

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 e submete-lo a 5 diferentes funções de densidade de probabilidade. Para a verificação da acurácia bem como das regiões de decisão são apresentados os resultados obtidos com três base de dados diferentes. A implementação foi feita no MatlabTM

Palavras-chaves: Bayes. Reconhecimento de padrões.

Introdução

0.1 Estimativa de função de densidade de probabilidade

A estimativa da distribuição de dados é importante pois permite utilizar uma representação compacta dos dados e ainda sim manter as informações mais relevantes da base. Existem basicamente três abordagens para estimar a função de densidade de probabilidade de um sinal: paramétrica, não-paramétrica e semiparamétrica. O sucesso destas representações dependem do modelo que tem sido definido.

0.1.1 Método não paramétrico

Os métodos não-paramétricos não fazem nenhuma consideração da distribuição de probabilidade dos dados. Em geral, estes métodos se caracterizam por conseguir uma estimativa adequada para qualquer conjunto de dados que recebem como entrada.

*cliftedavid@gmail.com

0.1.2 Método paramétrico

A abordagem paramétrica é geralmente usada quando a distribuição dos dados é conhecida antecipadamente ou quando os dados são simples de forma que permitam ser modelados usando uma distribuição conhecida, por exemplo gaussiana, Gamma, Laplace, etc

0.1.3 Método semi-paramétrico

A abordagem semiparamétrica combina a flexibilidade da abordagem não-paramétrica e a eficiência na avaliação dos parâmetros da abordagem paramétrica. Estes modelos utilizam um número de funções base que são sempre menores que o conjunto de treinamento. O uso dos modelos semiparamétricos baseados em gaussianas, GMM, tem se apresentado como uma ferramenta amplamente usada na estimativa da PDF de qualquer sinal.

0.2 Gaussian Mixture Models

0.2.1 Introdução

Considerando um conjunto de dados $X = x_1, x_2, \dots, x_n | x \in R$, a PDF dos dados pode ser aproximada por uma família F de funções de distribuição de probabilidades em R. Em algoritmos dedicados à estimativa da PDF, o problema é encontrar a função de distribuição $f(x) \in F$ que melhor gere os dados de entrada.

$$f(x, \Theta) = \sum_{k=1}^k P_k g(x, \mu_k, \sigma_k) \quad (1)$$

Θ é o conjunto de parâmetros do conjunto de funções que devem ser estimados durante a fase de treinamento. Desta forma para gaussiana temos

$$\Theta = \begin{bmatrix} \mu_1 & \sigma_1 \\ \dots & \dots \\ \mu_k & \sigma_k \end{bmatrix} \quad (2)$$

Θ pode ser estimado utilizando o Algoritmo Maximização da Expectância (EM). O algoritmo EM é um procedimento iterativo para estimar os parâmetros de uma mistura de gaussianas. Cada iteração do algoritmo EM consiste em dois processos: Expectância e Maximização. Esta aproximação se consegue através do cálculo da probabilidade de pertinência de um ponto às funções de distribuições na fase de expectância. Na fase de maximização são estimados os parâmetros que maximizam cada função de distribuição, ponderadas com os valores calculados na fase de expectância.

Na figura 1 é apresentado o resultado da aproximação da base da íris. São exibidas três das quatro combinações possíveis das características da base, essas combinações são utilizadas para realizar o treinamento do GMM.

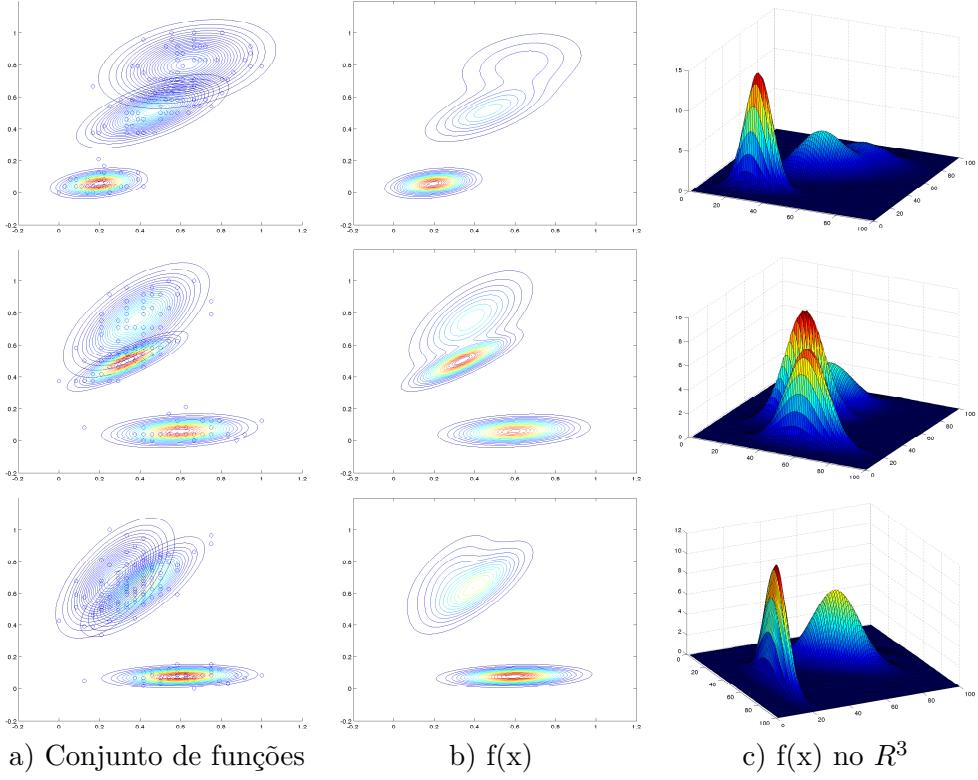


Figura 1 – Resultados obtidos ao utilizar a mistura de gaussianas para modelar a base de dados da íris. Na primeira linha temos o resultado do treinamento utilizando as informações comprimento da sepala e largura da pétala. Na linha seguinte temos largura da sépala e da pétala e na terceira linha o comprimento da sépala e largura da pétala

0.3 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 coluna vertebral, câncer de mama, diabetes e haberman.

1 Classificador de Bayes

1.1 Introdução

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

$$x \in w_i \iff P(w_i|x) \geq P(w_j|x) \forall i \neq j \quad (3)$$

1.2 Opção de Rejeição

Considerando o classificar de bayes um padrão é escolhido em detrimento a outro de acordo com sua probabilidade a posteriori, A opção de rejeição sugere que de acordo com este valor de probabilidade calculado a classificação pode ser rejeitada pois a mesma reflete também o grau de confiança na classificação. Desta forma podemos definir um limiar para este grau de confiança e assim caso a probabilidade a posteriori seja menor que este limiar a amostra pode ser classificada para a classe de rejeição. Temos a seguinte regra de decisão para um problema com duas classes:

$$x \in \begin{cases} w_1, & \text{se } P(w_1|x) > \beta \\ w_2, & \text{se } P(w_2|x) > \beta \\ w_r, & \text{caso contrário} \end{cases} \quad (4)$$

1.3 Avaliação de um classificador com Opção de Rejeição

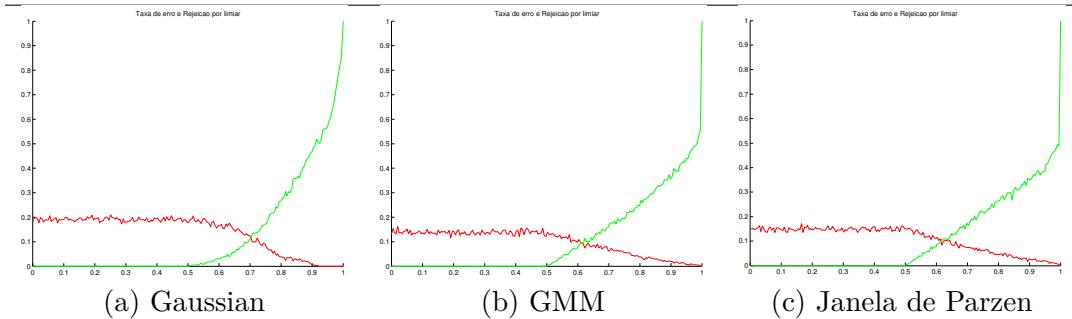
Algumas das métricas importantes ao utilizarmos um classificador com opção de rejeição são taxa de erro e taxa de acurácia em função do limiar de rejeição. A análise destas duas métricas permite identificar qual o grau de confiança ao realizar uma classificação, bem como evitar erros ao rejeitar amostras duvidosas. Na figura 2, são apresentadas as curvas obtidas ao variar o valor desse limiar para as quatro bases em análise neste trabalho.

Na figura 2, os limiares de rejeição β variam de 0 a 1. Podemos perceber que a taxa de rejeição aumenta somente a partir de 0,5. Isso ocorre devido a tomada de decisão ser feita apenas para duas classes, dessa forma, para um limiar inferior a 50% sempre haverá uma classe com uma probabilidade a posteriori maior.

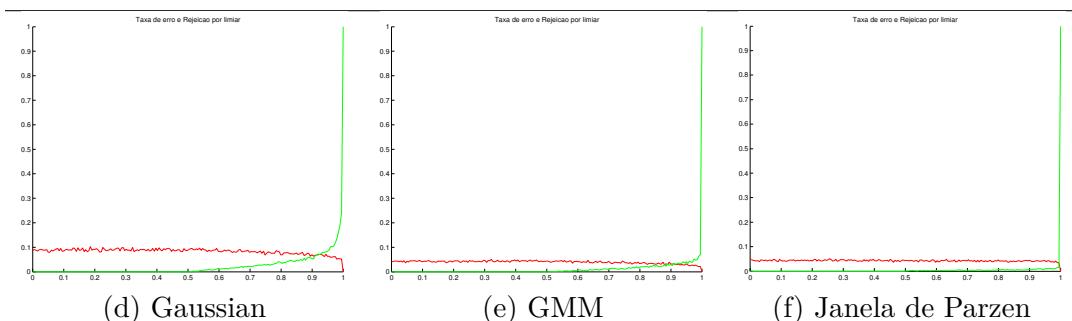
1.3.1 Impacto na região de decisão

Ao optarmos por um classificador com opção de rejeição criamos uma nova classe que acolherá os dados que não puderam ser discriminados para alguma das outras duas classes, dessa forma, a classe de rejeição também define uma região de decisão. Essa região tem sua área controlada pelo valor do limiar

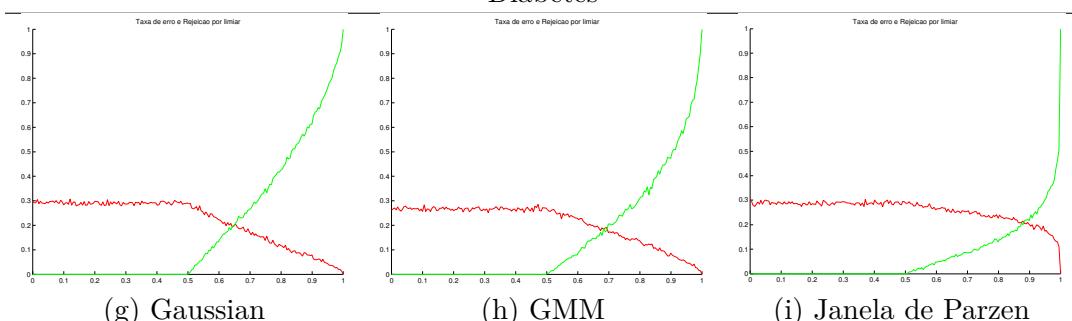
Coluna Vertebral



Câncer de mama



Diabetes



Haberman

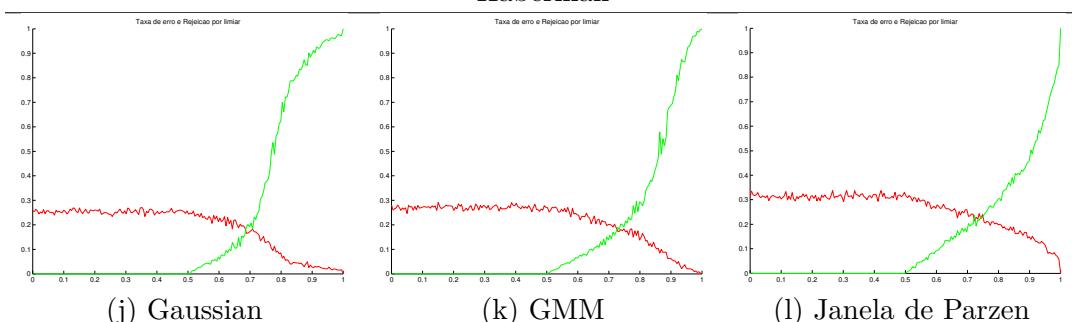


Figura 2 – Taxa de Erro e Taxa de Rejeição em função do limiar de rejeição utilizando a função gaussiana como PDF. Em verde a taxa de rejeição em vermelho a taxa de erro.

de rejeição e sempre inicializa na região de interseção das outras duas classes. Nas figuras 3,4 e 5 são exibidas as regiões de decisões controladas com o limiar de rejeição utilizando 3 diferentes funções de densidade de probabilidade.

1.4 Metodologia

Na sessão 1.5 são apresentados os resultados obtidos ao aplicar o classificador de bayes utilizando como função de densidade probabilidade a Gaussiana e as distâncias euclidiana e de mahalanobis. Para todos os conjuntos de dados os testes foram realizados sobre uma partição de 25% do total de dados, escolhida de forma aleatória para cada repetição dos algoritmos.

Os três primeiros testes usam a função gaussiana como função de densidade de probabilidade. No primeiro teste foram realizados os testes utilizando médias e variancias diferentes para cada classe com esses valores calculados apenas com dados da mesma classe. Assim estimamos um μ_i e um σ_i^2 para uma classe i utilizando apenas os valores de $x \in W_i$, onde W_i são os padrões da classe i .

O segundo teste realizado leva em consideração que todas as matrizes de covariância das classes possuem a seguinte forma: $\Sigma_i = \sigma^2 I \forall i$ portanto são matrizes diagonais com o mesmo valor de variancia, σ^2 e covariâncias nula.

O terceiro teste considera que todas as matrizes de covariâncias são diferentes porem os elementos que não estão na diagonal principal tem seus valores atualizados para zero, assim temos a matriz de covariância da classe i é $\Sigma_i : \Sigma_i(a, b) = 0 \forall a \neq b$.

O quarto e quinto teste leva em consideração que as classes são equiprováveis. Para o quarto a distinção entre as classes é feita através do cálculo da distância Euclidiana enquanto para o quinto é feito o cálculo da distância de Mahalanobis. Diferentemente das abordagens anteriores, onde a decisão por uma classe é feita a partir da maximização da probabilidade a posteriori, o uso da distância Euclidiana ou Mahalanobis age de forma diferente. Afim de manter a mesma ideia de maximização o resultado da distância é normalizado exponencialmente entre 0 e 1, onde 1 ocorre quando a distância entre uma amostra e a média da classe é zero.

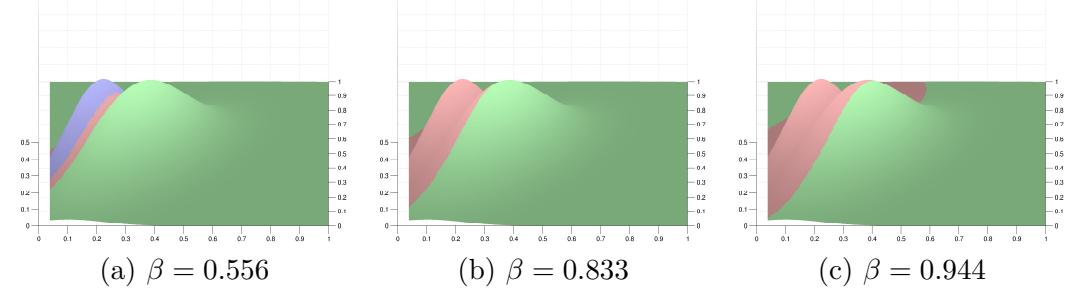
1.5 Resultados obtidos

Na tabela ?? 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, além disso são exibidos os resultados obtidos para diferentes matrizes de covariância.

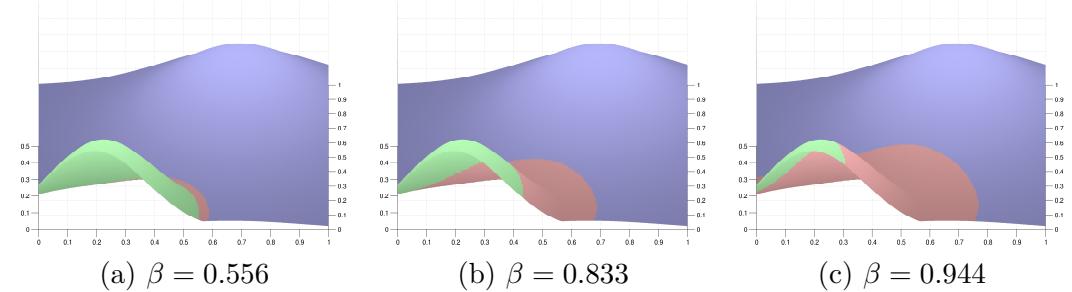
Na primeira tabela temos o classificador utilizando matrizes de covariâncias diferentes para cada classe. Isso permite que a região de decisão possa

PDF Gaussiana com RejOption

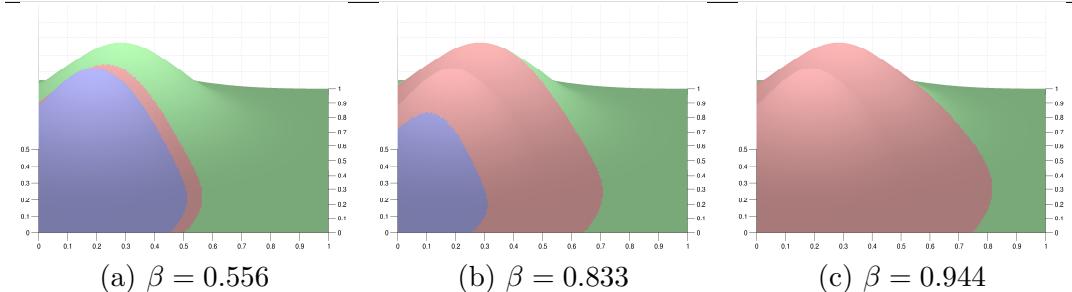
Coluna vertebral



Câncer de Mama



Diabetes



Haberman

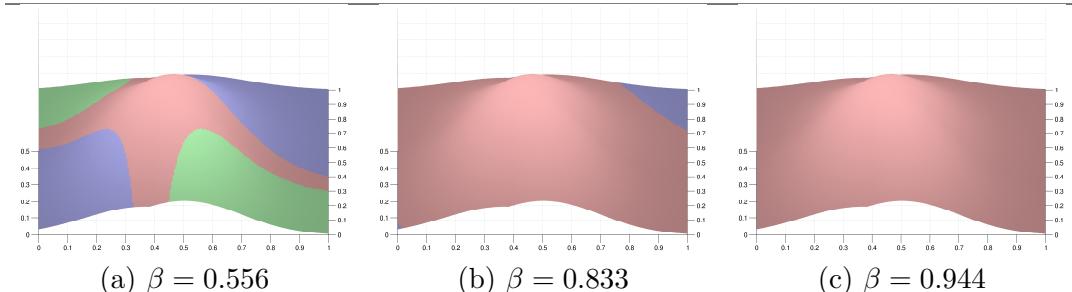
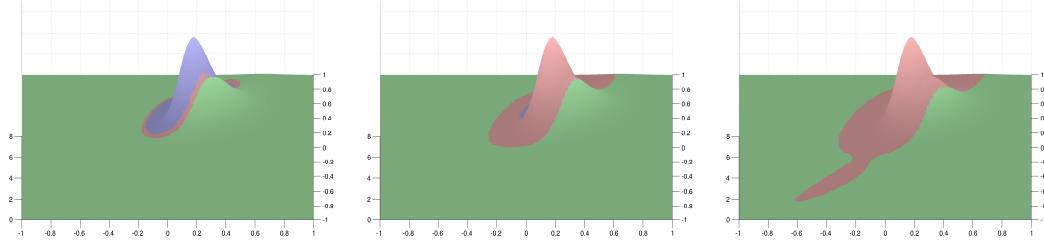


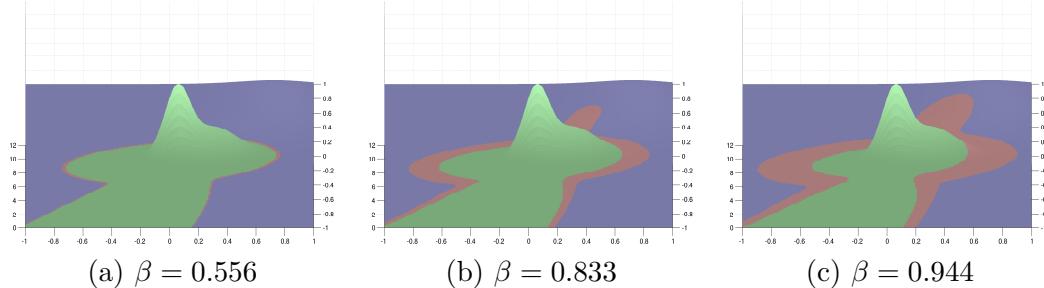
Figura 3 – Região de decisão formada utilizando a gaussiana como função de densidade de probabilidade. Foram testados três valores de limiar de rejeição β para as três bases.

PDF GMM com RejOption

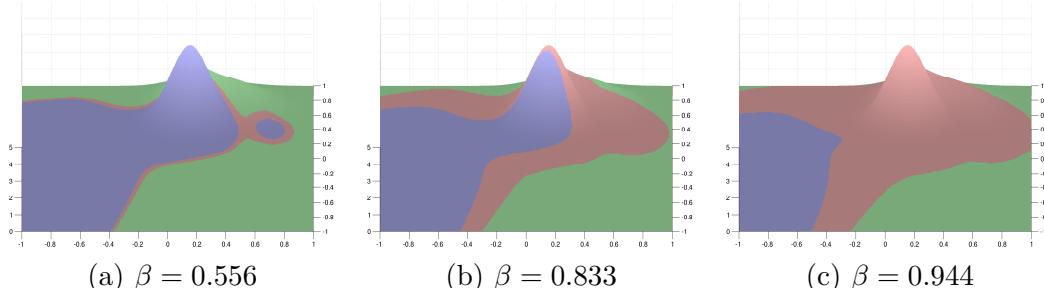
Coluna vertebral



Câncer de Mama



Diabetes



Haberman

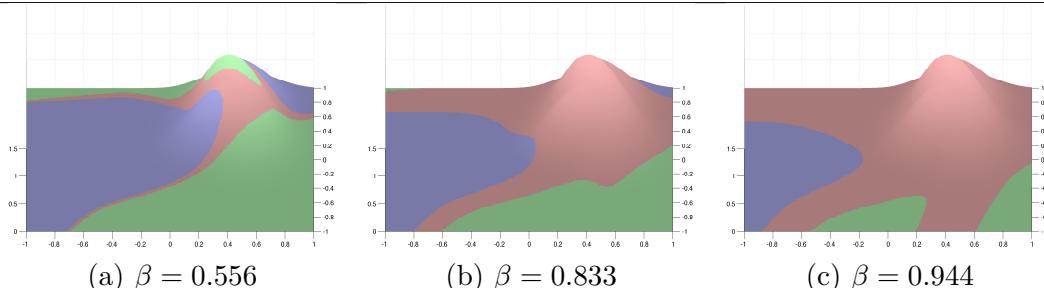
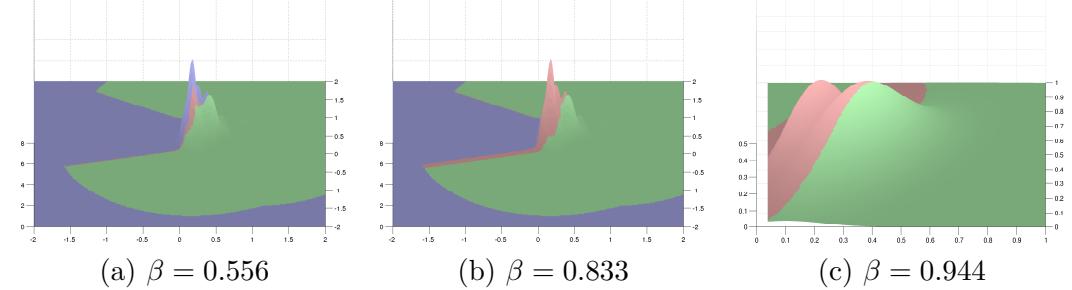


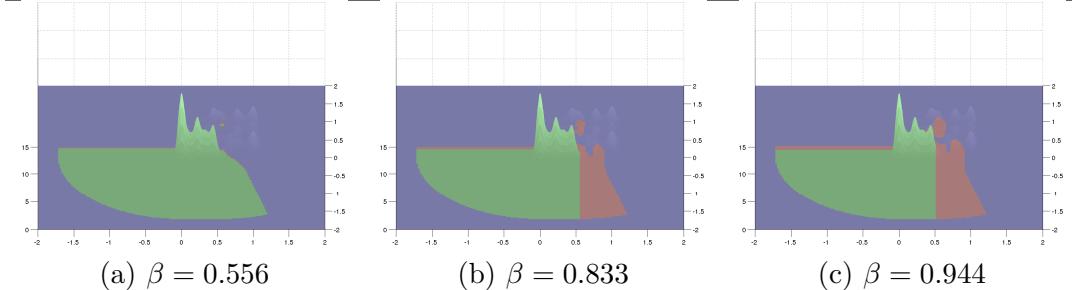
Figura 4 – Região de decisão formada utilizando mistura de gaussianas como função de densidade de probabilidade. Para todos os exemplos foram utilizadas 3 gaussianas para modelar cada classe. Foram testados 3 valores de limiar de rejeição β para as três bases.

PDF Janela de Parzen com RejOption

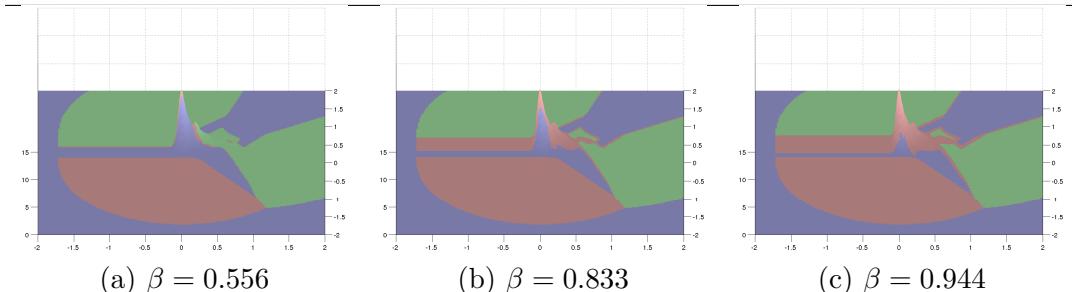
Coluna vertebral



Câncer de Mama



Diabetes



Haberman

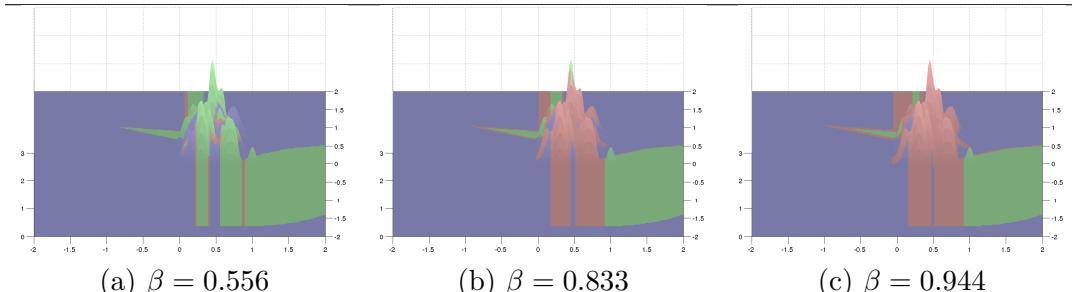


Figura 5 – Região de decisão formada utilizando janela de Parzen como função de densidade de probabilidade. Foram testados três valores de limiar de rejeição β para as três bases.

modelar os dados de uma forma quadrática e portanto espera-se um resultado melhor que a tabela seguinte.

Na tabela seguinte, ao fazer a escolha por apenas uma matriz de covariância para todas as classes o classificador se comporta de maneira linear, restringindo ainda mais a sua capacidade de discriminação dos dados, resultando portanto em uma acurácia média menor.

No terceiro teste, ao optar por matrizes de covariâncias diferentes obtemos novamente uma região de decisão não linear, neste caso quadrática, por isso temos uma melhora nos resultados em comparação com a tabela anterior.

No quarto teste, ao utilizarmos probabilidades a priori iguais e a distância Euclidiana como função de densidade de probabilidade temos novamente um classificador linear. Esse classificador também é conhecido como DMC(Distância mínima ao centróide).

No quinto e último teste temos a distância de Mahalanobis como função de densidade de probabilidade. Este tipo de distância leva em consideração a forma como os pontos estão dispostos no espaço para calcular a distância. Como cada classe possui uma matriz de covariância diferente, este classificador também é não linear. Ao realizarmos a normalização exponencial, como dito na sessão anterior 1.4, temos um resultado semelhante ao da primeira tabela. Uma pequena diferença existe pois neste quinto teste as probabilidades a priori foram definidas com valores iguais.

1.6 Análise da região de decisão

Abaixo são apresentadas as regiões de decisões obtidas para os 3 problemas, íris, vertebra e dermatologia.

Para o problema da íris as regiões de decisões foram calculadas utilizando os atributos largura da pétala x comprimento da pétala e largura da sépala x comprimento da pétala.

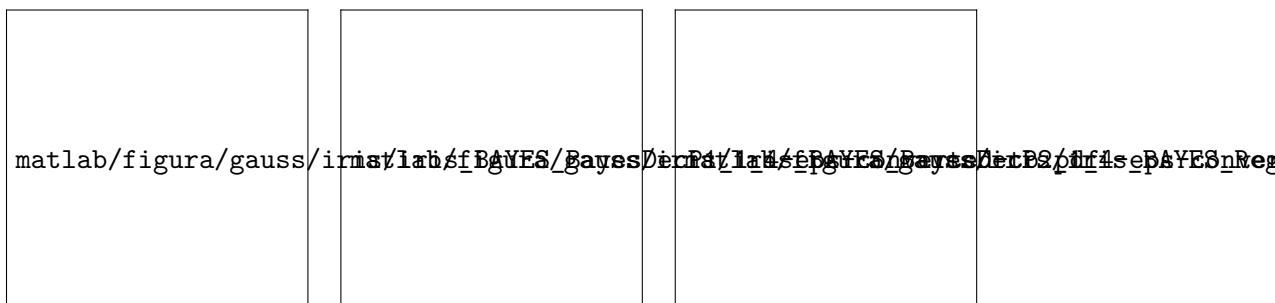
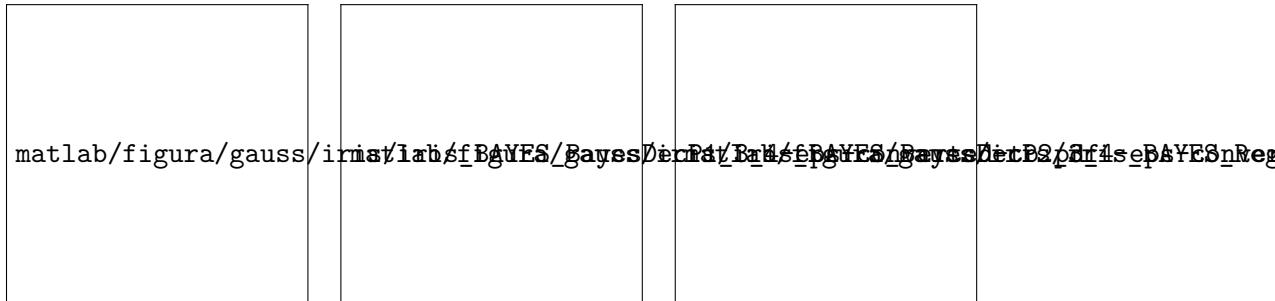
Para o problema da coluna vertebral as regiões de decisões foram calculadas utilizando os atributos pelvic incidence x pelvic tilt e pelvic incidence x sacral slope.

Para o problema da dermatologia as regiões de decisões foram calculadas utilizando os atributos erythema x exocytosis e erythema x acanthosis.

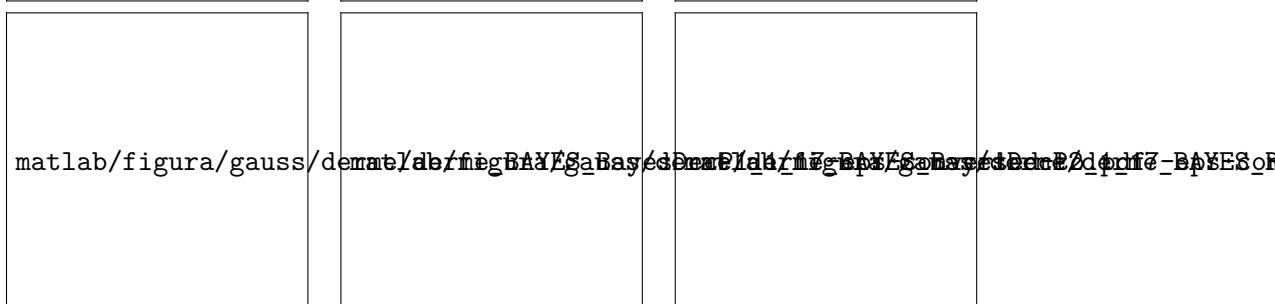
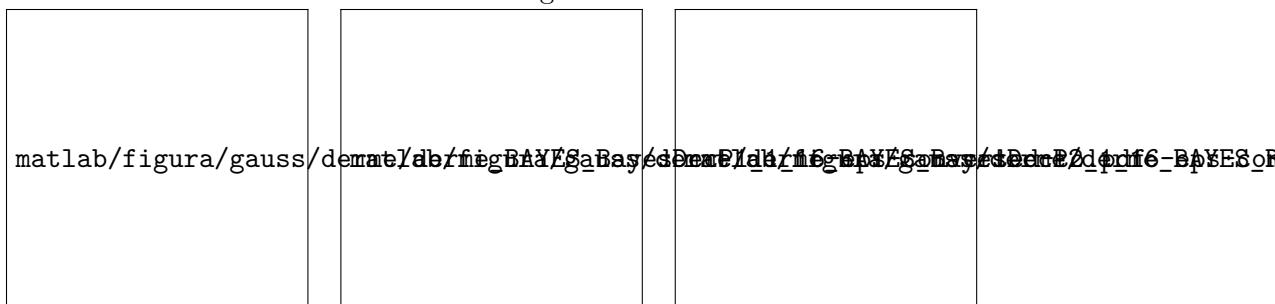
1.6.1 Gaussiana como PDF e matrizes de covariância distintas. $\Sigma_i \neq \Sigma_j \forall i \neq j$

Com matrizes de covariâncias distintas temos uma região de decisão quadrática desta forma podemos notar várias elipses ao longo das regiões, figura 6. Podemos perceber na terceira coluna como os dados foram classificados sobre a região de decisão. Poucas amostras foram classificadas de forma errada, em especial para a base da íris. Já para a base da dermatologia o número de amostras erradas foi ligeiramente maior.

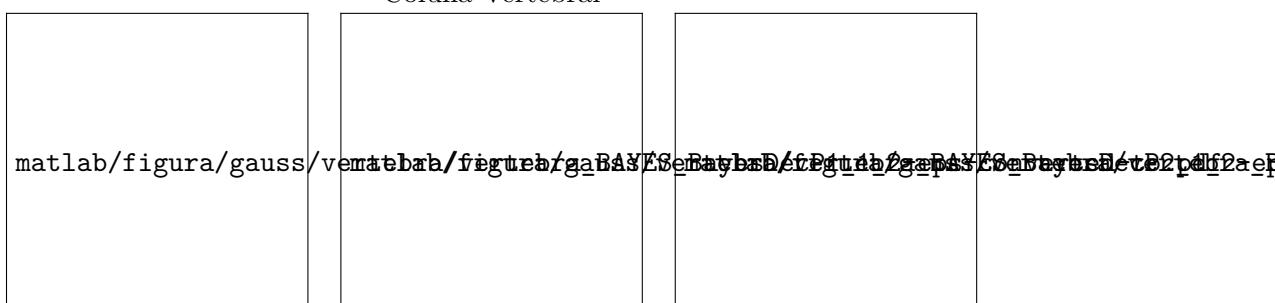
Íris



Dermatologia



Coluna Vertebral



Considerações finais

Neste trabalho foram apresentados os resultados obtidos ao testar várias funções de densidade de probabilidade junto ao classificador bayessiano. Os resultados podem ser visualizados na tabela ???. Podemos observar uma superioridade dos métodos não lineares frente aos lineares, bem como uma maior acurácia para métodos que consideram melhor a modelagem dos dados, quando se utiliza matrizes de covariância diferentes para cada classe de problema.