

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 com opção de rejeição a 4 bases de dados públicas, são elas: Base de doenças coluna vertebral ([NETO; BARRETO, 2009](#)), base de diagnóstico de diabetes ([SMITH et al., 1988](#)), base de diagnóstico de câncer de mama([MANGASARIAN; STREET; WOLBERG, 1995](#)) e base de haberman ([LO, 1993](#)).

Ao fim do trabalho é apresentado uma aplicação do classificador bayesiano combinado com um função de densidade de probabilidade do tipo janela de parzen à segmentação de imagens. É apresentado também a aplicação de Mistura de Gaussianas para segmentação de imagens. 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 ([PATTERN..., 2001](#)).

*cliftedavid@gmail.com

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. O uso de métodos não paramétricos para a estimação das funções de densidade de probabilidade vem da falta de informações a priori sobre a função de densidade de probabilidade dos dados. Como exemplo de método não paramétrico temos a janela de Parzen.

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 Janela de Parzen–Rosenblatt

A janela de Parzen, Parzen-Rosenblatt, é um método utilizado para estimar a função de densidade de probabilidade $p(x)$ com base nas amostras presentes na base de treinamento. A ideia básica da janela de Parzen é contar quantas amostras influenciam em determinada região R (Janela). A influência de cada amostra é definida a priori com a definição do Kernel. A utilização de kernels do tipo hipercubos ou Gaussianas multivariadas são as mais comuns e ao longo deste trabalho são utilizados apenas kernels de gaussianas multivariadas.

Definimos a influência de uma amostra x como $p(x)$. Para o caso da gaussiana multivariada temos:

$$p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu)^t \Sigma^{-1} (x - \mu)\right] \quad (1)$$

Portanto a influência total $P(x)$ é computada como:

$$P(x) = \frac{1}{N} \sum_{i=1}^N p(x_n) \quad (2)$$

Apesar da gaussiana multivariada aceitar qualquer matriz de covariância, contanto que esta seja regular e simetrica, neste trabalho são utilizadas apenas matriz que possuem uma configuração $\sigma^2 I$ ou seja, uma matriz diagonal de covariancia igual a σ .

0.3 Gaussian Mixture Models

0.3.1 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.

$$f(x, \Theta) = \sum_{k=1}^k P_k g(x, \mu_k, \sigma_k) \quad (3)$$

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

$$\Theta = \begin{bmatrix} \mu_1 & \sigma_1 \\ \dots & \dots \\ \mu_k & \sigma_k \end{bmatrix} \quad (4)$$

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.

1 Classificador de Bayes com opção de rejeição

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 ,

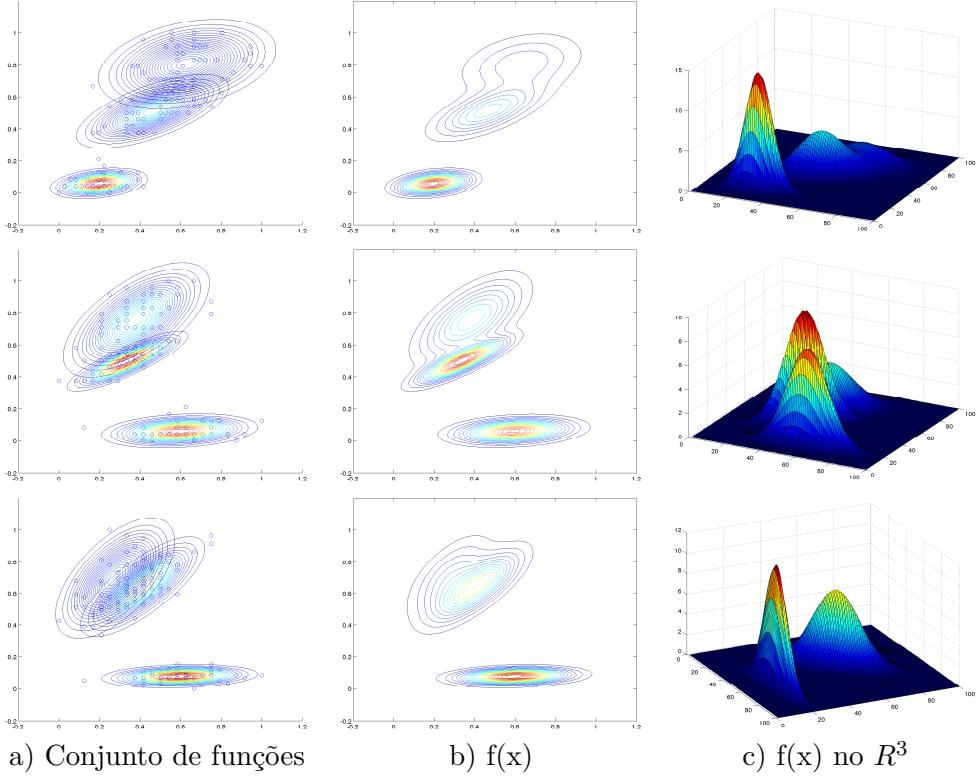


Figura 1 – Resultados obtidos ao utilizar a mistura de gaussianas para modelar a base de dados da íris. Na primeira linha temos o resultado do treinamento utilizando as informações comprimento da sepala e largura da pétala. Na linha seguinte temos largura da sépala e da pétala e na terceira linha o comprimento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

$P(w_i|x)$. Assim temos:

$$x \in w_i \iff P(w_i|x) \geq P(w_j|x) \forall i \neq j \quad (5)$$

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:

$$x \in \begin{cases} w_1, & \text{se } P(w_1|x) > \beta \\ w_2, & \text{se } P(w_2|x) > \beta \\ w_r, & \text{caso contrário} \end{cases} \quad (6)$$

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 portiereira maior.

O desempenho de um classificador com capacidade de rejeição pode ser descrito através de uma curva que leva em consideração a taxa de classificação (acurácia) em relação à sua taxa de rejeição. Esta representação é denominada curva A-R (Accuracy-Reject Curve), em que cada valor correspondente a uma taxa de erro e a uma taxa de rejeição que depende do custo de rejeição W_r ([SINPATCO..., 2011](#)). Esses valores são obtidos à partir da minimização do erro empírico que é calculado através da soma ponderada do erro e da taxa de rejeição obtidos para um valor de limiar de rejeição. Temos o erro empírico definido a seguir.

$$E_{empírico}(\beta) = E(\beta) + W_r R(\beta) \quad (7)$$

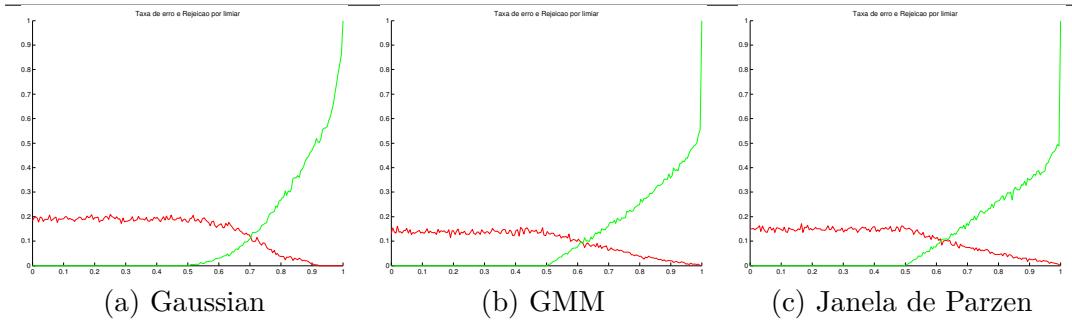
Dessa forma podemos obter o erro empírico para cada uma das bases, figura 3 em função do valor de β e de W_r . Na figura são exibidos apenas os W_r que geram um par $(E(\beta), R(\beta))$ distintos. Esse par ordenado é utilizado para gerar os gráficos da curva A-R exibido na figura 4.

Os mínimos de cada W_r são exibidos na tabela 1

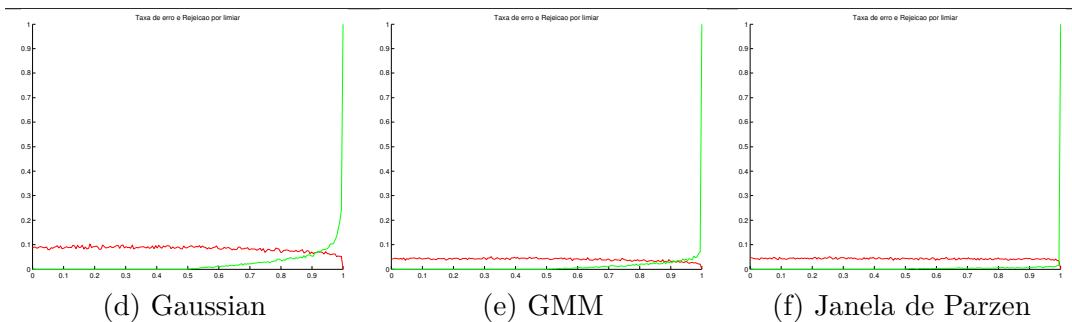
1.3.1 Impacto na região de decisão

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 de rejeição e sempre inicializa na região de interseção das outras duas classes. Nas figuras 5, 6 e 7 são exibidas as regiões de decisões controladas com o limiar de rejeição utilizando 3 diferentes funções de densidade de probabilidade.

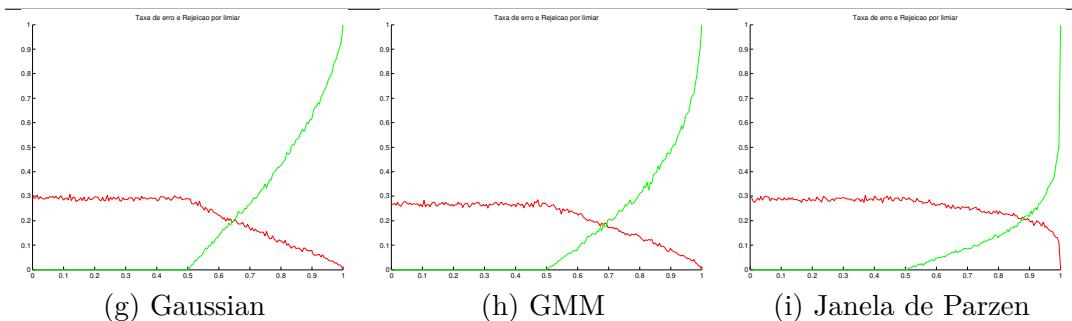
Coluna Vertebral



Câncer de mama



Diabetes



Haberman

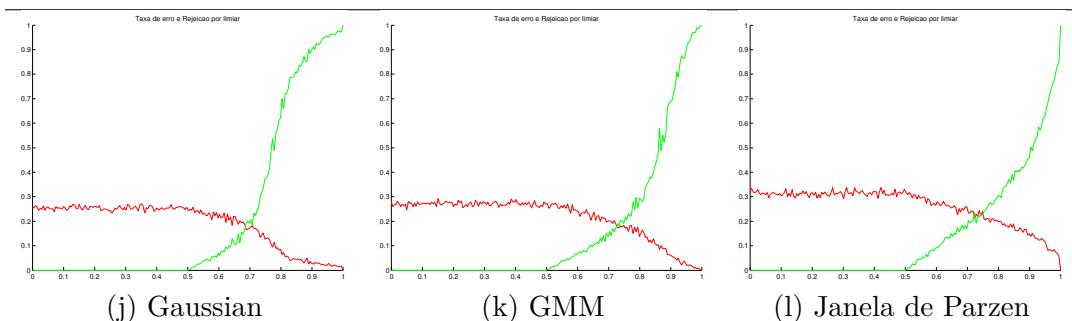


Figura 2 – Taxa de Erro e Taxa de Rejeição em função do limiar de rejeição utilizando a função gaussiana como PDF. Em verde a taxa de rejeição em vermelho a taxa de erro.

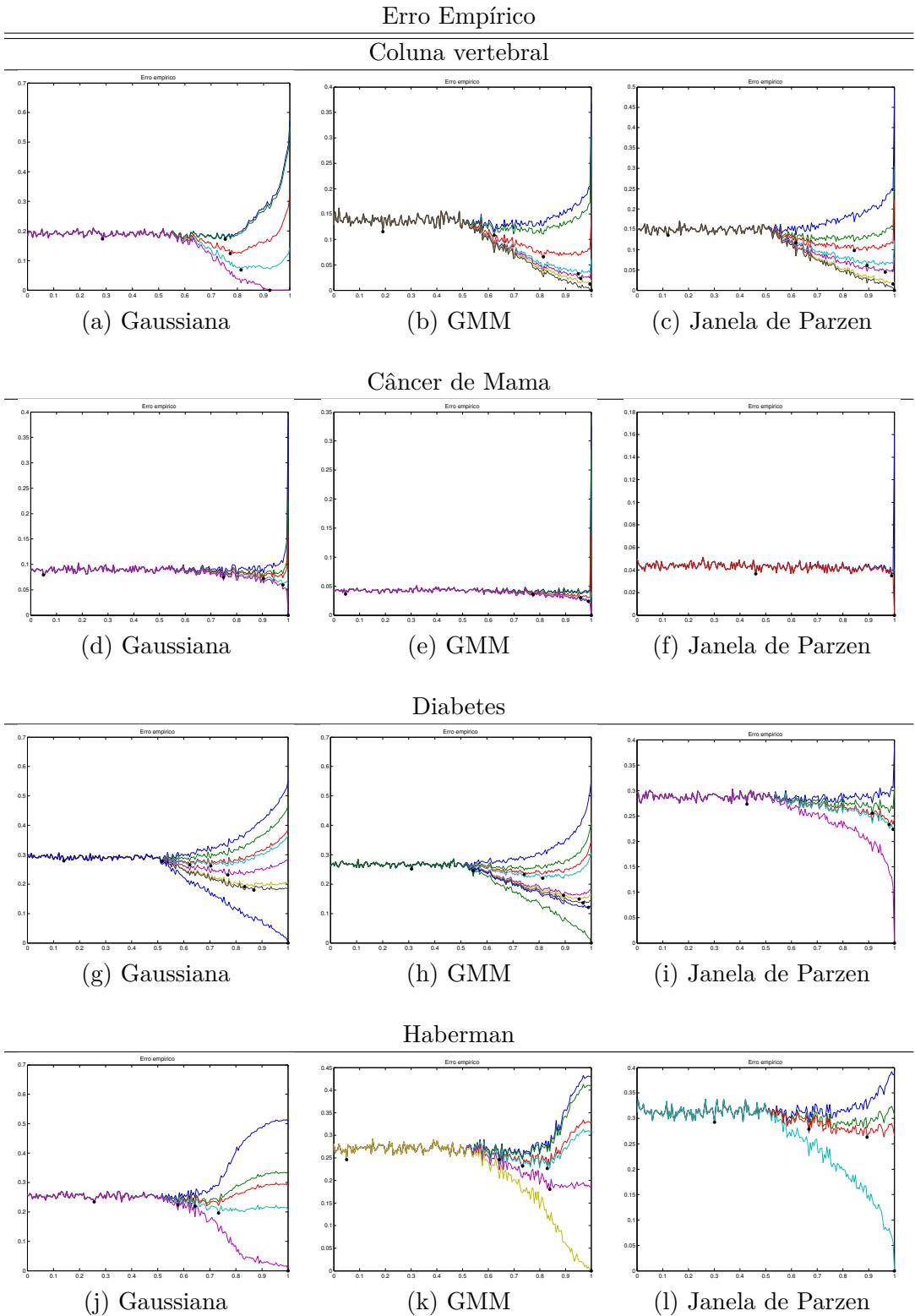
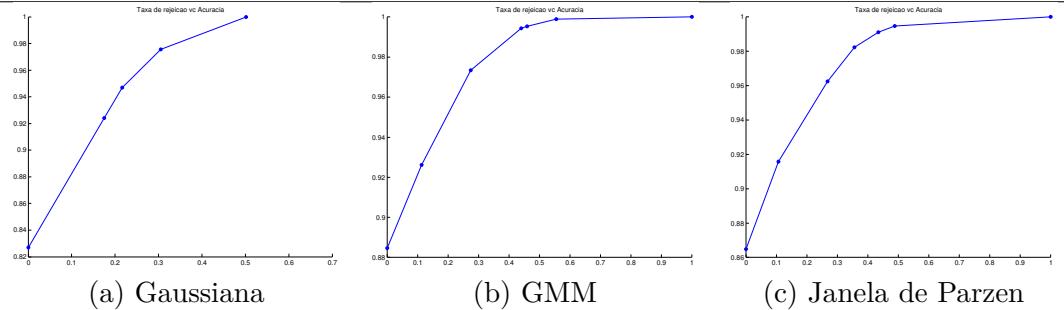


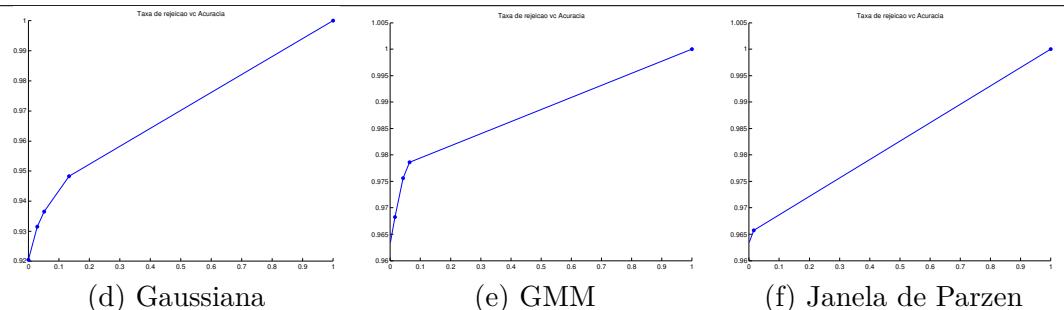
Figura 3 – Cada curva em cada um dos gráficos reflete o erro empírico em função de algum W_r . O valor de cada W_r utilizado para gerar as curvas podem ser visualizado com mais detalhes na tabela 1.

Curva A-R

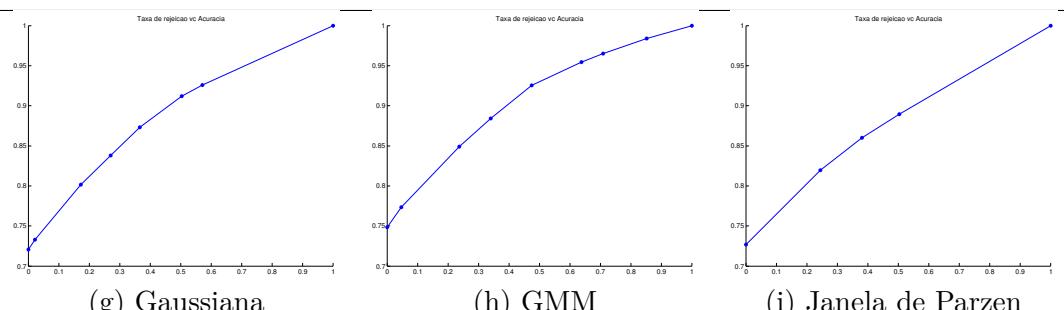
Coluna vertebral



Câncer de Mama



Diabetes



Haberman

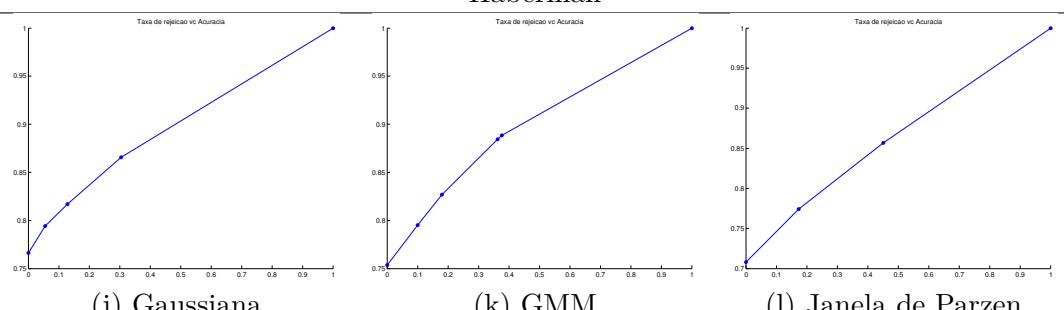
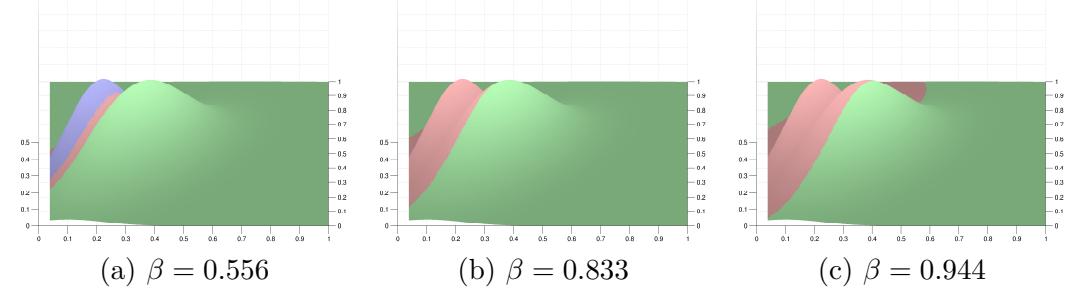


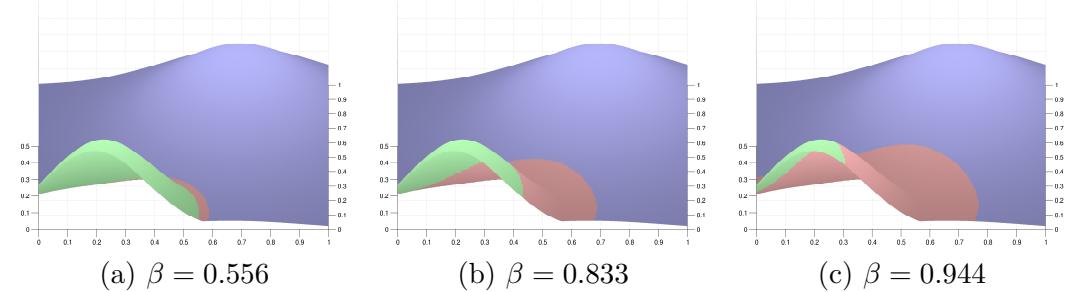
Figura 4 – Cada curva em cada um dos gráficos reflete o erro empírico em função de algum W_r . O valor de cada W_r utilizado para gerar as curvas podem ser visualizado com mais detalhes na tabela 1.

PDF Gaussiana com RejOption

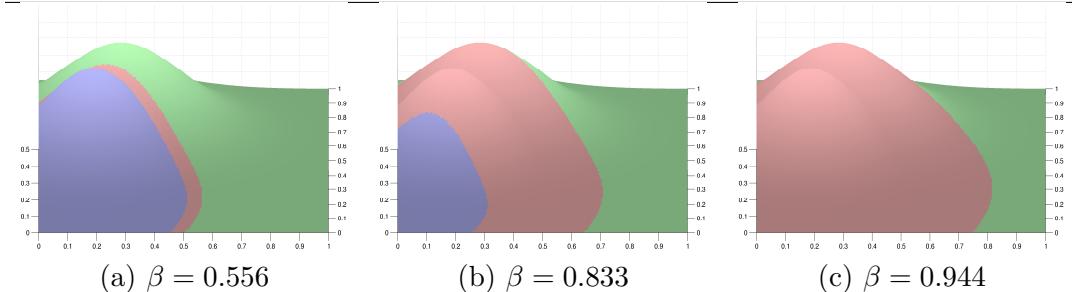
Coluna vertebral



Câncer de Mama



Diabetes



Haberman

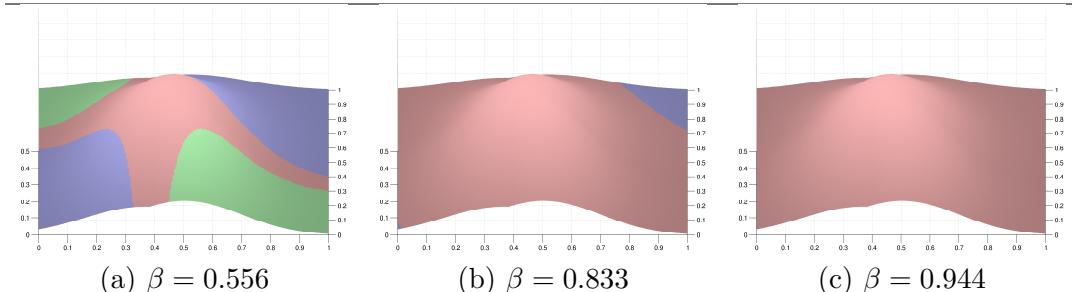
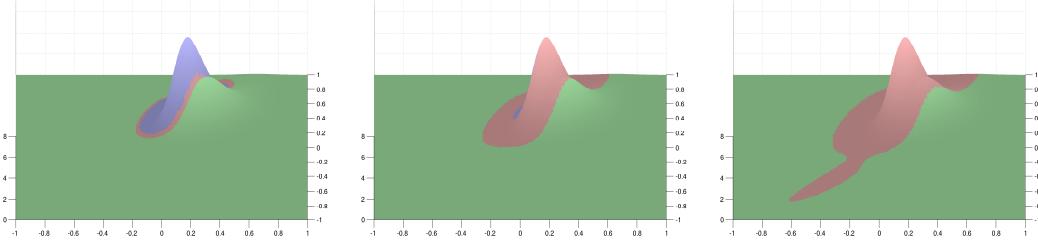


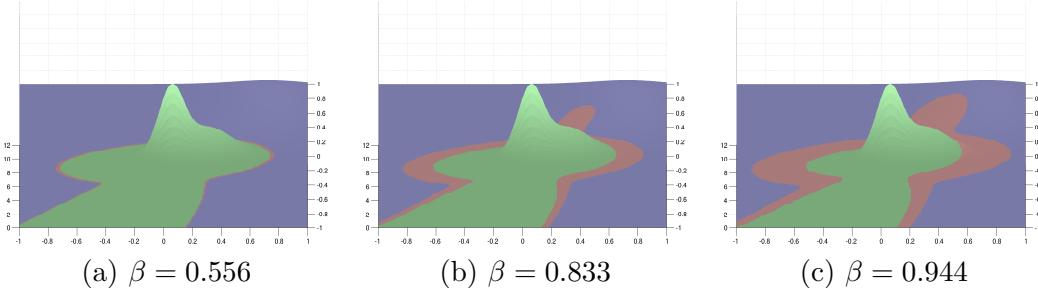
Figura 5 – Região de decisão formada utilizando três diferentes funções de densidade de probabilidade. Foram testados três valores de limiar de rejeição β para as três bases.

PDF GMM com RejOption

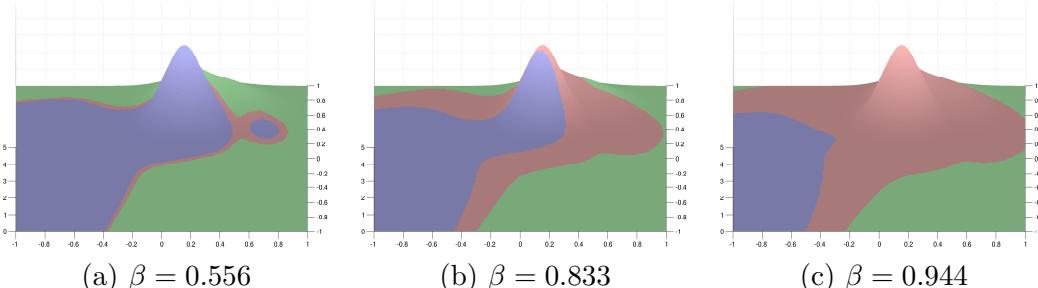
Coluna vertebral



Câncer de Mama



Diabetes



Haberman

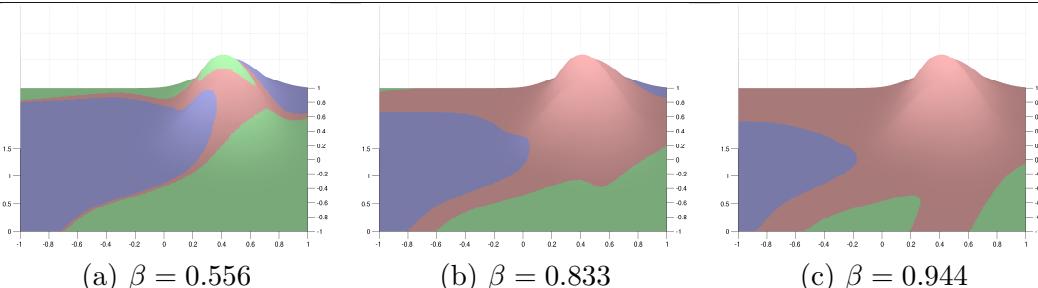
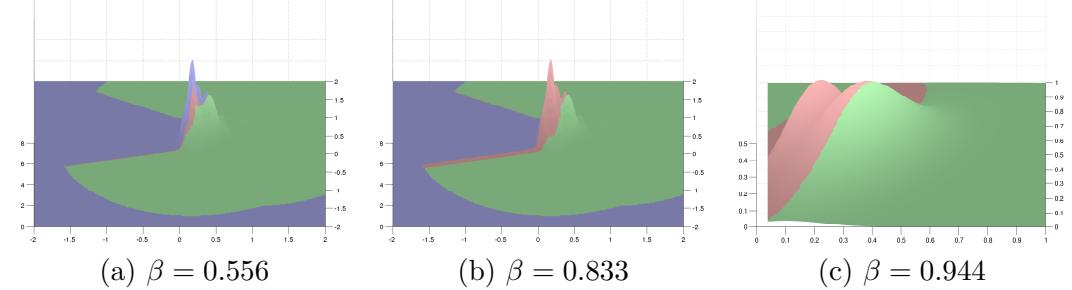


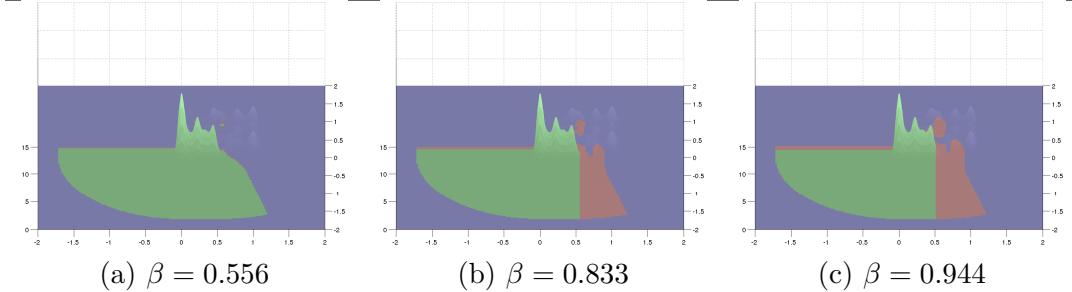
Figura 6 – Região de decisão formada utilizando mistura de gaussianas como função de densidade de probabilidade. Para todos os exemplos foram utilizadas 3 gaussianas para modelar cada classe. Foram testados 3 valores de limiar de rejeição β para as três bases.

PDF Janela de Parzen com RejOption

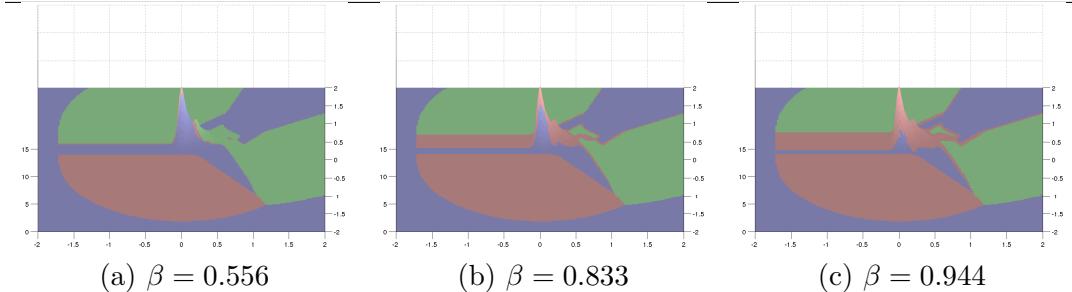
Coluna vertebral



Câncer de Mama



Diabetes



Haberman

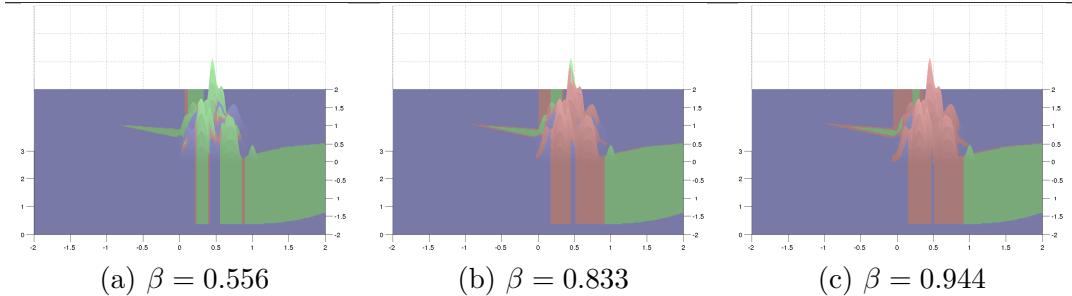


Figura 7 – Região de decisão formada utilizando janela de Parzen como função de densidade de probabilidade. Foram testados três valores de limiar de rejeição β para as três bases.

Erro Empírico								
Coluna vertebral				Câncer de Mama				
Diabetes				Haberman				
W_r	β_o	$A(\beta_o)$	$R(\beta_o)$	W_r	β_o	$A(\beta_o)$	$R(\beta_o)$	
(a) Gaussiana	0.571	0.286	0.827	0	0.367	0.191	0.885	0
	0.551	0.754	0.924	0.175	0.306	0.623	0.926	0.113
	0.327	0.774	0.947	0.217	0.143	0.814	0.974	0.274
	0.143	0.814	0.976	0.305	0.0612	0.95	0.994	0.44
	0	0.925	1	0.502	0.0408	0.96	0.995	0.459
					0.0204	0.995	0.999	0.556
					0	1	1	1
(b) GMM								
(c) Janela de Parzen	0.49	0.121	0.865	0	0.306	0.618	0.916	0.107
	0.224	0.844	0.963	0.268	0.122	0.894	0.982	0.356
	0.0816	0.965	0.991	0.435	0.0204	0.995	0.995	0.489
	0	1	1	1				
(d) Gaussiana	0.388	0.0503	0.92	0	0.327	0.0452	0.963	0
	0.224	0.749	0.931	0.0297	0.286	0.774	0.968	0.0156
	0.163	0.905	0.937	0.0524	0.143	0.96	0.976	0.0423
	0.0612	0.98	0.948	0.134	0.0408	0.99	0.979	0.0639
	0	1	1	1	0	1	1	1
(e) GMM								
(f) Janela de Parzen	0.163	0.462	0.963	0	0.0408	0.99	0.966	0.0156
	0	1	1	1				
(g) Gaussiana	0.551	0.141	0.721	0	0.551	0.312	0.749	0
	0.469	0.518	0.733	0.0219	0.408	0.548	0.774	0.0462
	0.388	0.623	0.802	0.172	0.347	0.744	0.849	0.237
	0.367	0.704	0.838	0.27	0.306	0.814	0.884	0.34
	0.286	0.769	0.873	0.366	0.184	0.894	0.925	0.474
	0.204	0.834	0.912	0.504	0.163	0.955	0.955	0.638
	0.184	0.869	0.926	0.572	0.143	0.97	0.965	0.709
	0	1	1	1	0.122	0.99	0.984	0.852
(h) GMM	0	1	1	1				
(i) Janela de Parzen	0.388	0.427	0.727	0	0.306	0.915	0.819	0.244
	0.245	0.98	0.86	0.381	0.224	0.995	0.89	0.504
	0	1	1	1				
(j) Gaussiana	0.429	0.0503	0.754	0	0.388	0.302	0.708	0
	0.327	0.578	0.794	0.0553	0.408	0.643	0.795	0.101
	0.286	0.643	0.817	0.128	0.327	0.734	0.827	0.181
	0.204	0.734	0.866	0.305	0.306	0.829	0.884	0.363
	0	1	1	1	0.184	0.839	0.888	0.377
(k) GMM	0	1	1	1				
(l) Janela de Parzen	0.388	0.668	0.774	0.173	0.265	0.894	0.857	0.452
	0	1	1	1				

Tabela 1 – Cada tabela exibe a acurácia e a taxa de rejeição ótimos encontrado para algum limiar β_o em função de W_r para cada uma das bases utilizando três diferentes PDFs.

2 Segmentação de imagem utilizando o classificador de bayes

Nas figuras 8 e 8 são apresentados os resultados da segmentação de duas imagens utilizando o classificador Bayessiano com janela de parzen como função de densidade de probabilidade. Para ambas as imagens foram testados os valores, $0.0100 * 10^{-4}$, $0.1733 * 10^{-4}$, $0.3367 * 10^{-4}$ e $0.5000 * 10^{-4}$ para a variancia da janela de Parzen.

A segmentação de uma imagem é realizada levando em consideração a intensidade normalizada entre 0 e 1 dos pixels nas três componentes RGB. A definição dos dados de uma classe é feita a partir da seleção de k regiões da imagem, não necessariamente de mesma área A_i . A área é utilizada para calcular a probabilidade a priori por isso classes com uma diferença de área

Bandeira dos EUA		Bandeira do Japão	
<i>h</i>	Acurácia	<i>h</i>	Acurácia
1e-06	0.619	1e-06	0.960
1.73e-05	0.820	1.73e-05	0.999
3.37e-05	0.936	3.37e-05	0.998
5e-05	0.947	5e-05	0.998

Tabela 2 – Acurácia obtida em função do tamanho da janela para as imagens da bandeira do Japão e dos Estados Unidos

muito grande podem impactar na classificação. Assim temos:

$$P(w_i) = \frac{A_i}{\sum_{i=1}^k A_J} \quad (8)$$

erro

Para o cálculo da acurácia da segmentação, foi definida uma máscara para cada uma das imagens. Foi realizada a comparação pixel a pixel e o resultado obtido para cada imagem é exibido na tabela 2.

3 Segmentação de imagem utilizando GMM

A utilização da mistura de gaussianas como ferramenta de segmentação autmática se dá na escolha de K regiões semelhantes, onde o valor de K é inicialmente acertado. O atributo levado em consideração para o calculo da semelhança é a intensidade dos 3 canais, R, G e B da imagem ([FARNOOSH; ZARPAK, 2008](#)). Os resultados obtidos com a segmentação utilizando GMM são exibidos na figura 10.

Referências

FARNOOSH, R.; ZARPAK, B. Image segmentation using gaussian mixture model. *IUST International Journal of Engineering Science*, 2008. Citado na página [13](#).

LO, W. D. *Logistic regression trees*. Tese (Doutorado) — University of WisconsinMadison, Dept. of Statistics, 1993. Citado na página [1](#).

MANGASARIAN, O. L.; STREET, W. N.; WOLBERG, W. H. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, v. 43, p. 570–577, Aug 1995. Citado na página [1](#).

NETO, A.; BARRETO, G. On the application of ensembles of classifiers to the diagnosis of pathologies of the vertebral column: A comparative analysis.

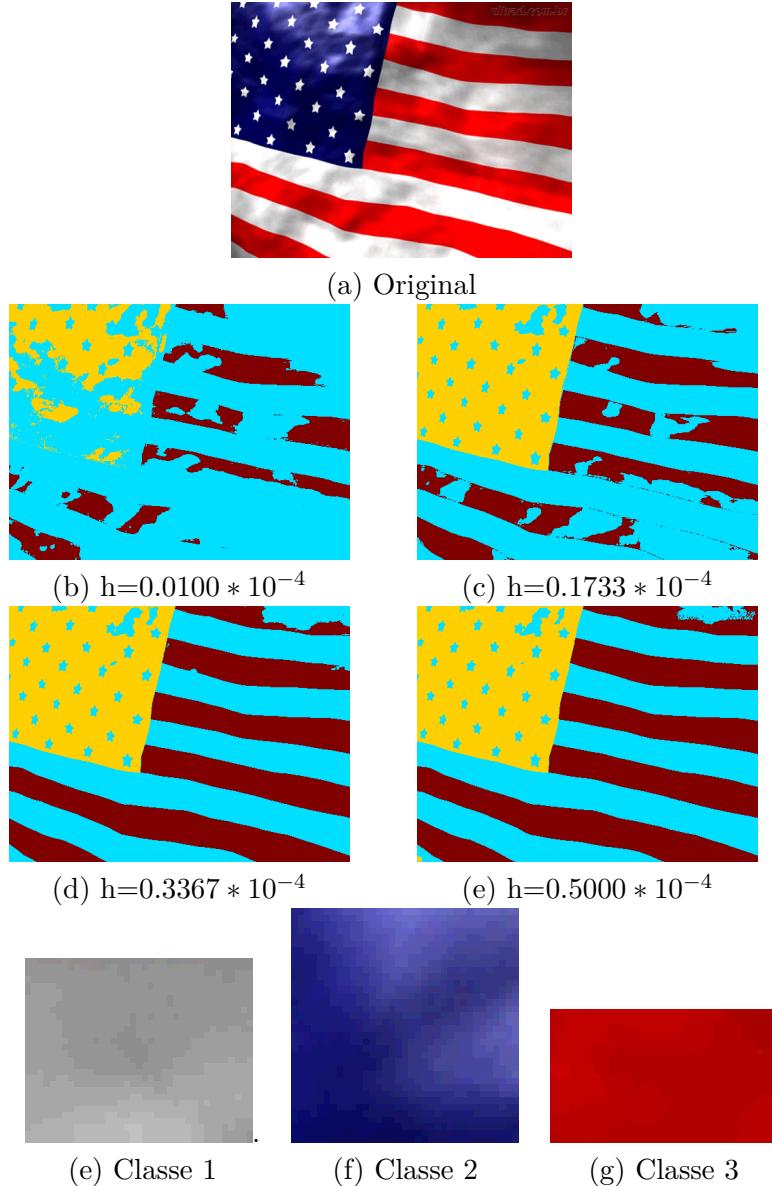


Figura 8 – Resultado da segmentação utilizando o classificador Bayessiano com janela de parzen. Com o aumento do valor de h pode-se notar que a segmentação se torna mais suave.

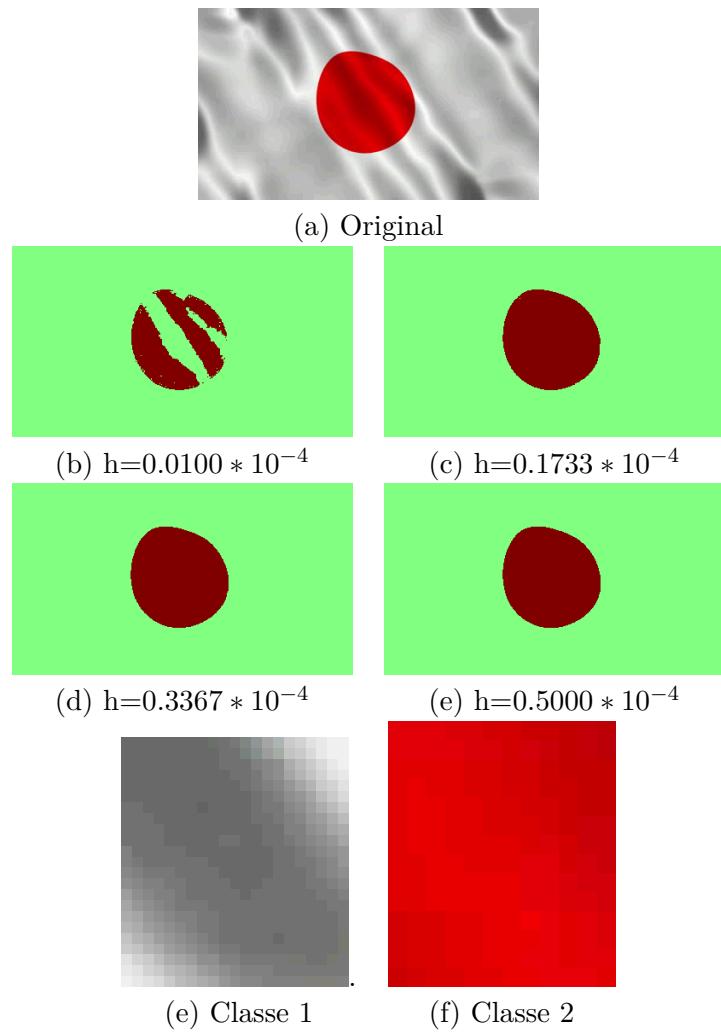


Figura 9 – Resultado da segmentação utilizando o classificador Bayessiano com janela de parzen. Com o aumento do valor de h pode-se notar que a segmentação se torna mais suave.

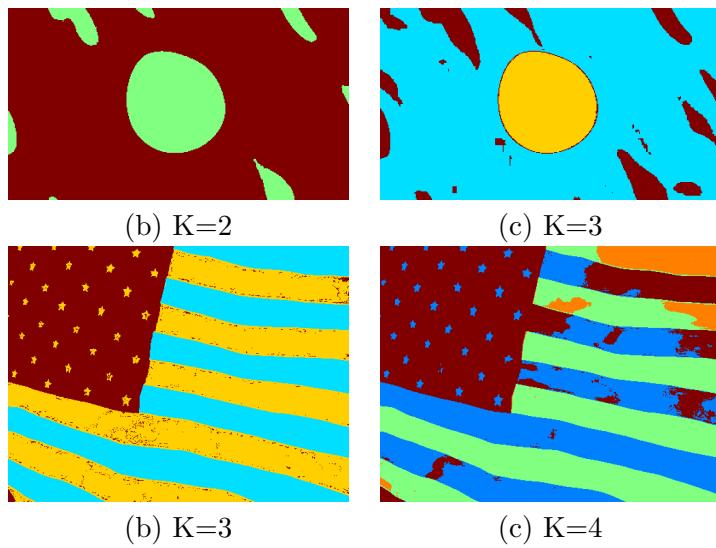


Figura 10 – Resultado da segmentação utilizando GMM. Na bandeira do Japão com o aumento do valor de K foi possível separar as sombras da bandeira. O mesmo ocorreu com a bandeira dos EUA. Para o valor K correto, 2 para a bandeira do Japão e 3 para os EUA, foi possível segmentar bem as imagens para a bandeira dos EUA. Para a bandeira do Japão houve uma confusão entre o centro vermelho e algumas regiões de sombra

Latin America Transactions, IEEE (Revista IEEE America Latina), v. 7, n. 4, p. 487–496, Aug 2009. ISSN 1548-0992. Citado na página 1.

PATTERN Classification. [S.l.]: Wiley, 2001. Citado na página 1.

SINPATCO II: NOVAS ESTRATÉGIAS DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DE PATOLOGIAS DA COLUNA VERTEBRAL. [S.l.]: UNIVERSIDADE FEDERAL DO CEARÁ, 2011. Citado na página 5.

SMITH, J. W. et al. Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Johns Hopkins APL Technical Digest*, v. 10, p. 262–266, 1988. Citado na página 1.