Relatório de Implementação de Classificadores Bayesianos Gaussianos

Angélica Viana

May 2019

1 Introdução

Este relatório visa descrever a implementação de dois classificadores bayesianos, o primeiro em sua forma pura (Naive) utilizando os cálculos das probabilidades a posteriori de cada classe na regra de decisão e a segunda abordagem utilizando discriminantes quadráticos e lineares. Após a implementação, esses algoritmos foram testados nas seguintes bases de dados: Íris, Coluna Vertebral, Artificial I, Dermatologia e Câncer de Mama.

2 Classificador Bayesiano Gaussiano

2.1 Naive

A primeira etapa desse trabalho consistiu na implementação do classificador bayesiano Gaussiano puro, cujo treinamento consiste em encontrar os parâmetros: $\boldsymbol{\mu_i}$ e Σ_i, em que

$$\boldsymbol{\mu} = \sum_{k=1}^{N_i} \frac{\boldsymbol{x_k}}{N_i}$$

Onde $\boldsymbol{\mu_i}$ representa a média dos padrões da classe i e N_i o número total de padrões da classe i.

E

$$\Sigma_i = E[(\boldsymbol{x} - \boldsymbol{\mu_i})(\boldsymbol{x} - \boldsymbol{\mu_i})^T]$$

Onde Σ_i denota a matriz de covariância dos dados da classe i e E é o símbolo de esperança.

Além disso, ainda na etapa de treinamento são computadas as probabilidades a priori $P(w_i)$ de cada classe, que é dada por

$$P(\omega_i) \approx \frac{N_i}{N}$$

Em que N_i novamente denota o número de elementos da classe i e N o número total de elementos do conjunto de treinamento.

A regra de decisão do classificador é dada pelo comparativo entre as probabilidades a posteriori, que é obtida a partir da regra de bayes

$$P(\omega_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_i)P(\omega_i)}{P(\boldsymbol{x})}$$

Onde a Função Densidade de Probabilidade Gaussiana é dada por

$$P(\boldsymbol{x}) = \frac{1}{2\pi^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(x-\boldsymbol{\mu})\right)$$

E l é um parâmetro representando a dimensionalidade dos dados.

2.2 Baseado em Discriminantes

A segunda etapa desse trabalho foi a implementação do classificador bayesiano gaussiano baseado em discriminantes quadráticos e lineares.

Primeiramente, para a implementação baseada em um discriminante quadrático, no treinamento foram calculados os mesmos parâmetros encontrados no treinamento do classificador puro, ou seja, $\boldsymbol{\mu}$, Σ_i e $P(w_i)$.

A regra de decisão do classificador foi dada pela fórmula abaixo:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_i})^T{\Sigma_i}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_i}) + lnP(\omega_i) + c_i$$

Onde c_i é uma constante igual a $-\frac{l}{2}ln2\pi - \frac{1}{2}ln|\Sigma_i|$

já as regras de decisão dos classificadores bayesianos gaussianos baseados em discriminantes lineares foram determinadas conforme algumas suposições:

- **Disc. Linear 1: Matrizes de covariância iguais para todas as classes:** Fazendo-se essa suposição, o discriminante linear pode ser obtido por

  $$g_i(\boldsymbol{x}) = {\boldsymbol{w_i}}^T x + \boldsymbol{w_{i0}}$$

  Onde $\boldsymbol{w_i} = \Sigma^{-1}\boldsymbol{\mu_i}$ e $\boldsymbol{w_{i0}} = lnP(\omega_i) - \frac{1}{2}{\boldsymbol{\mu_i}}^T\Sigma^{-1}\boldsymbol{\mu_i}$

- **Disc. Linear 2 - Matrizes de covariância Diagonal com elementos iguais:** Fazendo-se essa suposição, o discriminante linear pode ser obtido pela fórmula ainda mais simplificada

  $$g_i(\boldsymbol{x}) = \frac{1}{\sigma^2}{\boldsymbol{\mu_i}}^T\boldsymbol{x} + \boldsymbol{w_{i0}}$$

  Onde $\boldsymbol{w_{i0}} = lnP(\omega_i) - \frac{1}{2\sigma^2}{\boldsymbol{\mu_i}}^T\boldsymbol{\mu_i}$

- **Disc. Linear 3 - Matrizes de covariância iguais e classes equiprováveis:** Sob essas condições o compto do discriminante pode ser simplificado para

  $$g_i(\boldsymbol{x}) = \frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu_i})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu_i})$$

- **Disc. Linear 4 - Matriz de covariância equivalente a $\sigma^2 I$:** Sob as condições de classes equiprováveis e matriz de covariância iguais, neste caso, maximizar $g_i(\boldsymbol{x})$ é o mesmo que minimizar a distância euclidiana dada por

$$d_e = ||\boldsymbol{x} - \boldsymbol{\mu_i}||$$

- **Disc. Linear 5 - Matriz de covariância não-diagonal:** Sob as condições de classes equiprováveis e matriz de covariância iguais, para este, maximizar $g_i(\boldsymbol{x})$ equivale a minimizar a distância de Mahalanobis dada por

$$d_m = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu_i})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i})}$$

Na seção seguinte serão apresentados os resultados de alguns testes feitos com as bases de dados anteriormente citadas utilizando os classificadores aqui descritos.

3 Testes

Como descrito na seção anterior, foram implementados três classificadores, sendo eles o classificador bayesiano gaussiano em sua versão pura, o classificador bayesiano com discriminante quadrático e com discriminantes lineares. Para cada classificador, foram testadas cinco base de dados. 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 algoritmo referente as bases de dados da Íris, Coluna, Artificial, Dermatologia e Câncer respectivamente analisados a partir das tabelas apresentadas. Além disso, também são mostrados os resultados dos classificadores KNN e DMC obtidos no relatório anterior para efeitos de comparação.

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

As realizações cujos valores de acurácia nos testes mais se aproximaram da média das acurácias para cada classificador são representadas por suas matrizes de confusão.

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

Classificador	Treinamento			Teste		
	A.M.	D.P.	T.M. (s)	A.M.	D.P.	T.M. (s)
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.

	1	2	3
1	8	0	0
2	0	5	0
3	0	1	16

(a) Bayesiano Puro

	1	2	3
1	13	0	0
2	0	11	0
3	0	1	5

(b) Disc. Quadrático

	1	2	3
1	9	0	0
2	0	7	2
3	0	3	9

(c) Disc. Linear 1

	1	2	3
1	6	0	0
2	0	14	0
3	0	3	7

(d) Disc. Linear 2

	1	2	3
1	9	0	0
2	0	5	0
3	0	4	1

(e) Disc. Linear 3

	1	2	3
1	6	0	0
2	0	14	1
3	0	1	8

(f) Disc. Linear 4

	1	2	3
1	13	0	0
2	0	7	1
3	0	3	6

(g) Disc. Linear 5

	1	2	3
1	11	0	0
2	0	10	1
3	0	0	8

(h) KNN

	1	2	3
1	10	0	0
2	0	12	1
3	0	1	6

(i) DMC

Figure 1: Matrizes de confusão para a base da Iris.

As imagens abaixo representam as superfícies de decisão dos dois primeiros atributos da base da iris para o classificador bayesiano gaussiano puro, baseado em discriminante quadrático, baseado em discriminante linear 1, baseado em discriminante linear 2 e baseado em discriminante linear 3 respectivamente.

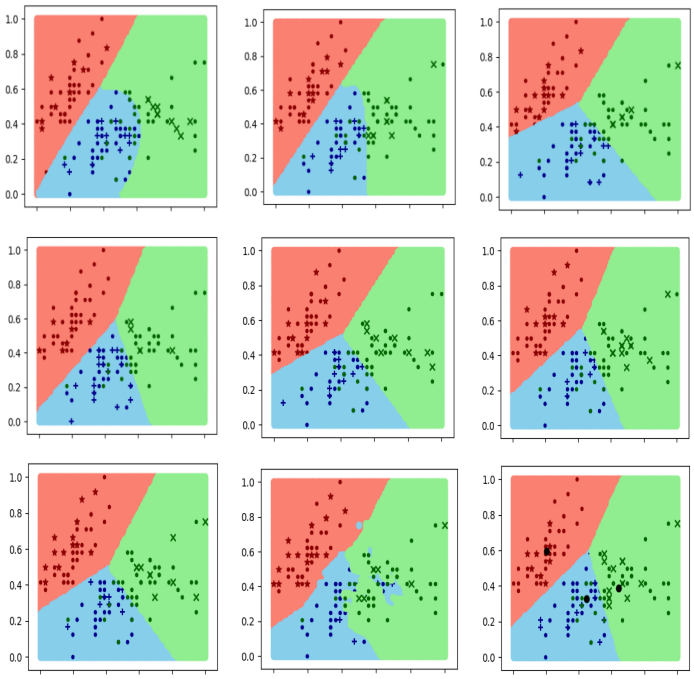

Figure 2: Superfícies de decisão dos classificadores para a base da iris (atributos 1 e 2).

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

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

Classificador	Treinamento			Teste		
	A.M.	D.P.	T.M. (s)	A.M.	D.P.	T.M. (s)
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.

As realizações cujos valores de acurácia nos testes mais se aproximaram da

média das acurácias para cada classificador são representadas por suas matrizes de confusão.

	1	2	3
1	4	1	7
2	0	21	7
3	0	3	18

(a) Bayesiano Puro

	1	2	3
1	5	1	4
2	0	38	0
3	3	2	8

(b) Disc. Quadrático

	1	2	3
1	5	0	4
2	1	30	4
3	1	2	14

(c) Disc. Linear 1

	1	2	3
1	8	0	0
2	0	5	0
3	0	1	16

(d) Disc. Linear 2

	1	2	3
1	10	0	2
2	4	27	1
3	5	1	11

(e) Disc. Linear 3

	1	2	3
1	12	2	0
2	1	19	10
3	2	2	13

(f) Disc. Linear 4

	1	2	3
1	8	0	1
2	7	30	2
3	4	0	9

(g) Disc. Linear 5

	1	2	3
1	8	0	3
2	7	27	4
3	3	3	13

(h) KNN

	1	2	3
1	9	1	2
2	1	18	3
3	6	3	18

(i) DMC

Figure 3: Matrizes de confusao para a base da coluna.

As imagens abaixo representam as superfícies de decisão dos dois primeiros atributos da base da coluna para o classificador bayesiano gaussiano puro, baseado em discriminante quadrático, baseado em discriminante linear 1, baseado em discriminante linear 2 e baseado em discriminante linear 3 respectivamente.

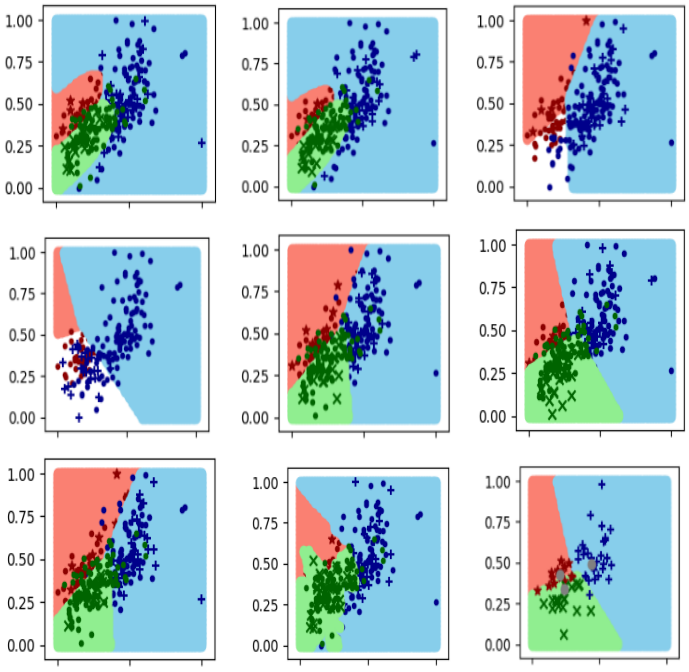

Figure 4: Superfícies de decisão dos classificadores para a base da coluna (atributos 1 e 2).

- 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 150 dados (50 para cada classe).

Table S3: 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 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.

As realizações cujos valores de acurácia nos testes mais se aproximaram da

média das acurácias para cada classificador são representadas por suas matrizes de confusão.

	1	2	3
1	11	0	0
2	0	9	0
3	0	0	10

(a) Bayesiano Puro

	1	2	3
1	11	0	0
2	0	10	0
3	0	0	9

(b) Disc. Quadrático

	1	2	3
1	9	0	0
2	0	13	0
3	0	0	8

(c) Disc. Linear 1

	1	2	3
1	8	0	0
2	0	12	0
3	0	0	10

(d) Disc. Linear 2

	1	2	3
1	11	0	0
2	0	12	0
3	0	0	7

(e) Disc. Linear 3

	1	2	3
1	13	0	0
2	0	10	0
3	0	0	7

(f) Disc. Linear 4

	1	2	3
1	14	0	0
2	0	9	0
3	0	0	7

(g) Disc. Linear 5

Figure 5: Matrizes de confusão para a base artificial1.

As imagens abaixo representam as superfícies de decisão da base artificial1 para o classificador bayesiano gaussiano puro, baseado em discriminante quadrático, baseado em discriminante linear 1, baseado em discriminante linear 2 e baseado em discriminante linear 3 respectivamente.

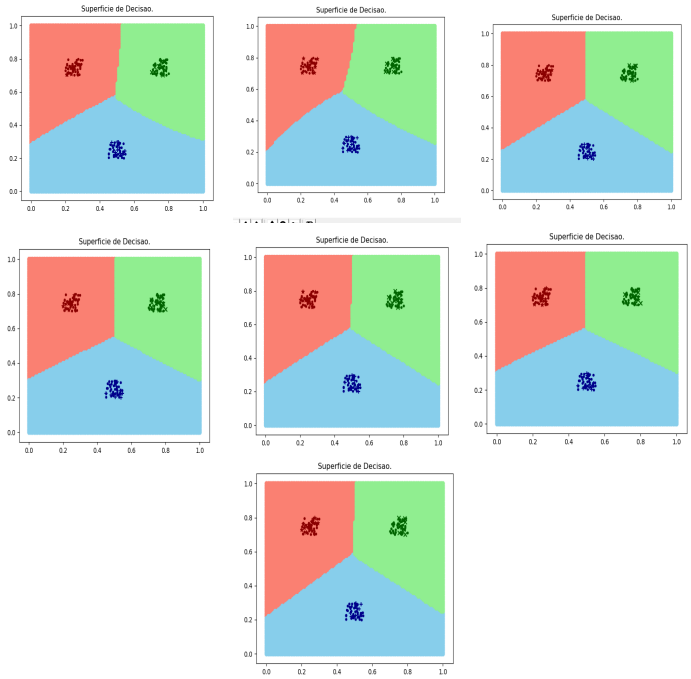

Figure 6: Superfícies de decisão dos classificadores para a base artificial1.

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

Table S4: 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 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.

As realizações cujos valores de acurácia nos testes mais se aproximaram da média das acurácias para cada classificador são representadas por suas matrizes de confusão.

	1	2	3	4	5	6
1	20	0	0	0	0	0
2	0	7	0	1	0	0
3	0	0	17	0	0	0
4	0	0	0	11	0	0
5	0	0	0	0	10	0
6	4	1	0	0	0	0

(a) Bayesiano Puro

	1	2	3	4	5	6
1	24	0	0	0	0	0
2	0	9	0	2	0	0
3	0	0	11	0	0	0
4	0	1	0	11	0	0
5	0	0	0	0	8	0
6	0	0	0	0	0	3

(b) Disc. Quadrático

	1	2	3	4	5	6
1	24	0	0	0	0	0
2	0	8	0	2	0	0
3	0	0	16	0	0	0
4	0	0	0	7	0	0
5	0	0	0	0	9	0
6	0	0	0	0	0	5

(c) Disc. Linear 1

	1	2	3	4	5	6
1	16	0	0	0	0	0
2	0	8	0	1	0	0
3	0	0	16	0	0	0
4	0	1	0	10	0	0
5	0	0	0	0	13	0
6	0	0	0	0	0	6

(d) Disc. Linear 2

	1	2	3	4	5	6
1	20	0	0	0	0	0
2	0	11	0	1	0	0
3	0	0	12	0	0	0
4	0	1	0	10	0	0
5	0	0	0	1	11	0
6	0	0	0	0	0	5

(e) Disc. Linear 3

	1	2	3	4	5	6
1	26	0	0	0	0	0
2	0	2	0	3	0	0
3	0	0	17	0	0	0
4	0	0	0	5	0	0
5	0	0	0	0	12	0
6	0	0	0	0	0	6

(f) Disc. Linear 4

	1	2	3	4	5	6
1	23	0	0	0	0	0
2	0	7	0	2	0	0
3	0	0	13	0	0	0
4	0	0	0	13	0	0
5	0	0	0	0	12	0
6	0	0	0	0	0	1s

(g) Disc. Linear 5

Figure 7: matrizes de confusão para a base dermatology.

- 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 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.

As realizações cujos valores de acurácia nos testes mais se aproximaram da média das acurácias para cada classificador são representadas por suas matrizes de confusão.

	1	2
1	78	4
2	3	51

(a) Bayesiano Puro

	1	2
1	78	4
2	2	52

(b) Disc. Quadrático

	1	2
1	80	2
2	6	48

(c) Disc. Linear 1

	1	2
1	82	1
2	5	48

(d) Disc. Linear 2

	1	2
1	82	4
2	2	48

(e) Disc. Linear 3

	1	2
1	76	2
2	2	56

(f) Disc. Linear 4

	1	2
1	86	2
2	4	44

(g) Disc. Linear 5

Figure 8: Matrizes de confusão para a base Breast Canser.

4 Conclusão

A partir de uma breve análise dos resultados obtidos nos testes que os valores obtidos com os classificadores lineares se assemelham com os obtidos com o DMC, em especial àquele cuja suposição de igualdade das matrizes de covariância é feita, bem como a suposição de que a mesma é diagonal, pois esses são equivalentes. Além disso, as performances dos classificadores bayesianos gaussianos baseados em discriminantes quadráticos e do próprio classificador gaussiano puro podem ser comparadas as do classificador KNN, que obtiveram tanto no treinamento como no teste médias de 97,33% de acurácia para a base da Iris, valores esses bem próximos aos obtidos com a mesma base utilizando o classificador puro e o quadrático. No entanto, para a base da coluna vertebral os resultados desses classificadores foram superiores aos do KNN. Outro fator interessante a se observar é que as bases de dados tanto do câncer de mama como a de dermatologia obtiveram elevados valores de acurácia e baixo desvio

padrão para todos os classificadores, o que nos permite inferir que as classes pertencentes a essas bases são fáceis de serem separadas.

5 Referências

THEODORIDIS, S; KOUTROUMBAS, K. **Pattern Recognition**. Academic Press, USA, 4th edition, 2009.