Relatório de implementação de um classificador bayesiano com base em Janela de Parzen

Angélica Viana

Junho 2019

1 Introdução

Este relatório visa descrever a implementação de um classificador bayesiano com função densidade de probabilidade obtida com base em janela de parzen. Após a implementação do algoritmo foram testadas as seguintes bases de dados: Iris, Coluna Vertebral, Breast Cancer, Dermatology e Artificial I. Na sequência, os resultados foram comparados com os demais classificadores implementados anteriormente.

2 Classificador com Janela de Parzen

O classificador bayesiano com função densidade de probabilidade obtida com base em janela de parzen pode ser implementado a partir da substituição da função densidade de probabilidade gaussiana do classificador naive pela seguinte função:

$$p(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(2\pi)^{\frac{l}{2}}h^l}exp\left(-\frac{(x-x_i)^T(x-x_i)}{2h^2}\right) \quad (1)$$

Ou seja, a pdf não conhecida é aproximada como uma média de N gaussianas, cada uma centrada em um ponto diferente do conjunto de treinamento. À medida que o parâmetro h se torna menor, a forma das Gaussianas se torna mais estreita e a influência de cada Gaussiana individual é mais localizada no espaço de característica ao redor da área de seu valor médio.

Neste trabalho, o parâmetro h ótimo foi encontrado a partir de uma busca em grade para cada uma das vinte realizações, com h variando no intervalo [0,1] com passo de 0,1. Os resultados obtidos da aplicação desse classificador as bases anteriormente citadas podem ser vistos na seção seguinte.

3 Testes

Como descrito anteriormente, foi implementado um classificado com função densidade de probabilidade baseada em Janela de Parzen e os testes foram realizados

sobre as bases já mencionadas. 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.

- **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 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-Parzen é representada a partir de sua matriz de confusão abaixo:

	1	2	3
1	9	0	0
2	0	8	0
3	0	1	12

(a) Bayesiano Parzen: Iris

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

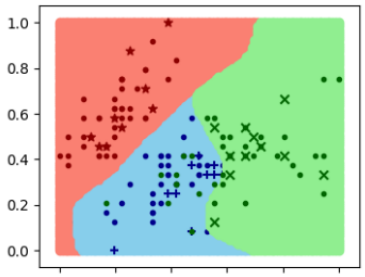

Figure 1: Superfície de decisão - Iris (2 primeiros atributos)

- 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 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-Parzen utilizando a base de dados da Coluna é representada a partir de sua matriz de confusão abaixo:

	1	2	3
1	6	0	5
2	0	29	3
3	4	1	13

(a) Bayesiano Parzen: Coluna

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

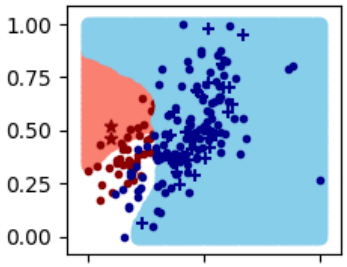

Figure 2: Superfície de decisão - Coluna (2 primeiros atributos)

- 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 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	4	0
2	0	4

(a) Bayesiano Parzen: Artificial

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

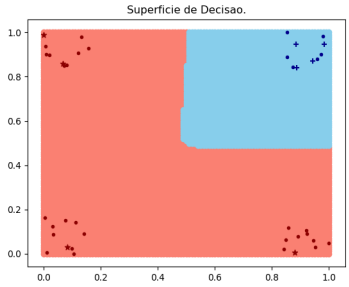

Figure 3: Superfície de decisão - Artificial I

- 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 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 Parzen utilizando a base de dados Dermatology é representada a partir de sua matriz de confusão abaixo:

	1	2	3	4	5	6
1	22	0	0	0	0	0
2	0	12	0	3	0	0
3	0	0	14	0	0	0
4	0	0	0	6	0	0
5	0	0	0	0	8	0
6	0	0	0	0	0	6

(a) Bayesiano Parzen: 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 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	85	3
2	3	45

(b) Bayesiano Parzen: Breast Cancer

4 Conclusão

A partir de uma análise breve dos resultados obtidos com o classificador com função distribuição de probabilidade baseada em janela de Parzen, podemos observar que ele apresenta resultados semelhantes aos obtidos com os demais classificadores para as mesmas bases de dados, no entanto demanda um pouco mais de tempo que os demais devido ao procedimento e busca em grade para otimização do parâmetro h.