Relatório de Implementação do Classificador com Mistura de Gaussianas

Angélica Viana

June 2019

1 Introdução

Este relatório visa descrever a implementação do classificador bayesiano com função de verosimilhança baseada em mistura de gaussiana. Após a implementação do modelo de misturas de gaussianas, foram realizadas duas etapas, a primeira consistiu na introdução desse modelo no cálculo da função de verossimilhança no classificador bayesiano e testes posteriores com as bases de dados da Iris, Coluna, Artificial I, Breast Canser e Dermaatology. Já a segunda etapa consistiu na utilização desse modelo de misturas de gaussianas para a segmentação de cores das bandeiras do Brasil, Japáo e EUA.

2 Misturas de Gaussianas

Este modelo é não supervisionado e consiste em um método iterativo para encontrar aproximações para os parâmetros correspondentes as gaussianas utilizadas na mistura, são eles μ, π e Σ, que correspondem respectivamente as medias de cada grupo, as ponderações de cada função e as matrizes de covariância de cada uma. Este algoritmo consiste basicamente em duas etapas que serão descritas nas subseções a seguir.

2.1 E-Step

É possível descrever a distribuição de mistura de gaussianas como uma combinação de gaussianas com pesos iguais a π como a seguir.

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

Onde K é o número de gaussianas que queremos modelar.

Tomando a equação acima, podemos calcular a distribuição posterior das responsabilidades que cada gaussiana tem para cada ponto de dados usando a

fórmula abaixo.

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j, \Sigma_j)}$$

Essa equação é apenas a regra de Bayes, onde π é o peso anterior e a probabilidade é normal.

2.2 M-Step

Depois de calcular todas as probabilidades a posteriori, é preciso conseguir uma estimativa dos parâmetros de cada gaussiana definida pelas equações abaixo e então avaliar a verossimilhança. Estes dois passos são então repetidos até a convergência.

- **Média de cada grupo:**

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

- **Covariância de cada grupo:**

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T$$

- **Pesos de cada grupo:**

$$\pi_k^{new} = \frac{N_k}{N}$$

  Onde N_k é a soma das responsabilidades em cada gaussiana k, dada por:

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

3 Implementação: Parte I

A primeira etapa consistiu da implementação do classificador bayesiano com função densidade de probabilidade obtida com base em modelos de misturas. Seja $P(\omega_i|x)$ a probabilidade a posteriori de um parão x pertencer a classe ω_i. Neste classificador, essa probabilidade é dada por:

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{\sum_{j=1^N} P(x|\omega_j)P(\omega_j)}$$

Onde,

$$P(x|\omega_i) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

Em que $N(x|\mu_k, \Sigma_k)$ é a distribuição normal para o grupo k, como foi visto na seção anterior.

Com a implementação desse classificador, os testes a seguir foram realizados com as 5 bases já mencionadas anteriormente. Os resultados de acurácia média (A.M.), desvio padrão (D.P.) e tempo médio (T.M.) obtidos após 20 realizações de treinamento e teste (Divisão fixa de 80% para treinamento e 20% para teste) para cada base são apresentados a seguir, além disso são apresentados os resultados dos algoritmos anteriormente implementados para essas mesmas bases para efeito de comparação. Foi realizado um grid search para encontrar o valor de k para cada base de dados referente ao número de agrupamentos. Foram testados 3 valores de k, k = 2, k = 3 e k = 4.

- **Base de Dados da Íris:** Contém 150 padrões, 4 atributos e 3 classes.

Table S1: Tabela contendo as métricas de avaliação para os dez algoritmos.

Classificador	Treinamento			Teste		
	A.M.	D.P.	T.M. (s)	A.M.	D.P.	T.M. (s)
Bayesiano GMM	98,29	0,61	0,5945	96,83	2,46	0,1483
Bayesiano Parzen	98,41	0,91	0,0004	95,33	3,23	0,3014
Bayesiano Puro	98,25	0,69	0,0015	97,5	2,07	0,0127
Disc. Quadrático	98,21	0.71	0,0029	97,33	3,26	0,0147
Disc. Linear 1	86,54	1,56	0,0020	83,67	7.22	0,0078
Disc. Linear 2	92,67	1,66	0,0063	91,16	6.43	0,0047
Disc. Linear 3	86,46	1,64	0,0031	85,33	5.31	0,0093
Disc. Linear 4	93,46	1,06	0,0096	92,67	3,59	0,0038
Disc. Linear 5	86,75	1,23	0,0056	85,50	5,60	0,0101
KNN	97,33	0,77	0,0000	97,33	3,05	0,0974
DMC	92,59	3,97	0,0064	92,59	3,97	0,0064

Fonte: Autoria Própria.

A realização cujo valor de acurácia mais se aproximou da acurácia média de teste para o classificador Bayesiano GMM é representada a partir de sua matriz de confusão abaixo:

	1	2	3
1	7	0	0
2	0	11	0
3	0	1	11

(a) Bayesiano GMM: Iris

Abaixo vê-se a superfície de decisão do classificador Bayesiano GMM para os dois primeiros atributos da base da Iris.

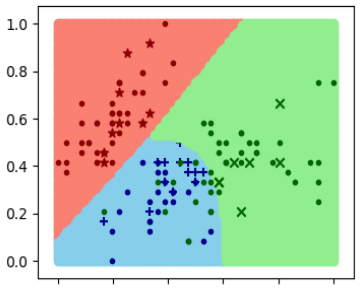

Figure 1: Superfície de decisão - Iris (2 primeiros atributos)

Para a base da iris, a variação no valor de k de 2 a 4 não surtiu muito feito nos resultados de acurácia, resultando nos mesmos valores de acurácia no treinamento.

- Base da Coluna vertebral: 310 padrões, 6 atributos e 3 classes.

Table S2: Tabela contendo as métricas de avaliação para os dez algoritmos.

Classificador	Treinamento			Teste		
	A.M.	D.P.	T.M. (s)	A.M.	D.P.	T.M. (s)
Bayesiano GMM	84,41	1,22	1,9270	81,72	4,31	0,2685
Bayesiano Parzen	83,02	1,12	0,0002	77,95	5,85	1,2792
Bayesiano Puro	86,88	1,12	0,0039	83,52	4,03	0,0257
Disc. Quadrático	87,56	1,04	0,0009	82,95	3,57	0,0322
Disc. Linear 1	83,06	1,41	0,0070	80,90	5,68	0,0195
Disc. Linear 2	69,70	1,11	0,0056	70,49	5,61	0,0063
Disc. Linear 3	79,86	2,18	0,0055	78,03	6,27	0,0165
Disc. Linear 4	75,02	1,47	0,0735	72,05	5,06	0,0053
Disc. Linear 5	79,41	1,80	0,0039	76,56	5,05	0,0187
KNN	78,90	1,10	0,0000	78,90	4,37	0,4690
DMC	74,18	4,91	0,0057	74,18	4,91	0,0057

Fonte: Autoria Própria.

A realização cujo valor de acurácia mais se aproximou da acurácia média de teste para o classificador Bayesiano GMM utilizando a base de dados da Coluna é representada a partir de sua matriz de confusão abaixo:

	1	2	3
1	6	0	2
2	0	22	3
3	5	1	22

(a) Bayesiano GMM: Coluna

Abaixo vê-se a superfície de decisão do classificador Bayesiano GMM para os dois primeiros atributos da base da Coluna Vertebral.

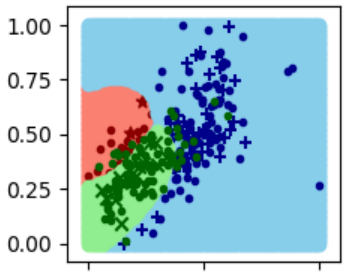

Figure 2: Superfície de decisão - Coluna (2 primeiros atributos)

Para a base de dados da Coluna Vertebral os valores de k que produziram os melhores resultados dentre as 20 realizações foram 3 e 4.

- Base Artificial I: Essa base foi criada a partir da geração de dados aleatoriamente ao redor dos pontos (0,25;0,75), (0,5; 0,5) e (0,75; 0,75). Totalizando 40 padrões (30 da classe 0 e 10 da classe 1).

Table S3: Tabela contendo as métricas de avaliação para os oito algoritmos.

Classificador	Treinamento			Teste		
	A.M.	D.P.	T.M. (s)	A.M.	D.P.	T.M. (s)
Bayesiano GMM	100,00	0,0	0,1920	100,00	0,0	0,2180
Bayesiano Parzen	100,00	0,0	0,0005	100,00	0,0	0,0250
Bayesiano Puro	100,00	0,0	0,0007	100,00	0,0	0,0118
Disc. Quadrático	100,00	0,0	0,0	100.00	0,0	0,0102
Disc. Linear 1	100,00	0,0	0,0015	100.00	0,0	0,0086
Disc. Linear 2	100,00	0,0	0,0024	100,00	0,0	0,0038
Disc. Linear 3	100,00	0,0	0,0023	100,00	0,0	0,0070
Disc. Linear 4	100,00	0,0	0,0032	100,00	0,0	0,0019
Disc. Linear 5	100,00	0,0	0,0022	100,00	0,0	0,0074

Fonte: Autoria Própria.

A realização cujo valor de acurácia mais se aproximou da acurácia média de teste para o classificador Bayesiano-Parzen utilizando a base de dados Artificial I é representada a partir de sua matriz de confusão abaixo:

	1	2
1	6	0
2	0	2

(a) Bayesiano GMM: Artificial

Abaixo vê-se a superfície de decisão do classificador Bayesiano GMM para os dois atributos da base Artificial I.

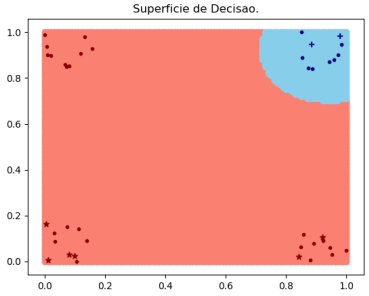

Figure 3: Superfície de decisão - Artificial I

Para a base de dados Artificial I, os valores de k não influenciaram nos resultados de acurácia e desvio padrão, pois independente dos mesmos, devido a fácil separabilidade dos dados, todas as acurácias deram 100% nos 20 testes e 0,0 de desvio padrão.

- Base Dermatology: 365 padrões, 34 atributos e 6 classes.

Table S4: Tabela contendo as métricas de avaliação para os oito algoritmos.

Classificador	Treinamento			Teste		
	A.M.	D.P.	T.M. (s)	A.M.	D.P.	T.M. (s)
Bayesiano GMM	93,71	2,72	1,7411	93,59	3,68	2,3007
Bayesiano Parzen	100,00	0,00	0,0005	96,50	2,25	1,3327
Bayesiano Puro	97,88	0,37	0,0064	91,27	2,77	0,1543
Disc. Quadrático	98,29	0,40	0,0117	92,68	3,47	0,1440
Disc. Linear 1	97,81	0.50	0,0146	96,69	2,28	0,1051
Disc. Linear 2	97,67	0,52	0,0116	97,04	1,66	0,0104
Disc. Linear 3	97,33	0,53	0,0085	96,55	1,63	0,0884
Disc. Linear 4	97,17	0,44	0,0138	95,98	1,68	0,0076
Disc. Linear 5	97,41	0,39	0,0012	96,90	1,81	0,1050

Fonte: Autoria Própria.

A realização cujo valor de acurácia mais se aproximou da acurácia média de teste para o classificador Bayesiano GMM utilizando a base de dados Dermatology é representada a partir de sua matriz de confusão abaixo:

	1	2	3	4	5	6
1	24	0	0	0	0	0
2	0	5	0	3	0	0
3	0	0	18	0	0	0
4	0	0	0	10	0	0
5	0	0	0	0	9	0
6	1	0	0	0	1	0

(a) Bayesiano GMM: Dermatology

Para a base de dados dermatology, o melhor valor de k encontrado foi k = 4.

- Base Breast Canser: 698 padrões, 10 atributos e 2 classes.

Table S5: Tabela contendo as métricas de avaliação para os sete algoritmos.

Classificador	Treinamento			Teste		
	A.M.	D.P.	T.M. (s)	A.M.	D.P.	T.M. (s)
Bayesiano GMM	95,89	0,61	0,8549	95,00	2,05	0,4881
Bayesiano Parzen	100,00	0,00	0,0004	95,51	0,89	5,42
Bayesiano Puro	95,75	0,45	0,0009	95,00	1,95	0,0404
Disc. Quadrático	95,50	0,41	0,0033	95,37	1,56	0,0424
Disc. Linear 1	94,55	0,50	0,0041	94,41	1,89	0,0247
Disc. Linear 2	96,11	0,43	0,0036	95,91	1,70	0,0097
Disc. Linear 3	96,52	0,36	0,0024	95,91	1,30	0,0234
Disc. Linear 4	96,30	0,29	0,0033	97,20	1,15	0,0054
Disc. Linear 5	96,45	0,42	0,0032	95,95	1,50	0,0283

Fonte: Autoria Própria.

A realização cujo valor de acurácia mais se aproximou da acurácia média de teste para o classificador Bayesiano Parzen utilizando a base de dados Breast Cancer é representada a partir de sua matriz de confusão abaixo:

	1	2
1	82	4
2	0	50

(b) Bayesiano GMM: Breast Cancer

Para a base de dados breast canser o melhor valor de k encontrado foi o de k = 4.

4 Implementação: Parte II

A segunda parte deste trabalho consistiu na utilização do modelo de mistura de gaussianas para a segmentação das bandeiras do Brasil, do japão e dos EUA. Para isso, as imagens foram convertidas em uma base de dados tridimensional, em que cada linha corresponde a um pixel da imagem e cada coluna o seu valor correspondente no canal R, G e B respectivamente.

Para cada bandeira foi selecionada uma cor a ser segmentada.

- Bandeira do Brasil: A cor azul.
- Bandeira do Japão: A cor vermelha.
- Bandeira dos EUA: A cor branca.

Os resultados de segmentação dessas cores podem ser visualizados a partir das imagens abaixo.

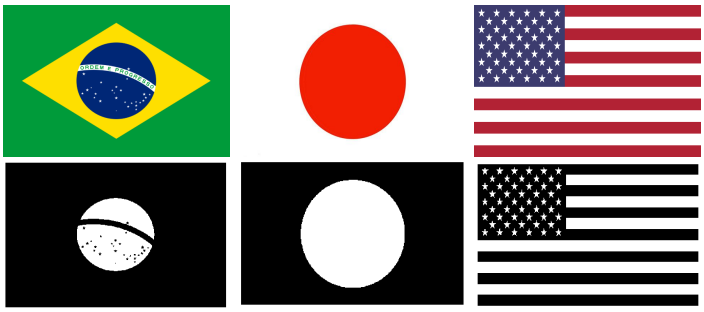

Figure 4: Cores segmentadas das bandeiras do Brasil, Japão e EUA.

Foram definidas para cada uma das segmentações um total de 100 iterações até a convergência e o número de agrupamentos passado para o modelo de misturas de gaussianas dependeu do número de cores contidas nas imagens das bandeiras. Para a bandeira do brasil o k = 4 (verde, amarela, azul e branco), para a do japão k = 2 (vermelha e branco) e para os EUA k = 3 (vermelho, azul e branco). As médias obtidas pelo classificador não supervisionado de cada das cores a serem segmentadas foram de 71,00, 209,67 e 82,01. Com essas médias, os valores dos pixels da imagem convertidas para tons de cinza foram comparados com essas médias e assim a cor de interesse é apresentada como branca após a segmentação e as demais como preta.

5 Conclusão

A partir de uma breve análise dos resultados obtidos, podemos observar que os resultados do classificador bayesiano gaussiano com pdf baseada em mistura de gaussianas aplicado as bases da iris, coluna, artificial, dermatology e breast canser foram relativamente bons e se equiparam aos obtidos com os demais classificadores, aproximando-se mais dos valores de acurácia e desvio padrão obtidos com o classificador bayesiano gaussiano puro e do baseado em discriminante quadrático. No entanto, o mesmo apresenta uma certa desvantagem quando comparado aos demais classificadores quanto a sua perfomance de tempo, pois o custo computacional para calcular os grupos de gaussianas e realizar o grid search para encontrar o melhor k é elevado. Podemos observar ainda que o modelo de mistura de gaussianas pode ser uma ótima alternativa para aplicações em segmentação de imagens, pois os resultados obtidos nos experimentos de segmentação de cores de bandeiras desse trabalho foram altamente relevantes.