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, (**??**). 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.

*cliftedavid@gmail.com

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.

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:

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

2.2 Metodologia

Na sessão 2.3 são apresentados os resultados obtidos ao aplicar o classificador de bayes utilizando como função de densidade probabilidade a Gaussiana. 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.

Inicialmente 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.

Covariancias distintas

DataSet	média	desvio Padrão	máximo	mínimo
iris	0.973563	0.029599	1.000000	0.896552
Vertebra	0.730108	0.052892	0.854839	0.612903
Dermatologia	0.973059	0.016281	1.000000	0.917808

Matriz de covariacia $\Sigma_i = \sigma^2 I \forall i$

DataSet	média	desvio Padrão	máximo	mínimo
iris	0.568966	0.141670	0.827586	0.241379
Vertebra	0.698925	0.049818	0.790323	0.580645
Dermatologia	0.963927	0.023170	1.000000	0.904110

Matriz de covariacia diagonal

DataSet	média	desvio Padrão	máximo	mínimo
iris	0.955172	0.034064	1.000000	0.896552
Vertebra	0.718280	0.041008	0.806452	0.629032
Dermatologia	0.973973	0.016622	1.000000	0.931507

Tabela 1 – TTTT

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

2.3 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, além disso são exibidos os resultados obtidos para diferentes matrizes de covariancia.

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

2.4 Análise da matriz de covariancia na definição 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 utilizandos os atributos largura da pétala x comprimento da pétala e largura da

Matrizes de covariancias diferentes

	setosa	versicolor	virginica
setosa	1.000000	0.000000	0.000000
versicolor	0.000000	0.922807	0.077193
virginica	0.000000	0.000000	1.000000
	setosa	versicolor	virginica
setosa	1.000000	0.000000	0.000000
versicolor	0.000000	0.921502	0.085185
virginica	0.000000	0.000000	1.000000

	Hernia discal	Espondilolise	Normal
Hernia discal	0.291536	0.077419	0.252046
Espondilolise	0.006270	0.924731	0.111293
Normal	0.090909	0.190323	0.662848

	psoriasis	seboreic	l. planus	pit rosea	c. dermat	pit rubra
psoriasis	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
seboreic	0.001506	0.932551	0.000000	0.070288	0.000000	0.000000
l. planus	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
pit rosea	0.000000	0.105572	0.000000	0.884984	0.000000	0.000000
c. dermat	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
pit rubra	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Matrizes de covariancias iguais

	setosa	versicolor	virginica
setosa	1.000000	0.000000	0.000000
versicolor	0.000000	0.924658	0.077193
virginica	0.000000	0.140411	0.856140
	setosa	versicolor	virginica
setosa	0.803509	0.170886	0.007435
versicolor	0.192982	0.382911	0.520446
virginica	0.010526	0.382911	0.539033

	Hernia discal	Espondilolise	Normal
Hernia discal	0.162319	0.057797	0.394649
Espondilolise	0.008696	0.862595	0.205686
Normal	0.028986	0.147219	0.757525

	psoriasis	seboreic	l. planus	pit rosea	c. dermat	pit rubra
psoriasis	0.992331	0.012887	0.000000	0.000000	0.000000	0.000000
seboreic	0.000000	0.884021	0.000000	0.154639	0.000000	0.000000
l. planus	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
pit rosea	0.000000	0.074742	0.000000	0.900344	0.000000	0.000000
c. dermat	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
pit rubra	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Matrizes de covariancias diferentes e diagonais

	setosa	versicolor	virginica
setosa	1.000000	0.000000	0.000000
versicolor	0.000000	0.920530	0.088561
virginica	0.000000	0.076159	0.915129
	setosa	versicolor	virginica
setosa	1.000000	0.000000	0.000000
versicolor	0.000000	0.921429	0.071197
virginica	0.000000	0.060714	0.944984

	Hernia discal	Espondilolise	Normal
Hernia discal	0.321023	0.077348	0.280265
Espondilolise	0.014205	0.895028	0.149254
Normal	0.099432	0.171271	0.684909

	psoriasis	seboreic	l. planus	pit rosea	c. dermat	pit rubra

sépala x comprimento da pétala.

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

Com matrizes de covariancias distintas temos uma região de decisão quadrática desta forma podemos notar várias elipses ao longo das regiões, figura 1. 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.

2.4.2 Matrizes de covariancia iguais e diagonal. $\Sigma_i = \sigma_i^2 I \forall i$

Ao utilizarmos a mesma matriz de covariancia para todas as classes temos uma região de decisão linear e portanto um maior erro na classificação dos dados. Isso pode ser verificado na terceira coluna da figura 2. Devido a natureza não linear dos dados das três classes uma superfície de decisão linear não consegue separar as classes de forma satisfatória.

2.4.3 Matrizes de covariancia distintas e diagonal.

Os valores da matriz de covariancia definem a forma do elipsóide utilizado na região de decisão. Desta forma uma matriz de covariancia diagonal define um elipsóide em que os pontos mais distantes encontram-se ortogonais aos eixos do espaço. Assim temos na figura 3 várias elipses ortogonais aos eixos do plano carteziano. A região de decisão continua quadrática e apenas limitada quanto a rotação dos dados.

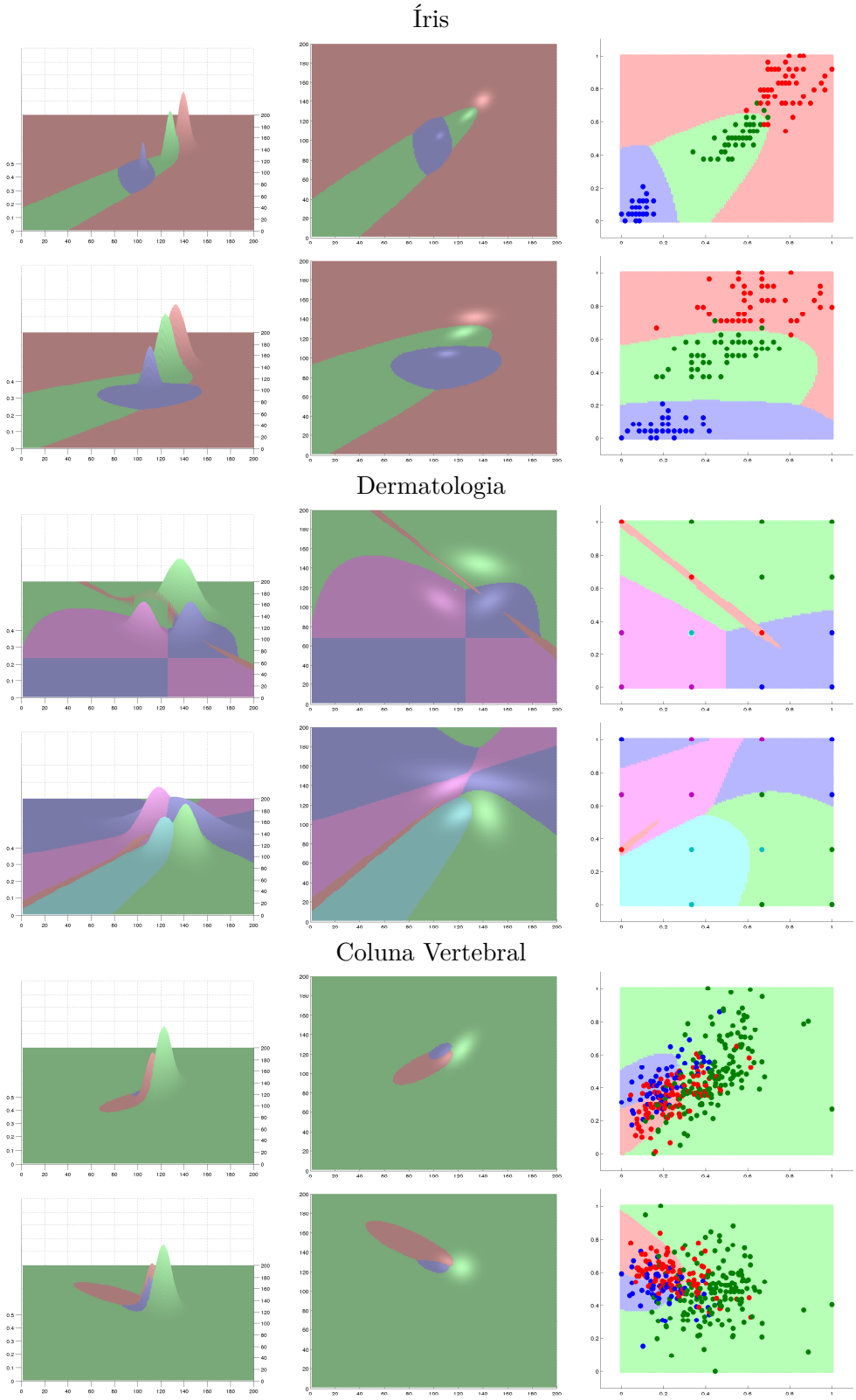
2.4.4 Região de decisão

2.4.5 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.

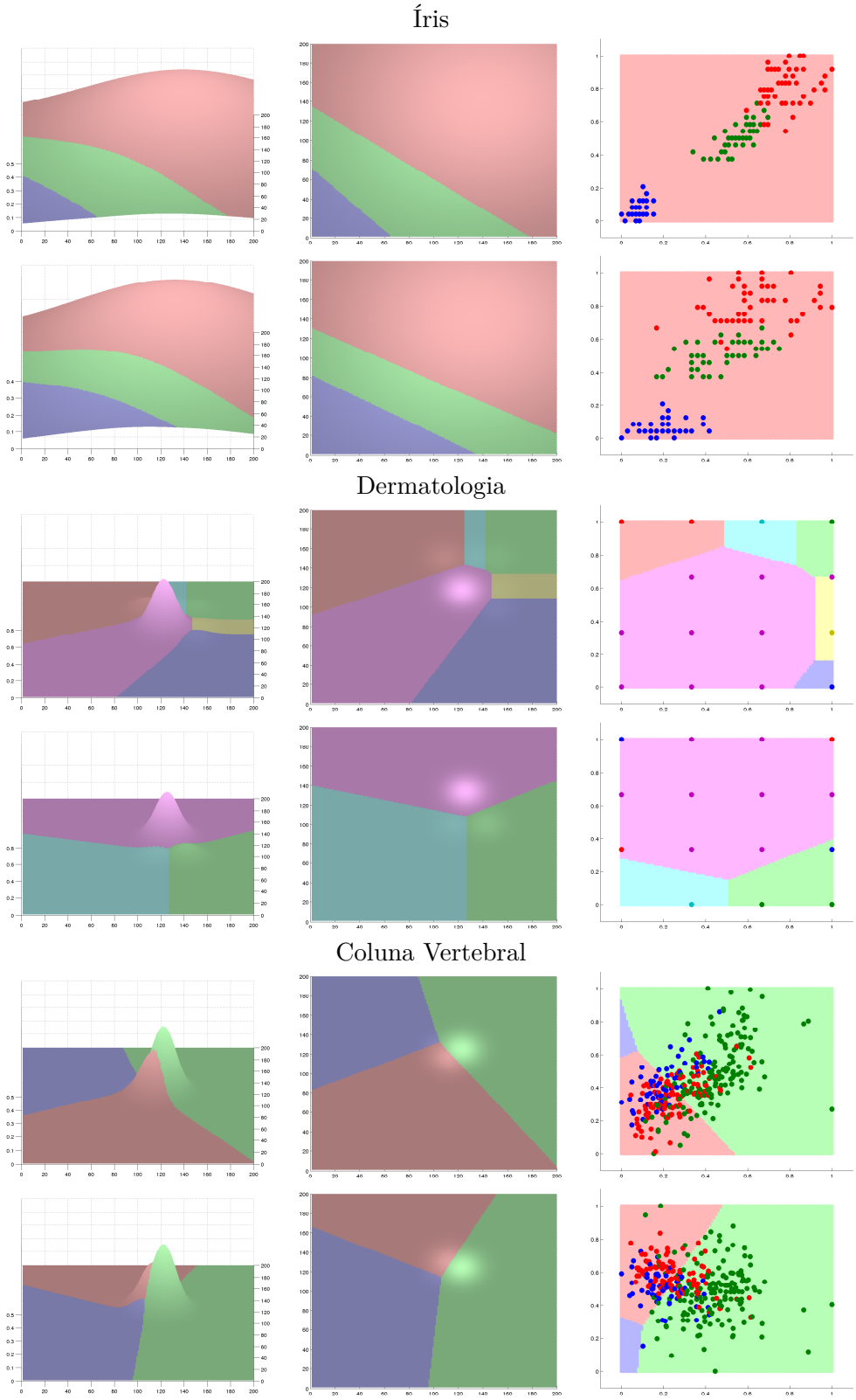
3 Segmentação uitlizando Classificador de Bayes

A segmentação de uma imagem é realizada levando em consideração a intensidade dos pixels nas três componentes RGB. A definição dos dados de 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 priore por isso classes com uma diferença de área muito grande podem impactar na classificação. Assim temos:



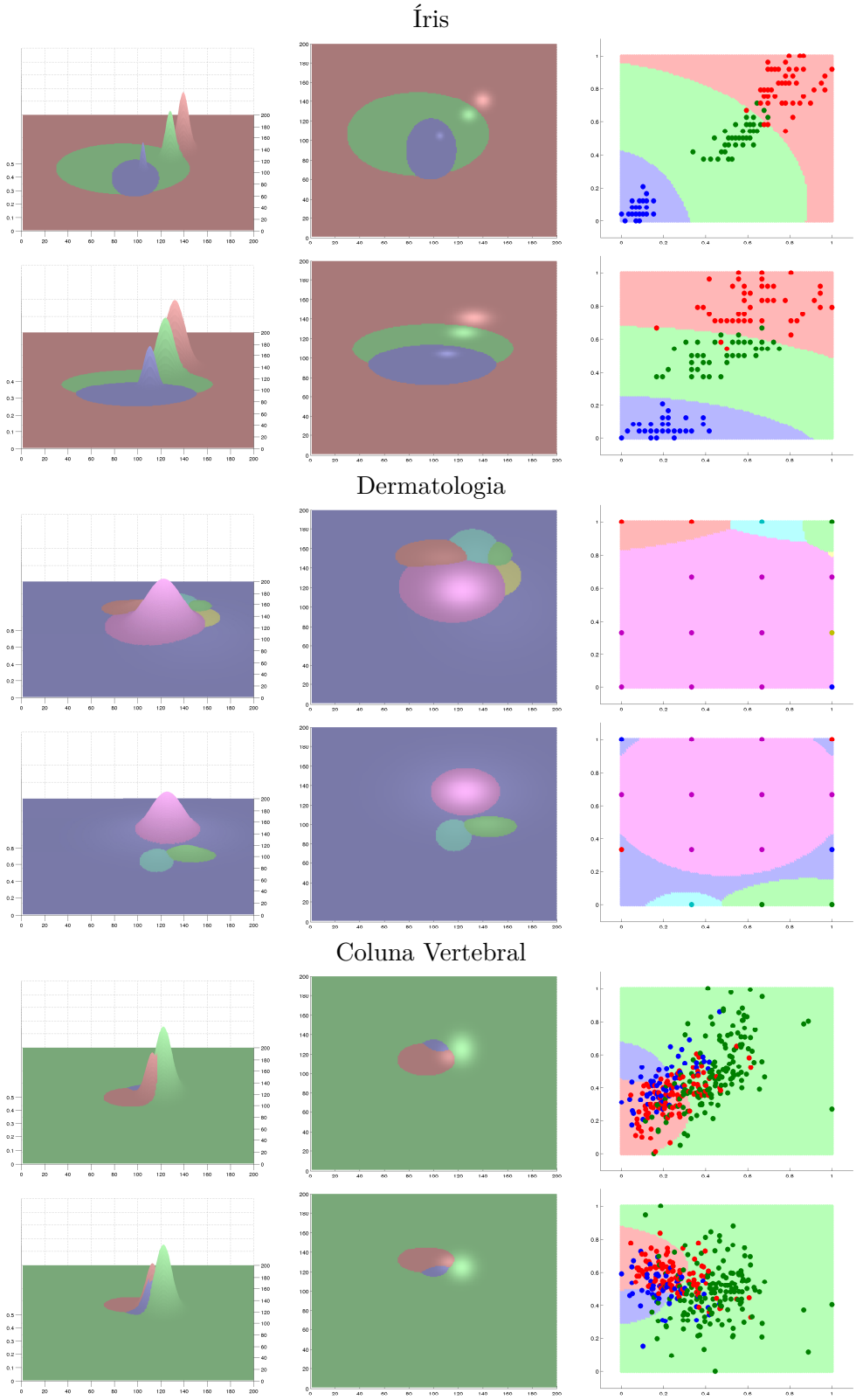
(a) $P(W_i|x)$ (b) Regiao de decisão (c) Resultado da classificação dos dados sobre a regiao de decisão

Figura 1 – Resultados utilizando matrizes de covariancias distintas.



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

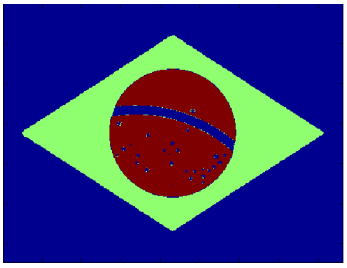
Figura 2 – Região de decisão calculada utilizando a mesma matriz de covariancia para todas as classes.



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

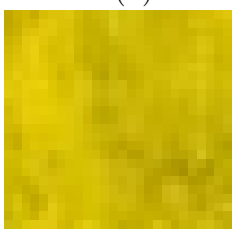
Figura 3 – Região de decisão calculada utilizando matrizes distintas porém diagonais.

(a) Original



(b)Resultado da Segmentação

(c) A_1



(d) A_2

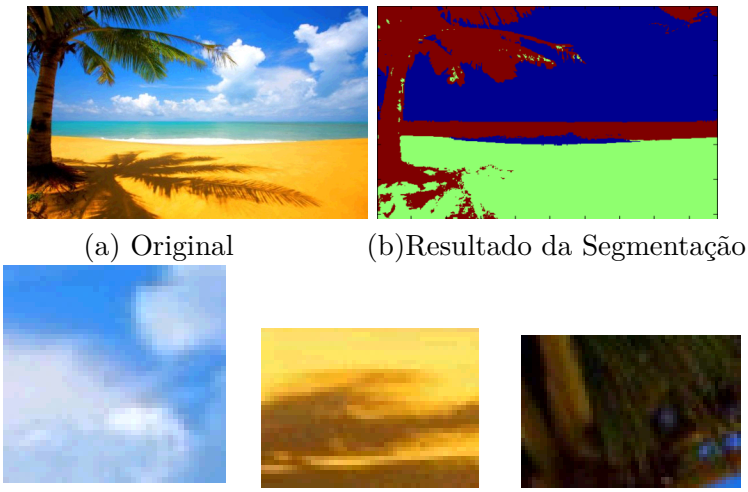
(e) A_3

Figura 4 – Resultado da segmentação utilizando o classificador Bayessiano. Na linha inferior são apresentados as amostras utilizadas para definir as classes. As áreas utilizadas são:(c) Azul, (d) Amarelo e (e) Verde

$$P(w_i) = \frac{A_i}{\sum_{i=1}^{k} A_J} \tag{2}$$

Nas figuras 4 e 5 são apresentado o resultado da segmentação de duas imagens utilizando o classificador Bayessiano. Os parâmetros σ^2 e μ são estimados de acordo com os dados de cada classe individualmente. Assim temos na figuras 4 e 5 o resultado deste classificador sobre 2 imagens distintas.

Referências



(a) Original (b)Resultado da Segmentação

(c) A_1 (d) A_2 (e) A_3

Figura 5 – Resultado da segmentação utilizando o classificador Bayessiano. Na linha inferior são apresentados as amostras utilizadas para definir as classes. As áreas utilizadas são:(c) céu, (d) areia e (e) coqueiro