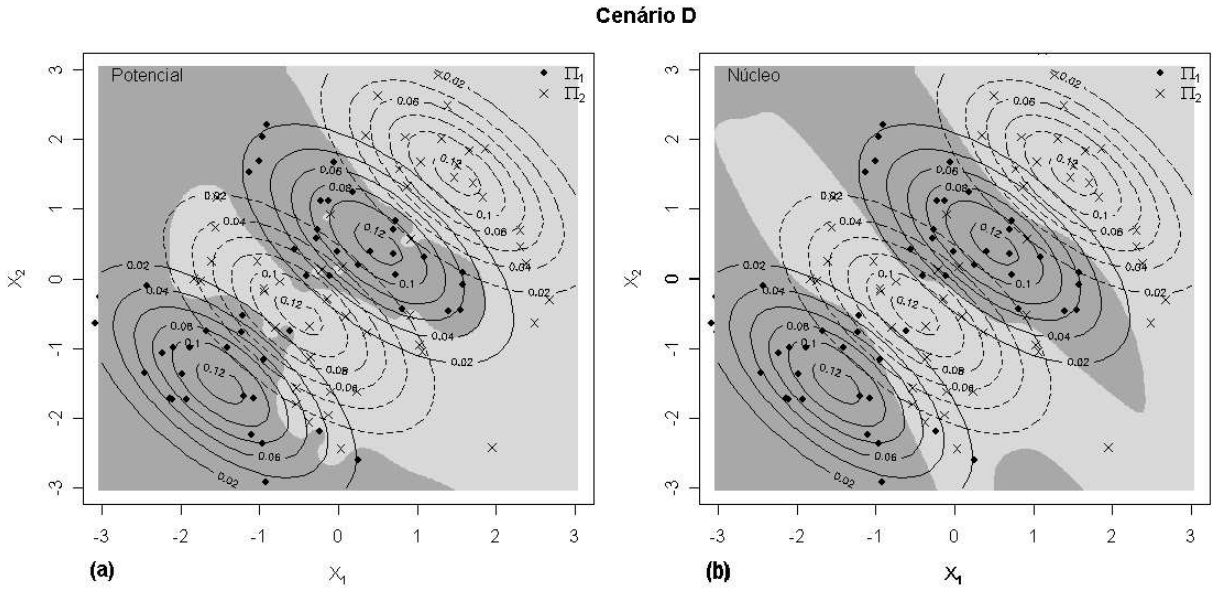

Figura 5.8: Visão próxima das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário D.

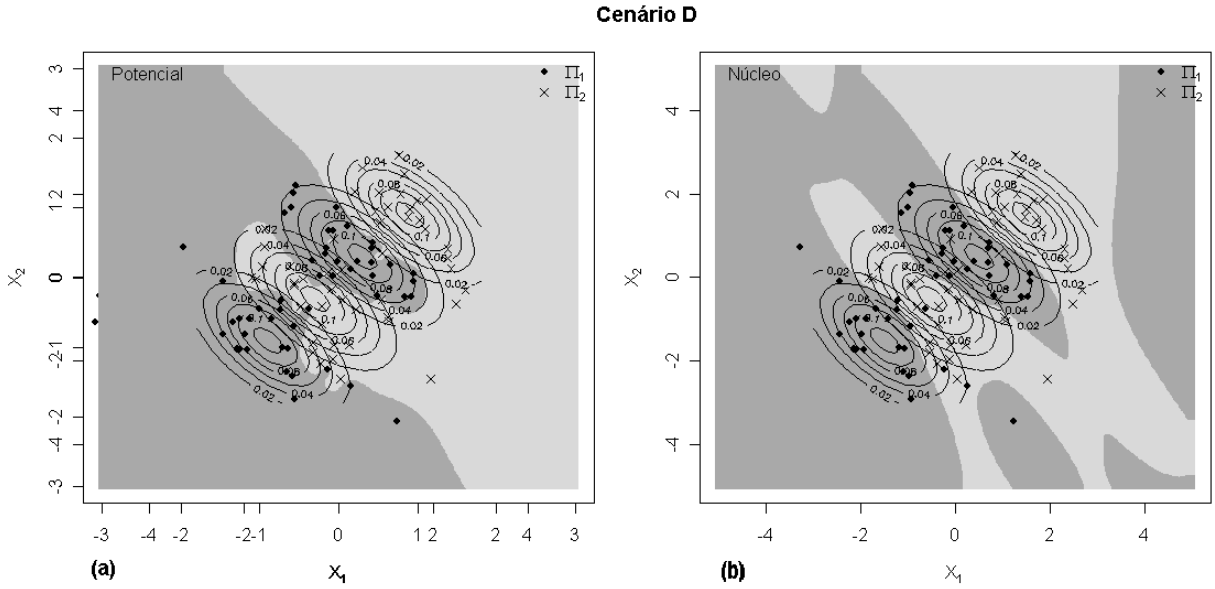

Figura 5.9: Visão afastada das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário D.

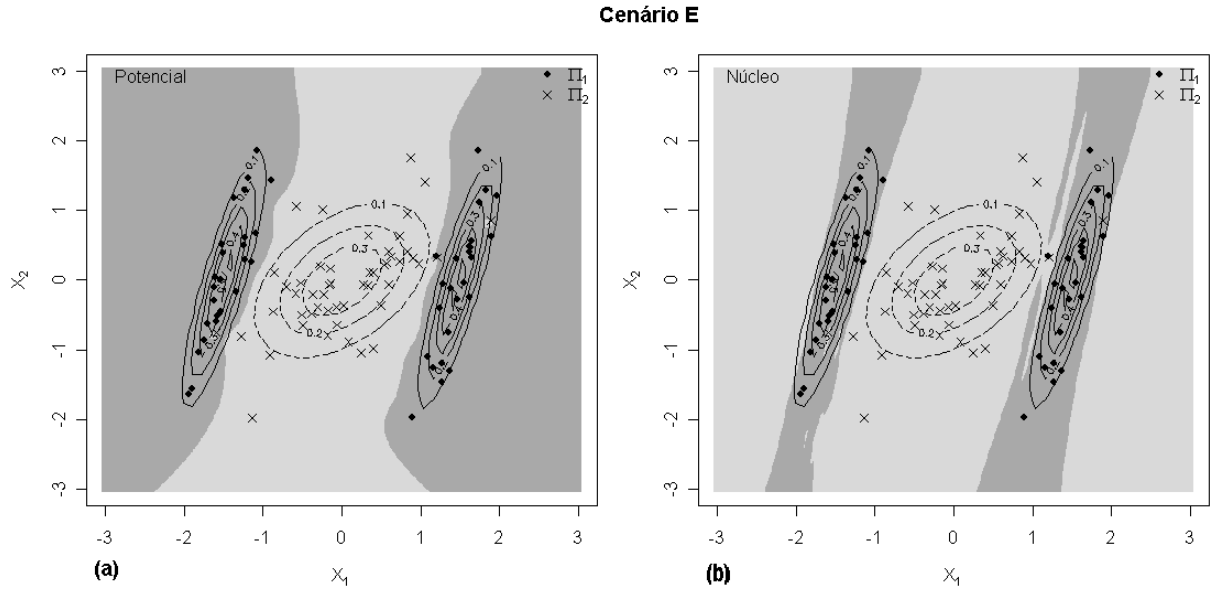

Figura 5.10: Visão próxima das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário E.

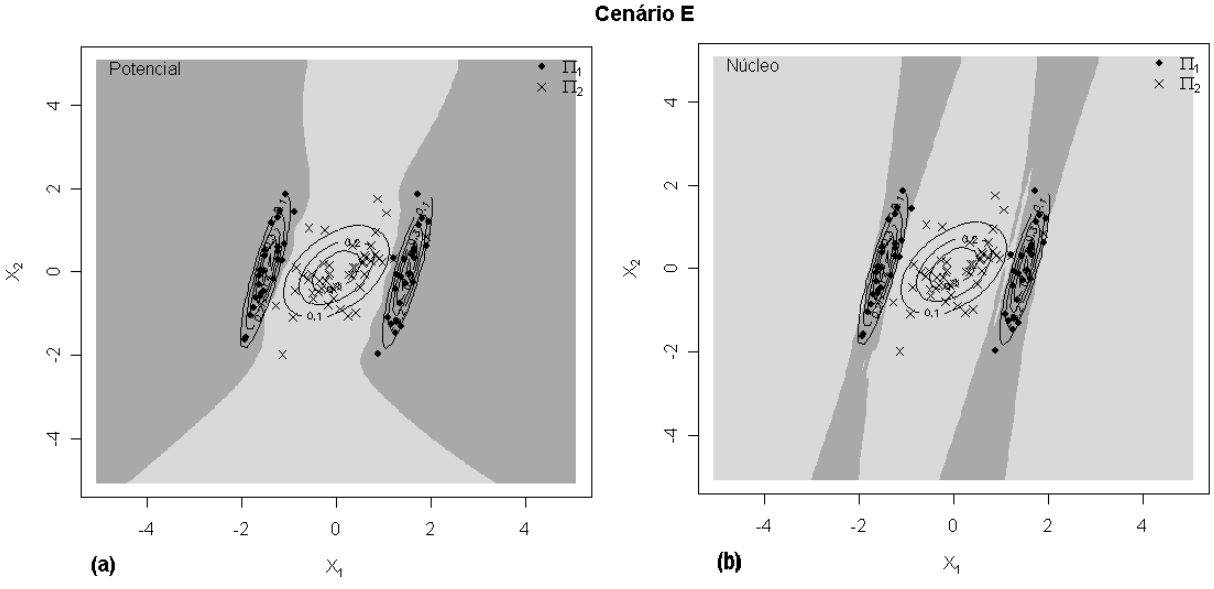

Figura 5.11: Visão afastada das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário E.

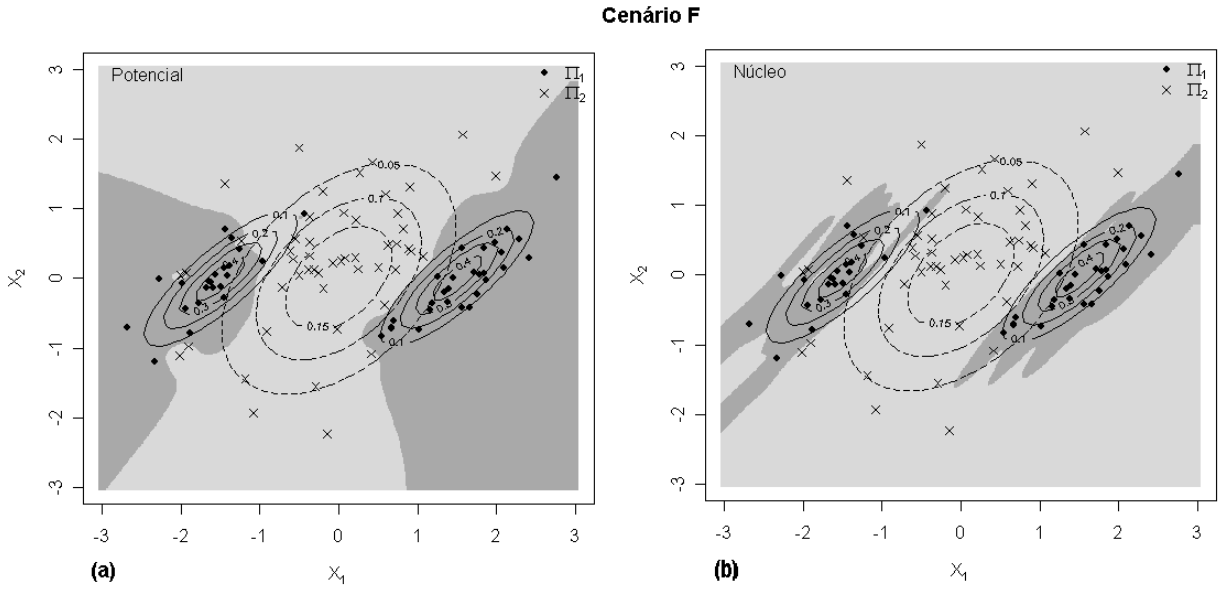

Figura 5.12: Visão próxima das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário F.

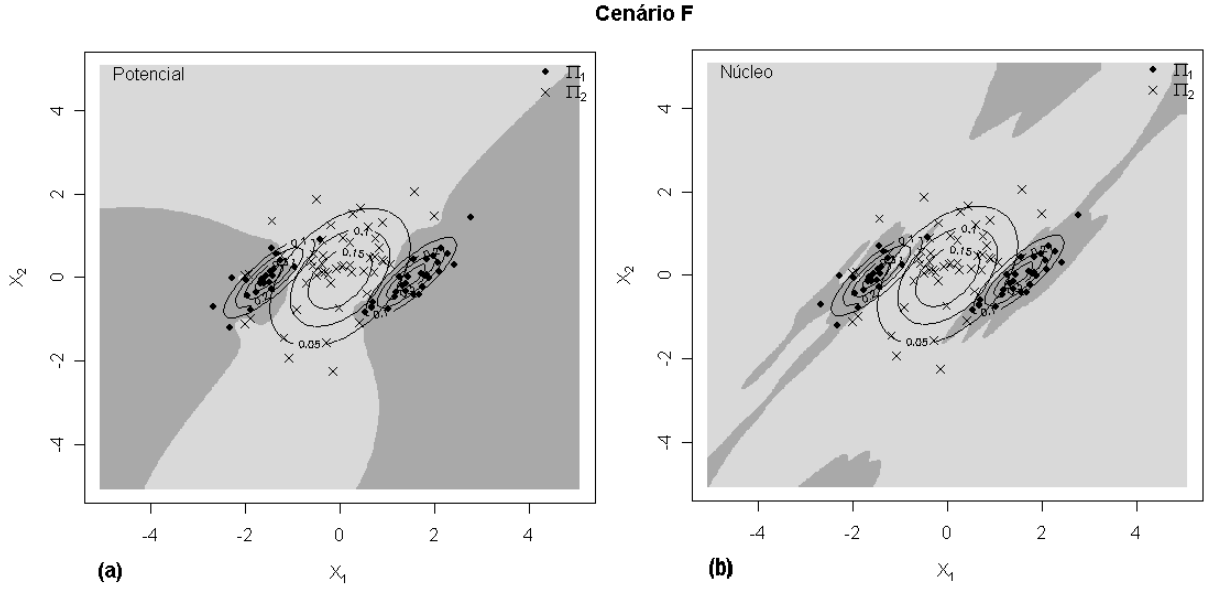

Figura 5.13: Visão afastada das regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b), com base em uma amostra treinamento ($n = 100$) das populações do cenário F.

Comparando os resultados das regiões de classificação mostradas nas figuras 5.2 a 5.13, pode-se concluir que o método de potencial oferece resultados intuitivamente mais aceitáveis do que o método de núcleo, em particular com aumento da escala, embora de fato seja difícil quantificar este efeito.

5.3 Comparação para Dados Reais

O conjunto de dados considerado é conhecido como Salmon [Johnson & Wichern (1998), Ferreira (2007)] e trata-se de medidas do diâmetro dos anéis de 100 salmões no primeiro ano de crescimento em água doce (X_1) e em água do mar (X_2) capturados em águas do estado americano do Alasca ou em águas do Canadá, com 50 observação para cada população. Na tabela 5.8 estão as taxas de erro, estimadas pelo procedimento bootstrap .632+, dos métodos de potencial e de núcleo. Assim, o método de potencial apresenta a menor taxa de erro, $6,30\%$, embora que a diferença para o método de núcleo seja pequena, $1,5\%$. A figura 5.14 apresenta o conjunto de dados e as regiões de classificação devidas aos dois métodos. Novamente, o método de potencial oferece resultados intuitivamente mais aceitáveis do que o método de núcleo na escala maior.

Tabela 5.8: Taxas de erro obtidas para os métodos de potencial e de núcleo através do procedimento .632+ para o conjunto de dados salmão.

Função discriminante	Taxa de Erro (%)
Potencial	6,30
Núcleo	7,80

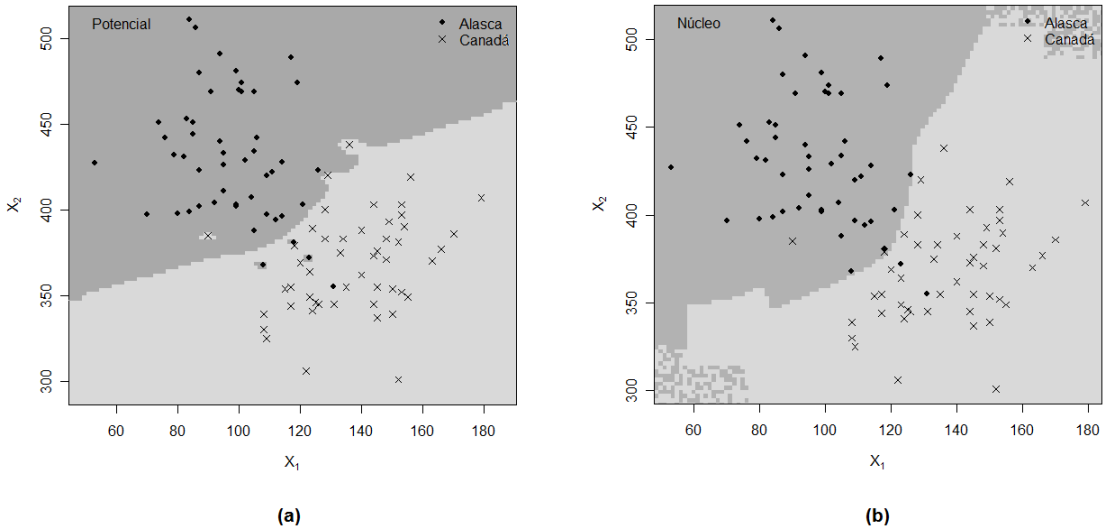

Figura 5.14: Regiões de classificação devidas aos métodos de potencial (a) e de núcleo (b) para o conjunto de dados salmão.

CAPÍTULO 6

Conclusões e Direções para Trabalhos Futuros

6.1 Conclusões

A proposição de uma nova técnica não é uma tarefa simples e conseqüentemente a difusão e o uso acontece de maneira lenta. Existe toda uma etapa de estudos para aceitação pela comunidade científica, até que a técnica se torne comum, tanto no meio acadêmico, como no meio profissional. Acreditamos que por este processo irá passar a análise discriminante de potencial. Por ser uma técnica nova, deve receber algumas críticas, que vão desde a formalização com base em sistemas físicos, até as densidades estimadas que podem não corresponder às densidades reais.

Em termos gerais, a análise discriminante de núcleo apresentou desempenho satisfatório e semelhante, em simulações e dados reais, a uma técnica já estabelecida e amplamente difundida – a análise discriminante de núcleo. Sendo que a principal diferença entre os dois métodos está na forma funcional utilizada para estimar as densidades e construir a regra de classificação. Isto é um indicativo de que o método de potencial tem uma boa formalização e conseqüentemente boas propriedades. Dentre as características observadas no método de potencial podemos citar: facilidade de implementação computacional; complexidade computacional de ordem $O(n)$; a regra de classificação é superajustada e mesmo assim tem uma boa capacidade de generalização; toda observação utilizada para construir a regra cria uma fronteira de classificação; embora sejam específicas para a classificação, as densidades estimadas não dependem de parâmetros de

suavização e a normalização pode ser feita numericamente; pode ser aplicado em problemas em que os conjuntos das classes não são linearmente separáveis, como também em problemas em que o vetor de observações pertence a uma dimensão muito alta. Finalmente, o método de potencial oferece regiões de classificação intuitivamente aceitáveis com aumento da escala.

Em suma, no contexto desta dissertação, a análise discriminante de potencial é recomendada como um método alternativo para problemas de classificação supervisionada, restando ao pesquisador decidir pela utilização da técnica.

6.2 Direções para Trabalhos Futuros

Nesta seção apresentamos direções para trabalhos futuros, alguns já estão sendo feitos, com o método de potencial. Uma vez que esta dissertação foi importante por propor, formalizar e mostrar que o método funciona. Assim, outras características do método devem ser avaliadas, e algumas são citadas a seguir:

- As densidades estimadas por diferentes transformações não-lineares;
- As regiões atípicas de classificação;
- A confiabilidade da regra de classificação;
- Comparação em relação ao classificador ótimo de Bayes;
- Desempenho em relação a outras técnicas (redes neurais, árvore de classificação, etc).

APÊNDICE A

Programas de Simulação

As rotinas em linguagem S utilizadas nesta dissertação para as simulações, gráficos e análise de dados são apresentadas neste apêndice.

A.1 Simulações

```
###################################
# Programa de simulacao com mis- #
# tura de normais para compara-  #
# cao entre os classificadores   #
# de nucleo e potencial          #
###################################


rm(list=ls(all=TRUE))


# Bibliotecas utilizadas

library(MASS)
library(ks)
library(class)


###################################
#          Parametros             #
###################################


R <- 1000 ## Numero de replicas de Monte Carlo
N <- 50 ## Tamanho da amostra de treinamento
n <- 1000 ## Tamanho da amostra de teste

y <- sort(rep(c(-1,1), N/2))
ytst <- sort(rep(c(-1,1), n/2))
```

```
##################################
#    Funcao para o potencial     #
#    resultante em um ponto      #
##################################


pot<-function(p,x,y){ #p=ponto onde se calcula o potencial,
p<-as.vector(p)       #x=amostra de treinamento, y=label{-1,1} de x
L1<-matrix(c(rep(1,length(p))),length(p),1)
Xl<-cbind(x,y)
m1<-Xl[1:length(y[y==-1]),1:length(p)]
m2<-Xl[(length(y[y==-1])+1):length(y),1:length(p)]
x1<-t(t(m1)-p)
x2<-t(t(m2)-p)
return(-sum(1/(sqrt((((x1)^2)%*%L1)))) + sum(1/(sqrt((((x2)^2)%*%L1)))))
}


##################################
#    Funcao para classificao     #
#    pelo metodo de potencial    #
##################################


classify<-function(x.tst,x.group,y.label){
#x.tst=vetor de observacoes a ser classificado,
#x.group=amostra de treinamento, y.label=label{-1,1} de x.group
res.cla<-numeric(nrow(x.tst))
for(i in 1:(nrow(x.tst))){
if(pot(x.tst[i,1:(ncol(x.tst))],x.group,y.label)< 0)
res.cla[i]<- -1
else
res.cla[i]<- 1
}
return(res.cla)
}



##################################
#      Cenarios simulados        #
##################################


## CENARIO 1 ##

mus1.1<-c(1, -1)
mus2.1<-c(-1, 1)

sigmas1.1<-matrix(c(4/9,14/45,14/45,4/9),2,2)
sigmas2.1<-matrix(c(4/9,0,0,4/9),2,2)

erro.kda.1 <- numeric(R)
erro.pm.1<- numeric(R)


## CENARIO 2 ##

mus1.2<-c(-1, 1)
mus2.2<-c( 1, -1/2)


sigmas1.2<-matrix(c(2/3,1/5,1/5,1/3),2,2)
sigmas2.2<-matrix(c(2/3,1/5,1/5,1/3),2,2)
```

```
erro.kda.2 <- numeric(R)
erro.pm.2<- numeric(R)


## CENARIO 3 ##

mus1.3<-c(-1, 1/2)
mus2.3<-c(1, -1/2)


sigmas1.3<-matrix(c(2/3,1/5,1/5,4/9),2,2)
sigmas2.3<-matrix(c(2/3,1/5,1/5,4/9),2,2)

erro.kda.3 <- numeric(R)
erro.pm.3<- numeric(R)


## CENARIO 4 ##

mus1.4<-rbind(c(-3/2, -3/2), c(1/2,1/2))
mus2.4<-rbind(c(3/2, 3/2), c(-1/2, -1/2))

sigmas1.4<-rbind(matrix(c(4/5,-1/2,-1/2,4/5),2,2),
                 matrix(c(4/5,-1/2,-1/2,4/5),2,2))
sigmas2.4<-rbind(matrix(c(4/5,-1/2,-1/2,4/5),2,2),
                 matrix(c(4/5,-1/2,-1/2,4/5),2,2))

props.4 <- c(1/2,1/2)

erro.kda.4 <- numeric(R)
erro.pm.4<- numeric(R)


## CENARIO 5##

mus1.5<-rbind(c(-3/2, 0), c(3/2, 0))
mus2.5<-c( 0, 0)

sigmas1.5<- rbind(matrix(c(1/12,1/4,1/4,1),2,2),
                  matrix(c(1/12,1/4,1/4,1),2,2))
sigmas2.5<- matrix(c(4/9,1/5,1/5,4/9),2,2)

props.5 <- c(1/2,1/2)

erro.kda.5 <- numeric(R)
erro.pm.5<- numeric(R)


## CENARIO 6 ##

mus1.6<- rbind(c(-3/2, 0), c(3/2, 0))
mus2.6<- c(0,0)

sigmas1.6<- rbind(matrix(c(3/10,1/4,1/4,3/10),2,2),
                  matrix(c(3/10,1/4,1/4,3/10),2,2))
sigmas2.6<- matrix(c(4/5,2/5,2/5,1),2,2)

props.6 <- c(1/2,1/2)

erro.kda.6 <- numeric(R)
erro.pm.6<- numeric(R)
```

```
###################################
#     Laco de Monte Carlo         #
###################################

set.seed(1234567)
inicio <- proc.time()[1]

##Laco de Monte Carlo
for (j in 1 : R){

##CENARIO 1 ##
x.1 <- rbind(rmvnorm.mixt(N/2, mus1.1, sigmas1.1,1),
               rmvnorm.mixt(N/2, mus2.1, sigmas2.1,1))

xtst.1<- rbind(rmvnorm.mixt(n/2, mus1.1, sigmas1.1,1),
                 rmvnorm.mixt(n/2, mus2.1, sigmas2.1,1))

## Kernel Method
H.1 <- Hkda(x.1, y, bw = "lscv")
fit.ke.1 <- kda(x.1, y, H.1, y=xtst.1)
erro.kda.1[j] <- compare(fit.ke.1, ytst)$error

## Potential Method
fit.p.1 <- classify(xtst.1,x.1,y)
erro.pm.1[j]<-compare(fit.p.1,ytst)$error



##CENARIO 2 ##
x.2 <- rbind(rmvnorm.mixt(N/2, mus1.2, sigmas1.2,1),
               rmvnorm.mixt(N/2, mus2.2, sigmas2.2,1))

xtst.2<- rbind(rmvnorm.mixt(n/2, mus1.2, sigmas1.2,1),
                 rmvnorm.mixt(n/2, mus2.2, sigmas2.2,1))

## Kernel Method
H.2 <- Hkda(x.2, y, bw = "lscv")
fit.ke.2 <- kda(x.2, y, H.2, y=xtst.2)
erro.kda.2[j] <- compare(fit.ke.2, ytst)$error

## Potential Method
fit.p.2 <- classify(xtst.2,x.2,y)
erro.pm.2[j]<-compare(fit.p.2,ytst)$error



##CENARIO 3 ##
x.3 <- rbind(rmvnorm.mixt(N/2, mus1.3, sigmas1.3,1),
               rmvnorm.mixt(N/2, mus2.3, sigmas2.3,1))

xtst.3<- rbind(rmvnorm.mixt(n/2, mus1.3, sigmas1.3,1),
                 rmvnorm.mixt(n/2, mus2.3, sigmas2.3,1))

## Kernel Method
H.3 <- Hkda(x.3, y, bw = "lscv")
fit.ke.3 <- kda(x.3, y, H.3, y=xtst.3)
erro.kda.3[j] <- compare(fit.ke.3, ytst)$error

## Potential Method
fit.p.3 <- classify(xtst.3,x.3,y)
erro.pm.3[j]<-compare(fit.p.3,ytst)$error



## CENARIO 4 ##
```

```
x.4 <- rbind(rmvnorm.mixt(N/2, mus1.4, sigmas1.4,props.4),
               rmvnorm.mixt(N/2, mus2.4, sigmas2.4,props.4))

  xtst.4<- rbind(rmvnorm.mixt(n/2, mus1.4, sigmas1.4,props.4),
                 rmvnorm.mixt(n/2, mus2.4, sigmas2.4,props.4))

## Kernel Method
H.4 <- Hkda(x.4, y, bw = "lscv")
fit.ke.4 <- kda(x.4, y, H.4, y=xtst.4)
erro.kda.4[j] <- compare(fit.ke.4, ytst)$error

## Potential Method
fit.p.4 <- classify(xtst.4,x.4,y)
erro.pm.4[j]<-compare(fit.p.4,ytst)$error



##CENARIO 5 ##
x.5 <- rbind(rmvnorm.mixt(N/2, mus1.5, sigmas1.5,props.5),
               rmvnorm.mixt(N/2, mus2.5, sigmas2.5,1))

xtst.5<- rbind(rmvnorm.mixt(n/2, mus1.5, sigmas1.5,props.5),
                 rmvnorm.mixt(n/2, mus2.5, sigmas2.5,1))

## Kernel Method
H.5 <- Hkda(x.5, y, bw = "lscv")
fit.ke.5 <- kda(x.5, y, H.5, y=xtst.5)
erro.kda.5[j] <- compare(fit.ke.5, ytst)$error

## Potential Method
fit.p.5 <- classify(xtst.5,x.5,y)
erro.pm.5[j]<-compare(fit.p.5,ytst)$error



##CENARIO 6 ##
x.6 <- rbind(rmvnorm.mixt(N/2, mus1.6, sigmas1.6,props.6),
               rmvnorm.mixt(N/2, mus2.6, sigmas2.6,1))

xtst.6<- rbind(rmvnorm.mixt(n/2, mus1.6, sigmas1.6,props.6),
                 rmvnorm.mixt(n/2, mus2.6, sigmas2.6,1))

## Kernel Method
H.6 <- Hkda(x.6, y, bw = "lscv")
fit.ke.6 <- kda(x.6, y, H.6, y=xtst.6)
erro.kda.6[j] <- compare(fit.ke.6, ytst)$error

## Potential Method
fit.p.6 <- classify(xtst.6,x.6,y)
erro.pm.6[j]<-compare(fit.p.6,ytst)$error

}
tempo <- proc.time()[1]- inicio[1]


result.cen1 <- matrix(c(mean(erro.kda.1),mean(erro.pm.1),
sd(erro.kda.1),sd(erro.pm.1)),2,2)
rownames(result.cen1) <- c("KDA","PM")
colnames(result.cen1) <- c("Média","Desvio Padrão")

result.cen2 <- matrix(c(mean(erro.kda.2),mean(erro.pm.2),
sd(erro.kda.2),sd(erro.pm.2)),2,2)
rownames(result.cen2) <- c("KDA","PM")
colnames(result.cen2) <- c("Média","Desvio Padrão")
```

```
result.cen3 <- matrix(c(mean(erro.kda.3),mean(erro.pm.3),
sd(erro.kda.3),sd(erro.pm.3)),2,2)
rownames(result.cen3) <- c("KDA","PM")
colnames(result.cen3) <- c("Média","Desvio Padrão")

result.cen4 <- matrix(c(mean(erro.kda.4),mean(erro.pm.4),
sd(erro.kda.4),sd(erro.pm.4)),2,2)
rownames(result.cen4) <- c("KDA","Knn-1","PM")
colnames(result.cen4) <- c("Média","Desvio Padrão")

result.cen5 <- matrix(c(mean(erro.kda.5),mean(erro.pm.5),
sd(erro.kda.5),sd(erro.pm.5)),2,2)
rownames(result.cen5) <- c("KDA","PM")
colnames(result.cen5) <- c("Média","Desvio Padrão")

result.cen6 <- matrix(c(mean(erro.kda.6),mean(erro.pm.6),
sd(erro.kda.6),sd(erro.pm.6)),2,2)
rownames(result.cen6) <- c("KDA","PM")
colnames(result.cen6) <- c("Média","Desvio Padrão")

sink("Resultado_Monte_Carlo.txt")

cat("Resultado da Simulação Monte Carlo\n\n")

cat("R = ", R, "Numero de replicas de Monte Carlo\n")
cat("N = ", N, "Tamanho da amostra de treinamento\n")
cat("n = ", n, "Tamanho da amostra de teste\n")

cat("\n")
cat("\n")
cat("\n Tempo de execução: ",tempo,"\n")
cat("\n")
cat("\n")
cat("\n -------------------------\n")

cat("\n Cenario 1 \n")
result.cen1
cat("\n")
cat("Summary erro.kda.1\n")
summary(erro.kda.1)
cat("\n")
cat("Summary erro.pm.1\n")
summary(erro.pm.1)
cat("\n -------------------------\n")

cat("\n Cenario 2 \n")
result.cen2
cat("\n")
cat("Summary erro.kda.2\n")
summary(erro.kda.2)
cat("\n")
cat("Summary erro.pm.2\n")
summary(erro.pm.2)
cat("\n -------------------------\n")

cat("\n Cenario 3 \n")
result.cen3
cat("\n")
cat("Summary erro.kda.3\n")
summary(erro.kda.3)
cat("\n")
cat("Summary erro.pm.3\n")
summary(erro.pm.3)
cat("\n -------------------------\n")
```

```
cat("\n Cenario 4 \n")
result.cen4
cat("\n")
cat("Summary erro.kda.4\n")
summary(erro.kda.4)
cat("\n")
cat("Summary erro.pm.4\n")
summary(erro.pm.4)
cat("\n -------------------------\n")


cat("\n Cenario 5 \n")
result.cen5
cat("\n")
cat("Summary erro.kda.5\n")
summary(erro.kda.5)
cat("\n")
cat("Summary erro.pm.5\n")
summary(erro.pm.5)
cat("\n -------------------------\n")

cat("\n Cenario 6 \n")
result.cen6
cat("\n")
cat("Summary erro.kda.6\n")
summary(erro.kda.6)
cat("\n")
cat("Summary erro.pm.6\n")
summary(erro.pm.6)
cat("\n -------------------------\n")

sink()

write.matrix(erro.kda.1, file = "erro.kda.1.txt")
write.matrix(erro.kda.2, file = "erro.kda.2.txt")
write.matrix(erro.kda.3, file = "erro.kda.3.txt")
write.matrix(erro.kda.4, file = "erro.kda.4.txt")
write.matrix(erro.kda.5, file = "erro.kda.5.txt")
write.matrix(erro.kda.6, file = "erro.kda.6.txt")

write.matrix(erro.pm.1, file = "Erro.pm.1.txt")
write.matrix(erro.pm.2, file = "Erro.pm.2.txt")
write.matrix(erro.pm.3, file = "Erro.pm.3.txt")
write.matrix(erro.pm.4, file = "Erro.pm.4.txt")
write.matrix(erro.pm.5, file = "Erro.pm.5.txt")
write.matrix(erro.pm.6, file = "Erro.pm.6.txt")
```

A.2 Aplicação

```
salmao <- read.table("salmon.dat")
attach(salmao)
y <- salmao[, 1]
x <- salmao[, 2:3]
(salmao.err.pda <- boot.632.pda(x, y, B))
(salmao.err.kda <- boot.632.kda(x, y, B))



################################################
#          Função  bootstrap .632+             #
################################################

boot.632plus.kda <- function(x, y, nboot) {
    x <- as.matrix(x)
    n <- length(y)
    saveii <- NULL
    priori <- as.vector(table(y) / n)
    H0 <- Hkda(x, y)
    yhat0 <- kda(x, y, H0, y=x, prior.prob=priori)
    err.meas <- function(y, yhat) {1 * (yhat != y)}
    app.err <- mean(err.meas(y, yhat0))
    err1 <- matrix(0, nrow = nboot, ncol = n)
    err2 <- numeric(nboot)
    set.seed(321)
    for (b in 1 : nboot) {
        ii <- sample(1 : n, replace = TRUE)
        saveii <- cbind(saveii, ii)
        Hb <- Hkda(x[ii, ], y[ii])
        yhat1 <- kda(x[ii, ], y[ii], Hb, y=x[ii, ], prior.prob=priori)
        yhat2 <- kda(x[ii, ], y[ii], Hb, y=x, prior.prob=priori)
        err1[b, ] <- err.meas(y, yhat2)
        err2[b] <- mean(err.meas(y[ii], yhat1))
    }
    optim <- mean(apply(err1, 1, mean) - err2)
    junk <- function(x, i) {sum(x == i)}
    e0 <- 0
    for (i in 1 : n) {
        o <- apply(saveii, 2, junk, i)
        if (sum(o == 0) == 0)
            cat("increase nboot for computation of the .632 estimator",
                fill = TRUE)
        e0 <- e0 + (1 / n) * (sum(err1[o == 0, i]) / sum(o == 0))
    }
    err.632 <- 0.368 * app.err + 0.632 * e0
    sum.ni <- 0
    for (i in 1:n){
    for (j in 1:n){
        if (y[i] == yhat0[j])
            (sum.ni <- sum.ni)
        else
        (sum.ni <- sum.ni + 1)
    }}

    hat.gamma <- sum.ni/(n*n) # taxa de erro de nao-informacao
    hat.R <- (e0 - app.err)/(hat.gamma - app.err) #taxa relativa de super-ajustamento
    hat.w <- 0.632/(1-0.368 * hat.R)
    err.632plus <- (1 - hat.w) * app.err + hat.w * e0
    result <- list(app.err, optim, err.632, hat.gamma, hat.R, hat.w, err.632plus)
    names(result) <- list("app.err", "optim", "err.632", "hat.gamma", "hat.R", "hat.w", "err.632+")
    return(result)
```

```
}


boot.632plus.pda <- function(x, y, nboot) {
    x <- as.matrix(x)
    n <- length(y)
    saveii <- NULL
    yhat0 <- classify(x,x,y)
    err.meas <- function(y, yhat) {1 * (yhat != y)}
    app.err <- mean(err.meas(y, yhat0))
    err1 <- matrix(0, nrow = nboot, ncol = n)
    err2 <- numeric(nboot)
    set.seed(321)
    for (b in 1 : nboot) {
        ii <- sample(1 : n, replace = TRUE)
        ii<-sort(ii)
  saveii <- cbind(saveii, ii)
         yhat1 <- classify(x[ii, ], x[ii, ], y[ii])
         yhat2 <- classify(x, x[ii, ], y[ii])
  err1[b, ] <- err.meas(y, yhat2)
        err2[b] <- mean(err.meas(y[ii], yhat1))
    }
    optim <- mean(apply(err1, 1, mean) - err2)
    junk <- function(x, i) {sum(x == i)}
    e0 <- 0
    for (i in 1 : n) {
        o <- apply(saveii, 2, junk, i)
        if (sum(o == 0) == 0)
            cat("increase nboot for computation of the .632 estimator",
                fill = TRUE)
        e0 <- e0 + (1 / n) * (sum(err1[o == 0, i]) / sum(o == 0))
    }
    err.632 <- 0.368 * app.err + 0.632 * e0
    sum.ni <- 0
    for (i in 1:n){
    for (j in 1:n){
        if (y[i] == yhat0[j])
            (sum.ni <- sum.ni)
        else
        (sum.ni <- sum.ni + 1)
    }}

    hat.gamma <- sum.ni/(n*n) # taxa de erro de nao-informacao
    hat.R <- (e0 - app.err)/(hat.gamma - app.err) #taxa relativa de super-ajustamento
    hat.w <- 0.632/(1-0.368 * hat.R)
    err.632plus <- (1 - hat.w) * app.err + hat.w * e0
    result <- list(app.err, optim, err.632, hat.gamma, hat.R, hat.w, err.632plus)
    names(result) <- list("app.err", "optim", "err.632", "hat.gamma", "hat.R", "hat.w", "err.632+")
    return(result)
}

(salmao.err.pda <- boot.632plus.pda(x, y, B))
(salmao.err.kda <- boot.632plus.kda(x, y, B))
```

A.3 Gráficos

```
source("potencia_discriminant.R")
desenha<-function(x,y,t,m){
x1 <- seq(-3, 3, length = t)
y1 <- seq(-3, 3, length = t)
#plot(x1, y1, type='n',xlab = expression(X[1]),ylab = expression(X[2]))
x1t<-numeric(t*t)
y1t<-numeric(t*t)
xy<-cbind(x1t,y1t)
ind <- 0
for (i in x1){
for (j in y1){
ind <- ind + 1
xy[ind,1] <- i
xy[ind,2] <- j
}
}
if (m == 1){
H <- Hkda(x, y, bw = "lscv")
fkres <- kda(x, y, H, y=xy)
ind <- 0
for (i in x1){
for (j in y1){
ind <- ind + 1
if (fkres[ind]==-1)
{points(i,j,col="darkgray",pch= 15)}
else
{points(i,j, col="gray85",pch= 15)}
}
}
for(i in 1: (length(y[y==-1]))){
points(x[i,1],x[i,2],pch= 16)
}
for(i in (length(y[y==-1])+1):(length(y))){
points(x[i,1],x[i,2],pch= 4)
}
}
if (m == 2){
fpres <- classify(xy,x,y)
ind <- 0
for (i in x1){
for (j in y1){
ind <- ind + 1
if (fpres[ind]==-1)
{points(i,j,col="darkgray",pch= 15)}
else
{points(i,j, col="gray85",pch= 15)}
}
}
for(i in 1: (length(y[y==-1]))){
points(x[i,1],x[i,2],pch= 16)
}
for(i in (length(y[y==-1])+1):(length(y))){
points(x[i,1],x[i,2],pch= 4)
}
}
}
```

Referências Bibliográficas

Cavalcante, C. d. N. (2004), Avaliação do núcleo estimador na estimação de funções de densidades multivariadas, Master's thesis, Universidade Federal de Minas Gerais. ICEx. Estatística.

Davison, A. C. & Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, New York.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001), *Pattern classification*, 2nd. edn, John Wiley & Sons, New york.

Duong, T. (2004), Bandwidth selectors for multivariate kernel density estimation, PhD thesis, University of Western Australia, School of Mathematics and Statistics.

Duong, T. (2007), 'ks: Kernel desnsity estimation and kernel discriminant analysis for multivariate data in R', *of Statistical Software* **21**.

Efron, B. (1979), 'Bootstrap methods: another look at the jackknife', *Annals of Statistics* **7**, 1–26.

Efron, B. (1983), 'Estimating the error rate of a prediction rule: improvements on cross-validation', *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Efron, B. & Tibshirani, R. (1997), 'Improvements on cross-validation: The .632+ bootstrap method', *Jornal of the American Statistical Association* **92**, 548 – 560.

Epanechnikov, V. A. (1969), 'Non-parametric estimation of a multivariate probability density', *Theory of Probability and its Applications* **14**, 153–158.

Ferreira, M. R. P. (2007), Análise discriminante clássica e de núcleo: avaliações e algumas contribuições relativas aos métodos boosting e bootstrap, Master's thesis, Universidade Federal de Pernambuco. CCEN. Estatística.

Fix, E. & Hodges, J. L. (1951), Discriminatory analysis — nonparametric discrimination: consistency properties, Technical report, US Air Force School of Aviation Medicine, Randolph Field, Texas.

Friedman, J., Hastie, T. & Tibshirani, R. (2000), *The Elements of Statistical Learning*, Springer, New York.

Fukunaga, K. (1972), *Introduction to statistical pattern recognition*, Orlando Academic Press.

Ghosh, A. K. & Chaudhuri, P. (2004), 'Optimal smoothing in kernel discriminant analysis', *Statistica Sinica* **14**, 457–483.

Hall, P. & Kang, K.-H. (2005), 'Bandwidth choice for nonparametric classification', *The Annals of Statistics* **33**, 284–306.

Halliday, D. (2006), *Fundamentos de física 7.ed.*, Rio de Janeiro : LTC.

Hoel, P. G., Port, S. C. & Stone, C. J. (1971), *Introduction to the Theory os Statistics*, Boston : Houghton Mifflin.

Jain, A. K., Duin, R. P. & Mao, J. (2000), 'Statistical pattern recognition: A review', *IEEE Transactions On Pattern Analysis and Machine Intelligence* **22**, 4–37.

Johnson, R. A. & Wichern, D. W. (1998), *Applied multivariate statistical analysis*, 4th. edn, Prentice-Hall, New York.

Kohn, A. F. (1998), *Reconhecimento de Padrões. Uma abordagem estatística.*, Sao Paulo : Universidade de Sao Paulo, Escola Politecnica.

Lamport, L. (1994), *A Document Preparation System*, 2nd. edn, Addison-Wesley, Massachusetts.

Mingoti, S. A. (2005), *Análise de dados através de métodos de estatística multivariada : uma abordagem aplicada*, Belo Horizonte : Editora UFMG.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974), *Introduction to the theory of statistics*, New York : McGraw-Hill.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *Annals of Mathematical Statistics* **33**, 1065–1076.

R Development Core Team (2006), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. *http://www.R-project.org

Rosenblatt, M. (1956), 'Remarks on some nonparametrics estimates of a density function', *The Annals of Mathematical Statistics.* **27**, 832–837.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York.

Silverman, B. W. (1982), 'Algorithm as 176: Kernel density estimation using the fast fourier transform', *Applied Statistics* **31**, 93–99.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, Springer-Verlag, New York.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions', *Journal of the Royal Statistical Society* **36**, 111–147.

Stone, M. (1977), 'An asymptotic equivalence of choice of model by cross-validation and akaike's criterion', *Journal of the Royal Statistical Society* **39**, 44–47.

Taylor, C. (1997), 'Classification and kernel density estimation', *Vistas in Astronomy* **41**, 411–417.

Wand, M. P. & Jones, M. C. (1993), 'Comparison of smoothing parametrizations in bivariate kernel density estimation', *Journal of the American Statistical Association* **88**, 520–528.

Wand, M. P. & Jones, M. C. (1995), *Kernel Smoothing*, Chapman & Hall, London.

Webb, A. (2002), *Statistical Pattern Recognition*, 2nd. edn, John Wiley & Sons, London.