Relatório de atividades: Uso do classificador Bayesiano

David Clifte*

2015, v-1.0

Resumo

Este trabalho apresenta os resultados obtidos ao aplicar o classificador bayessiano. A implementação foi feita no Matlab^{TM}

Palavras-chaves: Bayes. Reconhecimento de padrões.

Introdução

1 Preparação da base

1.1 Base de dados da flor de íris

A base de dados da flor de íris criado por Fisher (**??**). Nessa base de dados as informações obtidas das flores foram o comprimento e a largura das pétalas e sépalas de 3 tipos de flor de íris, virgínica, versicolor e setosa. Cada tipo de flor possue 50 instancias.

1.2 Base de dados da dermatologia

A base de dados da dermatologia foi criada por Altay, (GüVENIR; DEMIRöZ; ILTER, 1998). Nessa base foram coletadas informações de pacientes que possuiam sintomas de doenças de pele. As doenças são psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Foram coletadas 34 características de um total de 366 pacientes.

Lista de atributos.

Atributos clínicos:
1: erythema

*cliftedavid@gmail.com

2: scaling
3: definite borders
4: itching
5: koebner phenomenon
6: polygonal papules
7: follicular papules
8: oral mucosal involvement
9: knee and elbow involvement
10: scalp involvement
11: family history, (0 or 1)
34: Age (linear)

Atributos do tecido patológico:
12: melanin incontinence
13: eosinophils in the infiltrate
14: PNL infiltrate
15: fibrosis of the papillary dermis
16: exocytosis
17: acanthosis
18: hyperkeratosis
19: parakeratosis
20: clubbing of the rete ridges
21: elongation of the rete ridges
22: thinning of the suprapapillary epidermis
23: spongiform pustule
24: munro microabcess
25: focal hypergranulosis
26: disappearance of the granular layer
27: vacuolisation and damage of basal layer
28: spongiosis
29: saw-tooth appearance of retes
30: follicular horn plug
31: perifollicular parakeratosis
32: inflammatory monoluclear inflitrate
33: band-like infiltrate

1.3 Base de dados da coluna vertebral

Nessa base de dados as informações obtidas das flores foram o comprimento e a largura das pétalas e sépalas de 3 tipos de flor de íris, virgínica, versicolor e setosa. Cada tipo de flor possue 50 instancias.

1.4 Normalização e codificação

Após o carregamento da base foi realizado apenas a normalização dos dados e a codificação dos rótulos. A normalização foi realizada separadamente para cada atributo. Foi identificado o máximo e o mínimo do atributo p e todos os valores foram normalizados na faixa [0,1].

A codificação do rótulo foi feita no modelo 1-de-k(1-of-K ou one-hot encoding), nesse modelo o rótulo é codificado em um vetor onde cada posição do vetor representa uma classe. Nesse modelo para uma quantidade m de classes temos um vetor com m posições e a classe k é representada por um vetor onde todas as outras posições diferentes de k possuem o valor zero e a posição k possui o valor 1. Dessa forma as três classes possíveis da íris foram codificadas em um vetor de 3 posições, onde a posição 1, 2 e 3 representam a classe setosa, versicolor e virgínica respectivamente. O mesmo foi repetido para as bases da dermatologia e da coluna vertebral.

1.5 Análise das características

1.5.1 Base da íris

Na figura 1 é apresentada a matriz de características. Essa matriz consiste de gráficos formados pelos pares de características combinadas.

Na diagonal principal é apresentado o histograma do atributo. Devido a natureza continua dos atributos do problema, pois o comprimento e largura da sépala e pétala podem assumir qualquer valor real O histograma foi obtido após a realização da quantização dos atributos. Cada atributo foi quantizado em 15 possíveis valores com faixas de mesma largura. A largura da faixa foi obtida da seguinte forma $(max_p - min_p)/15$, onde max e min são os valores máximos e mínimos do atributo p.

Podemos perceber que a classe setosa pode facilmente ser separada das outras utilizando a largura ou o comprimento da pétala como variável. Nos histogramas localizados na parte inferior isso pode ser verificado, pétalas com comprimentos menores que 0,25 ou sépalas com largura menor que 0,3 de comprimento são claramente da classe setosa. Já para as duas outras características essa separação não é tão simples, perceba a sobreposição dos histogramas em todos os atributos bem como a mistura das classes nos gráficos de dispersão.

Figura 1 – Matriz de características.

1.5.2 Base da dermatologia

1.5.3 Base da coluna vertebral

2 Classificador de Bayes

2.1 Introdução

Dada uma classificação entre M classes, o classificador de Bayes faz a seleção de um dado x com base na probabilidade de w_i dado um x , $P(w_i|x)$. Assim temos:

$$x \in w_i \iff P(w_i|x) \geq P(w_j|x) \forall i \neq j \tag{1}$$

	Gaussian		Parzen		
DataSet	média	desvio Padrão	média	desvio Padrão	largura da janela
Íris	0.9568	0.0445	0.9486	0.0237	0.0010
Dermatologia	0.9725	0.0189	0.9659	0.0175	0.1846
Vertebra	0.7299	0.0489	0.7649	0.0607	0.0010

Tabela 1 – Acurácia média e desvio padrão em função do valor de k.

2.2 Função de densidade de probabilidade

2.3 Metodologia

A avaliação do classificar de Bayes foi feita de várias formas ao longo deste trabalho, veja a subssessão 2.4.

Na sessão 2.4 são apresentados os resultados obtidos ao aplicar o classificador de bayes utilizando como função de densidade probabilidade a Gaussiana e a janela de parzen. Para a janela de parzen foi exibido o melhor resultado para um determinada largura de janela do tipo gaussiana. O tamanho da janela foi determinado após uma pesquisa linear entre valores na faixa de 0.001 e 3 para o valor da variância.

2.4 Resultados obtidos

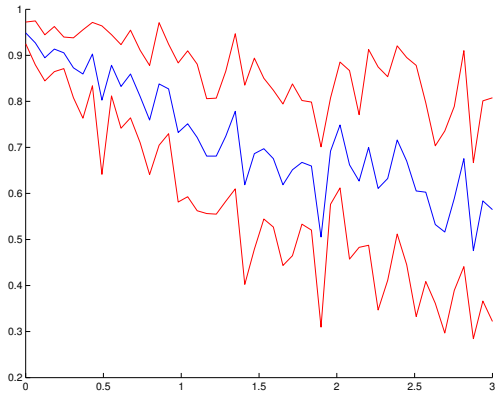
Na tabela 1 são exibidos os resultados obtidos ao realizar a classificação utilizando a regra de Bayes com funções de densidade probabilidade (PDF) diferentes. Para cada base de dados tem-se o resultado utilizando a Gaussiana como PDF e a janela de parzen.

2.4.1 Busca por tamanho de janela ótima de Parzen

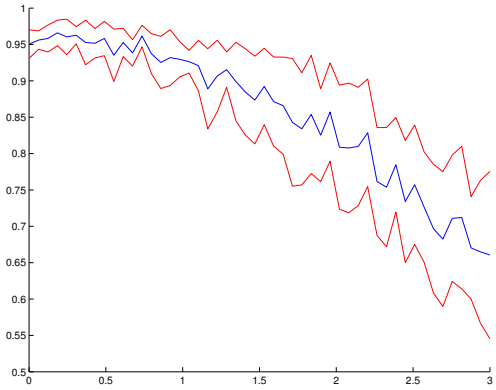
Abaixo, figura 2 são exibidos os resultados obtidos ao realizar a busca pelo tamanho da janela ótima de parzen para cada base de dados. Foram considerados todos os atributos disponíveis em cada base, estes foram normalizados e codificados apropriadamente assim como dito na subsessão 1.4.

A linha ao centro de cada gráfico é a acurácia as outras duas são os limites. Os limites são determinados a uma distância de um σ^2(desvio padrão) acima e abaixo da acurácia média. Pode-se perceber que indiferentemente da base de dados o aumento considerável da janela de parzen causa uma redução da acurácia do classificador. Isso é causado por uma maior interferência dos vizinhos no cálculo da probabilidade condicional o que faz com que seja levado em conta apenas as probabilidades a priore para a determinação da classe para um dado x. Isso pode ser observado quando a variância, largura da janela de parzen, é maior que 2 na base da íris ou maior que 0.7 na base da coluna vertebral.

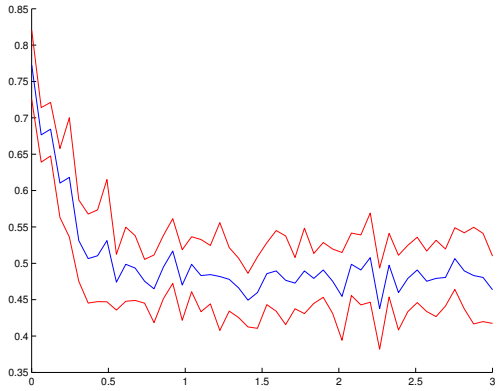
Os pontos máximos destas curvas nos indicam o valor ótimo da janela



(a) Íris



(b) Dermatologia



(c) Coluna Vertebral

Figura 2 – Resultados da busca pela largura ótima da janela de parzen.

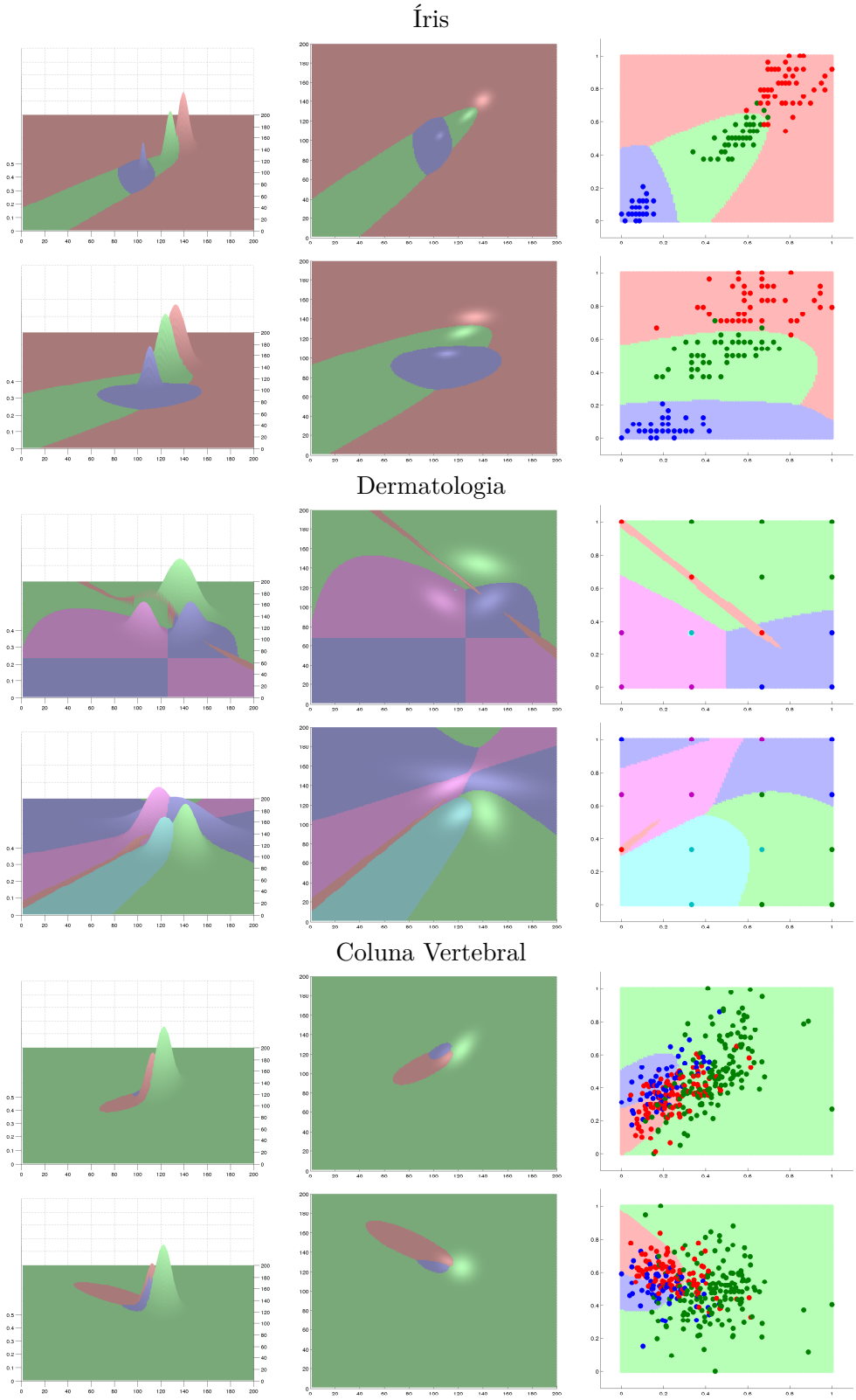
de parzen. Desta forma temos na tabela 1 o valor de acurácia e tamanho da janela para as três bases.

2.5 Análise da matriz da região de decisão

Abaixo são apresentadas as regiões de decisões

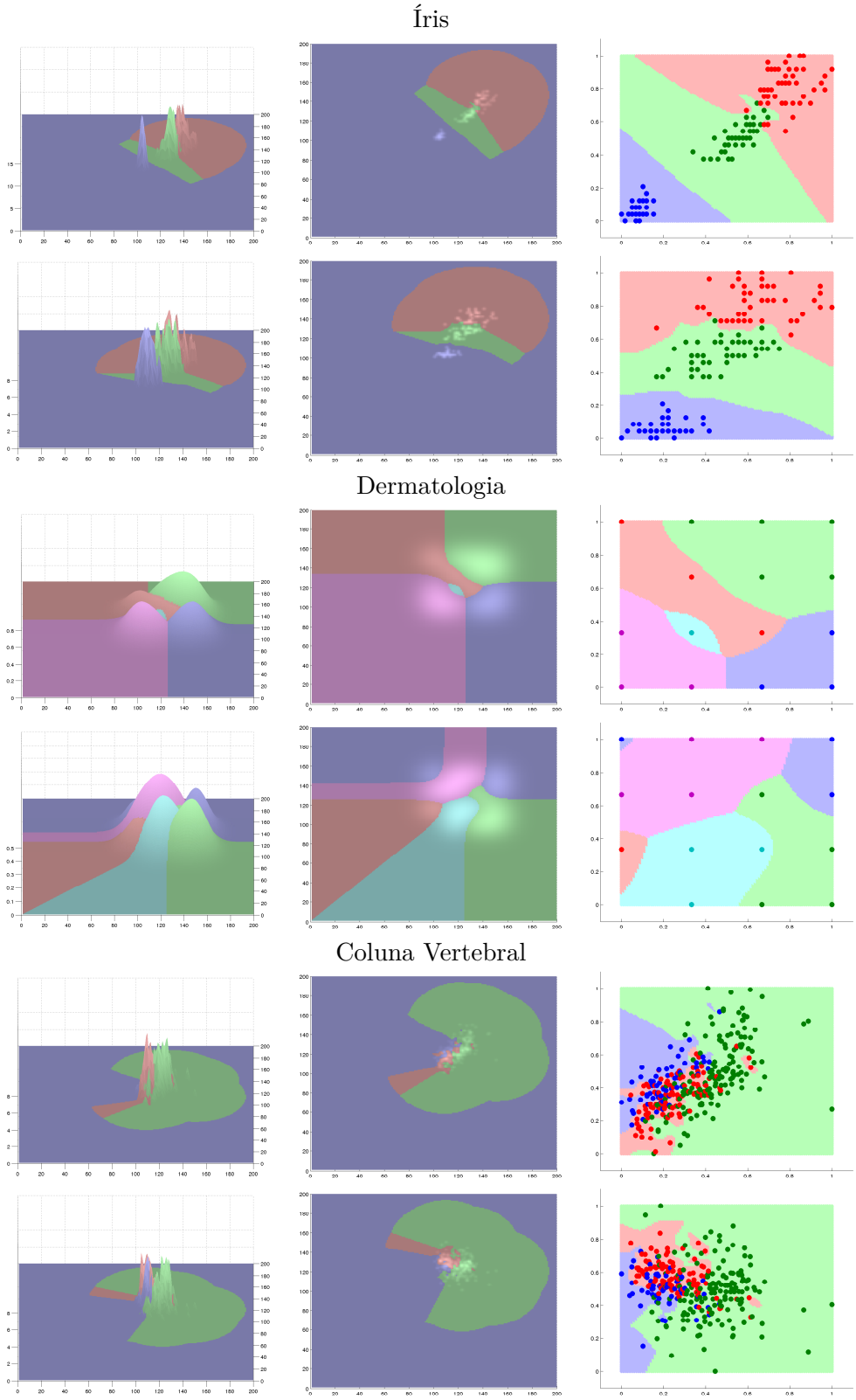
2.5.1 Matrizes de covariancia distintas. $\Sigma_i \neq \Sigma_i \forall i \neq j$

Com matrizes de covariancias distintas temos uma região de decisão quadrática.



(a) $P(W_i|x)$ (b) Regiao de decisão (c) Resultado da classificação dos dados sobre a regiao de decisão

Figura 3 – Resultados da busca pela largura ótima da janela de parzen.



(a) $P(W_i|x)$ (b) Regiao de decisão (c) Resultado da classificação dos dados sobre a regiao de decisão

Figura 4 – Região de decisão calculada utilizando janela de parzen com largura $\sigma^2 = 0.05I$.

2.5.2 Matrizes de covariancia iguais. $\Sigma_i \Sigma \forall i$

2.5.3 Matrizes de covariancia diagonais distintas. $\Sigma_i \neq \Sigma_i \forall i \neq j$

2.5.4 Matrizes de covariancia não diagonais distintas. $\Sigma_i \neq \Sigma_i \forall i \neq j$

2.5.5 Janela de Parzen

2.5.6 Região de decisão

Nas a figuras 5 e 6 são exibidas as regiões de decisões ao aplicar o KNN considerando apenas duas características. Para a figura 5 são consideradas apenas os atributos largura da sépala e pétala. Um importante ponto a ser levado em consideração é que ao descartar determinados atributos podem surgir informações duplicadas. É importante filtrar estas informações, podem surgir linhas com informações iguais porém rótulos diferentes o que pode causar um erro na geração região de decisão..

Diferentes regiões de decisões foram calculadas para diferentes valores de K. Para K=1 é notável que a existencia de ruídos podem gerar regiões que prejudicam a generalização. Para k=5 é possível perceber que pontos isolados não mais definem a classe de determinada região do espaço,veja na figura 5 a existência de 2 pontos da classe virgínica(em vermelho) na região em que foi classificada como da classe versícolor(em verde). A medida que o K aumenta podemos perceber a queda no valor da acurácia. Vários pontos são classificados de forma errada.

O mesmo se aplica para a figura 6, porém nesta foram consideradas apenas as larguras e comprimentos da sépala.

2.5.7 Matriz Confusão

Na tabela **??** é exibida a matriz confusão obtida para K igual a 10. Esse valor de k foi escolhido devido aos testes de acurácia em função de k mostrarem que com este valor é obtida a melhor acurácia. Além da matriz confusão a tabela **??** traz os resultados, falso-positivo, falso-negativo, verdadeiro-positivo e verdadeiro-negativo.

Considerações finais

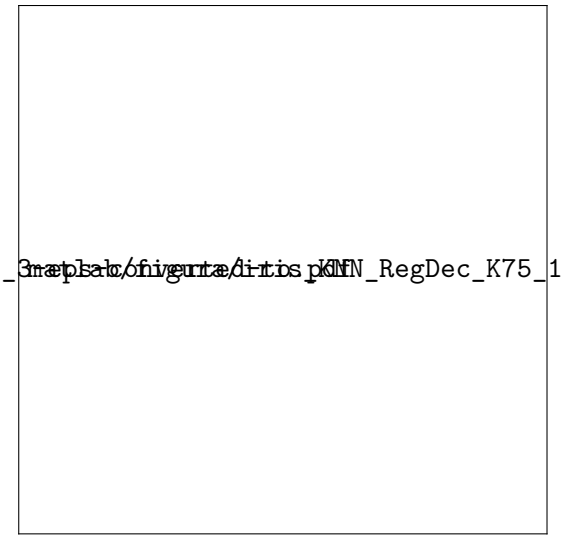
Podemos concluir que o KNN, indiferentemente da forma como os dados estão dispostos no espaço, é possível separar as regiões por mais irregulares que sejam. O KNN possui um parâmetro que permite regular a confiabilidade de determinado ponto no espaço, reduzindo assim a incidencia de ruídos.

Apesar de não ser apresentado nenhuma avaliação de desempenho neste relatório, é notável a diferença entre os tempos de execução de ambas as técnicas.

(a) k=1

(b) k=10

(c) k=50



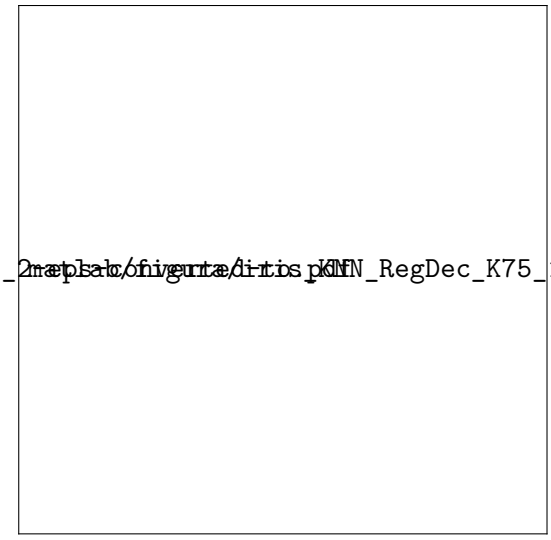
(d) k=75

Figura 5 – Região de decisão utilizando Largura da sépala e da pétala.

(a) k=1

(b) k=10

(c) k=50



(d) k=75

Figura 6 – Região de decisão utilizando comprimento e largura da sépala.

Referências

GüVENIR, H. A.; DEMIRöZ, G.; ILTER, N. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, v. 13, p. 147, 1998. Citado na página 1.