UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MARIA FERNANDA BARBOSA WANDERLEY

Estudos em Estimação de Densidade por Kernel: Métodos de Seleção de Características e Estimação do Parâmetro Suavizador

Prof. Dr. Antônio de Pádua Braga
Orientador

Prof. Docteur René Natowicz
Co-orientador

Belo Horizonte, Dezembro de 2013

Ficha Catalográfica

Barbosa Wanderley, Maria Fernanda

Estudos em Estimação de Densidade por Kernel: Métodos de Seleção de Características e Estimação do Parâmetro Suavizador / Maria Fernanda Barbosa Wanderley. – Belo Horizonte: UFMG EE, 2013.

95 f.: il.

Defesa de Tese para Obtenção do Título de Doutor em Engenharia Elétrica – Universidade Federal de Minas Gerais. Programa de Pós Graduação em Engenharia Elétrica, Belo Horizonte, BR–MG, 2013.

Orientador: Antônio de Pádua Braga; Co-orientador: René Natowicz.

1. Estimação não-paramétrica de Densidades. 2. Seleção de Características. 3. Estimação da Largura do Kernel. I. de Pádua Braga, Antônio. II. Natowicz, René. III. Título.

Estudos em Estimação de Densidade por Kernel: Métodos de Seleção de Características e Estimação do Parâmetro Suavizador

Maria Fernanda Barbosa Wanderley

Tese de doutorado submetida à banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Engenharia Elétrica.

Aprovado por:

Prof. Dr. Antônio de Pádua Braga (Orientador)

Prof. Docteur René Natowicz (Co-orientador)

Prof. Dr. Carlos Humberto Llanos Quintero

Prof. Dr. Eduardo Mazoni Andrade Marçal Mendes

Prof. Dr. Felipe Maia Galvão França

Prof. Dr. Marcelo Azevedo Costa

Belo Horizonte, Dezembro de 2013

À minha mãe Maria Lúcia (em memória):
"I never dreamt that I would get to be
The creature that I always meant to be
But I thought in spite of dreams
You'd be sitting somewhere here with me"
Being Boring - PetShopBoys

AGRADECIMENTOS

Ao meu marido, Tiago Mota, por todo apoio, suporte e paciência durante esses anos de sangue, suor e lágrimas. Esse trabalho é tão seu quanto meu.

Ao meu orientador, Prof. Antônio Braga, pelos conselhos, orientação e ideias. Boa parte do meu amadurecimento enquanto pesquisadora e futura professora se devem a você. Fico muito grata e orgulhosa de poder dizer que fui sua orientada.

Ao meu co-orientador, Prof. René Natowicz, pela recepção calorosa na ESIEE-Paris, pelas dicas pra aprimorar meu francês e pela garrafa de Guaraná Antarctica quando eu estava morrendo de saudade do Brasil.

Ao meu pai e minhas irmãs, que mesmo em outra cidade participaram dessa jornada. Em especial à minha irmã Ana Carolina, pelas conversas diárias, interesse e presença. Te amo, sis.

Aos amigos do LITC, pela troca de ideias, momentos de descontração e risadas. Vocês sem dúvida tornaram essa caminhada mais fácil e agradável.

Aos amigos do MACSIN, pelos almoços, cafés, força e amizade. Muito obrigada por tudo.

À UFMG, agora minha segunda casa.

Ao CNPq e à CAPES, pelo auxílio financeiro durante minha pesquisa no Brasil e na França.

RESUMO

Problemas de indução de funções são muitas vezes representados por meio de medidas de afinidade entre os elementos do conjunto indutivo de amostras, sendo as matrizes de kernel um método bastante difundido. O presente trabalho tem como objetivo obter informação das relações de afinidade entre os dados a partir da matriz de kernel calculada, partindo da hipótese que tais relações geométricas seriam coerentes com os rótulos conhecidos. Foram propostos métodos univariados e multivariados de seleção de características utilizando estimação de densidade por kernel (KDE), bem como métodos para estimar a largura do kernel baseados na coerência dos rótulos com a geometria do problema. Para avaliar a relação da estrutura dos dados com os rótulos foi utilizado um classificador baseado em estimação de densidade por kernel (KDE) e comparou-se o desempenho dos métodos propostos com outros conhecidos na literatura. Para as bases de dados testadas, o desempenho dos métodos propostos mostrou-se semelhante aos utilizados como base de comparação. Tais resultados indicam que é viável selecionar modelos através do cálculo direto das densidades e da geometria do problema de separação em questão.

Palavras-chave: Estimação não-paramétrica de Densidades, Seleção de Características, Estimação da Largura do Kernel.

Study on Kernel Density Estimation: Feature Selection Methods and Smoothing Parameter Estimation

ABSTRACT

Function induction problems are frequently represented by affinity measures between the elements of the inductive sample set, being kernel matrices a well known one. This work have as objective obtain information of the relations between data from the calculated kernel matrix, starting from the hypothesis that those geometric relations are coherent with known labels. Univariate and multivariate feature selection methods that use kernel density estimation (KDE) were proposed. Methods for perform estimation of kernel width, based at the geometric coherence between label and problem geometry, were also proposed. To assess the relation of data structure with the labels, a classifier based on kernel density estimation (KDE) was used and the performance of the proposed methods was compared with others known from literature. To the databases tested, the performance of the proposed methods were similar to the ones in the literature. Results indicates that is practicable selecting models through the direct calculation of densities and the geometry from the class separation.

Keywords: Non-parametric Density Estimation, Feature Selection, Kernel Width Estimation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

"E se eu fosse o primeiro a voltar
Pra mudar o que eu fiz,
Quem então agora eu seria?"

O Velho e o Moço, Los Hermanos

De modo geral, o problema do aprendizado de máquina pode ser descrito como o processo pelo qual, dado um vetor de entrada $\mathbf{x}$ obtido por meio de uma função geradora $p(x)$ desconhecida, uma máquina de aprendizado estima uma saída y_e, usando as relações por ela inferidas entre as entradas e saídas y correspondentes (CHERKASSKY; MULIER, 2007). A partir das possíveis soluções $y_e = f(\mathbf{x})$, fornecidos pela máquina de aprendizado, muitos métodos escolhem aquele que mais se aproxima da saída real y, visando à minimização do risco empírico.

Os fundamentos da teoria do Aprendizado Estatístico (VAPNIK, 2000) estabelecem condições de convergência para o risco empírico, as quais assumem a existência de um conjunto de dados suficientemente grande e representativo. Garantidas as condições de convergência, a indução de um modelo geral e que seja válido em todo o domínio da função geradora dos dados seria possível. No entanto, na maioria das situações reais, especialmente aquelas que envolvem problemas de alta dimensão, pode não

ser viável avaliar de antemão se as condições de convergência foram atingidas.

Embora existam trabalhos na literatura que tenham por objetivo estimar o tamanho mínimo da amostra para um determinado problema (ADCOCK, 1997; HWANG et al., 2002), não há nenhuma prova formal da generalidade desses métodos. Na prática, o processo de amostragem é considerado inerente ao problema e a indução de modelos é realizada assumindo-se que os dados sejam representativos e que as condições de convergência tenham sido atingidas. Não obstante, boa parte dos esforços de desenvolvimento da área visam à construção de modelos indutivos que sejam robustos a eventuais desvios de amostragem e escassez de informação.

A grande dificuldade no processo de indução de funções a partir de dados amostrais está, portanto, na premissa de que a função do oráculo gerador dos dados possa ser aproximada a partir de um conjunto reduzido de dados. Espera-se que o processo de indução seja capaz de resultar em uma função paramétrica universal que reproduza o seu comportamento. O Aprendizado Transdutivo (GAMMERMAN; VOVK; VAPNIK, 1998) apresenta uma alternativa à aproximação universal, ao fazer aproximações particulares para o conjunto de amostras ao invés de induzir uma função global de maior custo e muitas vezes inviável de ser aproximada.

O problema de indução de modelos a partir de dados escassos surge de várias formas em Aprendizado de Máquina e em particular na estimação paramétrica de densidades. Para este tipo de problema os parâmetros globais do modelo são induzidos a partir de um conjunto, muitas vezes restrito, de amostras. Similarmente ao Aprendizado Transdutivo, os estimadores de densidade não-paramétricos, particularmente os estimadores de densidade por *kernel* (KDE - *Kernel Density Estimator* (PARZEN, 1962)) visam a estimar densidades com base em informações locais ao invés de estimar parâmetros globais para modelos de dados. Esta abordagem, utilizada neste trabalho, é particularmente interessante quando o conjunto de dados é restrito, já

que a mesma não demanda que se assuma um modelo global para os dados nem que se façam considerações sobre a modalidade da função geradora.

Conhecer a relação estrutural dos dados pode fornecer informações iniciais importantes sobre os dados analisados. Em quais regiões eles se concentram mais? Existem sub-grupos dentro dos grandes grupos? É possível separar claramente os dados? Tal relação pode ser inferida a partir de medidas de afinidade ou similaridade, as quais frequentemente também representam problemas de indução de funções, uma vez que essas baseiam-se de certa forma em medidas de similaridade relativas às amostras de treinamento e seus rótulos.

A representação de afinidades por meio de matrizes de *kernel* tornou-se bastante difundida após a popularização das Máquinas de Vetores de Suporte (SVM - *Support Vector Machines*) (VAPNIK, 2000) em que funções de *kernel* são utilizadas para representar o conjunto indutivo no espaço de características (CORTES; VAPNIK, 1995; VAPNIK, 2000) por meio de mapeamentos não-lineares. A matriz de *kernel*, por meio da qual a transformação não-linear é realizada, contém também medidas de similaridade entre amostras e grupos de amostras para todos os elementos do conjunto indutivo.

O método não-paramétrico de estimação de densidade por *kernel* (KDE) utiliza tais matrizes para induzir funções a partir da informação estrutural contida nos dados, tendo como parâmetro a definir apenas a largura da função de *kernel*, sem necessidade de suposições *a priori* sobre a forma da função geradora. Tal parâmetro, também chamado de parâmetro suavizador do *kernel*, possui papel fundamental no KDE, sendo o valor escolhido o que define se a estimativa realizada consegue ou não representar de maneira adequada a relação entre os dados. A escolha incorreta do parâmetro pode encobrir uma estrutura multimodal ou perder informação mais geral por destacar exageradamente as relações locais.

1.1 Objetivos e Contribuições

O presente trabalho baseia-se na hipótese de que é possível induzir funções a partir da informação geométrica contida nos dados, obtida a partir da função de *kernel*. Busca-se explorar a relação entre a estrutura e os rótulos atribuídos aos dados, verificando se existe coerência entre essas duas informações.

A partir da hipótese inicial duas vertentes para a exploração dessa informação foram traçadas. A primeira faz uso da informação contida na matriz de *kernel* para selecionar características; a segunda busca oferecer métodos para estimar a largura do *kernel*, baseado na coerência entre a estrutura e os rótulos.

Tendo o KDE e o classificador KDE-Bayes (WANDERLEY et al., 2010) como centro de todos os métodos, são descritas a seguir as contribuições esperadas com esse trabalho:

- Explorar a informação obtida a partir do KDE para selecionar características, especialmente em casos onde os dados são escassos e esparsos. Foram propostos métodos univariados e multivariados, todos utilizando o classificador KDE-Bayes como base para selecionar as características. Nos métodos univariados cada uma das características é utilizada no KDE-Bayes como um classificador único e, em seguida, são ordenadas segundo uma métrica escolhida. No caso multivariado, o processo é o mesmo, porém utilizando um sub-conjunto de características. Como o espaço de busca dos sub-conjuntos de características pode ser muito grande, propôs-se utilizar algoritmos genéticos para realizar a busca de modo mais eficiente. Nesses algoritmos cada um dos indivíduos da população representam uma solução viável para o problema (neste caso um sub-conjunto de características) onde, baseado na teoria da Seleção Natural de Darwin, os mais adaptados sobrevivem. Para esta proposta os resultados

obtidos indicam que aproveitar a relação local entre os dados, tanto no caso univariado quanto no multivariado, pode ser promissor.

- Explorar a existência de coerência entre a geometria dos problemas e dos rótulos atribuídos a cada classe para propor novos métodos de estimação do parâmetro suavizador do *kernel*. Os dois métodos baseiam-se no conceito de que a região de separação entre classes deve ocorrer em um local de baixa densidade (CHAPELLE et al., 2006), o que levaria à minimização do erro do modelo. Porém, apenas a minimização do erro não é suficiente pois poderia levar a um estimador com *overfitting*, tal como na Figura 1.1. Embora o classificador representado pela curva rosa tenha um erro maior, ele é menos complexo do que o representado pela curva verde, que tem erro zero. Este último claramente tem *overfitting* e segue quase perfeitamente o contorno das classes, levando possivelmente a erros de generalização. Por isso, tem-se como objetivo simultaneamente minimizar o erro e controlar a complexidade do modelo, caracterizando o problema como bi-objetivo. Os métodos encontrados na literatura baseiam-se, em geral, na minimização do erro médio quadrático integrado (JONES; MARRON; SHEATHER, 1996), cujo cálculo depende do conhecimento da função geradora dos dados. Como essa informação é desconhecida, tais métodos assumem gaussianidade dos dados, o que pode levar a estimativas incorretas. Como alternativa, apresentam-se dois métodos que exploram a coerência entre a geometria do problema e os rótulos atribuídos aos dados.

1.2 Organização do Trabalho

O presente trabalho foi dividido em seis capítulos contendo entre eles uma apresentação dos conceitos de estimação não-paramétrica de densidades, os métodos desenvolvidos bem como seus resultados.

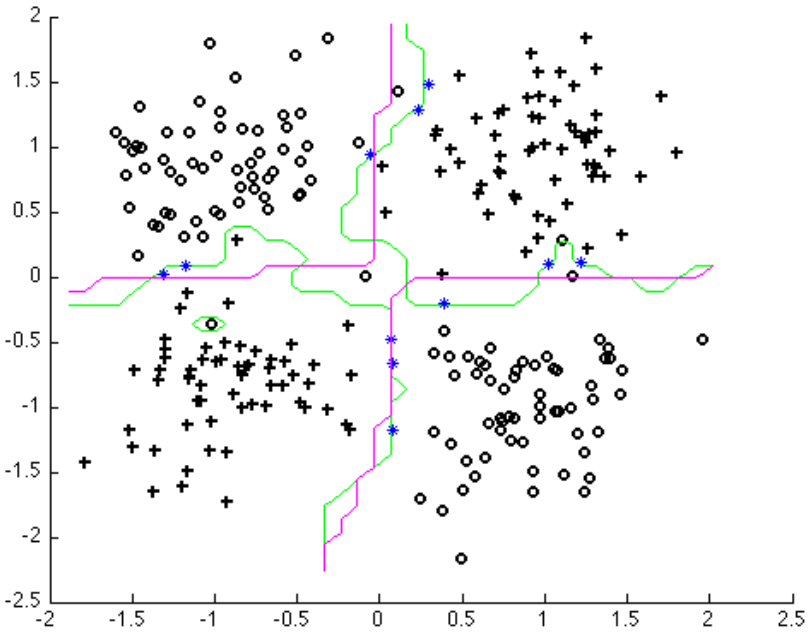

Figura 1.1: Comparação de complexidade entre classificadores. Na figura, o classificador representado pela curva verde é mais complexo do que o rosa, embora tenha erro zero.

A base teórica para os métodos desenvolvidos é introduzida no Capítulo 2, apresentando os conceitos de estimação não-paramétrica de densidades por *kernel*, univariada e multivariada, e sua relação com matrizes de afinidade. Neste mesmo capítulo é analisada a influência do parâmetro suavizador do *kernel* na estimativa realizada.

No Capítulo 3 estão os três métodos propostos para seleção de características, baseados em estimação não-paramétrica de densidades, divididos entre univariados e multivariados. O método em que todos os outros se baseiam é o KDE-Bayes (seção 3.1), que utiliza a estimação de densidade por *kernel* para calcular a verossimilhança do classificador bayesiano. A partir dele desenvolveu-se os métodos univariados de seleção por acurácia e pela área embaixo da curva ROC (FAWCETT, 2006). O AG-

KDE-Bayes, método multivariado, emprega um algoritmo genético para realizar a busca no espaço dos grupos de características do problema em questão.

No Capítulo 4 estão os métodos propostos para seleção da largura do *kernel*, cuja importância foi discutida no Capítulo 2. Os métodos propostos baseiam-se na hipótese de que a região de separação entre classes deve ser de baixa densidade para calcular o valor mais adequado.

Nos Capítulos 5 e 6 os resultados e as conclusões acerca dos métodos são apresentados. Para as bases testadas os resultados indicam que é viável selecionar modelos utilizando a informação da geometria do problema e os rótulos atribuídos às classes. Por fim, encontram-se os direcionamentos futuros para a continuação do presente trabalho.

2 REVISÃO DE LITERATURA

> "You have to know the past to understand the present."
>
> Carl Sagan

Esse Capítulo aborda o referencial teórico a partir do qual foram desenvolvidos os métodos propostos nesse trabalho. Na Seção 2.1 o conceito de matriz de afinidade e a caracterização da matriz de *kernel* como uma medida de similaridade entre amostras é apresentado. Nas Seções seguintes, o método de estimação de densidades por *kernel* é descrito, bem como sua forma multivariada, e é discutida a influência da escolha do parâmetro suavizador como determinante para uma estimação coerente.

2.1 Matrizes de Afinidade

Em geral, problemas que possuem dados escassos e esparsos representam um desafio de modelagem para estimadores paramétricos, baseados em densidade unimodal (THOMPSON; TAPIA, 1990). Por isso, a estimação não-paramétrica de densidades, que pode ser usada com qualquer tipo de distribuição, é comumente utilizada na modelagem de dados de problemas onde não há informação *a priori* sobre as

distribuições.

Os métodos de estimação paramétrica de densidades supõem que os dados possuem uma estrutura fixa, uma vez definida a estrutura, o problema passa a ser estimar os parâmetros (média e desvio padrão, supondo uma distribuição gaussiana) que melhor se ajustam aos dados. Porém, nos casos onde os dados são escassos e esparsos essa suposição pode levar a estimadores mais ajustados a uma determinada região do espaço de entrada, por exemplo à classe majoritária ou a uma região que possua mais informação, não refletindo a função de densidade geradora dos dados. Uma alternativa, neste caso, é utilizar a relação local entre os dados através, por exemplo, da afinidade ou similaridade entre eles.

Seja um conjunto $D_u = \{\mathbf{x}_i\}_{i=1}^{N}$, onde N é o número de amostras, a_{ij} os elementos da matriz de afinidades $\mathbf{A} = [a_{ij}]$ contêm uma medida de afinidade (ou similaridade) entre as amostras $(\mathbf{x}_i, \mathbf{x}_j)$ (SCOTT; LONGUET-HIGGINS, 1990). Como as medidas de similaridade são usualmente reflexivas, a matriz A é usualmente simétrica, ou seja, $a_{ij} = a_{ji}$. Há várias formas de representar as afinidades entre padrões, entre elas a representação por métricas de distância, comuns em métodos de agrupamento de dados (*clustering*) (JOHNSON, 1967) ou através de *kernels*, conforme representado na Equação 2.1 para um *kernel* Gaussiano.

$$k(\mathbf{x}_i, \frac{1}{\sqrt{2\pi}}\mathbf{x}_j) = e^{-\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{h}\right)^2} \tag{2.1}$$

onde h é o raio ou desvio padrão da função Gaussiana e $k(\mathbf{x}_i, \mathbf{x}_j) = a_{ij}$.

A matriz de *kernel* N x N resultante da Equação 2.1 contém, para um determinado valor de h, as relações reflexivas para todos os pares $(\mathbf{x}_i, \mathbf{x}_j)$ e pode ser representada na forma diagonal em blocos, conforme mostrado na Equação 2.2.

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} & \dots & \mathbf{K}_{1c} \\ \mathbf{K}_{21} & \mathbf{K}_{22} & \dots & \mathbf{K}_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_{c1} & \mathbf{K}_{c2} & \dots & \mathbf{K}_{cc} \end{bmatrix} \tag{2.2}$$

onde c é o número de partições (*clusters*) do conjunto de amostras $D_u = \{x_i\}_{i=1}^{N}$.

Cada uma das submatrizes $\mathbf{K}_{ij}$ da Equação 2.2 contém as afinidades entre os elementos dos grupos i e j do conjunto de amostras. Assim, a representação ao nível de agrupamentos, ou *clusters*, permite também extrair da matriz de kernel outras informações importantes sobre as distribuições dos dados e sobre as relações entre amostras e grupos de amostras. Na Figura 2.1 são apresentados 150 vetores bi-dimensionais amostrados de 5 distribuições distintas. A Figura 2.2 representa a matriz de kernel Gaussiano resultante das amostras da Figura 2.1, a qual é ordenada de acordo com as distribuições geradoras, conhecidas de antemão para efeitos deste exemplo. Pode-se ver claramente na Figura 2.2 que as relações de afinidade entre os elementos de um mesmo grupo e entre elementos de grupos diferentes são visualmente perceptíveis nesta forma de representação, exemplificando o alcance da informação contida na matriz de *kernel*.

Assim, a representação das afinidades através da matriz de *kernel* permite não somente induzir modelos de classificação e regressão a partir dos dados (CORTES; VAPNIK, 1995), como também contém informação para a indução de um estimador para $f_y(x)$, a função de densidade geradora do conjunto $D_u = \{x_i\}_{i=1}^{N}$.

Os estimadores de densidade por *kernel*, que serão descritos na seção seguinte, baseiam-se nas relações reflexivas $k(\mathbf{x}_i, \mathbf{x}_j) = k_{ij}$ para fazer estimativas locais da função de densidade $f_y(x)$, geradora dos dados. Estes estimadores não-paramétricos possuem apenas um parâmetro a ser ajustado, usualmente relacionado à suavidade do *kernel* utilizado, como por exemplo o parâmetro h na Equação 2.1. Apesar de os

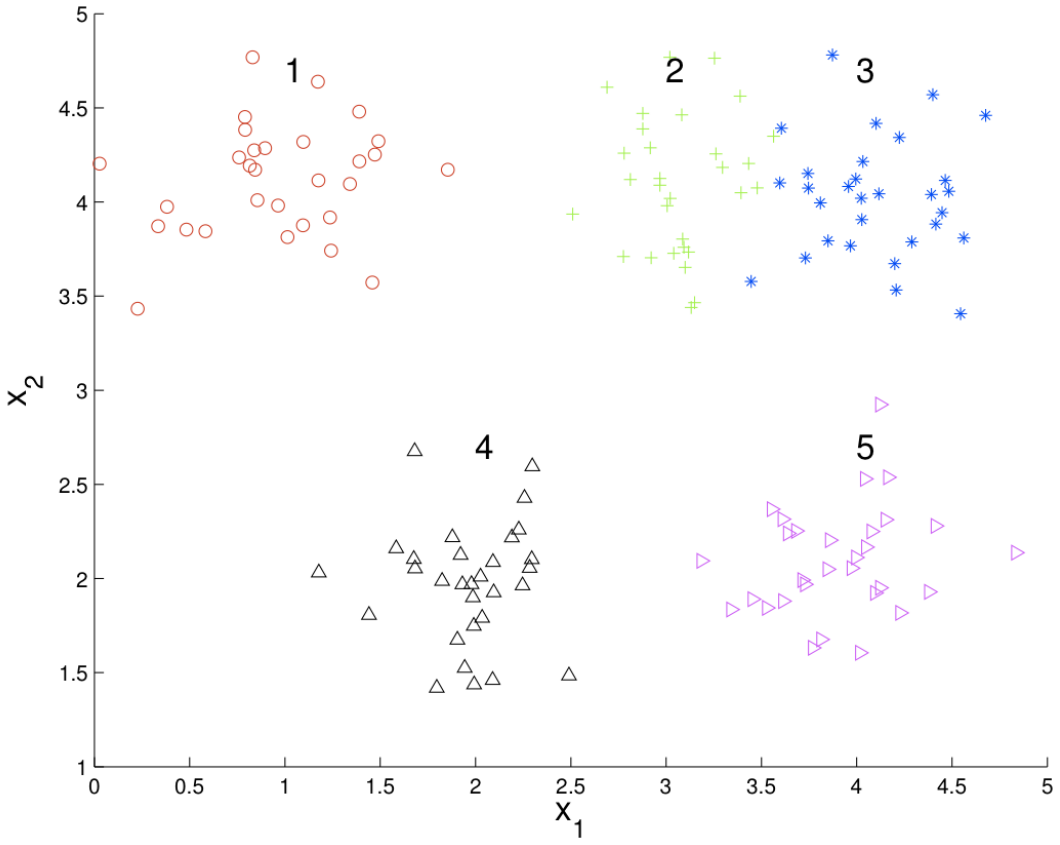

Figura 2.1: Dados bi-dimensionais amostrados de cinco distribuições distintas.

estimadores de densidade por *kernel* dependerem de apenas um parâmetro global, eles não possuem limitações quanto ao número de variáveis ou quanto à aproximação de funções de densidade multimodais, apresentando-se assim como potencialmente atrativos para aplicações onde há escassez de amostras e de informações *a priori* sobre as densidades geradoras dos dados, característica frequente em problemas de bioinformática.

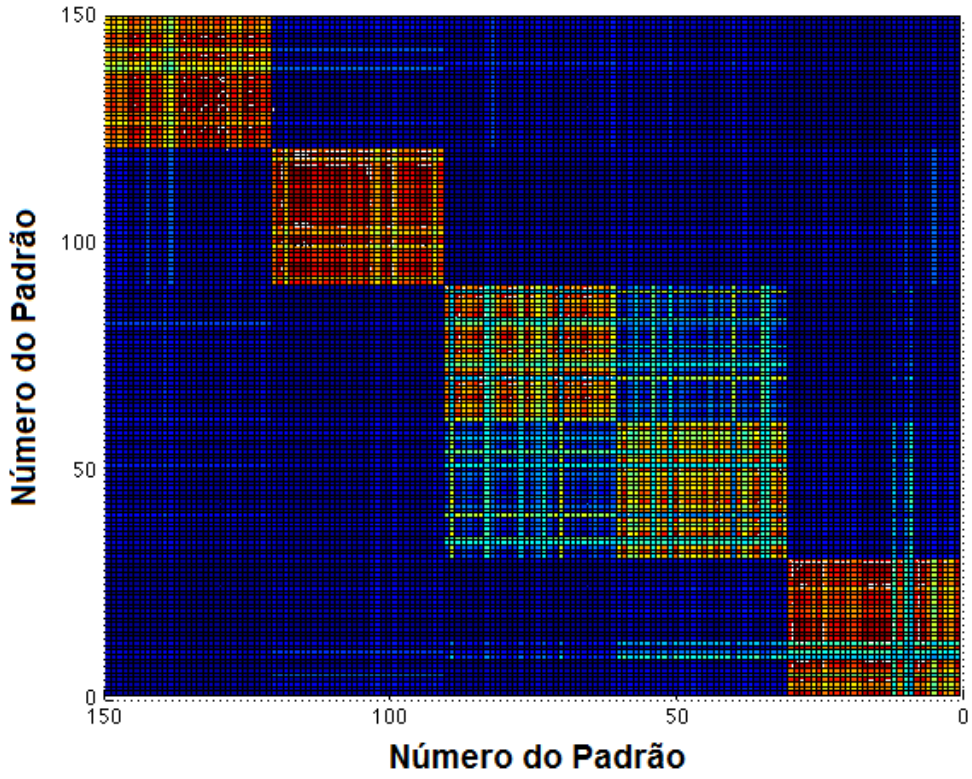

Figura 2.2: Matriz de Proximidade para os cinco agrupamentos mostrados na Figura 2.1.

2.2 Estimador de Densidade por *Kernel* - KDE

Um estimador de densidade por *kernel*, ou KDE (PARZEN, 1962), é obtido através da superposição de funções de *kernel*, como descrito na Equação 2.1, centralizadas em cada um dos elementos $\mathbf{x}_i (i = 1 \ldots N)$ do conjunto de amostras. Assim, a estimativa de densidade $\hat{f}(x_t)$ no ponto x_t depende apenas da relação espacial entre x_t e os elementos da amostra $\mathbf{x}_i (i = 1 \ldots N)$, quantificada pela métrica embutida na função de *kernel*. De uma maneira geral, a Equação 2.3 descreve um estimador univariado de densidade por *kernel*.

$$\hat{f}(x_t) = \frac{1}{Nh} \sum K(x_t, x_i) \tag{2.3}$$

onde N é o número de amostras, h é o parâmetro de suavização do *kernel* e $K(x_t, x_i)$ é o operador de *kernel*, cuja integral $\int K(u)du$ deve ser unitária. O argumento da função $K(.)$ é na verdade o ponto onde se deseja fazer a estimação, já que as amostras $x_i(i = 1 \ldots N)$ são fixas e fornecidas de antemão.

Um exemplo de estimativa com KDE é mostrado na Figura 2.3, na qual são apresentadas a representação por histograma e a estimação contínua resultante da Equação 2.3, para dados amostrados de duas distribuições normais com médias em -4 e 4. É interessante observar que a estimativa do KDE representa a distribuição conjunta dos dois modos da função geradora. Caso fosse feita a modelagem paramétrica desta distribuição bi-modal, seria necessário encontrar as duas partições geradoras, modelar cada uma individualmente e misturá-las para então obter a distribuição conjunta. Seria necessário, também, estimar pelo menos os parâmetros do algoritmo de agrupamento, tipicamente o número de partições, e os parâmetros de cada distribuição individualmente. A estimação com o KDE requer somente a determinação do parâmetro h, associado à abertura das funções gaussianas.

2.3 KDE Multidimensional

A estimativa multivariada de funções de densidade com o KDE, conforme descrito na Equação 2.3, pode ser obtida diretamente ao se considerar funções de *kernel* multidimensionais, conforme Equação 2.1, caso as variáveis de entrada sejam consideradas independentes. No entanto, para o caso de não haver independência, a estimativa com o KDE considera também a utilização de valores diferentes de h para cada uma das dimensões do vetor x.

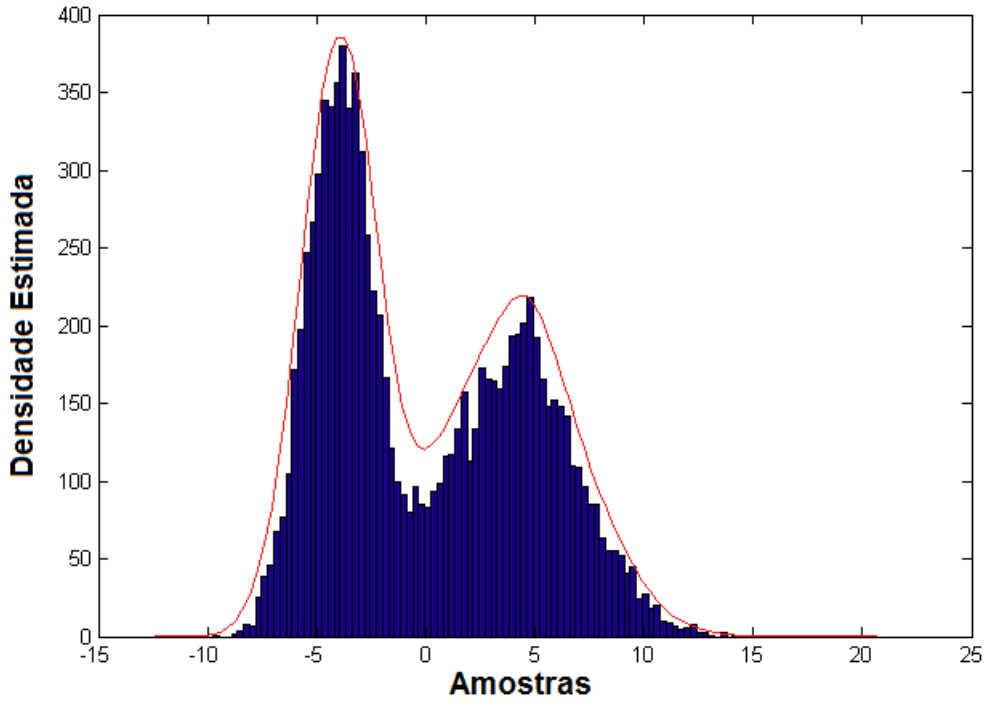

Figura 2.3: Estimação de densidade utilizando um histograma e estimação de densidade por *kernel*. As funções geradoras possuem médias -4 e 4.

Considere que o vetor arbitrário $\mathbf{x}_j$ seja representado com as suas n dimensões como $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$. Assim, a forma geral do KDE multidimensional é apresentada na Equação 2.4.

$$\hat{f}(x_t) = \frac{1}{N \prod_{i=1}^{n} h_i} \sum_{i=1}^{N} K\left(\frac{x_{t1} - x_{i1}}{h_1}, \ldots, \frac{x_{tn} - x_{Nn}}{h_n}\right) \tag{2.4}$$

Uma alternativa ao uso de uma função de *kernel* multidimensional é o *kernel* multiplicativo (SCOTT, 1992). Neste caso, um *kernel* unidimensional é usado para cada uma das dimensões, cada uma com a sua respectiva largura h. Assim, o *kernel* n-

dimensional é representado pelo produto dos *kernels* em cada uma das n dimensões univariadas, resultando na Equação 2.5.

$$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N}\left[\prod_{j=1}^{n}\frac{1}{h_j}K\left(\frac{\mathbf{x}_i - x_{ij}}{h_j}\right)\right] \tag{2.5}$$

Assumindo independência e o mesmo raio h para todas as dimensões, a estimativa de densidade pelo KDE Gaussiano em um determinado ponto arbitrário x_i pode ser obtida através da soma dos produtos acumulados em todas as dimensões para todos os padrões do conjunto de amostras. Finalmente, reescrevendo o produtório e considerando que o somatório da Equação 2.5 corresponde à soma de todos os elementos de uma linha (ou coluna) da matriz de *kernel* Gaussiano com raio h, chega-se a Equação 2.6.

$$\hat{f}_h(x_i) = \frac{1}{Nh^n}\sum_{k=1}^{N}K(x_i, x_k) \tag{2.6}$$

É importante ressaltar neste ponto que o *kernel* Gaussiano para estimativas de densidade pelo método multiplicativo e aquele utilizado para construir modelos indutivos, como SVMs, possuem a mesma forma, podendo-se diferir apenas pelo parâmetro h. Esta constatação abre caminho para especulações de que o mesmo parâmetro h poderia satisfazer a ambos os problemas (QUEIROZ; BRAGA; PEDRYCZ, 2009) quando a densidade é estimada pelo KDE através de um *kernel* multiplicativo.

Assim, conforme a Equação 2.6, a estimativa da densidade $\hat{f}_h(x_i)$ se resume em encontrar o valor de h que satisfaça a alguma restrição ou função-objetivo. No entanto, a caracterização de objetivos para a estimativa de funções de densidade não é tão direta, já que o problema possui uma natureza não-supervisionada. Na

Seção 2.4, métodos de estimação de h presentes na literatura são apresentados, e no Capítulo 4, são propostos dois novos métodos.

2.4 Métodos de Estimação da Largura do *Kernel*

Conforme apresentado na Seção 2.2, a largura h do *kernel* é o único parâmetro a ser determinado no KDE, sendo este o responsável pela suavização da curva escolhida para realizar a estimação. Na Figura 2.4 a influência do valor de h utilizado fica bastante clara e demonstra a importância de se realizar uma escolha apropriada. No referido exemplo tinha-se como objetivo aproximar uma função $N(0, 1)$ (curva preta) e foram utilizados os valores de largura iguais a 0,1, 0,4, 1 e 10. A curva vermelha é sub-suavizada e, devido ao valor de largura muito pequeno ($h = 0, 1$), apresenta todos os contornos dos pontos onde a estimação foi realizada. Esta situação gera um modelo com baixa capacidade de generalização devido ao *overfitting* nos dados de treinamento. A curva verde ($h = 10$) representa a situação diametralmente oposta, a super-suavização, que leva a perda de informação sobre a relação local entre os dados. Por fim, os valores $h = 0, 4$ (curva magenta) e $h = 1$ (curva azul), se aproximam melhor do objetivo, sendo o primeiro o mais apropriado para este caso. Analisando a Figura 2.4 fica clara a importância da estimação correta da largura do *kernel*, cujo alguns métodos serão apresentados a seguir.

O problema da estimação da largura do *kernel* é de natureza não-supervisionada, uma vez que não se conhece de antemão a função geradora dos dados e, por isso, não seria em princípio possível minimizar uma função de erro para obter o referido parâmetro. Apesar disso, o primeiro método de estimação da largura (SILVERMAN, 1986) baseia-se na minimização do erro médio quadrático integrado (MISE), com o objetivo de encontrar o h que obtenha o melhor $\hat{f}(x)$ para a função f geradora. O MISE é dado pela Equação 2.7

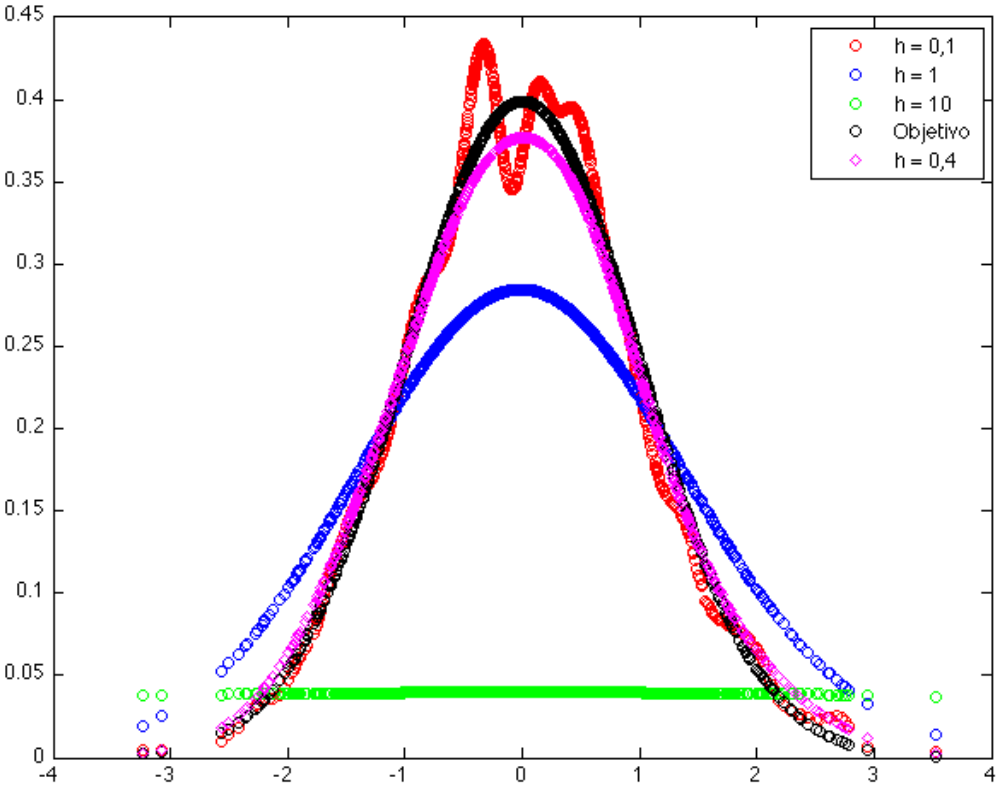

Figura 2.4: Exemplo da influência do valor escolhido para a largura na estimação de densidade da função.

$$MISE(\hat{f}(x)) = \int E\{\hat{f}(x) - f(x)\}^2 dx \tag{2.7}$$

$$= \int \{E\hat{f}(x) - f(x)\}^2 dx + \int var\hat{f}(x) dx, \tag{2.8}$$

ou seja, o erro médio quadrático integrado pode ser expresso em termos da soma do viés integrado e da variância integrada. De acordo com (SILVERMAN, 1986), o viés pode ser representado por $\frac{1}{2}h^2 f''(x)k_2$ e a variância por $\frac{1}{nh}\int K(t)^2 dt$, sendo a Equação 2.7 reescrita para

$$MISE(\hat{f}(x)) = \overbrace{\frac{1}{4}h^4 k_2^2 \int f''(x)^2 dx}^{\text{viés}} + \overbrace{\frac{1}{nh}\int K(t)^2 dt}^{\text{variância}}, \tag{2.9}$$

onde h é a largura do *kernel*, k_2 é uma constante proveniente do segundo termo da expansão da Série de Taylor, $f''(x)$ é a derivada segunda da função geradora, n é o tamanho da amostra dos dados e K a função de *kernel* usada. Silverman propõe que o valor de h ótimo seria aquele que minimiza o MISE, porém, como tal cálculo depende da derivada segunda da função geradora dos dados que é desconhecida e do tipo de função de *kernel* utilizado, o autor assume que os dados foram gerados por uma distribuição normal e que o *kernel* é Gaussiano. Assim, após manipulações algébricas, o valor de h proposto é dado pela Equação 2.10

$$h_1 = 1,06\ \sigma\ n^{-\frac{1}{5}}. \tag{2.10}$$

Para dados normalizados, propõe-se, como equivalente à Equação 2.10, a Equação 2.11:

$$h_{11} = \left(\frac{4}{n+2}\right)^{\frac{1}{n+4}} * \left(N^{\frac{-1}{n+4}}\right), \tag{2.11}$$

onde n é o número de dimensões e N o número de amostras.

Supondo que a função geradora dos dados seja realmente uma Gaussiana, o valor dado por h_1 proverá uma boa estimação de densidade dos dados porém, essa afirmação nem sempre é verdadeira. Em tais casos, a estimativa pode ser suavizada, como mostra a Figura 2.5.

Silverman sugere que uma medida de espalhamento mais robusta seja utilizada e substitui, então, a variância dos dados pelo intervalo inter-quartis (IQR), propondo assim h_2 dado por:

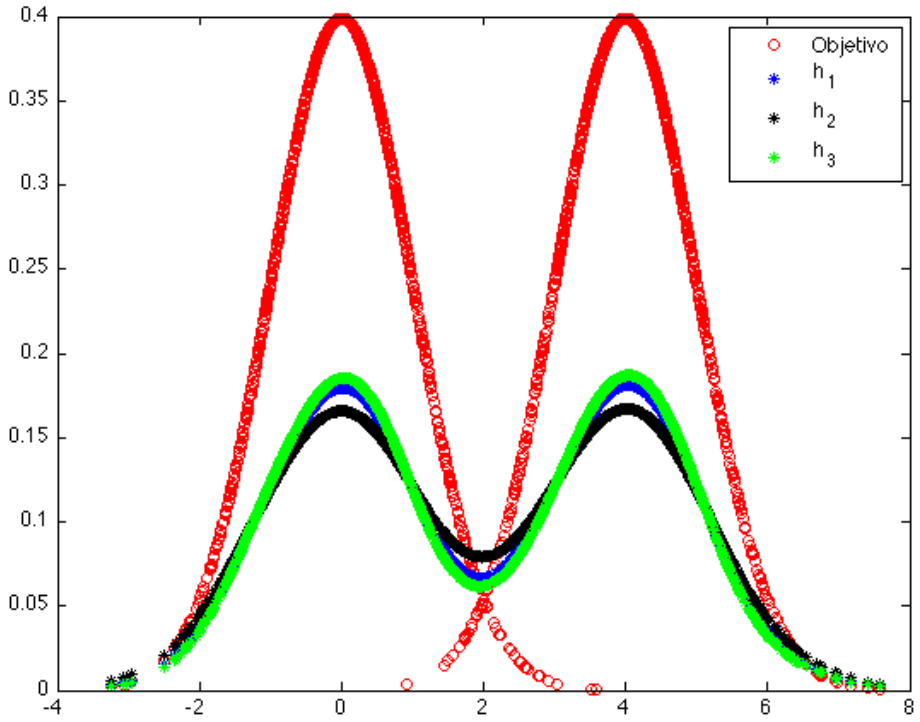

Figura 2.5: Comportamento das larguras propostas por (SILVERMAN, 1986) com relação à uma função geradora bimodal.

$$h_2 = 0,79\ IQR\ n^{-\frac{1}{5}}, \tag{2.12}$$

onde $R = Q_3 - Q_1$, ou seja, a diferença entre o terceiro e primeiro quartis. Porém, de acordo com a Figura 2.5, h_2 suaviza ainda mais a estimativa de bimodais, gerando assim a terceira proposta de largura de *kernel*:

$$h_3 = 0,9\ A\ n^{-\frac{1}{5}}, \tag{2.13}$$

onde $A = \min(\sigma,\ \frac{IQR}{1,34})$.

As larguras de *kernel* propostas nas Equações 2.10, 2.12 e 2.13 dependem de duas variáveis, a medida de espalhamento (desvio padrão ou intervalo inter-quartil) e o tamanho da amostra. A variação da largura com relação às variáveis que a compõem pode ser vista na Figura 2.6, na qual variou-se a medida de espalhamento entre 0,1 e 4, e o tamanho da amostra entre 1 e 200. Observando o exemplo algumas informações podem ser obtidas: o limite superior do valor de h é determinado pelo valor máximo do espalhamento; dado um valor de espalhamento e, em aproximadamente 84% dos casos $h \leq \frac{e}{2}$; o limite inferior de h é dado por $(e * 0,34)$. Assim, pode concluir-se que o valor de h está intimamente ligado ao espalhamento dos dados, ressaltando a importância da escolha da medida apropriada para esse fim.

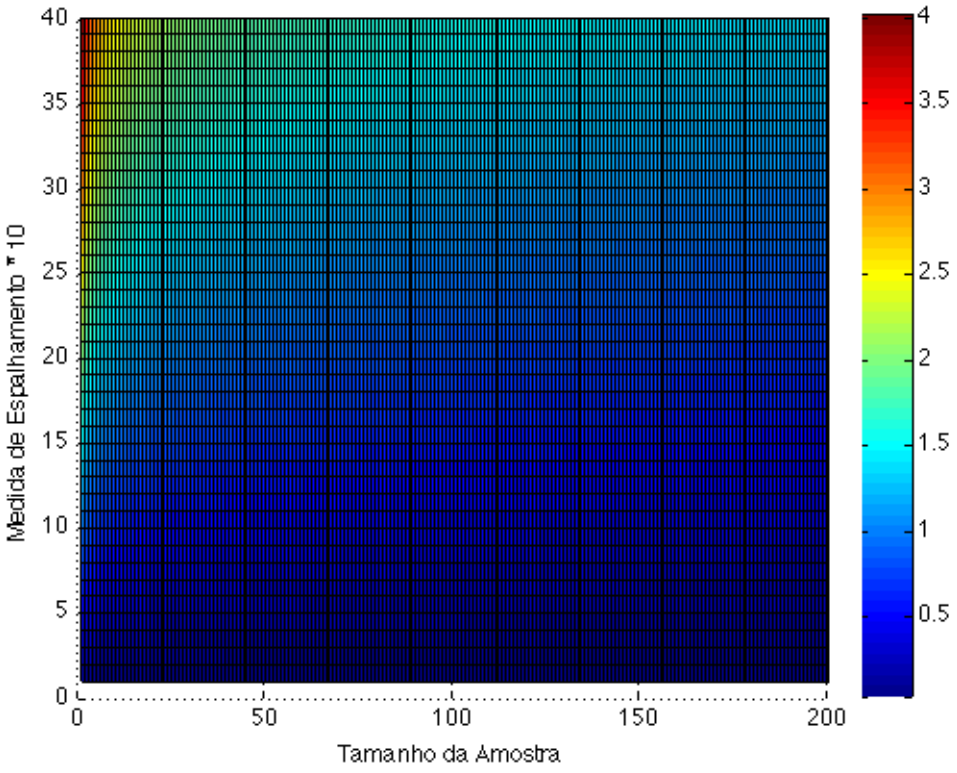

Figura 2.6: Variação da largura h com relação às variáveis que a compõem, a medida de espalhamento e o tamanho da amostra.

Em (SCOTT, 1992), encontra-se outra proposta de cálculo da largura h, também baseada na minimização do erro da função de densidade estimada com relação a função geradora dos dados, desconhecida. Neste caso, utilizou-se o erro médio quadrático assintótico (AMISE), composto pela soma da variância integrada (IV) e do quadrado do viés integrado (ISB), representado na Equação 2.14

$$AMISE = \overbrace{\frac{R(K)}{nh}}^{IV} + \overbrace{\frac{1}{4}\sigma^4 h^4 R(f'')}^{ISB}, \tag{2.14}$$

onde $R(g) = \int_{-\infty}^{\infty} g(u)^2 \, du$ é a rugosidade da função. Assim, o h ideal seria aquele que minimiza a AMISE, cujo cálculo depende do conhecimento prévio da função geradora. Após manipulações algébricas (para mais detalhes ver (SCOTT, 1992), p.165) propõe-se utilizar

$$h_4 = 3\left[\frac{R(K)}{35\sigma_k^4}\right]^{\frac{1}{5}} \sigma \; n^{-\frac{1}{5}} \tag{2.15}$$

$$= 1,144 \; \sigma \; n^{-\frac{1}{5}}, \tag{2.16}$$

onde $R(K) = \frac{0,5}{\sqrt{\pi}}$ e $\sigma_k^2 = 1$, assumindo-se *kernel* gaussiano. A menos de uma constante a Equação 2.15 se iguala às demais, sendo composta por uma medida de espalhamento e uma função do tamanho da amostra e seguindo a mesma análise feita na Figura 2.6.

Outros métodos de estimação da largura do *kernel* podem ser encontradas na literatura, em geral baseados na minimização de alguma função de erro (MOLANES-LÓPEZ; CAO, 2008; LIAO; WU; LIN, 2010; GAJEK; LENIC, 1993). Embora não seja o escopo deste trabalho, é importante mencionar também trabalhos que propõem estimação de uma largura variável para os casos de *kernels* multidimensionais, frequentemente baseados em noções de *clustering* (WU; CHEN; CHEN, 2007;

ZHANG; KING; HYNDMAN, 2006; LAKHDAR; SBAI, 2012).

2.5 Conclusão do Capítulo

Neste Capítulo foram apresentados os conceitos de estimação não-paramétrica de densidade, além de uma análise da importância da escolha adequada do parâmetro h, que desempenha o papel de suavizador da função de *kernel* empregada. Na Seção 2.4 foram apresentados os métodos clássicos de estimação do parâmetro h, baseados na minimização de uma função de erro. Embora intuitivos, a necessidade de suposições sobre qual seria a função geradora dos dados para obter a largura pode exercer influência no resultado encontrado.

3 MÉTODOS DE SELEÇÃO DE CARACTERÍSTICAS

Grail Knight: "You must choose. But choose wisely, for while the true Grail will bring you life, the false Grail will take it from you."

Indiana Jones and the Last Crusade

Neste Capítulo serão apresentados os métodos de seleção de características propostos nesse trabalho. Na Seção 3.1 encontra-se o método KDE-Bayes, classificador bayesiano baseado em estimação não paramétrica de densidades por *kernel*. Por fim, na Seção 3.2, estão os métodos univariados e multivariado de seleção de características, todos baseados no desempenho do KDE-Bayes segundo alguma métrica.

3.1 O Classificador KDE-Bayes

A construção de classificadores generativos, com base em estimativas de densidade das funções geradoras, se baseia no princípio da existência de coerência entre a rotulação dos dados e as funções que os geraram. Assim, com base neste princípio,

as funções geradoras estimadas, com o KDE, por exemplo, devem ser coerentes com os rótulos y_i presentes no conjunto de dados $D = \{x_i, y_i\}_{i=1}^{N}$. Para o caso dos classificadores bayesianos, a probabilidade *a posteriori* $P\left(C_j|x_i\right)$ de um padrão pertencer a uma determinada classe C_j deve ser maior para a classe com a qual este padrão foi rotulado. Este princípio garante não somente a minimização do Risco Empírico (VAPNIK, 2000) do conjunto de dados, mas também a robustez do modelo perante o conjunto de testes. Para o caso de problemas de classificação binária com duas classes C_1 e C_2, a razão entre as verossimilhanças resulta no classificador representado na Equação 3.1.

$$Classe(x) = \begin{cases} C_1 & \text{se } \frac{P(x|C_1)}{P(x|C_2)} > \frac{N_2}{N_1}, \\ C_2 & \text{caso contrário,} \end{cases} \tag{3.1}$$

onde $P(x|C_1)$ e $P(x|C_2)$ são as verossimilhanças para as classes com respeito ao vetor x, $P(C_1)$ e $P(C2)$ são as probabilidades *a priori* para cada classe e N_1 e N_2 o tamanho das amostras das classes C_1 e C_2.

Com base na informação de rotulação, as verossimilhanças $P(x|C_1)$ e $P(x|C_2)$ para as classes C_1 e C_2 podem ser estimadas com o KDE e a classificação final ser realizada. Conforme discutido anteriormente, a estimativa coerente das densidades pelo KDE será dependente da escolha do valor de h para *kernels* gaussianos. Não obstante, o conhecimento dos rótulos y_i permite uma análise do problema com base no princípio da coerência entre a função geradora das densidades e as rotulações atribuídas aos pontos.

3.2 Métodos de Seleção de Características Baseada em Estimação não Paramétrica

O uso de modelos indutivos de previsão, por terem como objetivo induzir os parâmetros de um modelo geral, nem sempre representa uma boa escolha, especialmente nos casos em que os dados são escassos e esparsos. Tais modelos baseiam-se no princípio de que toda a informação necessária para a estimação dos parâmetros está contida nos dados e, em casos em que o número de amostras é baixo, podem levar a resultados enviesados.

A partir da estimativa de densidade provida pelo KDE e da classificação realizada pelo KDE-Bayes, propõe-se três métodos de seleção de características, dois univariados e um multivariado.

3.2.1 Seleção por Acurácia

Um dos métodos mais simples e intuitivos de seleção de características utiliza a acurácia da resposta do método de classificação utilizado. Dado um classificador qualquer e um conjunto de treinamento, os parâmetros do classificador são ajustados e em seguida novos padrões são apresentados a ele. O método consiste em, dadas as características ordenadas decrescentemente de acordo com o número de previsões corretas, selecionar as que estão acima de um determinado limiar de acurácia.

Embora simples, o método não possui bom desempenho em determinados casos. Quando as classes do problema apresentam desbalanceamento, característica comum em problemas reais, os atributos selecionados tendem a ser enviesados com relação a classe maior. No caso em que a distribuição das classes é balanceada (ou próximo a isso) o método seleciona boas características, apresentando uma alternativa viável.

3.2.2 Seleção por AUC

A análise do gráfico ROC (do inglês *Receiver Operating Characteristic*) é um método de avaliação e seleção de um classificador binário baseado em seu desempenho segundo determinadas métricas (FAWCETT, 2006). Através de tal gráfico é possível visualizar a relação entre os erros do classificador para ambas as classes, diferentemente da análise de acurácia, que pode priorizar a classe majoritária.

Para o cálculo da curva ROC o desempenho do classificador deve ser medido de acordo com a sua sensibilidade e especificidade, que levam em consideração a taxa de acerto por classe. A sensibilidade pode ser definida como a medida da capacidade do classificador identificar corretamente as ocorrências da classe minoritária, enquanto a especificidade tem papel análogo para a classe majoritária.

A sensibilidade é calculada utilizando a expressão a seguir:

$$Se = \frac{VP}{VP + FN}, \tag{3.2}$$

onde VP são os casos onde as amostras positivas foram corretamente classificadas e FN representa o número de casos onde amostras positivas foram classificadas como negativas.

Já a especificidade é calculada por:

$$Es = \frac{VN}{VN + FP}, \tag{3.3}$$

onde VN representa o número de casos em que as amostras negativas foram corretamente classificadas e FP são os casos em que amostras negativas foram classificadas como positivas.

A curva ROC é obtida a partir do gráfico cujo eixo das abscissas representa os casos de falso positivo (FP) e os das ordenadas os casos de verdadeiro positivo (VP) de

um sistema de classificação binária variando o limiar de discriminação entre zero e infinito. A variação do limiar é equivalente a percorrer o espaço dos classificadores conservadores (especificidade 1 e sensibilidade 0, em que o classificador acerta somente a classificação das amostras da classe maior) para os classificadores liberais (especificidade 0 e sensibilidade 1, em que o classificador acerta somente a classificação das amostras da classe menor). É importante ressaltar que a qualidade da curva ROC está diretamente ligada à acurácia do método de classificação utilizado.

Um dos métodos de avaliação de desempenho de classificadores a partir da curva ROC é a medida da área abaixo da curva (AUC) para cada um dos classificadores. Quanto maior a AUC de um dado classificador, melhor o seu desempenho médio para o conjunto de exemplos avaliado. Embora os valores da AUC estejam sempre no intervalo $[0, 1]$, dado que a curva está contida dentro de um quadrado de lado 1, os classificadores cuja AUC está abaixo da diagonal do quadrado (ou seja, AUC = 0,5) são descartados, uma vez que seu desempenho é pior do que classificadores que escolhem as respostas aleatoriamente(BRADLEY, 1997).

A análise da curva ROC representa um bom método para seleção de características, dado que é possível escolher a característica que apresente o classificador mais balanceado entre sensibilidade e especificidade.

O método de seleção de características por AUC é composto pelos seguintes passos:

1. Obtenção das probabilidades *a posteriori* de cada classe utilizando o KDE-Bayes;
2. Cálculo das curvas ROC;
3. Análise da AUC;

4. Ordenação das características de acordo com a AUC, sendo as de maior valor as que geram os melhores classificadores.

3.2.3 AG-KDE-Bayes: Selecionando subconjuntos de características

Em problemas cuja dimensionalidade é muito alta torna-se inviável testar todas as combinações de elementos para selecionar os melhores grupos de características. Por exemplo, para uma base de dados cujo número de características é 30 são necessários $2^{30} = 1073741824$ testes para avaliar todas as possibilidades de agrupamento entre elas. Para contornar esse problema um método evolucionário foi escolhido de forma que a busca pelo espaço de soluções viáveis fosse realizado de forma mais eficiente.

Algoritmos genéticos são uma classe de algoritmos baseados na teoria de evolução de Darwin. A idéia básica é encontrar a melhor solução para um dado problema através de um processo evolutivo que seleciona a mais adaptada dentre as soluções (GOLDBERG, 1989).

No algoritmo, inicialmente, um conjunto de possíveis soluções para o problema é gerado aleatoriamente. Cada solução gerada é chamada de indivíduo ou cromossomo e o conjunto de indivíduos é chamado de população (MUNAKATA, 1998). Após a geração da população inicial uma série de iterações (gerações) ocorrem até que a condição de parada seja alcançada. Nessas gerações os seguintes passos ocorrem: primeiro os melhores indivíduos são selecionados de acordo com uma função de avaliação. Após a seleção, segundo uma certa probabilidade, os indivíduos escolhidos sofrem recombinação e, em seguida, uma mutação. Ao fim da iteração uma nova população está formada e o algoritmo segue até que um critério de parada seja satisfeito (MICHALEWICZ, 1996).

No presente trabalho a população de subconjuntos de características, codificado

binariamente, é gerado aleatoriamente. Através das gerações os seguintes passos ocorrem: os indivíduos são selecionados de acordo com uma função de avaliação, usando o método da roleta, e, então, os operadores de recombinação e mutação são aplicados de acordo com probabilidades (p_r e p_m, respectivamente). Ao final de uma iteração uma nova população é criada e o algoritmo continua até que uma condição de parada seja atingida (GOLDBERG, 1989).

O operador de mutação aqui utilizado, responsável por aumentar a variabilidade genética da população, desempenha três funções no algoritmo: aumentar o número de características, diminuir o número de características ou alterar a posição do indivíduo no espaço de busca, mudando algum gene. O tipo de operação a ser realizada também é definido por uma probabilidade e, independentemente de ser uma operação de crescimento ou decrescimento do tamanho do indivíduo, o gene é escolhido aleatoriamente.

Para medir a adaptação de um subconjunto de características duas métricas foram utilizadas, a especificidade (es) e a sensibilidade (se), dadas pelo desempenho do classificador KDE-Bayes no indivíduo que está sendo avaliado. Para obter um equilíbrio entre as métricas, a função de avaliação f_{aval} utilizada foi a média geométrica,

$$f_{aval} = \sqrt{es * se}.$$

O valor de f_{aval} é alto quando o valor de cada métrica individualmente é alto e quando a diferença entre eles é pequena (KUBAT; HOLTE; MATWIN, 1997). Assim, os melhores indivíduos são aqueles cujas métricas são equilibradas, enquanto indivíduos que são tendenciosos para uma classe ou outra recebem uma avaliação pior.

Outro importante operador utilizado foi o elitismo, que preserva o melhor indivíduo para a próxima geração, evitando que ele seja perdido durante o processo de recombinação e mutação.

3.3 Conclusão do Capítulo

Neste Capítulo foi apresentado o método KDE-Bayes, classificador bayesiano baseado em estimação não-paramétrica de densidades por *kernel*. A seguir, os métodos de seleção de características, baseados no KDE-Bayes, foram mostrados. No Capítulo 5 encontram-se os resultados dos experimentos realizados com os métodos deste Capítulo.

4 MÉTODOS DE SELEÇÃO DA LARGURA DO *KERNEL*

"If you choose not to decide, you still have made a choice"

Freewill, Rush

Nesse Capítulo serão apresentados os métodos de seleção de largura do *kernel* propostos nesse trabalho. Como ressaltado na Seção 2.4, a determinação adequada do parâmetro suavizador do *kernel* está diretamente relacionada à qualidade da estimativa realizada, justicando a importância do tema.

4.1 O Problema Bi-Objetivo da Seleção da Largura do *Kernel*

Considere o conjunto de amostras apresentado na Figura 4.1 e a sua matriz $\mathbf{K}$ para $h = 1$ apresentada na Figura 4.2. Este exemplo, apesar de ser sintético e bem controlado, representa o problema de maneira geral.

A visualização da matriz de kernel correspondente aos dados nos permite identifi-

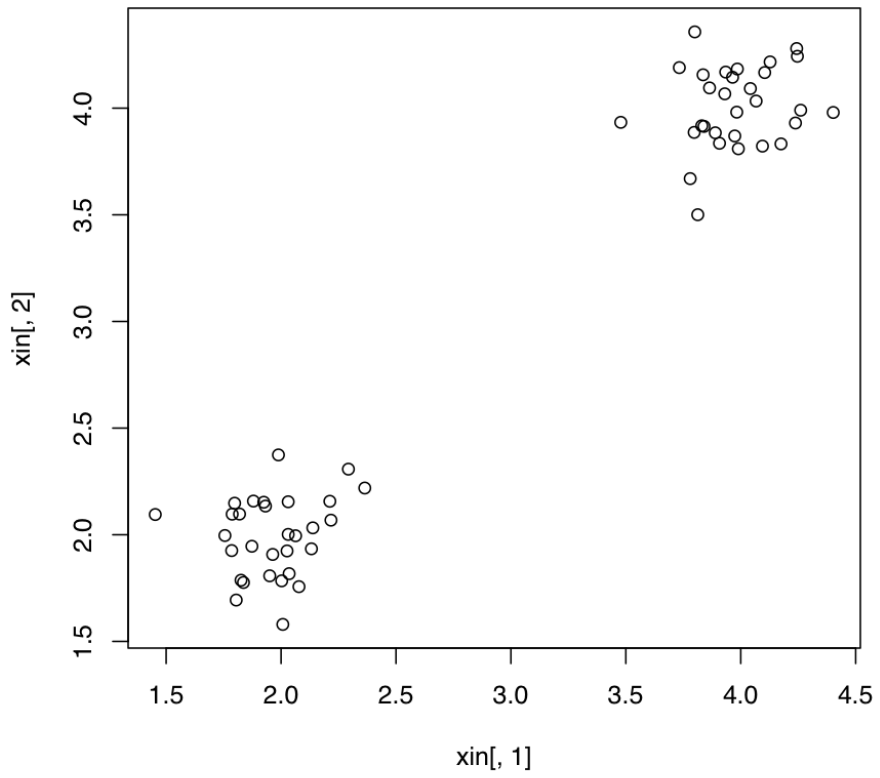

Figura 4.1: Dados amostrados de duas distribuições Gaussianas com média $m_1 = [2, 2]^T$ e $m_2 = [4, 4]^T$.

car claramente quatro submatrizes distintas que compõem o *kernel*, as quais serão caracterizadas aqui como $\mathbf{K}_{11}$, $\mathbf{K}_{12}$, $\mathbf{K}_{21}$ e $\mathbf{K}_{22}$. Considerando-se que os dois agrupamentos de dados caracterizam duas classes distintas, as submatrizes $\mathbf{K}_{11}$ e $\mathbf{K}_{22}$ contêm as relações intra-classes e as submatrizes $\mathbf{K}_{12}$ e $\mathbf{K}_{21}$ contêm as relações entre-classes. Desse modo, a estimativa de densidade de acordo com a Equação 2.6 pode ser reescrita através da composição das densidades estimadas para matrizes adjacentes, conforme Equações 4.1 e 4.2, nas quais os termos $P(x_i, y_i = -1|C_1)$, $P(x_i, y_i = -1|C_2)$, $P(x_i, y_i = +1|C_1)$ e $P(x_i, y_i = +1|C_2)$ representam as estimativas

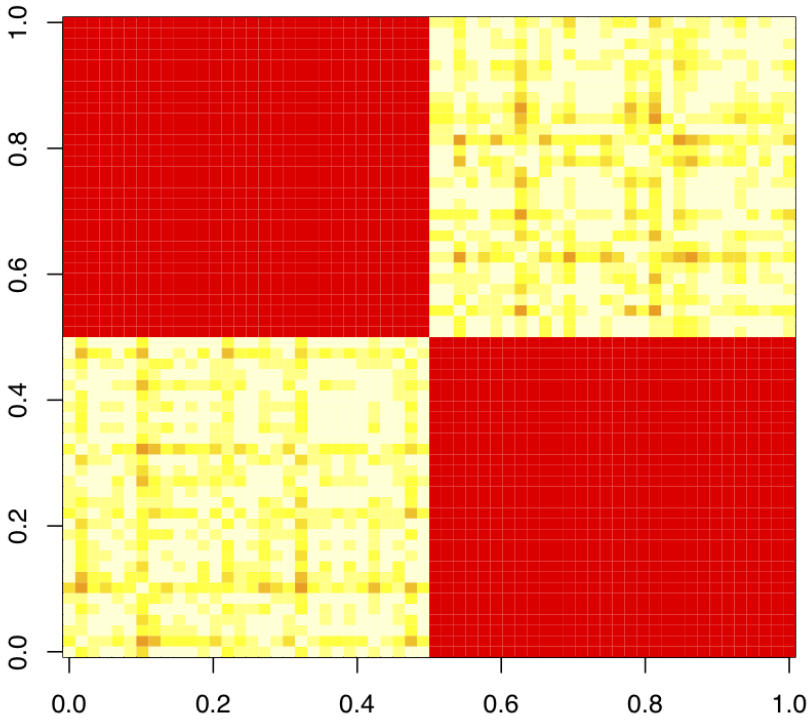

Figura 4.2: Kernel Gaussiano K para o exemplo da Figura 4.1 com $h = 1$.

$P(x_i|C_i)$ de acordo com os rótulos y_i.

$$\hat{f}(x_i \in C_1) = \overbrace{\frac{1}{N_1 h^n} \sum_{k=1}^{N_1} \mathbf{K}_{11}(x_i, x_k)}^{P(\{x_i, y_i=-1\}|C_1)P(C_1)} + \overbrace{\frac{1}{N_2 h^n} \sum_{p=1}^{N_2} \mathbf{K}_{12}(x_i, x_p)}^{P(\{x_i, y_i=-1\}|C_2)P(C_2)} \tag{4.1}$$

$$\hat{f}(x_i \in C_2) = \overbrace{\frac{1}{N_1 h^n} \sum_{k=1}^{N_1} \mathbf{K}_{21}(x_i, x_k)}^{P(\{x_i, y_i=+1\}|C_1)P(C_1)} + \overbrace{\frac{1}{N_2 h^n} \sum_{p=1}^{N_2} \mathbf{K}_{22}(x_i, x_p)}^{P(\{x_i, y_i=+1\}|C_2)P(C_2)} \tag{4.2}$$

Sendo conhecidos $P(C_1)$ e $P(C_2)$ é possível estimar a verossimilhança para os padrões de cada uma das classes C_1 e C_2 de acordo com as Equações 4.3 a 4.6. Em um problema de classificação binária, espera-se que as probabilidades estimadas pelas Equações 4.3 e 4.5 sejam maximizadas e aquelas estimadas pelas Equações 4.4 e 4.6 sejam minimizadas para cada um dos padrões $x_i \in D$. De fato, a maximização das diferenças entre estas quantidades fornece um caminho para a minimização do erro de aproximação do conjunto de dados com base somente na coerência entre rotulação e densidades estimadas.

$$P(\{x_i, y_1 = -1\}|C_1) = \frac{1}{N_1 h^n} \sum_{k=1}^{N_1} \mathbf{K}_{11}(x_i, x_k) \tag{4.3}$$

$$P(\{x_i, y_1 = -1\}|C_2) = \frac{1}{N_2 h^n} \sum_{k=1}^{N_2} \mathbf{K}_{12}(x_i, x_k) \tag{4.4}$$

$$P(\{x_i, y_1 = +1\}|C_1) = \frac{1}{N_1 h^n} \sum_{k=1}^{N_1} \mathbf{K}_{21}(x_i, x_k) \tag{4.5}$$

$$P(\{x_i, y_1 = +1\}|C_2) = \frac{1}{N_2 h^n} \sum_{k=1}^{N_2} \mathbf{K}_{22}(x_i, x_k) \tag{4.6}$$

Considerando-se que os rótulos $y_i, \forall x_i \in D$ são conhecidos, espera-se que a estimativa de densidade pelo KDE conforme Equação 2.6 seja capaz de maximizar as

probabilidades posteriores $P(C_1|x_i \in C_1)$ e $P(C_2|x_i \in C_2)$ e ao mesmo tempo minimizar as probabilidades cruzadas $P(C_1|x_i \in C_2)$ e $P(C_2|x_i \in C_1)$, ou seja, encontrar o máximo das funções representadas nas Equações 4.7 e 4.8.

$$f_{C_1} = \frac{1}{N_1 h^n} \sum_{k=1}^{N_1} \mathbf{K}_{11}(x_i, x_k) - \frac{1}{N_1 h^n} \sum_{k=1}^{N_1} \mathbf{K}_{21}(x_i, x_k) \tag{4.7}$$

$$f_{C_2} = \frac{1}{N_2 h^n} \sum_{k=1}^{N_2} \mathbf{K}_{22}(x_i, x_k) - \frac{1}{N_2 h^n} \sum_{k=1}^{N_2} \mathbf{K}_{12}(x_i, x_k) \tag{4.8}$$

Assim, a largura h utilizada deve maximizar as funções de custo representadas nas Equações 4.7 e 4.8, porém, na prática, a maximização das diferenças resulta em uma faixa de valores de h. Este comportamento era esperado, já que o problema geral de aproximação requer não somente a minimização do erro, mas também a minimização da complexidade do modelo, caracterizando o problema como bi-objetivo do ponto de vista de Otimização ((TEIXEIRA et al., 2000), (OKABE; JIN; SENDHOFF, 2003)). De maneira semelhante ao procedimento adotado por (SILVERMAN, 1986), deseja-se aqui também não somente a maximização das funções de custo representadas pelas Equações 4.7 e 4.8, mas também a minimização do Risco Estrutural (VAPNIK, 2000), através da maximização da margem de separação entre as classes. Assim, neste trabalho, a seleção do valor de h é realizada em duas etapas. Inicialmente, de acordo com um valor de erro permitido obtido através das funções de custo representadas pelas Equações 4.7 e 4.8, obtém-se um conjunto de valores de h. Na etapa seguinte, o valor de h que resulta na maior margem de separação entre as classes é selecionado entre aqueles valores obtidos na etapa anterior.

A metodologia adotada neste trabalho para a seleção de h através de duas funções-objetivo é análoga àquelas descritas para outros modelos de aprendizado, como redes

neurais artificiais, SVMs ou mesmo aproximadores polinomiais, em que uma função de erro e de complexidade são minimizadas. O método aqui descrito, por ser baseado na estrutura dos dados e no princípio da separação em região de baixa densidade, não requer uma busca exaustiva para o parâmetro h, já que funções de custo são descritas para o erro e para a suavização da resposta do modelo. Além do mais, como a função de erro é limitada por ϵ, o problema de natureza bi-objetivo ((TEIXEIRA et al., 2000), (OKABE; JIN; SENDHOFF, 2003)) é descrito como sub-problemas mono-objetivo. É claro que, por ser baseado em uma consideração *a priori* sobre uma característica da margem de separação, o desempenho do modelo dependerá da validade desta consideração para o problema em questão. Não obstante, mesmo outras máquinas de aprendizado, como as SVMs, se baseiam em algum princípio *ad-hoc*, como a maximização da margem de separação. Não é objetivo deste trabalho apresentar um método geral para a construção de classificadores, mesmo porque muitos dos resultados da literatura estão no limite do desempenho dos conjuntos de dados disponíveis. O que se pretende explorar é a coerência entre a geometria do problema e a indução de máquinas de aprendizado, particularmente para o caso de classificadores binários.

Com o objetivo de identificar os pontos da região da margem de separação em que serão calculadas as densidades, utilizou-se neste trabalho o método descrito por (TORRES; CASTRO; BRAGA, 2012), o qual se baseia no Grafo de Gabriel (DE BERG et al., 2008). Este método, proposto originalmente visando à seleção de modelos neurais de margem larga em aprendizado multi-objetivo (TEIXEIRA et al., 2000) e (TORRES; CASTRO; BRAGA, 2012), tem como uma de suas etapas a identificação de pontos médios entre as amostras das duas classes. Neste trabalho, os pontos médios obtidos serão utilizados como pontos de referência da região de separação nos quais as densidades devem ser avaliadas, visando à seleção do parâmetro h.

4.2 Seleção da Largura do *Kernel*: Baseada em Derivadas e Pontos Médios

De posse dos pontos médios, obtidos a partir do Grafo de Gabriel, as densidades são calculadas individualmente utilizando-se o KDE com valores de h que satisfaçam à restrição imposta às Equações 4.7 e 4.8. O problema geral de otimização resultante da combinação destas duas equações pode ser descrito como o problema de minimização do erro, caracterizado pela maximização da função-objetivo, apresentada na Equação 4.9

$$J_1 = \frac{1}{N_1 h^n}\sum_{k=1}^{N_1}\mathbf{K}_{11}(x_i,x_k) - \frac{1}{N_1 h^n}\sum_{k=1}^{N_1}\mathbf{K}_{21}(x_i,x_k) + \frac{1}{N_2 h^n}\sum_{k=1}^{N_2}\mathbf{K}_{22}(x_i,x_k) - \frac{1}{N_2 h^n}\sum_{k=1}^{N_2}\mathbf{K}_{12}(x_i,x_k), \quad (4.9)$$

sujeito a uma condição de suavização descrita como uma restrição a uma segunda função J_2 conforme a Equação 4.10 que se segue:

$$\begin{gathered}\arg_{max}\, J_1 \\ \text{sujeito a } J_2 < \delta\end{gathered} \quad (4.10)$$

A forma geral apresentada para a Equação 4.10 se assemelha àquela descrita para muitos outros métodos indutivos como SVMs ou redes neurais (VAPNIK, 2000; HAYKIN, 1994), nos quais uma função de erro empírico é minimizada sujeita a uma condição que de alguma forma impõe uma restrição à capacidade efetiva do modelo (ou ao erro no caso de SVMs). Na formulação geral de treinamento de SVMs

e também no Aprendizado Multi-objetivo de redes neurais, a função de custo que representa a complexidade do modelo, como a função J_2, está relacionada à norma do vetor de pesos, que garante maximização da margem de separação (TEIXEIRA et al., 2000). Não obstante, em ambas as abordagens uma etapa decisória é necessária para a escolha do modelo final. Um exemplo de decisor para redes neurais é o decisor de margem larga baseado no Grafo de Gabriel (TORRES; CASTRO; BRAGA, 2012); para SVMs a prática mais comum é a busca exaustiva por validação cruzada ou *grid-search* (VAN GESTEL et al., 2004). Apesar de o problema de Programação Quadrática (QP) que caracteriza o aprendizado de SVMs ter solução única global, ele é resolvido para um determinado valor de constante de regularização, o qual é selecionado *a priori*. Portanto, de maneira análoga a outros modelos de aprendizado, a função J_2 da Equação 4.10 representará o modelo de seleção ao qual será incorporado algum critério definido *a priori*.

Ao minimizar a função J_1 obtém-se um intervalo de valores de h que satisfazem à tolerância de erro da Equação 4.10, $[h_{min}, h_{erro \leq \epsilon}]$, sendo h_{min} a menor largura que minimiza J_1, e h_{erro}, a maior largura que minimiza J_1 sujeito à variável de folga ϵ. Qualquer valor de h no intervalo satisfaz à restrição de J_1, no entanto, a restrição a J_2 determinará o valor de h a ser escolhido.

Seja **PM** a matriz de coordenadas dos pontos médios calculados de acordo com o método de (TORRES; CASTRO; BRAGA, 2012) e **D** a matriz das densidades estimadas de acordo com a equação para o *kernel* Gaussiano (Equação 2.6), nos pontos médios, para todo h pertencente ao intervalo. Partindo-se do princípio de que as densidades devem ser minimizadas na região de separação (CHAPELLE et al., 2006), o valor de h a ser selecionado no intervalo deve garantir a condição de minimização em **PM**. Como o critério de decisão deve ser tal que haja coerência no comportamento da densidade em todos os pontos de **PM**, adotou-se aqui o critério de decisão descrito pela Equação 4.11, que garante a direção de minimização para

todos os pontos médios.

$$\frac{dD}{dh} < 0, \forall p \in \mathbf{PM} \tag{4.11}$$

Como pode ser visto na Figura 4.3, o critério de decisão da Equação 4.11 visa à busca da coerência no comportamento das densidades em relação a h. A minimização direta da soma das densidades em todos os pontos, por exemplo, pode não resultar em um bom critério de decisão já que valores de densidade podem ter valores discrepantes como pode ser observado no gráfico. O decisor da Equação 4.11 garante uma condição mínima de coerência para os valores de densidade nos pontos médios, ou seja, que a função aproximadora está tendendo para suavização em todos os pontos médios.

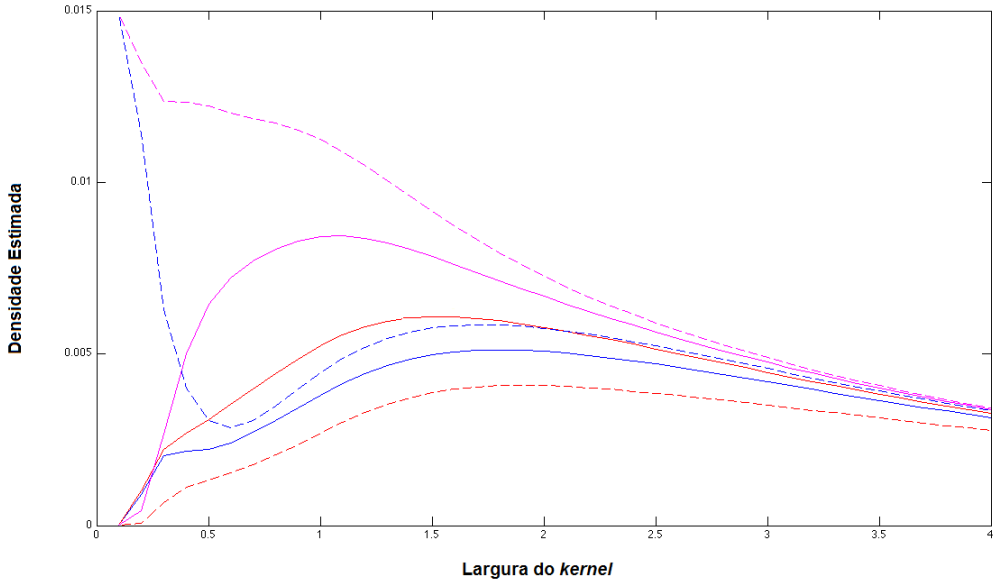

Figura 4.3: Variação da densidade nos pontos da margem com relação à variação da largura do kernel.

4.3 Seleção da Largura do *Kernel*: Baseada na Diferença de Comportamento da Densidade na Margem e fora dela

A partir da hipótese de que a região de separação entre classes deve ser de baixa densidade (CHAPELLE et al., 2006), o segundo método proposto estuda o efeito da largura do *kernel* na densidade dos pontos na margem e dentro das classes.

Para satisfazer à função-objetivo J_1, deseja-se que a densidade acumulada dentro das classes seja alta e que a densidade nos pontos da margem seja baixa. Além disso, tal como mencionado no início deste Capítulo, é necessário controlar a complexidade do modelo, representada aqui pelas situações de *overfitting* (alta complexidade) e *underfitting* (baixa complexidade).

Na Figura 4.4 o efeito da variação da largura h nas densidades na margem e fora dela pode ser visto. A região hachurada representa o espaço de larguras h que satisfazem a função-objetivo J_1. A partir do ponto de cruzamento entre as duas curvas, a densidade na margem torna-se mais alta do que nas classes, contrariando a hipótese da baixa densidade na fronteira de separação entre as classes.

Assim, uma restrição J_2 é necessária para escolher qual dos valores de largura, entre no intervalo $[h_{min}, h_{cruzamento}]$, modela melhor a relação geométrica entre os pontos de cada classe. Neste método, utilizou-se validação cruzada, sendo o h com melhor desempenho nos testes o escolhido. Inicialmente, a métrica escolhida foi a acurácia de um classificador binário baseado em KDE, à qual não obteve bons resultados. Tal comportamento pode ser justificado observando as métricas apresentadas na Figura 4.5. Como a acurácia mede o número de acertos em relação ao número total de amostras, em caso de bases com classes desbalanceadas, frequentemente o maior valor está relacionado a um maior acerto da classe majoritária. Com isso, a largura h selecionada tenderia a privilegiar tal classe. Para escolher a largura h com mais

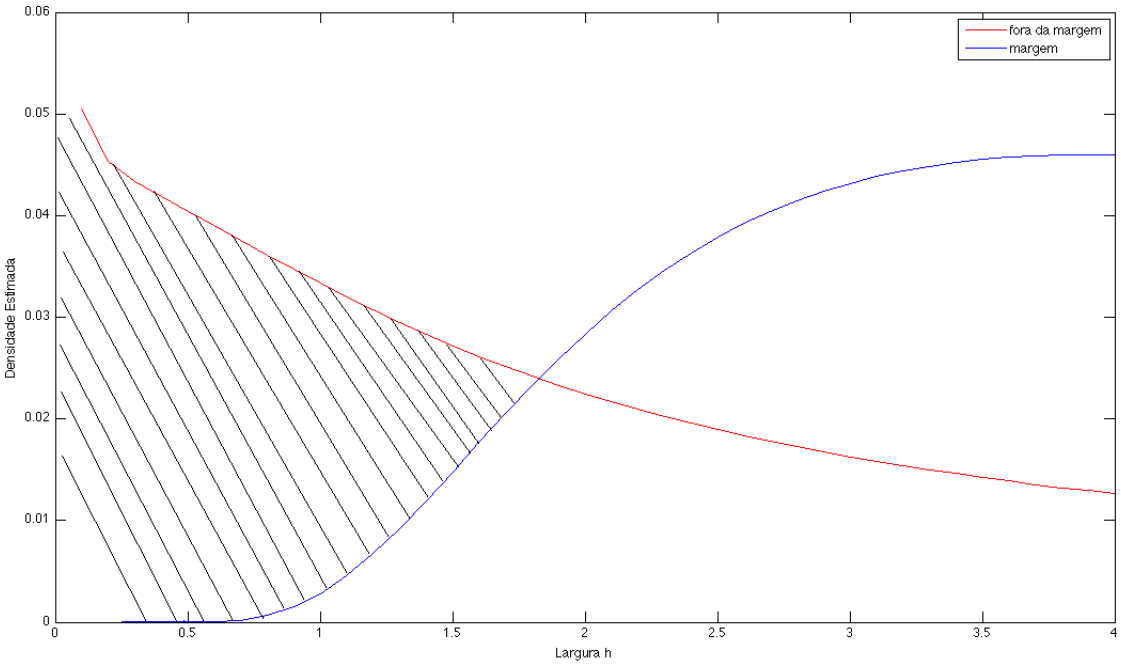

Figura 4.4: Variação da densidade na margem e dentro das classes com relação à variação da largura do kernel.

equilíbrio entre as classes utilizou-se a média geométrica tal como na Seção 3.2.3.

4.4 Conclusão do Capítulo

Neste Capítulo foram apresentados os dois métodos propostos para a estimação do parâmetro suavizador do *kernel*. Com tais métodos tem-se por objetivo obter um valor de largura para o *kernel* que, uma vez utilizado no KDE, represente de forma apropriada a relação geométrica entre os dados, bem como a coerência com os rótulos.

No Capítulo 5 serão apresentados os resultados obtidos pelo KDE-Bayes com as larguras estimadas de acordo com os métodos aqui expostos e outros encontrados

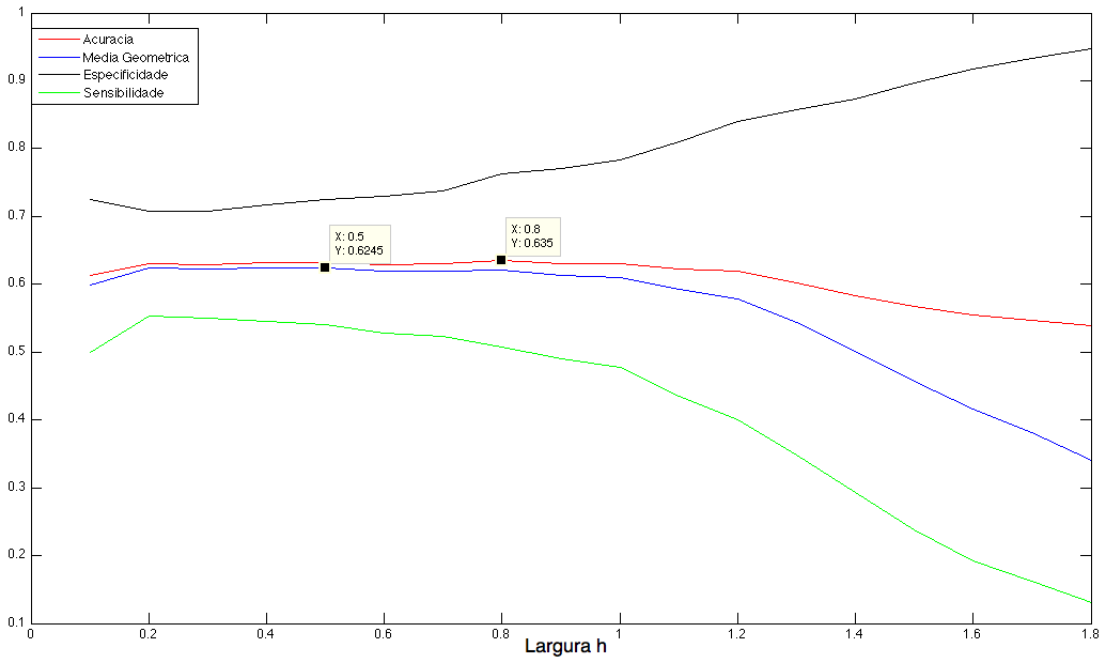

Figura 4.5: Comportamento das métricas de um classificador binário (acurácia, média geométrica, especificidade e sensibilidade) de acordo com a variação da largura h.

na literatura (descritos na Seção 2.4).

5 EXPERIMENTOS

> "'The Answer to the Great Question... Of Life, the Universe and Everything... Is... Forty-two,' said Deep Thought, with infinite majesty and calm."
>
> *The Hitchhiker's Guide to the Galaxy*, Douglas Adams

Neste Capítulo são apresentados os seis experimentos realizados para avaliar o desempenho dos métodos propostos nos Capítulos 3 e Capítulo 4. As Seções a seguir estão organizadas em duas sub-seções, uma com as bases de dados utilizadas no experimento e outra com a metodologia e resultados obtidos.

5.1 Experimento 1: KDE-Bayes Univariado

5.1.1 Base de Dados: Quimioterapia Neoadjuvante

Em 2008, baseado nos dados internacionais disponíveis mais recentes, foram estimados 12,4 milhões de novos casos e 7,6 milhões de mortes por câncer no mundo.

Os tipos mais comuns em termos de incidência são o de pulmão (1,52 milhões de casos), de mama (1,29 milhões de casos) e de cólon (1,15 milhões de casos) (BOYLE; LEVIN, 2008).

O câncer de mama, foco do estudo do projeto CAPES-COFECUB no qual este trabalho se insere, pode ser tratado através de quimio e radioterapia e/ou cirurgia, que pode ser parcial (quadrantectomia) ou radical (com a retirada total da mama). No caso do câncer operável, pode-se fazer uso da quimioterapia antes da cirurgia (quimioterapia neoadjuvante) para diminuir o tamanho do tumor e evitar que o mesmo se espalhe por outros órgãos.

A quimioterapia, porém, apresenta muitos efeitos colaterais uma vez que age não somente nas células cancerosas mas também em outras células do corpo que possuem a mesma característica de crescimento e multiplicação acelerados que os tumores. Dentre os efeitos colaterais estão anemia e diminuição da resistência a infecções causadas pela ação nas células produtoras dos glóbulos sanguíneos vermelhos e brancos, aumentando a vulnerabilidade do paciente. A previsão da resposta a quimioterapia neoadjuvante pode levar à seleção de tratamentos apropriados para cada paciente, reduzindo o sofrimento deste durante os procedimentos.

Em geral, prever a eficiência do tratamento utilizando características clínicas dos pacientes não funciona adequadamente e, por isso, utiliza-se a informação baseada na expressão de mRNA para obter perfis de diversos tumores e assim realizar a previsão (NATOWICZ et al., 2008). Os perfis obtidos através de biópsias podem ser correlacionados com características como o tamanho do tumor, o estágio em que se encontra, recorrência do tumor e sensibilidade ao tratamento. Uma resposta patológica completa (PCR) na cirurgia está correlacionada com um excelente resultado, enquanto uma resposta incompleta (NoPCR) está associada a um resultado ruim. É importante prever corretamente se uma paciente é PCR ou NoPCR pois no segundo

caso outras alternativas de tratamento podem ser buscadas.

Os dados deste problema são escassos e esparsos, compostos de 133 amostras cada uma com 22283 características, dificultando o uso de modelos paramétricos de estimação de densidade, que poderiam ser enviesados pelas amostras. Além disso, devido ao alto número de variáveis, uma abordagem multidimensional também é prejudicada, aumentando a necessidade de métodos de seleção de características apropriados. Outra característica dos dados é o desbalanceamento das duas classes, que inspira cuidados quanto à escolha do método de seleção, para evitar que esta seja tendenciosa com relação a classe majoritária.

Uma triagem clínica com 133 pacientes com câncer de mama em estágio entre I - III, foi conduzida no Nellie B. Connaly Breast Center no M.D. Anderson Cancer Center da Universidade do Texas (HESS et al., 2006). Desse total, 82 pacientes são de Houston, Estados Unidos (dados americanos) e 51 são de Villejuif, França (dados franceses).

Para cada um dos 133 pacientes existem dados de 22283 sondas, obtidas através da expressão gênica feita com microarrays. Os pacientes estão divididos entre as classes da seguinte maneira:

- Dos 82 pacientes de Houston, EUA:
  - 61 pacientes noPCR e 21 pacientes PCR
- Dos 51 pacientes de Villejuif, França
  - 38 pacientes noPCR e 13 pacientes PCR.

Estes dados estão disponíveis publicamente em `http://bioinformatics.mdanderson.`

org/pubdata.html.

5.1.2 Metodologia e Resultados

O primeiro experimento realizado tinha como objetivo determinar se o método KDE-Bayes (Capítulo 3) obtinha bons resultados como seletor de características. Para isto, utilizou-se um subconjunto da base de dados de quimioterapia neoadjuvante (sub-seção 5.1.1), com as 30 características selecionadas pelo trabalho de (NATOWICZ et al., 2008).

A partir desse conjunto reduzido foram selecionadas para estudo as três sondas que apresentaram acurácia maior do que 50%, 213134_x_at, 205548_s_at e 209604_s_at. Esta segunda redução foi feita para que fosse possível testar as características individualmente e também em conjunto, sem um custo computacional alto.

Para efeitos de comparação as mesmas características foram testadas utilizando regressão logística, método amplamente utilizado para classificação de dados médicos com saídas dicotômicas (BAGLEY; WHITE; GOLOMB, 2001). Métodos de regressão são frequentemente utilizados para descrever a relação entre uma variável de saída e uma ou mais componentes de entrada. Na regressão logística, essa variável de saída é binária, ou seja, o vetor de características de entrada é classificado em duas classes opostas (HOSMER; LEMESHOW, 2000).

A função logística é dada por

$$f(z) = \frac{1}{1 + e^{-z}}$$

onde z é um conjunto de características e $f(z)$ é a probabilidade de uma determinada saída, dadas as entradas. A função logística aceita como entrada qualquer valor entre mais e menos infinito, porém sua saída é limitada por valor no intervalo $[0, 1]$.

Duas abordagens foram propostas: na primeira cada uma das sondas é utilizada como um classificador e na segunda as três são utilizadas em conjunto. No primeiro caso, a densidade de cada classe (PCR e noPCR) é estimada de acordo com o valor de expressão gênica de cada sonda, utilizando os dados coletados em Houston. Após isso, o classificador bayesiano utiliza a densidade estimada pelo KDE. Para o regressor logístico o procedimento é o mesmo, e a clasificação é realizada utilizando somente uma das sondas.

Na segunda abordagem o classificador bayesiano utiliza um critério de voto da maioria para decidir a qual classe cada uma das entradas deve ser designada. Para o regressor logístico as três sondas são utilizadas ao mesmo tempo, isto é, o regressor é ajustado de acordo com a informação contida nas três sondas.

Plotando a estimativa das densidades é possível notar que a distribuição da segunda sonda utilizada é bastante diferente da primeira e terceira sondas. As figuras 5.1, 5.2 e 5.3 mostram a diferença entre elas.

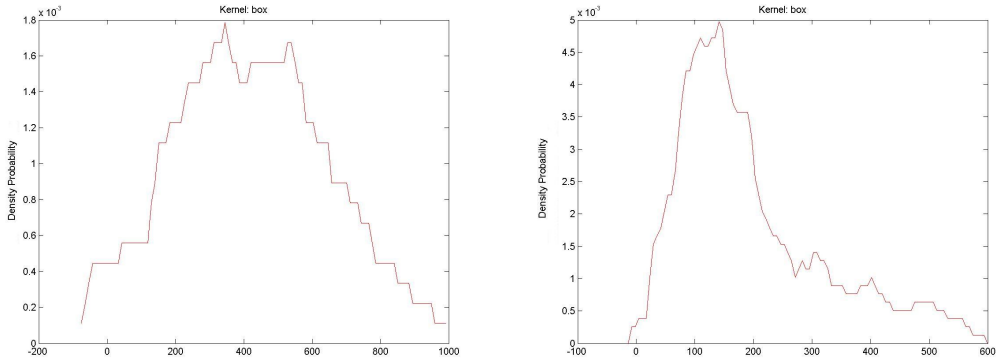

Figura 5.1: Função de densidade estimada para a classe PCR (à esquerda) e noPCR (à direita) para a sonda 1.

Os melhores resultados foram alcançados quando a sonda 2 foi utilizada sozinha como um classificador. O fato de sua distribuição ser diferentes das outras pode indicar que o modelo global seria enviesado por ela, com o classificador utilizando

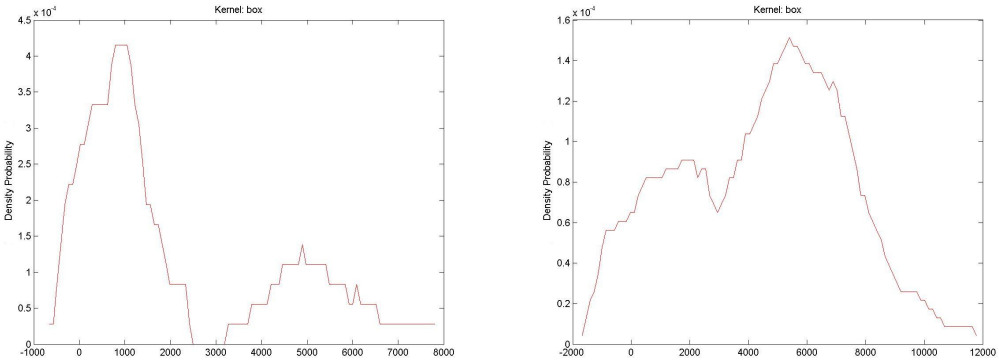

Figura 5.2: Função de densidade estimada para a classe PCR (à esquerda) e noPCR (à direita) para a sonda 2.

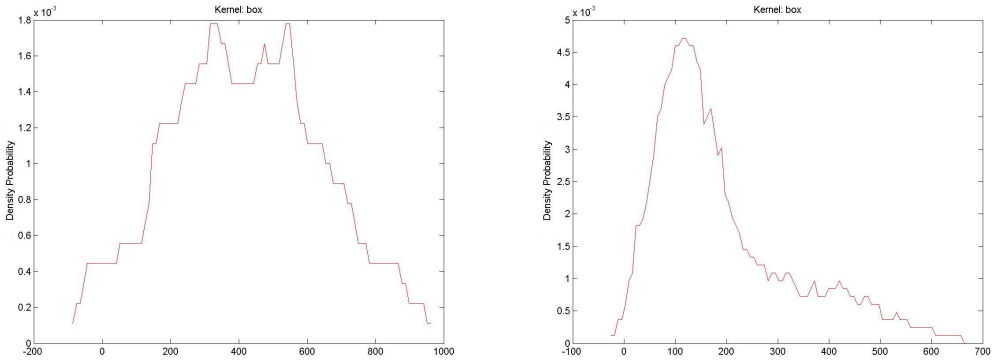

Figura 5.3: Função de densidade estimada para a classe PCR (à esquerda) e noPCR (à direita) para a sonda 3.

menos a informação gerada pelas duas outras sondas.

Comparando as tabelas 5.1, 5.2, 5.3 e 5.4 observa-se que a sonda 2 (205548_s_at) gera o melhor classificador, mesmo no caso em que as três sondas são utilizadas em conjunto. É importante enfatizar que o valor de sensibilidade obtido pela sonda 2 é excelente, pois indica um número baixo de falsos negativos. Neste problema é muito importante identificar corretamente os pacientes que teriam sucesso com a quimioterapia neodjuvante, evitando assim que procedimentos desnecessários e mais radicais sejam utilizados.

Tabela 5.1: Resultados dos dois métodos para a sonda 1 (213134_x_at).

Método	Classificador Bayesiano	Regressão Logística
Sonda	213134_x_at	213134_x_at
Ac	0,8627	0,7450
Se	0,6153	0,3846
Es	0,8780	0,8048

Tabela 5.2: Resultados dos dois métodos para a sonda 2 (205548_s_at).

Método	Classificador Bayesiano	Regressão Logística
Sonda	205548_s_at	205548_s_at
Ac	0,9019	0,7450
Se	0,9230	0,3076
Es	0,8994	0,8994

Tabela 5.3: Resultados dos dois métodos para a sonda 3 (209604_s_at).

Método	Classificador Bayesiano	Regressão Logística
Sonda	209604_s_at	209604_s_at
Ac	0,8627	0,7647
Se	0,6153	0,1538
Es	0,8780	0,9024

Apesar dos resultados para a sonda 2 serem melhores do que os com as outras, os resultados obtidos com as sondas 1 e 3 e as três sondas juntas são melhores do que os resultados obtidos usando regressão logística. Os resultados apresentados para estas sondas são similares aos apresentados na literatura e superiores em alguns

Tabela 5.4: Resultados dos dois métodos com as três sondas agrupadas.

Método	Classificador Bayesiano	Regressão Logística
Ac	0,8823	0,7647
Se	0,6923	0,4615
Es	0,8780	0,8048

casos (NATOWICZ et al., 2008; HESS et al., 2006). Isso indica que a metodologia proposta é apropriada para o problema, e se estendida possivelmente dará melhores resultados do que os conhecidos atualmente. Outro resultado interessante é que a performance das sondas 1 e 3 é praticamente a mesma, refletindo a similaridade na estrutura dos dados (figura 5.1 e figura 5.3).

Os resultados obtidos com regressão logística foram os piores de todas abordagens. A acurácia está em torno de 75 %, com a melhor sensibilidade em torno de 46%. Este método impõe uma estrutura fixa, ajustando os dados a uma curva logística, que dificilmente acomoda qualquer informação que não tenha este formato.

Pode-se observar neste experimento que o agrupamento de características boas não necessariamente gera um classificador com uma boa performance. Na seleção univariada o comportamento de uma característica na presença das demais não é avaliada, assim uma determinada característica pode ser muito boa quando analisada isoladamente mas quando combinada com outra pode resultar em um classificador pior. Isso ocorre pois as características agrupadas podem ser redundantes ou até mesmo fornecerem informações conflitantes, casos que um seletor univariado não detectaria.

Além do ponto de vista do classificador é importante também fazer uma análise do genes correspondentes às sondas selecionadas. As sondas 205548_s_at e 213134_x_at mapeiam o mesmo gene, BTG3, que possui propriedades antiproliferativas e inibe a expressão do gene E2F3 (OU et al., 2007). Este último é super expressado em casos de câncer de pulmão e atua no ciclo celular de células que estão

se proliferando. Ou seja, o gene BTG3 está diretamente relacionado a proliferação de células (cancerígenas ou não), sendo relevante para este estudo. Além disso, ele possui uma sonda que se comporta como um bom classificador. A outra sonda estudada, 209604_s_at, mapeia o gene GATA3, um ativador da transcrição produzido pelas células luminais das mamas, consideradas como uma das responsáveis por originar certos tipos de cânceres de mama. Sua análise imunohistoquímica pode ser base para um novo teste clínico para prever a recorrência de tumor em casos de câncer de mama (MEHRA et al., 2005).

5.2 Experimento 2: AG-KDE-Bayes Multivariado

O segundo experimento realizado teve como objetivo aprimorar os resultados obtidos no experimento da sub-seção 5.1, usando uma abordagem multivariada de seleção de características para a base de quimioterapia neoadjuvante (Sub-seção 5.1.1).

Para este experimento foi utilizado o mesmo subconjunto de trinta sondas do teste anterior, assim é possível analisar a diferença entre os resultados. O método empregado na seleção de características foi o apresentado na sub-seção 3.2.3, sendo os parâmetros utilizados no algoritmo genético a seleção por roleta, cruzamento de dois pontos com probabilidade 0,8, mutação com probabilidade 0,1, elitismo, população de duzentos indivíduos (com tamanho limitado de no máximo cinco sondas) e cem gerações.

No que concerne à estimação não-paramétrica da densidade o valor da largura h utilizado foi o descrito em (SILVERMAN, 1986) para dados normalizados, dado por:

$$h_1 = A(K) * \ n^{\frac{-1}{d+4}} \tag{5.1}$$

onde n é o número de amostras, d o número de dimensões e $A(K)$ para o kernel

gaussiano multivariado de dimensão d

$$A(k) = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}}.$$

Após trinta iterações foram analisados os melhores indivíduos de cada população de acordo com sua sensibilidade e especificidade. Além dessa análise os resultados também foram comparados com os obtidos por (HESS et al., 2006), utilizando DLDA (*diagonal linear discriminant analysis*) (DUDOIT; FRIDLYAND; SPEED, 2002) e as trinta sondas selecionadas no mesmo trabalho. Para tal comparação foram utilizados os dados apresentados por (TABCHY et al., 2010), composto por 91 pacientes, 19 da classe PCR e 72 da classe noPCR.

De acordo com o algoritmo genético o melhor indivíduo foi o composto por quatro características, apresentado na tabela 5.5, com sensibilidade e especificidade acima de 0,92.

Tabela 5.5: Sensibilidade, Especificidade, Acurácia e Matriz de Confusão para o melhor conjunto de sondas indicado pelo Algoritmo Genético.

Sondas		Matriz de Confusão		
203693_s_at	se = 0,9231		P	N
213134_x_at	es = 0,9211	P	12	3
214053_at	ac = 0,9216			
217542_at		N	1	35

Comparando os resultados desse grupo de características com os resultados de (HESS et al., 2006) utilizando DLDA (tabela 5.6) percebe-se que o desempenho dos dois é bem próximo, com a sensibilidade maior naquele. Porém, o grupo selecionado pelo método AG-KDE-Bayes utiliza somente quatro sondas, enquanto o método de Hess utiliza trinta.

O segundo melhor indivíduo selecionado pelo genético é composto por quatro características, tendo sensibilidade igual a 1. Na tabela 5.7, são apresentados os valores

Tabela 5.6: Comparativo utilizando DLDA entre o grupo de sondas selecionadas pelo AG e as 30 sondas de (HESS et al., 2006)

Alg. Genético				Hess - 30			
Trein.		Val.		Trein.		Val.	
18	15	12	22	20	14	13	20
3	46	7	50	1	47	6	52

da métricas utilizadas, além da matriz de confusão.

Tabela 5.7: Sensibilidade, Especificidade, Acurácia e Matriz de Confusão para o segundo conjunto de sondas indicado pelo Algoritmo Genético.

Sondas		Matriz de Confusão		
203693_s_at	se = 1		P	N
206401_s_at	es = 0,8684	P	13	5
214053_at	ac = 0,9020			
217542_at		N	0	33

Comparando os resultados obtidos pelo método AG-KDE-Bayes com os obtidos por Hess (tabela 5.8), também pode-se observar que o desempenho também foi similar com os dois apresentando a mesma acurácia.

Tabela 5.8: Comparativo utilizando DLDA entre o grupo de sondas selecionadas pelo AG e as 30 sondas de (HESS et al., 2006)

Alg. Genético				Hess - 30			
Trein.		Val.		Trein.		Val.	
21	20	10	17	20	14	13	20
0	41	9	55	1	47	6	52

Dos genes mapeados pelas sondas selecionadas temos, além dos dois já apresentados na sub-seção 5.1, outros três cujas funções estão diretamente ligadas ao comportamento anormal da célula com câncer. O gene associado a sonda 214053_at, ERBB4, está associado com a apoptose de células tumorais em tumores primários de mama (NARESH et al., 2006), enquanto o gene MDM2 (sonda 217542_at) atua como regulador do P53, supressor de tumores. Entre o segundo conjunto de sondas selecionado, que obteve sensibilidade igual a um nos dados de treinamento, está a

206401_s_at que mapeia o gene MAPT. A baixa expressão desse gene contribui para o aumento da sensibilidade do tumor a quimioterapia neodjuvante (ROUZIER et al., 2005) e, por isso, é observada nos pacientes PCR. Tal propriedade fez da sonda para MAPT uma boa característica para identificar pacientes PCR, vide o resultado de sensibilidade apresentado.

Observando os resultados apresentados nesta sub-seção pode-se concluir que é possível obter um desempenho bom utilizando um número menor de características, validando, assim, o método de seleção AG-KDE-Bayes multivariado. Além disso, a seleção de características multivariada mostra-se mais eficiente do que a utilização de agrupamentos a partir de uma seleção univariada.

5.3 Experimento 3: Seleção por AUC

5.3.1 Base de Dados: Leucemia Aguda

O conjunto de dados de leucemia foi apresentado por (GOLUB et al., 1999), que utilizava medição de expressão gênica em leucemia aguda.

Os casos de leucemia foram classificados em dois tipos:

- ALL - *Acute Lymphoblastic Leukemia* (leucemia linfoblástica aguda), de precursores linfóides e,
- AML - *Acute Myeloid Leukemia* (leucemia mielóide aguda), de precursores mielóides.

Apesar de as duas classes de leucemia serem bem conhecidas, não há um método

estabelecido para realizar diagnóstico, sendo necessária a experiência de cada médico para realizá-lo. Dado que o tratamento para cada um dos tipos de câncer é diferente, é importante distinguir entre eles.

Os dados experimentais (que podem ser obtidos em `http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43`), foram divididos em dois grupos, para realização de treinamento e teste. O grupo de treinamento consiste em 38 amostras de medula óssea coletadas no momento do diagnóstico. Dessas, 27 eram de pacientes com ALL e 11 de pacientes com AML. O grupo de testes consiste de 34 amostras, sendo 24 de pacientes com ALL e 10 de pacientes com AML. Para cada um dos pacientes foram utilizadas 7129 sondas para a obtenção dos dados de expressão gênica.

5.3.2 Metodologia e Resultados

Os resultados mostrados a seguir serão divididos em duas partes: comparação das AUCs obtidas a partir de análise discriminante, linear e quadrática, e um regressor logístico (Sub-seção 5.1.2) e comparação das sondas selecionadas com as apresentadas em (GOLUB et al., 1999).

Funções de discriminação linear são definidas pela combinação linear dos componentes do vetor de argumentos, dado por

$$f(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i$$

onde w_i são os pesos das entradas e w_0 define o limiar (GUYON et al., 2006). Se um vetor $\mathbf{x}$ satisfaz a condição de $f(\mathbf{x}) > 0$, então o modelo atribui o rótulo da classe positiva para ele, caso contrário o rótulo da classe negativa é atribuído. A equação $f(\mathbf{x}) = 0$ define uma superfície de decisão que separa os pontos atribuídos a classe 1 dos da classe 2. Quando $f(\mathbf{x})$ é linear, tal superfície de decisão é um hiperplano.

Com a adição de termos envolvendo produtos dos pares das componentes de x, uma função discriminante quadrática é obtida:

$$f(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij} x_i x_j$$

Dado que $x_i x_j = x_j x_i$, é possível assumir que $w_{ij} = w_{ji}$ sem perda de generalidade. Assim, a função de discriminação quadrática tem $d(d+1)/2$ coeficientes a mais do que a função linear, sendo possível produzir superfícies de separação mais complexas. A superfície de separação definida por $g(\mathbf{x}) = 0$, neste caso, é uma hiperquádrica (DUDA; HART; STORK, 2000).

Como esperado os resultados obtidos pelo KDE-Bayes foram muito superiores aos dos outros métodos. Isso se deve ao fato da informação da densidade das classes estimada pelo KDE incorporar a informação local dos dados. O desempenho dos outros foi similar, sendo o discriminante linear ligeiramente superior em relação ao regressor logístico e ao discriminante quadrático. Na tabela 5.9 apresentamos o resultado quantitativo para AUC igual a um, maior ou igual a $0,95$ e maior ou igual a $0,8$.

Tabela 5.9: Resultado quantitativo dos métodos para as AUCs. Os valores da tabela se referem ao número de sondas com AUC igual ao valor das linhas

AUC	KDE	Log.Reg.	DL	DQ
1	959	1	1	1
≥0,95	1996	14	18	12
≥0,80	5764	240	323	244

Nas figuras 5.4 e 5.5 apresentamos as AUCs para o método do KDE para uma sonda com AUC $= 1$ e AUC $= 0,95$.

Comparando os resultados obtidos pelos métodos aplicados no presente trabalho, somente a sonda 6855 - TCF3 *Transcription factor* 3 (E2A *immunoglobulin enhancer*

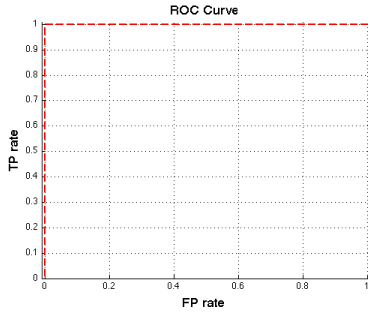

Figura 5.4: Gráfico ROC com AUC = 1. No eixo x a taxa de falsos positivos e no eixo y a taxa de verdadeiros positivos.

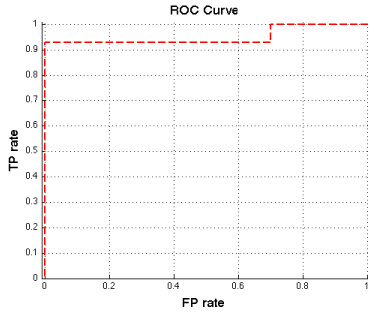

Figura 5.5: Gráfico ROC com AUC = $0,95$. No eixo x a taxa de falsos positivos e no eixo y a taxa de verdadeiros positivos.

binding factors E12/E47) também foi apresentada em (GOLUB et al., 1999). Tal sonda apresenta AUC = 1 e foi apontada por todos os métodos, seu gene está envolvido em rearranjos cromossomiais recorrentes, associados com leucemia linfoblástica aguda em crianças (DP, 2003).

Outras dez sondas também com AUC = 1 apontadas pelo KDE-Bayes foram:

- Sonda 21 - AFFX-DapX-M_at (*endogenous control*)
- Sonda 40 - AFFX-HUMRGE/M10098_M_at (*endogenous control*)

- Sonda 41 - AFFX-HUMRGE/M10098_3_at (*endogenous control*)
- Sonda 52 - AFFX-M27830_5_at (*endogenous control*)
- Sonda 53 - AFFX-M27830_M_at (*endogenous control*)
- Sonda 55 - AFFX-HSAC07/X00351_3_st (*endogenous control*)
- Sonda 65 - VRK1 AB000449_at
- Sonda 119 - RAS-RELATED PROTEIN RAB-11A AF000231_at
- Sonda 123 - "TTF-I *interacting peptide* 20 mRNA, *partial* cds"AF000560_at
- Sonda 126 - *Transmembrane protein* mRNA AF000959_at

Uma vez que as 50 sondas em (GOLUB et al., 1999) foram escolhidas arbitrariamente, as sondas acima poderiam ser boas escolhas para integrar o classificador apresentado pelo trabalho supra-citado.

5.4 Experimento 4: AG-KDE-Bayes Multivariado - Comparação com Outros Seletores de Características

5.4.1 Base de Dados: Oncologia

Para mesurar o desempenho do método proposto na Seção 3.2.3 em comparação a outros métodos de seleção de características, seis bancos de dados de oncologia, publicamente disponíveis, foram escolhidos: Colon (`http://genomics-pubs.princeton.edu/oncology/affydata/index.html`), Lymphoma (`http://www.gems-system.org/`), Leukemia, Brain (os dois últimos disponíveis em `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`), Prostate (`http://www.gems-system.org/`), e Ovarian

(`http://data.cgt.duke.edu/clinicalcancerresearch.php`). Tais bancos de dados tem entre 200 e 22283 características e menos de 103 amostras. Desse modo, são bons candidatos a seleção de características. Em adição às bases supra-mencionadas, a base apresentada na Sub-seção 5.1.1 também foi utilizada.

5.4.2 Metodologia e Resultados

Para as bases de dados de oncologia, os resultados obtidos pelo AG-KDE-Bayes foram comparados com dois outros métodos de seleção: *δ-test*, um método de filtro baseado na otimização de uma função bi-objetivo que visa a maximização da distância entre-classes e a minimização do número de características, e ABEUS, um *wrapper* baseado na otimização da performance de um classificador (GARDEUX et al., 2013). Foi realizada validação cruzada *3-fold* e o desempenho dos métodos foram medidos em relação à sua acurácia, especificidade, sensibilidade e valores preditivos positivo e negativo.

Os parâmetros do AG-KDE-Bayes foram: largura h apresentada na Equação 2.10, 150 indivíduos, 100 gerações, $0,8$ como probabilidade de recombinação, $0,7$ como probabilidade de mutação e elitismo de um indivíduo. A probabilidade de mutação é mais alta do que o usual porque um grande números de mudanças no indivíduo é desejado neste caso.

Na Tabela 5.10, são exibidos os resultados para os seis bancos de dados de oncologia. Em geral, o AG-KDE-Bayes seleciona um número menor de características e tem desempenho semelhante ao obtidos utilizando-se os métodos *δ-test* e ABEUS.

Após avaliado o desempenho do AG-KDE-Bayes nos bancos de dados de oncologia foram utilizados os dados de expressão gênica de pacientes de câncer de mama (Subseção 5.1.1). Diferentemente da redução de características realizada na Seção 5.2, o

Tabela 5.10: Validação cruzada *3-fold* do AG-KDE-Bayes, δ-KDE-Bayes e ABEUS-KDE-Bayes. Ac = acurácia, Se = sensibilidade, Es = especificidade, PPV, NPV: valores preditivos positivo e negativo.

dados	colon	lymphoma	leukemia	prostate	brain	ovaries
AG-KDE-Bayes						
#características	6,9 ±0,11	6,5 ±0,04	8,8 ±0,17	6,5 ± 0,05	5,9 ±0,05	8,8 ±0,14
Ac	0,80 ±0,0	0,83 ±0,0	0,86 ±0,0	0,80 ±0,0	0,65 ±0,01	0,74 ±0,01
Se	0,83 ±0,0	0,72 ±0,01	0,80 ±0,01	0,80 ±0,0	0,63 ±0,02	0,73 ±0,01
Sp	0,76 ±0,01	0,87 ±0,0	0,90 ±0,0	0,81 ±0,01	0,65 ±0,02	0,75 ±0,01
PPV	0,86 ±0,0	0,64 ±0,0	0,80 ±0,01	0,84 ±0,0	0,48 ±0,01	0,65 ±0,01
NPV	0,705 ±0,01	0,90 ±0,0	0,90 ±0,00	0,76 ±0,01	0,78 ±0,01	0,79 ±0,01
δ-KDE-Bayes						
#caracteristicas	13,1 ± 6,72	4,2 ± 1,71	3,8 ± 1,77	7,0 ± 4,47	17,8 ± 5,55	10,0 ± 3,6
Ac	0,81 ± 0,08	0,86 ± 0,1	0,95 ± 0,0	0,90 ± 0,0	0,66 ± 0,1	0,67 ± 0,1
Se	0,88 ± 0,1	0,80 ± 0,2	0,89 ± 0,1	0,90 ± 0,1	0,47 ± 0,24	0,6 ± 0,2
Sp	0,68 ± 0,2	0,88 ± 0,1	0,98 ± 0,0	0,91 ± 0,1	0,76 ± 0,2	0,71 ± 0,2
PPV	0,84 ± 0,1	0,70 ± 0,1	0,97 ± 0,1	0,91 ± 0,1	0,62 ± 0,2	0,68 ± 0,2
NPV	0,77 ± 0,2	0,94 ± 0,1	0,95 ± 0,1	0,91 ± 0,1	0,73 ± 0,1	0,72 ± 0,1
ABEUS-KDE-Bayes						
#caracteristicas	6,8 ± 4,2	4,1 ± 1,2	3,2 ± 0,8	12,5 ± 9,44	14,8 ± 6,3	4,6 ± 1
Ac	0,78 ± 0,1	0,86 ± 0,1	0,89 ± 0,1	0,85 ± 0,1	0,61 ± 0,1	0,62 ± 0,1
Se	0,84 ± 0,1	0,81 ± 0,2	0,86 ± 0,1	0,83 ± 0,1	0,40 ± 0,1	0,59 ± 0,2
Sp	0,65 ± 0,2	0,88 ± 0,1	0,91 ± 0,1	0,87 ± 0,1	0,73 ± 0,1	0,64 ± 0,2
PPV	0,83 ± 0,1	0,71 ± 0,2	0,84 ± 0,1	0,86 ± 0,1	0,47 ± 0,2	0,59 ± 0,1
NPV	0,71 ± 0,13	0,94 ± 0,0	0,93 ± 0,05	0,85 ± 0,1	0,69 ± 0,1	0,66 ± 0,1

método de seleção parte do conjunto de amostras de treinamento com todas as 22283 características. Os dados foram divididos em um conjunto de treinamento (82 casos americanos, 61 amostras PCR e 21 amostras NoPCR) e um conjunto de teste (51 casos franceses, 38 amostras PCR e 13 amostras NoPCR) e os resultados comparados com os métodos DLDA-30 e Preditores Clínicos, de acordo com as mesmas métricas do experimento com as bases de oncologia.

Na Tabela 5.11 encontram-se os resultados do AG-KDE-Bayes, DLDA-30 (HESS et al., 2006) e dos preditores clínicos. Embora com menos um terço das características usadas pelo DLDA-30, o AG-KDE-Bayes tem desempenho em torno de 10% melhor do que o método mencionado anteriormente. Quando comparado aos preditores clínicos o AG-KDE-Bayes também tem desempenho superior, porém utiliza três vezes mais características.

Tabela 5.11: Desempenho dos preditores no conjunto de testes (Villejuif, 51 casos) independente do conjunto de treinamento (Houston, 82 casos). Os preditores clínicos são baseados em idade, status do receptor de estrogênio e grau do núcleo da célula cancerígena. PPV, NPV: valores preditivos positivo e negativo.

	AG-KDE-Bayes	DLDA-30	Preditores Clínicos
#Características	9	30	3
Acurácia	**0,86**	0,76	0,78
Sensibilidade	**0,92**	**0,92**	0,61
Especificidade	**0,84**	0,71	**0,84**
PPV	0,67	0,52	0,57
NPV	0,97	0,96	0,86

Tabela 5.12: Resumo das Bases de Dados do UCI Utilizadas.

Nome da Base	N. de Características	Classe 1	Classe 2
ACR	14	383	307
BLD	6	145	200
ION	33	225	126
SNR	60	97	111
TTT	9	626	332
WBC	9	444	239
HEA	13	150	120

5.5 Experimento 5: Estimação da Largura h Baseada em Derivadas

5.5.1 Bases de Dados do UCI

Os testes iniciais, utilizando 7 bases de dados públicas (Tabela 5.12) da Universidade de Irvine (http://archive.ics.uci.edu/ml/), foram realizados com o objetivo de verificar o comportamento do método proposto.

5.5.2 Metodologia e Resultados

Para cada uma das bases de dados foram realizadas 10 repetições da validação cruzada *3-fold*, analisando o desempenho para a largura do kernel proposta por (SILVERMAN, 1986) e aquela obtida segundo o método descrito no presente trabalho, que considera a margem máxima entre as classes. Para realizar o treinamento do modelo, para as 7 bases públicas, foram utilizados 2/3 dos dados, sendo o 1/3 restante para teste. No caso dos dados da quimioterapia neo-adjuvante os dados de Houston foram utilizados para treinamento e os de Villejuif para teste. Os resultados foram avaliados de acordo com média e o desvio padrão da Acurácia, Sensibilidade e Especificidade, calculados a partir dos resultados de cada uma das repetições da validação cruzada. Para comparar estatisticamente o resultado obtido pelo método KDE-Bayes para cada um dos valores de h propostos, foi realizada uma análise de variância (ANOVA) nas médias dos critérios utilizados na avaliação de desempenho das larguras h.

Na Tabela 5.13, são apresentados os resultados do classificador KDE-Bayes para a largura proposta na Seção 4.2 (Ac, Es, Se, Ac Teste, Es Teste, Se Teste) e por Silverman (Acs, Ess, Ses, Acs Teste, Ess Teste, Ses Teste). Os valores indicados na tabela são a média e o desvio padrão dos resultados obtidos nas execuções da validação cruzada, para o conjunto de treinamento e teste.

De modo semelhante, a Tabela 5.14 apresenta os resultados para os dados do problema da predição da eficácia da quimioterapia neo-adjuvante em pacientes de câncer de mama. As métricas e os valores indicados são os mesmos utilizados nos experimentos apresentados na Tabela 5.13.

De acordo com a análise de variância (ANOVA) realizada nos resultados apresentados nas Tabelas 5.13 e 5.14, os valores obtidos pelos dois métodos de cálculo da

Tabela 5.13: Bases de Dados da UCI: resultados da validação cruzada *3-fold* para a largura de *kernel* definida pelo método apresentado na Seção 4.2 (erro limitado em 0,05) e a definida por Silverman. Ac = Acurácia, Es = Especificidade, Se = Sensibilidade, Acs = Acurácia Silverman, Ess = Especificidade Silverman, Ses = Sensibilidade Silverman.

dados	ACR	BLD	ION	SNR	TTT	WBC	HEA
Seleção Baseada em Derivadas - Treinamento							
Ac	0,993 ± 0,002	0,980 ± 0,006	0,997 ± 0,001	1,000 ± 0,000	1,000 ± 0,000	0,996 ± 0,002	1,000 ± 0,000
Es	0,999 ± 0,001	0,975 ± 0,010	0,996 ± 0,001	1,000 ± 0,000	1,000 ± 0,000	0,998 ± 0,004	1,000 ± 0,000
Se	0,986 ± 0,004	0,983 ± 0,009	0,999 ± 0,003	1,000 ± 0,000	1,000 ± 0,000	0,993 ± 0,004	1,000 ± 0,000
Silverman - Treinamento							
Acs	0,986 ± 0,002	0,901 ± 0,010	0,996 ± 0,002	1,000 ± 0,000	1,000 ± 0,000	0,991 ± 0,001	0,999 ± 0,002
Ess	0,999 ± 0,002	0,812 ± 0,027	0,998 ± 0,001	1,000 ± 0,000	1,000 ± 0,000	1,000 ± 0,000	1,000 ± 0,000
Ses	0,970 ± 0,006	0,966 ± 0,006	0,992 ± 0,000	1,000 ± 0,000	1,000 ± 0,000	0,975 ± 0,003	0,997 ± 0,004
Seleção Baseada em Derivadas - Teste							
Ac Teste	0,993 ± 0,002	0,631 ± 0,016	0,862 ± 0,012	0,847 ± 0,013	0,883 ± 0,010	0,957 ± 0,005	0,782 ± 0,019
Es Teste	0,999 ± 0,001	0,529 ± 0,043	0,983 ± 0,007	0,792 ± 0,035	1,000 ± 0,000	0,977 ± 0,005	0,803 ± 0,026
Se Teste	0,986 ± 0,004	0,705 ± 0,027	0,646 ± 0,024	0,895 ± 0,030	0,662 ± 0,030	0,921 ± 0,014	0,757 ± 0,033
Silverman - Teste							
Acs Teste	0,986 ± 0,002	0,636 ± 0,025	0,882 ± 0,009	0,842 ± 0,013	0,921 ± 0,007	0,961 ± 0,005	0,788 ± 0,018
Ess Teste	0,999 ± 0,002	0,442 ± 0,049	0,988 ± 0,006	0,785 ± 0,031	1,000 ± 0,000	0,976 ± 0,004	0,807 ± 0,025
Ses Teste	0,970 ± 0,006	0,777 ± 0,028	0,694 ± 0,032	0,892 ± 0,035	0,771 ± 0,020	0,933 ± 0,014	0,764 ± 0,022

Tabela 5.14: Problema do Câncer de Mama: resultados da validação cruzada *3-fold* para a largura de *kernel* definida pelo método apresentado na Seção 4.2 (erro limitado em 0,05) e a definida por Silverman. Ac = Acurácia, Es = Especificidade, Se = Sensibilidade, Acs = Acurácia Silverman, Ess = Especificidade Silverman, Ses = Sensibilidade Silverman.

Ac	Es	Se	Acs	Ess	Ses
Treinamento					
1,000 ± 0,000	1,000 ± 0,000	1,000 ± 0,000	1,000 ± 0,000	1,000 ± 0,000	1,000 ± 0,000
Teste					
0,753 ± 0,039	0,783 ± 0,032	0,668 ± 0,064	0,755 ± 0,024	0,814 ± 0,030	0,582 ± 0,063

largura do *kernel*, o proposto na Seção e o proposto por (SILVERMAN, 1986), h_1 (Equação 2.10) são estatisticamente equivalentes. Tais resultados corroboram com a hipótese assumida na Seção 4, indicando a existência de coerência entre a rotulação dos dados e a função geradora dos mesmos. Embora os resultados sejam equivalentes, o método proposto apresenta maior equilíbrio no desempenho tanto na classe majoritária quanto na minoritária, enquanto o valor da largura do kernel proposto por (SILVERMAN, 1986) tem melhor desempenho para a classe majoritária.

No contexto do problema da previsão da eficácia da quimioterapia neo-adjuvante

(Tabela 5.14), apesar dos dois métodos terem desempenho igual no treinamento, o presente método é ligeiramente superior no conjunto de teste. Uma vez que o referido problema trata da opção de oferecer ou não um tratamento anterior à cirurgia a um paciente, é desejável que o número de falsos negativos seja o menor possível, ainda que o número de falsos positivos seja um pouco maior. Assim, o valor de h proposto por Silverman erra em 36% dos casos, contra 30% para o valor apresentado neste trabalho.

Uma vez que o método ora apresentado baseia-se na região de baixa densidade localizada entre classes, o mesmo possui algumas limitações. O método aplica-se principalmente aos problemas de classificação binária, devido à dificuldade em se determinar a margem de separações para três ou mais classes. Advindo do mesmo conceito, encontra-se também a dificuldade em utilizar aprendizado não-supervisionado, nos casos em que uma classe é composta por dois ou mais *clusters*, o que poderia ser interpretado como um problema de múltiplas classes.

5.6 Experimento 6: Estimação da Largura h Baseada na Diferença de Comportamento da Densidade na Margem e fora dela

5.6.1 Bases de Dados do Keel

Além das bases de dados apresentadas na Seção 5.5.1, das 30 sondas selecionadas por (HESS et al., 2006) e das 18 sondas selecionadas por (HORTA, 2008) para o problema do câncer de mama 5.1.1, também foram utilizadas bases com classes desbalanceadas. Para testar o comportamento do método de seleção de largura proposto na Seção 4.3 em bases desbalanceadas, quatro bases de dados do repositório Keel (http://sci2s.ugr.es

Tabela 5.15: Resumo das Bases de Dados do Keel Utilizadas.

Nome da Base	# de Características	Classe 1	Classe 2	c_1/c_2
Ecoli	5	77	143	0,54
Vehicle	18	628	218	2,88
Cleveland Heart	13	160	13	12,31
Page Blocks	10	4913	559	8,79

/keel/datasets.php) foram utilizadas (Tabela 5.15).

5.6.2 Metodologia e Resultados

O presente experimento foi dividido em duas etapas. Na primeira, o método descrito na Seção 4.3 foi testado e comparado com os apresentados na Seção 2.4, pretendendo por fim observar a magnitude dos valores de h selecionados por cada método. A seguir, foram realizadas 10 repetições de validação cruzada *3-fold* em cada uma das bases.

Na Tabela 5.16, são apresentados os resultados do classificador KDE-Bayes para as larguras mencionadas acima. Os resultados são apresentados de acordo com as métricas de acurácia (Ac), média geométrica (Mgeo), especificidade (Es) e sensibilidade (Se).

Dentre os resultados, para a base TTT, destacam-se as larguras de h_1 a h_4, por apresentar sensibilidade igual a zero. Os valores estimados para algumas dimensões da classe minoritária é zero, como por exemplo [0, 26 0, 28 0, 26 0, 29 0, 00 0, 30 0, 26 0, 29 0, 27 0, 00], diminuindo o valor da verossimilhança e, por consequência, atribuindo a amostra da classe c_2 para a classe c_1. O mesmo comportamento observa-se para as bases WBC e HEART CLEV. Outro comportamento digno de nota é o na base Câncer(Hess), que apresenta sensibilidade muito próxima a zero. Nesse caso nota-se uma grande disparidade entre os valores de h para cada classe, também alterando a

verossimilhança e posterior classificação das amostras.

Tabela 5.16: Bases de Dados da UCI e da Keel: resultados do classificador KDE-Bayes para a largura de *kernel* apresentada nas Seções 2.4 e 4.3. Ac = Acurácia, Mgeo = Média Geométrica, Es = Especificidade, Se = Sensibilidade.

	BLD	SNR	TTT	WBC	Câncer(Euler)	Câncer(Hess)	ECOLI	VEHICLE	HEART CLEV.	PG_BLKS
Seleção Baseada na Diferença										
Ac	0,614	0,913	0,929	0,935	0,702	0,803	0,958	0,972	0,982	0,923
Mgeo	0,600	0,909	0,908	0,915	0,694	0,791	0,948	0,959	0,866	0,522
Es	0,541	0,875	0,972	0,972	0,602	0,815	0,920	0,920	1,000	0,996
Se	0,666	0,945	0,850	0,862	0,802	0,769	0,978	1,000	0,750	0,274
Silverman Normalizado - h_{11} (Equação 2.11)										
Ac	0,543	0,898	0,927	0,929	0,706	0,647	0,972	0,953	0,964	0,940
Mgeo	0,471	0,896	0,888	0,908	0,697	0,606	0,959	0,904	0,857	0,673
Es	0,312	0,875	1,000	0,972	0,597	0,684	0,920	1,000	0,981	0,995
Se	0,712	0,918	0,790	0,850	0,815	0,538	1,000	0,819	0,750	0,456
Silverman Não Normalizado - h_1 (Equação 2.10)										
Ac	0,631	0,768	0,654	0,933	0,823	0,784	0,986	0,957	0,929	0,939
Mgeo	0,602	0,737	0,000	0,947	0,775	0,391	0,988	0,956	0,000	0,75
Es	0,500	0,593	1,000	0,898	0,868	1,000	1,000	0,956	1,000	0,98
Se	0,727	0,918	0,000	1,000	0,692	0,153	0,978	0,958	0,000	0,575
Silverman IQR - h_2 (Equação 2.12)										
Ac	0,649	0,782	0,654	0,652	0,823	0,764	0,986	0,953	0,929	0,938
Mgeo	0,573	0,757	0,000	0,000	0,804	0,275	0,988	0,949	0,000	0,884
Es	0,395	0,625	1,000	1,000	0,842	1,000	1,000	0,956	1,000	0,951
Se	0,833	0,918	0,000	0,000	0,769	0,076	0,978	0,944	0,000	0,822
Silverman Min - h_3 (Equação 2.13)										
Ac	0,657	0,782	0,654	0,652	0,823	0,784	0,986	0,957	0,929	0,941
Mgeo	0,613	0,757	0,000	0,000	0,804	0,392	0,988	0,952	0,000	0,869
Es	0,479	0,625	1,000	1,000	0,842	1,000	1,000	0,961	1,000	0,958
Se	0,787	0,918	0,000	0,000	0,769	0,153	0,978	0,944	0,000	0,79
Scott - h_4 (Equação 2.15)										
Ac	0,631	0,782	0,654	0,938	0,823	0,784	1,000	0,968	0,929	0,941
Mgeo	0,602	0,757	0,000	0,951	0,775	0,392	1,000	0,964	0,000	0,751
Es	0,500	0,625	1,000	0,9054	0,868	1,000	1,000	0,971	1,000	0,982
Se	0,727	0,918	0,000	1,000	0,692	0,153	1,000	0,958	0,000	0,575

Nas Tabelas 5.17 e 5.18 são apresentados os resultados para a validação cruzada *3-fold* para o conjunto de treinamento e teste.

Para a maioria dos experimentos apresentados na Tabela 5.18 os resultados obtidos com a largura proposta neste trabalho (Seção 4.3) superam ou se equivalem aos demais, à exceção das bases Câncer(Euler) e HEART CLEV. Comparando os resultados para o conjunto de treinamento (Tabela 5.17), para a base SNR, há uma perda de aproximadamente 20% para as larguras h_1 a h_4, sugerindo que a estimativa utilizada estava super-suavizada e consequentemente com *overfitting* com relação aos dados de treinamento. O mesmo ocorre para a base Câncer(Hess), incluindo também a largura h_{11}.

Quanto ao comportamento com relação às bases desbalanceadas, o método proposto encontra dificuldade apenas para a base HEART CLEV. Embora o valor da acurácia seja alto, a média geométrica está em torno de $0,50$, devido ao resultado obtido para sensibilidade, apontando que o uso da acurácia enquanto métrica não é adequado. Tal resultado é justificado não pela quantidade relativa de amostras das classes c_1 e c_2, mas sim pela quantidade absoluta de amostras da classe minoritária. Usando a divisão de 2/3 para treinamento e 1/3 para teste, tem-se aproximadamente 9 casos de treinamento e 4 para teste, dificultando a estimativa correta da densidade da classe c_2. Por esta razão também observa-se um desvio padrão entre 30 e 45% aproximadamente para os resultados de sensibilidade.

5.7 Conclusão do Capítulo

Nesse Capítulo foram apresentados os experimentos realizados para avaliar o desempenho dos métodos de seleção de características (Capítulo 3) e de estimação da largura do *kernel* (Capítulo 4), propostos neste trabalho. Os experimentos iniciais de seleção de características, com métodos univariados (seleção por acurácia (WANDERLEY et al., 2010) e por AUC) indicaram ser viável realizar seleção de características utilizando métricas simples e o classificador KDE-Bayes. Em seguida, sua evolução natural, o AG-KDE-Bayes mostrou-se eficiente para seleção multivariada (WANDERLEY et al., 2013). Por fim, foram testados os dois métodos de estimação da largura do *kernel* baseado na informação geométrica dos dados, sem suposição de normalidade dos dados. Os resultados mostram que tanto o método baseado em derivadas (WANDERLEY et al., 2014) quanto o baseado na diferença entre a densidade nas classes e nos pontos da margem são comparáveis aos encontrados na literatura, fornecendo uma boa alternativa.

No Capítulo 6 encontram-se as conclusões do trabalho e propostas de continuidade.

Tabela 5.17: Bases de Dados da UCI e da Keel: resultados do classificador KDE-Bayes para a largura de *kernel* apresentada nas Seções 2.4 e 4.3 para o conjunto de treinamento. Ac = Acurácia, Mgeo = Média Geométrica, Es = Especificidade, Se = Sensibilidade.

	BLD	SNR	TTT	WBC	Câncer(Euler)	Câncer(Hess)	ECOLI	VEHICLE	HEART CLEV.	PG_BLKS
Seleção Baseada na Diferença										
Ac	0,9843 ± 0,0307	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9771 ± 0,0069	0,9864 ± 0,0202	0,9504 ± 0,0291	0,9946 ± 0,0037	0,9997 ± 0,0007	1,0000 ± 0,0000	0,9875 ± 0,0023
Mgeo	0,9832 ± 0,0339	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9726 ± 0,0080	0,9721 ± 0,0417	0,9612 ± 0,0237	0,9922 ± 0,0053	0,9998 ± 0,0006	1,0000 ± 0,0000	0,9515 ± 0,0095
Es	0,9775 ± 0,0500	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9875 ± 0,0050	1,0000 ± 0,0000	0,9386 ± 0,0357	0,9846 ± 0,0105	0,9996 ± 0,0008	1,0000 ± 0,0000	0,9965 ± 0,0008
Se	0,9892 ± 0,0179	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9579 ± 0,0116	0,9467 ± 0,0790	0,9846 ± 0,0164	1,0000 ± 0,0000	0,9999 ± 0,0005	1,0000 ± 0,0000	0,9085 ± 0,0179
Silverman Normalizado - h_{11} (Equação 2.11)										
Ac	0,8918 ± 0,0068	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9912 ± 0,0011	0,9933 ± 0,0034	1,0000 ± 0,0000	0,9864 ± 0,0023	0,9887 ± 0,0015	1,0000 ± 0,0000	0,9978 ± 0,0005
Mgeo	0,8746 ± 0,0103	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9874 ± 0,0016	0,9868 ± 0,0067	1,0000 ± 0,0000	0,9803 ± 0,0034	0,9794 ± 0,0027	1,0000 ± 0,0000	0,9904 ± 0,0028
Es	0,7962 ± 0,0257	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9610 ± 0,0067	0,9984 ± 0,0005	1,0000 ± 0,0000	0,9997 ± 0,0001
Se	0,9611 ± 0,0130	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9749 ± 0,0031	0,9738 ± 0,0132	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9608 ± 0,0051	1,0000 ± 0,0000	0,9811 ± 0,0056
Silverman Não Normalizado - h_1 (Equação 2.10)										
Ac	0,9839 ± 0,0038	1,0000 ± 0,0000	0,6534 ± 0,0004	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9955 ± 0,0015	1,0000 ± 0,0000	0,9923 ± 0,0232	0,9588 ± 0,0008
Mgeo	0,9853 ± 0,0036	1,0000 ± 0,0000	0,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9935 ± 0,0022	1,0000 ± 0,0000	0,9000 ± 0,3015	0,7724 ± 0,0053
Es	0,9946 ± 0,0049	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9870 ± 0,0044	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000
Se	0,9761 ± 0,0061	1,0000 ± 0,0000	0,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9000 ± 0,3015	0,5967 ± 0,0082
Silverman IQR - h_2 (Equação 2.12)										
Ac	0,9961 ± 0,0019	1,0000 ± 0,0000	0,6534 ± 0,0004	0,6501 ± 0,0005	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9960 ± 0,0020	1,0000 ± 0,0000	0,9249 ± 0,0027	0,9877 ± 0,0010
Mgeo	0,9967 ± 0,0017	1,0000 ± 0,0000	0,0000 ± 0,0000	0,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9942 ± 0,0029	1,0000 ± 0,0000	0,0000 ± 0,0000	0,9379 ± 0,0050
Es	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9884 ± 0,0058	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000
Se	0,9933 ± 0,0034	1,0000 ± 0,0000	0,0000 ± 0,0000	0,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,0000 ± 0,0000	0,8797 ± 0,0095
Silverman Min - h_3 (Equação 2.13)										
Ac	0,9916 ± 0,0022	1,0000 ± 0,0000	0,6534 ± 0,0004	0,6501 ± 0,0005	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9955 ± 0,0015	1,0000 ± 0,0000	0,9249 ± 0,0027	0,9844 ± 0,0010
Mgeo	0,9924 ± 0,0018	1,0000 ± 0,0000	0,0000 ± 0,0000	0,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9935 ± 0,0022	1,0000 ± 0,0000	0,0000 ± 0,0000	0,9205 ± 0,0055
Es	0,9977 ± 0,0035	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9870 ± 0,0044	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000
Se	0,9872 ± 0,0050	1,0000 ± 0,0000	0,0000 ± 0,0000	0,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,0000 ± 0,0000	0,8473 ± 0,0102
Scott - h_4 (Equação 2.15)										
Ac	0,9762 ± 0,0064	1,0000 ± 0,0000	0,6534 ± 0,0004	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9955 ± 0,0015	1,0000 ± 0,0000	0,9923 ± 0,0232	0,9572 ± 0,0010
Mgeo	0,9781 ± 0,0055	1,0000 ± 0,0000	0,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9935 ± 0,0022	1,0000 ± 0,0000	0,9000 ± 0,3015	0,7625 ± 0,0062
Es	0,9908 ± 0,0067	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9870 ± 0,0044	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000
Se	0,9656 ± 0,0124	1,0000 ± 0,0000	0,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	1,0000 ± 0,0000	0,9000 ± 0,3015	0,5814 ± 0,0094

Tabela 5.18: Bases de Dados da UCI e da Keel: resultados do classificador KDE-Bayes para a largura de *kernel* apresentada nas Seções 2.4 e 4.3 para o conjunto de teste. Ac = Acurácia, Mgeo = Média Geométrica, Es = Especificidade, Se = Sensibilidade.

	BLD	SNR	TTT	WBC	Câncer(Euler)	Câncer(Hess)	ECOLI	VEHICLE	HEART CLEV.	PG_BLKS
Seleção Baseada na Diferença										
Ac	**0,6397 ± 0,0864**	0,8540 ± 0,0842	1,0000 ± 0,0000	0,9655 ± 0,0248	0,7926 ± 0,0998	0,8126 ± 0,0948	0,9807 ± 0,0237	0,9697 ± 0,0223	0,9595 ± 0,0311	0,9679 ± 0,0050
Mgeo	**0,6154 ± 0,0920**	0,8412 ± 0,0948	1,0000 ± 0,0000	0,9576 ± 0,0326	0,6840 ± 0,1751	0,7915 ± 0,1250	0,9713 ± 0,0349	0,9640 ± 0,0242	0,5070 ± 0,4633	0,8830 ± 0,0217
Es	0,5518 ± 0,1344	0,7788 ± 0,1523	1,0000 ± 0,0000	0,9820 ± 0,0144	0,8720 ± 0,1102	0,8232 ± 0,1146	0,9452 ± 0,0663	0,9752 ± 0,0291	0,9958 ± 0,0159	0,9882 ± 0,0031
Se	0,7030 ± 0,1172	0,9193 ± 0,0730	1,0000 ± 0,0000	0,9347 ± 0,0565	0,5725 ± 0,2302	0,7842 ± 0,2184	0,9993 ± 0,0071	0,9537 ± 0,0436	0,4667 ± 0,4536	0,7895 ± 0,0385
Silverman Normalizado - h_{11} (Equação 2.11)										
Ac	0,6267 ± 0,0679	0,8561 ± 0,0633	1,0000 ± 0,0000	0,9618 ± 0,0267	0,8129 ± 0,0669	0,6860 ± 0,1039	0,9773 ± 0,0229	0,9610 ± 0,0185	0,8559 ± 0,0649	0,9357 ± 0,0134
Mgeo	0,5611 ± 0,0831	0,8491 ± 0,0651	1,0000 ± 0,0000	0,9548 ± 0,0276	0,7001 ± 0,0917	0,2771 ± 0,2870	0,9658 ± 0,0345	0,9310 ± 0,0359	0,4056 ± 0,4180	0,6900 ± 0,0650
Es	0,4243 ± 0,1370	0,7944 ± 0,0891	1,0000 ± 0,0000	0,9774 ± 0,0270	0,8978 ± 0,0810	0,8600 ± 0,1119	0,9339 ± 0,0667	0,9905 ± 0,0128	0,8938 ± 0,0797	0,9868 ± 0,0065
Se	0,7750 ± 0,1173	0,9106 ± 0,0689	1,0000 ± 0,0000	0,9330 ± 0,0337	0,5583 ± 0,1404	0,1833 ± 0,2145	1,0000 ± 0,0000	0,8766 ± 0,0675	0,4000 ± 0,4381	0,4863 ± 0,0872
Silverman Não Normalizado - h_1 (Equação 2.10)										
Ac	0,6203 ± 0,0740	0,8121 ± 0,0850	0,6534 ± 0,0034	0,9370 ± 0,0309	0,8049 ± 0,1314	0,8249 ± 0,1006	0,9776 ± 0,0298	0,9704 ± 0,0161	0,8216 ± 0,0689	0,9485 ± 0,0052
Mgeo	0,5964 ± 0,0875	0,7733 ± 0,1091	0,0000 ± 0,0000	0,9499 ± 0,0252	0,7352 ± 0,1675	0,6735 ± 0,2779	0,9756 ± 0,0315	0,9581 ± 0,0257	0,7761 ± 0,2710	0,7644 ± 0,0350
Es	0,5333 ± 0,1410	0,6267 ± 0,1751	1,0000 ± 0,0000	0,9030 ± 0,0477	0,8567 ± 0,1171	0,8978 ± 0,0926	0,9732 ± 0,0540	0,9825 ± 0,0151	0,8187 ± 0,0908	0,9890 ± 0,0040
Se	0,6850 ± 0,0928	0,9727 ± 0,0419	0,0000 ± 0,0000	1,0000 ± 0,0000	0,6500 ± 0,2333	0,5917 ± 0,3121	0,9795 ± 0,0434	0,9351 ± 0,0487	0,8500 ± 0,3218	0,5921 ± 0,0553
Silverman IQR - h_2 (Equação 2.12)										
Ac	**0,6463 ± 0,0564**	0,8071 ± 0,0727	0,6534 ± 0,0034	0,6501 ± 0,0045	0,7907 ± 0,1491	0,7738 ± 0,0778	0,9733 ± 0,0353	0,9704 ± 0,0134	0,9255 ± 0,0241	0,9399 ± 0,0103
Mgeo	**0,6134 ± 0,0772**	0,7580 ± 0,1001	0,0000 ± 0,0000	0,0000 ± 0,0000	0,7765 ± 0,1769	0,4372 ± 0,3079	0,9724 ± 0,0366	0,9692 ± 0,0184	0,0000 ± 0,0000	0,8757 ± 0,0238
Es	0,5190 ± 0,1207	0,5844 ± 0,1583	1,0000 ± 0,0000	1,0000 ± 0,0000	0,7956 ± 0,1347	0,9289 ± 0,0650	0,9732 ± 0,0540	0,9713 ± 0,0157	1,0000 ± 0,0000	0,9554 ± 0,0091
Se	0,7400 ± 0,0667	1,0000 ± 0,0000	0,0000 ± 0,0000	0,0000 ± 0,0000	0,7750 ± 0,2542	0,3083 ± 0,2569	0,9729 ± 0,0449	0,9675 ± 0,0365	0,0000 ± 0,0000	0,8031 ± 0,0408
Silverman Min - h_3 (Equação 2.13)										
Ac	0,6230 ± 0,0705	0,8074 ± 0,0780	0,6534 ± 0,0034	0,6501 ± 0,0045	0,7907 ± 0,1491	0,7892 ± 0,0962	0,9776 ± 0,0298	0,9740 ± 0,0174	0,9255 ± 0,0241	0,9466 ± 0,0077
Mgeo	0,6040 ± 0,0828	0,7678 ± 0,1012	0,0000 ± 0,0000	0,0000 ± 0,0000	0,7632 ± 0,1682	0,4940 ± 0,3507	0,9756 ± 0,0315	0,9637 ± 0,0269	0,0000 ± 0,0000	0,8719 ± 0,0262
Es	0,5462 ± 0,1286	0,6167 ± 0,1614	1,0000 ± 0,0000	1,0000 ± 0,0000	0,8056 ± 0,1455	0,9189 ± 0,0756	0,9732 ± 0,0540	0,9841 ± 0,0160	1,0000 ± 0,0000	0,9646 ± 0,0054
Se	0,6800 ± 0,0752	0,9727 ± 0,0419	0,0000 ± 0,0000	0,0000 ± 0,0000	0,7417 ± 0,2440	0,4000 ± 0,3304	0,9795 ± 0,0434	0,9446 ± 0,0503	0,0000 ± 0,0000	0,7888 ± 0,0464
Scott - h_4 (Equação 2.15)										
Ac	0,6202 ± 0,0779	0,8121 ± 0,0850	0,6534 ± 0,0034	0,9399 ± 0,0291	0,8049 ± 0,1314	0,8332 ± 0,0899	0,9909 ± 0,0183	0,9704 ± 0,0178	0,8448 ± 0,0612	0,9472 ± 0,0052
Mgeo	0,6029 ± 0,0858	0,7733 ± 0,1091	0,0000 ± 0,0000	0,9523 ± 0,0237	0,7352 ± 0,1675	0,6946 ± 0,2725	0,9861 ± 0,0280	0,9566 ± 0,0268	0,7901 ± 0,2766	0,7529 ± 0,0380
Es	0,5533 ± 0,1242	0,6267 ± 0,1751	1,0000 ± 0,0000	0,9075 ± 0,0449	0,8567 ± 0,1171	0,8978 ± 0,0926	0,9732 ± 0,0540	0,9841 ± 0,0160	0,8438 ± 0,0807	0,9896 ± 0,0036
Se	0,6700 ± 0,0985	0,9727 ± 0,0419	0,0000 ± 0,0000	1,0000 ± 0,0000	0,6500 ± 0,2333	0,6250 ± 0,3002	1,0000 ± 0,0000	0,9305 ± 0,0488	0,8500 ± 0,3218	0,5743 ± 0,0585

6 CONCLUSÕES

> "It always seems impossible until it is done."
>
> Nelson Mandela

Este trabalho, tendo como base a estimação não-paramétrica de densidades por *kernel* (KDE), apresentou um estudo acerca de dois aspectos do tema. Após uma revisão bibliográfica sobre o tema (Capítulo 2), destacando-se a importância da escolha adequada do parâmetro suavizador do *kernel*, nos Capítulos seguintes são apresentados os métodos propostos.

Inicialmente, o KDE foi utilizado enquanto ferramenta, na construção de um classificador generativo (KDE-Bayes - Capítulo 3), aproveitando-se da capacidade do método de inferir relações locais entre os dados, em oposição aos modelos indutivos, que precisam induzir parâmetros de um modelo geral a partir dos dados, frequentemente escassos, como em alguns dos bancos de dados utilizados no Capítulo 5, podendo levar a classificadores enviesados com relação à amostra.

Analisando os resultados do primeiro experimento (Seção 5.1) é possível ver a diferença de desempenho entre o método paramétrico e o não-paramétrico. A imposição

de uma estrutura fixa a dados escassos e esparsos reflete-se diretamente nos resultados, nos quais o método de regressão logística mostra um bom desempenho para a classe maior, em detrimento da classe menor. O desbalanceamento das classes também dificulta uma boa performance do método paramétrico que fica enviesado para a classe maior.

Ao ser apresentado aos mesmos dados, o método não-paramétrico mostrou um desempenho superior, inclusive para a classe menor. Nesse experimento também foi possível obter indícios de que a estrutura dos dados representada por cada uma das características relaciona-se diretamente com os resultados. Duas das características (figuras 5.1 e 5.3) possuem estrutura e desempenho bem semelhantes, o que indicaria que elas seriam redundantes fazendo parte do mesmo subconjunto de características. Ou seja, a partir da informação estrutural das características é possível eliminar de agrupamentos características redundantes, diminuindo assim o espaço de busca na seleção multidimensional de características. A utilização de um algoritmo evolucionário proporciona uma exploração mais eficiente do espaço de subconjuntos de características, gerando assim a proposta do AG-KDE-Bayes. Esse *wrapper* possibilita que um grupo de características seja analisado em conjunto, em oposição aos métodos de filtro, que fazem um *ranking* das características.

O segundo aspecto do KDE abordado foi a influência do parâmetro suavizador para a estimação de densidade e consequentemente para o classificador binário KDE-Bayes. No Capítulo 4 dois métodos para estimação da largura do *kernel* h foram propostos, partindo da hipótese de que haveria coerência geométrica entre os dados e os rótulos das classes. Por basearem-se na ideia advinda do aprendizado semi-supervisionado de que entre as classes deve haver uma região de baixa densidade, por ora os métodos propostos limitam-se a problemas de classificação binária.

Os resultados corroboram com a hipótese assumida, sendo a contribuição do pre-

sente trabalho uma alternativa para a seleção de modelos, baseada na geometria do problema e nos rótulos conhecidos para cada classe.

O princípio de que a superfície de separação se localiza em uma região de baixa densidade tem sido utilizado na literatura como norteador para a construção de classificadores de margem larga. Este princípio sugere que o separador de margem máxima e que minimiza também o erro do conjunto indutivo de dados deve se localizar em uma região de baixa densidade. Apesar de ser este o princípio geral dos classificadores de margem larga, como as SVMs, por exemplo, as densidades nos pontos de separação não são diretamente calculadas. Usualmente a identificação da região de baixa densidade é obtida como resultado da maximização da margem de separação através de uma função-objetivo associada à magnitude dos parâmetros (pesos) do modelo.

Neste trabalho, no entanto, foi apresentada uma abordagem que visa primeiro a identificar a região de baixa densidade para, através de um critério de seleção, obter uma suavização adequada da superfície de separação que resulte em um classificador com uma margem larga de separação. Através da construção de matrizes de kernel apropriadas e da identificação geométrica dos pontos médios de separação utilizando-se o Grafo de Gabriel foram descritas funções-objetivo para a minimização do erro de classificação. A suavização da resposta do modelo de erro mínimo é obtida através de dois métodos de seleção: um que se baseia no cálculo das densidades nos pontos médios e outro que se baseia na diferença do somatório das densidades nas classes e nos pontos da margem. O modelo final avaliado em várias bases de dados se mostrou robusto aos dados de teste, sugerindo a existência de um bom equilíbrio entre viés e variância obtido indiretamente através do seletor baseado no cálculo das densidades.

Os resultados obtidos são compatíveis com aqueles obtidos por métodos que fazem o controle do viés e da variância de maneira explícita, como aquele proposto por (SIL-

VERMAN, 1986). Muitos dos resultados obtidos estão dentro de faixas limites de *benchmarking* das bases de dados usadas como testes. Assim, não era nosso objetivo propor uma nova metodologia que superasse o desempenho de modelos correntes, mesmo porque isto talvez não seja possível para as bases utilizadas. Foram, assim, descritos neste trabalho dois novos métodos através dos quais foi possível mostrar a viabilidade da seleção de modelos através do cálculo direto das densidades e da geometria do problema de separação. O foco nas densidades dos pontos e não no cálculo direto da margem de separação se apresenta como uma alternativa viável para a construção de modelos generativos de separação.

6.1 Trabalhos Futuros

6.1.1 Inserção de Informação *a priori*

Os resultados dos métodos de estimação não-paramétrica KDE e a análise das funções biológicas das sondas no problema da quimioterapia neoadjuvante forneceram novos tipos de informação, que devem ser agregadas aos dados para facilitar a seleção de características. Utilizando o KDE a estrutura dos dados passa a ser conhecida, assim, tal informação poderia ser utilizada para detecção e eliminação de características redundantes dentro do mesmo agrupamento. Já com a análise biológica é possível determinar, em alguns casos, quais características estão relacionadas a que classes, podendo receber um peso diferente no agrupamento, dependendo dos demais elementos deste. Embora (VAPNIK; VASHIST, 2009) tenham proposto algo semelhante, sua abordagem visava à construção de um novo modelo de aprendizado, não tendo por objetivo seleção de características.

6.1.2 Método de Agrupamento a partir da Informação Estrutural dos Dados

Seja um problema de duas ou mais classes, dado a estimativa de densidade das classes utilizando KDE multidimensional a ideia é obter a separação dos dados em grupos. Uma vez que a estrutura dos dados é conhecida, a informação sobre a qual *cluster* pertence cada uma das amostras está presente, sendo necessário estudar um modo de extrair essa informação.

6.1.3 Estimação da Largura do *Kernel* a partir da Matriz de *Kernel* SVM

O *kernel* SVM realiza um mapeamento do espaço de entrada para o espaço das características, utilizando uma função não-linear ϕ, que em geral possui uma dimensão maior do que o espaço original, possibilitando a separação por um hiperplano.

Dado o mapeamento não-linear do espaço de entrada a nova função de decisão do SVM é dada pela Equação 6.1:

$$d(x) = \sum_{i=1}^{p} \alpha_i y_i \phi(x_i) \cdot \phi(x) + b, \tag{6.1}$$

onde o produto escalar pode ser substituído pela função de kernel $K(x_i, x)$.

Utilizando a Equação 2.6, supondo um *kernel* Gaussiano e, para efeitos deste exemplo, supondo que $n = 3$, temos:

$$\begin{aligned} \hat{f}_h(x) &= \frac{e^{-0.5*\left(\frac{x_1 - x_{i1}}{h_1}\right)^2}}{\sqrt{2\pi}} * \frac{e^{-0.5*\left(\frac{x_2 - x_{i2}}{h_2}\right)^2}}{\sqrt{2\pi}} * \frac{e^{-0.5*\left(\frac{x_3 - x_{i3}}{h_3}\right)^2}}{\sqrt{2\pi}} \\ &= \frac{1}{\sqrt{2\pi}^3} * e^{-0.5\left[\left(\frac{x_1 - x_{i1}}{h_1}\right)^2 + \left(\frac{x_2 - x_{i2}}{h_2}\right)^2 + \left(\frac{x_3 - x_{i3}}{h_3}\right)^2\right]} \end{aligned}$$

Para o kernel gaussiano do SVM:

$$k(u,v) = e^{-0.5*\left[\left(\frac{u_1-v_{i1}}{\sigma}\right)^2+\left(\frac{u_2-v_{i2}}{\sigma}\right)^2+\left(\frac{u_3-v_{i3}}{\sigma}\right)^2\right]}$$

logo o *kernel* SVM é equivalente a um *kernel* KDE cujas larguras das gaussianas são idênticas e utiliza um fator normalizador. A partir dessa equivalência surgem duas questões:

- Seria o h que melhor reflete a estrutura dos dados um bom parâmetro para o *kernel* SVM?
- Seria o σ que gera o classificador SVM de margem máxima coerente com a relação local e geometria dos dados?

REFERÊNCIAS

ADCOCK, C. Sample size determination: a review. **Journal of the Royal Statistical Society: Series D (The Statistician)**, [S.l.], v.46, n.2, p.261–283, 1997.

BAGLEY, S. C.; WHITE, H.; GOLOMB, B. A. Logistic Regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. **Journal of Clinical Epidemiology**, [S.l.], v.54, n.10, p.979–985, 2001.

BOYLE, P.; LEVIN, B. **World Cancer Report 2008**. Lyon: International Agency for Research on Cancer, 2008.

BRADLEY, A. The use of the area under the roc curve in the evaluation of machine learning algorithms. **Pattern Recognition**, [S.l.], v.30, p.1145 – 1159, 1997.

CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. et al. **Semi-supervised learning**. [S.l.]: MIT press Cambridge, 2006. v.2.

CHERKASSKY, V.; MULIER, F. M. **Learning from data**: concepts, theory, and methods. [S.l.]: Wiley. com, 2007.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, [S.l.], v.20, n.3, p.273–297, 1995.

DE BERG, M.; CHEONG, O.; KREVELD, M. van; OVERMARS, M. **Computational geometry**. [S.l.]: Springer, 2008.

DP, L. E2A basic helix-loop-helix transcription factors in human leukemia. **Front Biosci.**, [S.l.], v.8, p.206–222, 2003.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 2nd.ed. [S.l.]: Wiley-Interscience, 2000.

DUDOIT, S.; FRIDLYAND, J.; SPEED, T. Comparison of discrimination methods for the classification of tumors using gene expression data. **Journal of the American statistical association**, [S.l.], v.97, n.457, p.77–87, 2002.

FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, [S.l.], v.27, n.8, p.861–874, 2006.

GAJEK, L.; LENIC, A. An approximate necessary condition for the optimal bandwidth selector in kernel density estimation. **Applicationes Mathematicae**, [S.l.], v.22, n.1, p.123–138, 1993.

GAMMERMAN, A.; VOVK, V.; VAPNIK, V. Learning by Transduction. In: IN UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 1998. **Anais...** Morgan Kaufmann, 1998. p.148–155.

GARDEUX, V.; NATOWICZ, R.; WANDERLEY, M.; CHELOUAH, R. Optimization for Feature Selection in DNA Microarrays. In: SIARRY, P. (Ed.). **Heuristics**: theory and applications. New York, USA: Nova Publishers, 2013. p.287–310.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization and learning**. Massachusetts: Addison-Wesley, 1989.

GOLUB, T. R.; SLONIM, D. K.; TAMAYO, P.; HUARD, C.; GAASENBEEK, M.; MESIROV, J. P.; COLLER, H.; LOH, M. L.; DOWNING, J. R.; CALIGIURI, M. A.; BLOOMFIELD, C. D.; ; LANDER, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. **Science**, [S.l.], v.286, p.531–537, 1999.

GUYON, I.; GUNN, S.; NIKRAVESH, M.; ZADEH, L. **Feature Extraction**: foundations and applications. [S.l.]: Springer, 2006.

HAYKIN, S. **Neural networks**: a comprehensive foundation. [S.l.]: Prentice Hall PTR, 1994.

HESS, K.; ANDERSON, K.; SYMMANS, W.; VALERO, V.; IBRAHIM, N.; MEJIA, J.; BOOSER, D.; THERIAULT, R.; BUZDAR, A.; DEMPSEY, P.; ROUZIER, R.; SNEIGE, N.; ROSS, J.; VIDAURRE, T.; GOMEZ, H.; HORTOBAGYI, G.; PUSZTAI, L. Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer. **Journal of Clinical Oncology**, [S.l.], v.24, n.26, p.4236–4244, 2006.

HORTA, E. G. **Previsores para a Eficiência da Quimioterapia Neoadjuvante no Câncer de Mama**. 2008. Dissertação (Mestrado em Ciência da Computação) — Escola de Engenharia, UFMG, Belo Horizonte, MG, Brasil.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. 2.ed. [S.l.]: Wiley Series in Probability ans Statistics, 2000.

HWANG, D.; SCHMITT, W.; STEPHANOPOULOS, G.; STEPHANOPOULOS, G. Determination of minimum sample size and discriminatory expression patterns in microarray data. **Bioinformatics**, [S.l.], v.18, n.9, p.1184–1193, 2002.

JOHNSON, S. C. Hierarchical clustering schemes. **Psychometrika**, [S.l.], v.32, n.3, p.241–254, 1967.

JONES, M.; MARRON, J.; SHEATHER, S. A brief survey of bandwidth selection for density estimation. **Journal of the American Statistical Association**, [S.l.], p.401–407, 1996.

KUBAT, M.; HOLTE, R.; MATWIN, S. Learning when negative examples abound. **Machine Learning: ECML-97**, [S.l.], p.146–153, 1997.

LAKHDAR, Y.; SBAI, E. H. Optimization of the smoothing parameter of variable kernel estimator. In: COMMUNICATIONS, COMPUTING AND CONTROL APPLICATIONS (CCCA), 2012 2ND INTERNATIONAL CONFERENCE ON, 2012. **Anais...** [S.l.: s.n.], 2012. p.1–5.

LIAO, J.; WU, Y.; LIN, Y. Improving Sheather and Jones' bandwidth selector for difficult densities in kernel density estimation. **Journal of Nonparametric Statistics**, [S.l.], v.22, n.1, p.105–114, 2010.

MEHRA, R.; VARAMBALLY, S.; DING, L.; SHEN, R.; SABEL, M.; GHOSH, D.; CHINNAIYAN, A.; KLEER, C. Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. **Cancer research**, [S.l.], v.65, n.24, p.11259, 2005.

MICHALEWICZ, Z. **Genetic Algorithms + Data Structures = Evolution Programs**. 3.ed. Berlim: Springer, 1996.

MOLANES-LÓPEZ, E. M.; CAO, R. Plug-in bandwidth selector for the kernel relative density estimator. **Annals of the Institute of Statistical Mathematics**, [S.l.], v.60, n.2, p.273–300, 2008.

MUNAKATA, T. **Fundamentals of the New Artificial Intelligence**. 3.ed. New York: Springer-Verlag, 1998.

NARESH, A.; LONG, W.; VIDAL, G.; WIMLEY, W.; MARRERO, L.; SARTOR, C.; TOVEY, S.; COOKE, T.; BARTLETT, J.; JONES, F. The ERBB4/HER4 intracellular domain 4ICD is a BH3-only protein promoting apoptosis of breast cancer cells. **Cancer research**, [S.l.], v.66, n.12, p.6412, 2006.

NATOWICZ, R.; INCITTI, R.; HORTA, E. G.; CHARLES, B.; GUINOT, P.; YAN, K.; COUTANT, C.; ANDRE, F.; PUSZTAI, L.; ROUZIER, R. Prediction of the outcome of preoperative chemotherapy in breast cancer by DNA probes that convey

information on both complete and non complete responses. **BMC Bioinformatics**, [S.l.], v.9, p.149, march 2008.

OKABE, T.; JIN, Y.; SENDHOFF, B. A critical survey of performance indices for multi-objective optimisation. In: EVOLUTIONARY COMPUTATION, 2003. CEC'03. THE 2003 CONGRESS ON, 2003. **Anais...** [S.l.: s.n.], 2003. v.2, p.878–885.

OU, Y.; CHUNG, P.; HSU, F.; SUN, T.; CHANG, W.; SHIEH, S. The candidate tumor suppressor BTG3 is a transcriptional target of p53 that inhibits E2F1. **The EMBO journal**, [S.l.], v.26, n.17, p.3968–3980, 2007.

PARZEN, E. On estimation of a probability density function and mode. **The annals of mathematical statistics**, [S.l.], v.33, n.3, p.1065–1076, 1962.

QUEIROZ, F.; BRAGA, A.; PEDRYCZ, W. Sorted Kernel Matrices as Cluster Validity Indexes. In: IFSA/EUSFLAT CONF., 2009. **Anais...** [S.l.: s.n.], 2009. p.1490–1495.

ROUZIER, R.; RAJAN, R.; WAGNER, P.; HESS, K.; GOLD, D.; STEC, J.; AYERS, M.; ROSS, J.; ZHANG, P.; BUCHHOLZ, T. et al. Microtubule-associated protein tau: a marker of paclitaxel sensitivity in breast cancer. **Proceedings of the National Academy of Sciences of the United States of America**, [S.l.], v.102, n.23, p.8315, 2005.

SCOTT, D. **Multivariate density estimation**. [S.l.]: Wiley Online Library, 1992. v.139.

SCOTT, G. L.; LONGUET-HIGGINS, H. C. Feature grouping by relocalisation of eigenvectors of the proximity matrix. In: BRITISH MACHINE VISION CONFERENCE, 1990. **Proceedings...** [S.l.: s.n.], 1990. p.103–108.

SILVERMAN, B. Density Estimation for Statistics and Data Analysis. **Monographs on Statistics and Applied Probability**, London, 1986.

TABCHY, A.; VALERO, V.; VIDAURRE, T.; LLUCH, A.; GOMEZ, H.; MARTIN, M.; QI, Y.; BARAJAS-FIGUEROA, L.; SOUCHON, E.; COUTANT, C. et al. Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. **Clinical Cancer Research**, [S.l.], v.16, n.21, p.5351, 2010.

TEIXEIRA, R. d. A.; BRAGA, A. P.; TAKAHASHI, R. H.; SALDANHA, R. R. Improving generalization of MLPs with multi-objective optimization. **Neurocomputing**, [S.l.], v.35, n.1, p.189–194, 2000.

THOMPSON, J. R.; TAPIA, R. A. **Nonparametric function estimation, modeling and simulation**. 1a.ed. [S.l.]: Ed. Siam - Society for Industrial and Applied Mathematics, 1990.

TORRES, L. C.; CASTRO, C. L.; BRAGA, A. P. A computational geometry approach for pareto-optimal selection of neural networks. In: **Artificial Neural Networks and Machine Learning–ICANN 2012**. [S.l.]: Springer, 2012. p.100–107.

VAN GESTEL, T.; SUYKENS, J. A.; BAESENS, B.; VIAENE, S.; VANTHIENEN, J.; DEDENE, G.; DE MOOR, B.; VANDEWALLE, J. Benchmarking least squares support vector machine classifiers. **Machine Learning**, [S.l.], v.54, n.1, p.5–32, 2004.

VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer, 2000.

VAPNIK, V.; VASHIST, A. A new learning paradigm: learning using privileged information. **Neural Networks**, [S.l.], v.22, n.5, p.544–557, 2009.

WANDERLEY, M. F. B.; BRAGA, A. P.; MENDES, E. M. A. M.; NATOWICZ, R.; ROUZIER, R. Non-Parametric Kernel Density Estimation for the Prediction of Neoadjuvant Chemotherapy Outcomes. In: ANNUAL INTERNATIONAL CON-

FERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY (EMBC'10), 32., 2010. **Proceedings...** [S.l.: s.n.], 2010.

WANDERLEY, M. F. B.; GARDEUX, V.; NATOWICZ, R.; BRAGA, A. P. GA-KDE-Bayes: an evolutionary wrapper method based on non-parametric density estimation applied to bioinformatics problems. In: ESANN 2013 PROCEEDINGS, EUROPEAN SYMPOSIUM ON ARTIFICIAL NEURAL NETWORKS, COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING, 2013. **Anais...** [S.l.: s.n.], 2013.

WANDERLEY, M. F. B.; TORRES, L. C. B.; NATOWICZ, R.; BRAGA, A. P. Um Estimador de Largura de Kernel Baseado em Margem Larga Aplicado à Previsão de Resposta à Quimioterapia Neoadjuvante. **Revista Brasileira de Engenharia Biomédica**, [S.l.], 2014. Aceito para publicação.

WU, T.-J.; CHEN, C.-F.; CHEN, H.-Y. A variable bandwidth selector in multivariate kernel density estimation. **Statistics & probability letters**, [S.l.], v.77, n.4, p.462–467, 2007.

ZHANG, X.; KING, M. L.; HYNDMAN, R. J. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. **Computational Statistics & Data Analysis**, [S.l.], v.50, n.11, p.3009–3031, 2006.