Trabalho 2 e 3
Classificador Naive Bayes e Bayesiano

Savio Lopes Rabelo

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Campus Fortaleza – CE – Brasil

saviorabelo.ti@gmail.com

Resumo. *Este relatório descreve a implementação do classificador Naive Bayes e Bayesiano aplicada à classificação de padrões. A metodologia utilizada para a implementação é constituída por duas fases: treinamento e teste, com cada conjunto sendo composto por 80% e 20% das bases de dados, respectivamente. Foram usadas quatro bases de dados disponíveis online no repositório UCI Machine Learning. Os resultados são bastante satisfatórios, chegando em taxas de acerto em 100% em algumas bases.*

1. Introdução

O Naive Bayes é um algoritmo probabilístico simples baseado no teorema de Bayes. Este utiliza dados de treino para formar um modelo probabilístico baseado na evidência das *features* nos dados. O algoritmo supõe que há uma independência entre as *features* do modelo. Isso significa que a presença de uma determinada *feature* não tem nenhuma relação com as outras. No caso de um texto, o classificador assume que as palavras não tem uma relação entre elas. A regra de Bayes é definida da seguinte forma:

$$P(W_j|x) = \frac{p(x|W_j)P(W_j)}{p(x)}, \tag{1}$$

onde $p(x|W_j)$ é a função de verossimilhança (*likelihood function*), $P(W_j)$ é a priori e $p(x)$ é a evidência (*evidence*).

A função densidade univariada Gaussiana continua denotada usualmente por $N(\mu, \sigma^2)$, com média μ e variância σ é dada por:

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2}\frac{(\mathbf{x}-\mu)^2}{\sigma^2}\right\}. \tag{2}$$

A função densidade multivariada Gaussiana continua denotada usualmente por $N(\mu, \Sigma)$, com média μ e matriz de covariância Σ é dada por:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}, \tag{3}$$

onde $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$ é o vetor de médias, Σ é uma matriz $d \times d$, Σ^{-1} é a inversa de Σ, $|\Sigma|$ é a determinante de Σ.

A matriz de covariância é dada por $\Sigma = (\sigma)_{ij=1}^{d}$ com $\sigma_{ii} = E[(x_i - \mu_i)^2]$ (variância da componente x_i) e $\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)], i \neq j$, a covariância entre as componentes x_i e x_j. Como $\sigma_{ij} = \sigma_{ji}$, a matriz de covariância é simétrica. Mais ainda, é também positiva semi-definida.

Em virtude da forma exponencial das distribuições envolvidas é preferível se trabalhar com as seguintes funções discriminantes, os quais envolvem a função logarítmo natural (ln), Discriminante Quadrático:

$$\begin{aligned} g_i(\mathbf{x}) &= ln\left(\frac{p(x|W_j)P(W_j)}{p(x)}\right) \\ &= ln\, p(x|W_j) + ln\, P(W_j) \\ &= -\frac{1}{2}(\mathbf{x} - \mu_i)^T\Sigma_i^{-1}(\mathbf{x} - \mu_i) + ln\, P(W_i) + c_i. \end{aligned} \tag{4}$$

Para o Discriminante Linear a matriz de covariância é igual para todas as classes, assim, a Equação é dada por:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T\Sigma^{-1}(\mathbf{x} - \mu_i) + ln\, P(W_i). \tag{5}$$

Para um problema linearmente separável, rotaciona os dados de modo a maximizar a distância entre as classes e minimizar a distância intra-classe.

2. Metodologia

No primeiro momento foi realizada a separação do conjunto de dados em dois subconjuntos: treinamento e teste. Os valores utilizados para os conjuntos equivalem a 80% do conjunto original para a fase de treinamento e 20% do conjunto original para a fase de teste. Logo depois, os dados foram normalizados para eliminação de redundâncias indesejadas e também foram embaralhados. Por ser um problema com mais de 2 classes, foi exigida uma codificação diferente, Um-versus-Todos (do inglês *One-v-All*, OvA).

Para a avaliação dos resultados alcançados na classificação, foram utilizados as seguintes métricas: a precisão ou valor preditivo positivo, taxa de sensibilidade ou taxa positiva verdadeira, especificidade ou taxa real negativa e acurácia. As Equações são apresentadas a seguir:

$$Precisao = \frac{VP}{VP + FP}, \tag{6}$$

$$Sensibilidade = \frac{VP}{VP + VN}, \tag{7}$$

$$Especificidade = \frac{VN}{N} = \frac{VN}{FP + VN}, \tag{8}$$

$$Acuracia = \frac{VP + VN}{P + N}, \tag{9}$$

onde P e N é o número de padrões de cada classe. VP é o verdadeiro positivo. VN é o verdadeiro negativo. FP é o falso positivo e FN é o falso negativo.

3. Conjuntos de Dados

. Para análise comparativas neste estudo, foram usados quatro conjuntos de dados: *Iris Flower Data Set*, *Vertebral Column Data Set*, *Dermatology Data Set* e *Breast Cancer Wisconsin Data Set*; todos disponíveis online no repositório *UCI Machine Learning* [Lichman 2013].

O banco de dados da Íris[1] é o conjunto mais conhecido que se encontra na literatura de reconhecimento de padrões. O conjunto de dados contém 3 classes de 50 instâncias cada, onde cada classe se refere a um tipo de planta de íris. Uma classe é linearmente separável das outras 2 classes.

Informações dos atributos:

1. Tamanho da sépala em cm
2. Largura da sépala em cm
3. Tamanho da pétala em cm
4. Largura da pétala em cm
5. Classe:
    (a) Iris Setosa
    (b) Iris Versicolour
    (c) Iris Virginica

Já o conjunto de dados da Coluna Vertebral[2] contém seis valores para características biomecânicas usadas para classificar pacientes ortopedistas em 3 classes (normal, hérnia de disco ou espondilolistese) ou 2 classes (normal ou anormal). Foi utilizado nessa prática o conjunto com três classes.

Informações dos atributos:

1. Incidência pélvica
2. Inclinação pélvica
3. Ângulo de lordose lombar
4. Inclinação sacra
5. Raio pélvico
6. Grau de espondilolistese
7. Classe:
    (a) Hérnia de Disco (DH)
    (b) Espondilolistese (SL)
    (c) Normal (NO)
    (d) Anormal (AB)

O banco de dados de Dermatologia[3] é constituído de 34 atributos. Esse banco é parte de um estudo que aponta o tipo de Eryhemato-Squamous Disease, uma doença de pele.

Informações dos atributos (valores de 0 a 3, exceto quando indicado):

1. Eritema
2. Escala

[1]Disponível em `https://archive.ics.uci.edu/ml/datasets/iris`
[2]Disponível em `http://archive.ics.uci.edu/ml/datasets/vertebral+column`
[3]Disponível em `http://archive.ics.uci.edu/ml/datasets/dermatology`

3. Fronteiras Definidas
4. Coceira
5. Fenômeno Koebner
6. Pápulas Poligonais
7. Pápulas Foliculares
8. Envolvimento da Mucosa Oral
9. Envolvimento no Joelho e no Cotovelo
10. Envolvimento do Couro Cabeludo
11. Histórico Familiar (0 ou 1)
12. Atributos Histopatológicos

    ⋮

33. Atributos Histopatológicos
34. Idade (Classe de 1 a 6)

E o banco de dados de Câncer de Mama[4] é constituído de 10 atributos. Informações dos atributos:

1. Número do código de amostra (número de identificação)
2. Clump Espessura (1 - 10)
3. Uniformidade do tamanho da célula (1 - 10)
4. Uniformidade da forma da Célula (1 - 10)
5. Adesão Marginal (1 a 10)
6. Tamanho Único de Células Epiteliais (1 - 10)
7. Núcleos Nus (1 - 10)
8. Cromatina Branda (1 a 10)
9. Nucleoli Normal (1 - 10)
10. Mitoses (1 - 10)
11. Classe (2 para benigno, 4 para maligno)

4. Simulações Computacionais

Para realizar os experimentos, foi utilizado um computador com a seguinte configuração: processador Intel(R) Core(TM) i7-6500U a 2.5 GHz com 8 GB de RAM e executando Windowns 10. Além disso, foi utilizado a linguagem de programação MATLAB. Todos os testes foram feitos com 50 realizações em cada base.

[4]Disponível em `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\%28Original\%29`

A Tabela 1 mostra os resultados do classificador Naive Bayes em todas as bases de dados, levando em consideração as métricas já mencionadas.

Métricas	Bases de Dados					
	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	97,87	85,00	100,00	92,32	100,00	100,00
Taxa Mínima	93,33	75,81	100,00	85,14	100,00	100,00
Taxa Máxima	100,00	98,81	100,00	98.65	100,00	100,00
Desvio Padrão	02,31	05,05	00,00	02.97	00,00	00,00
Sensibilidade	97,71	80,03	100,00	91,95	100,00	100,00
Especificidade	98,91	92,56	100,00	98,58	100,00	100,00
Precisão	97,98	81,41	100,00	89.26	100,00	100,00
Tempo (s)	00,52	00,63	00,44	02,07	00.78	00,26

Tabela 1. Resultados do classificador Naive Beyes.

Na Figura 1 é apresentada a superfície de decisão com o classificador Naive Bayes.

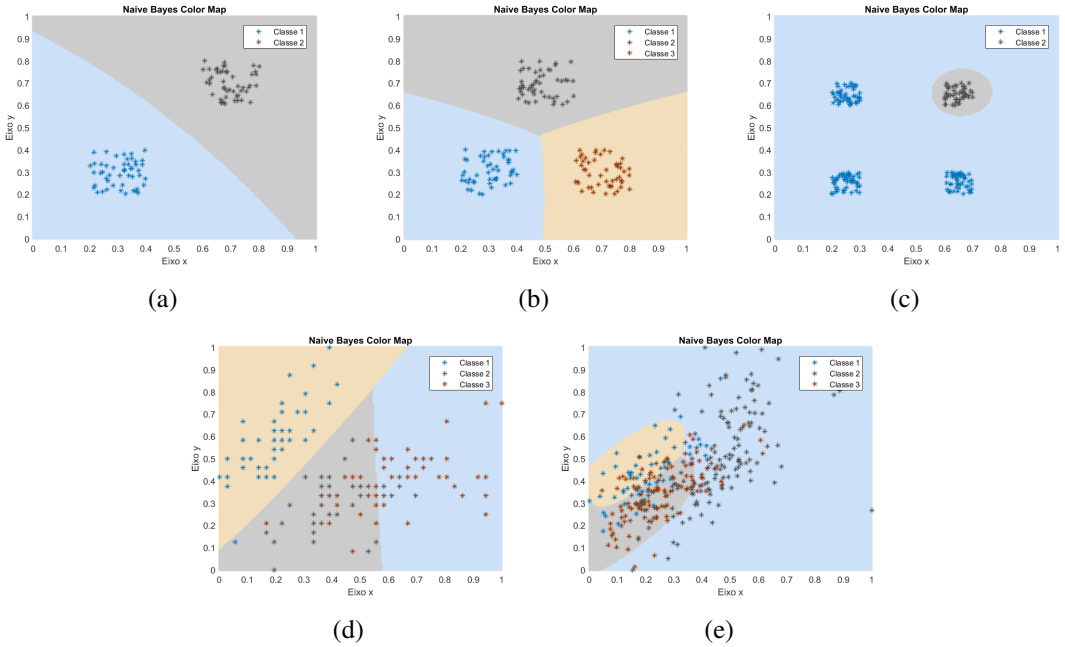

Figura 1. Superfície de decisão do classificador Naive Bayes. (a) Base artificial com 2 classes. (b) Base artificial com 3 classes. (c) Base artificial AND. (d) Íris com dois primeiros atributos. (e) Coluna com dois primeiros atributos.

A Figura 2 apresenta uma matriz de confusão de cada base de dados. Essa matriz é a matriz que ficou mais perto da acurácia.

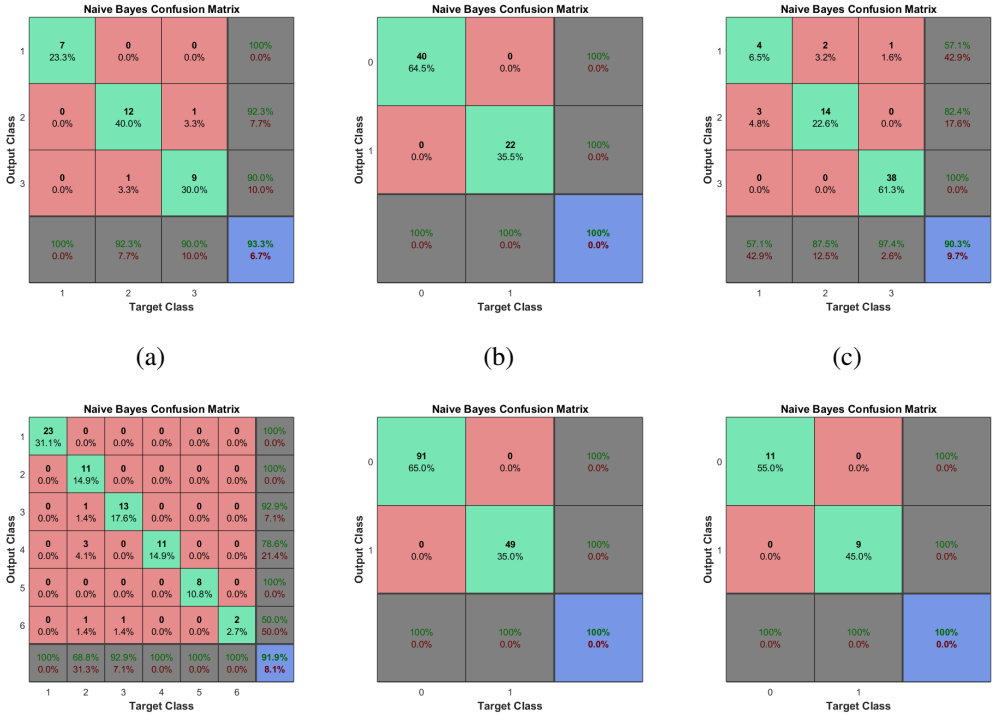

Figura 2. Matriz de Confusão para o classificador Naive Bayes. (a) Íris. (b) Coluna (2C). (c) Coluna (3C). (d) Dermatologia. (e) Câncer. (f) Artificial (2C).

A Tabela 2 e 3 mostram os resultados do classificador Bayesiano com discriminante linear e quadrático, respectivamente, em todas as bases de dados, levando em consideração as métricas já mencionadas.

	Bases de Dados					
Métricas	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	84,27	81,00	100,00	95,51	100,00	100,00
Taxa Mínima	73,33	67,74	100,00	87,84	100,00	100,00
Taxa Máxima	96,67	90,32	100,00	100,00	100,00	100,00
Desvio Padrão	06,53	05,56	00,00	02,49	00,00	00,00
Sensibilidade	84,91	76,50	100,00	94,82	100,00	100,00
Especificidade	92,14	90,22	100,00	99,09	100,00	100,00
Precisão	85,40	79,41	100,00	95,61	100,00	100,00
Tempo (s)	00,51	00,79	00,62	08,19	01,46	00,29

Tabela 2. Resultados do classificador Bayesiano com discriminante linear.

Métricas	Bases de Dados					
	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	97,53	84,81	100,00	92,78	100,00	100,00
Taxa Mínima	93,33	74,19	100,00	86,49	100,00	100,00
Taxa Máxima	100,00	95,16	100,00	97,30	100,00	100,00
Desvio Padrão	02,41	04,19	00,00	02,59	00,00	00,00
Sensibilidade	97,52	79,94	100,00	90,41	100,00	100,00
Especificidade	98,80	92,60	100,00	98,77	100,00	100,00
Precisão	97,46	80,82	100,00	93,35	100,00	100,00
Tempo (s)	00,56	00,90	00,61	07,10	01,62	00,30

Tabela 3. Resultados do classificador Bayesiano com discriminante quadrático.

A Figura 3 apresenta uma matriz de confusão de cada base de dados. Essa matriz é a matriz que ficou mais perto da acurácia.

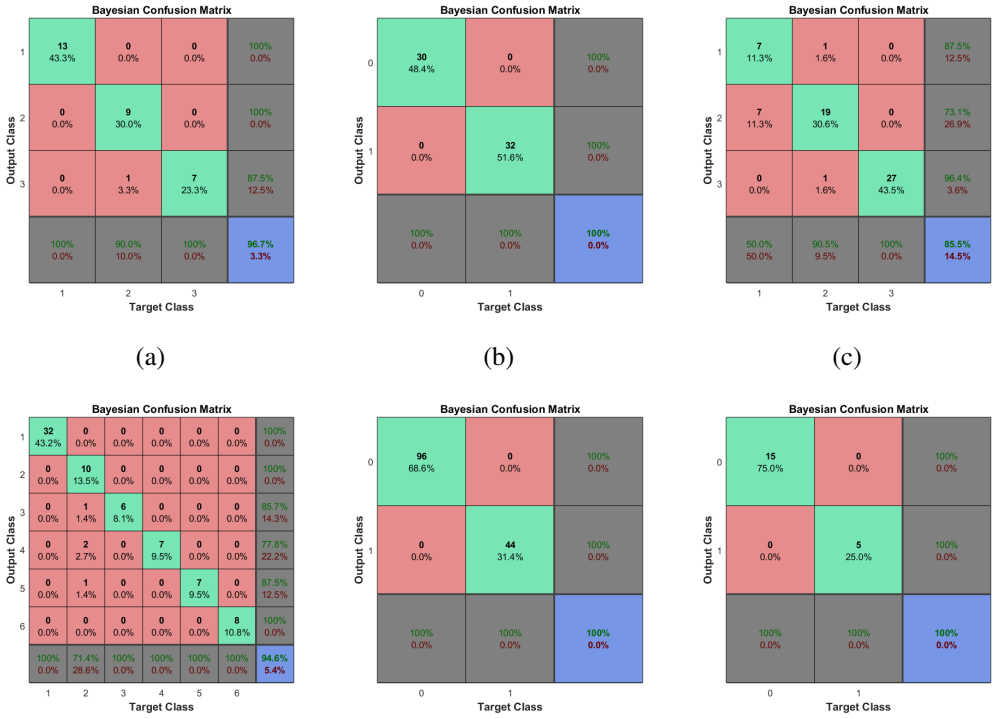

Figura 3. Matriz de Confusão para o classificador Bayesiano com discriminante quadrático. (a) Íris. (b) Coluna (2C). (c) Coluna (3C). (d) Dermatologia. (e) Câncer. (f) Artificial (2C).

Na Figura 4 e 5 é apresentada a superfície de decisão do classificador Bayesiano com discriminante linear e quadrático, respectivamente.

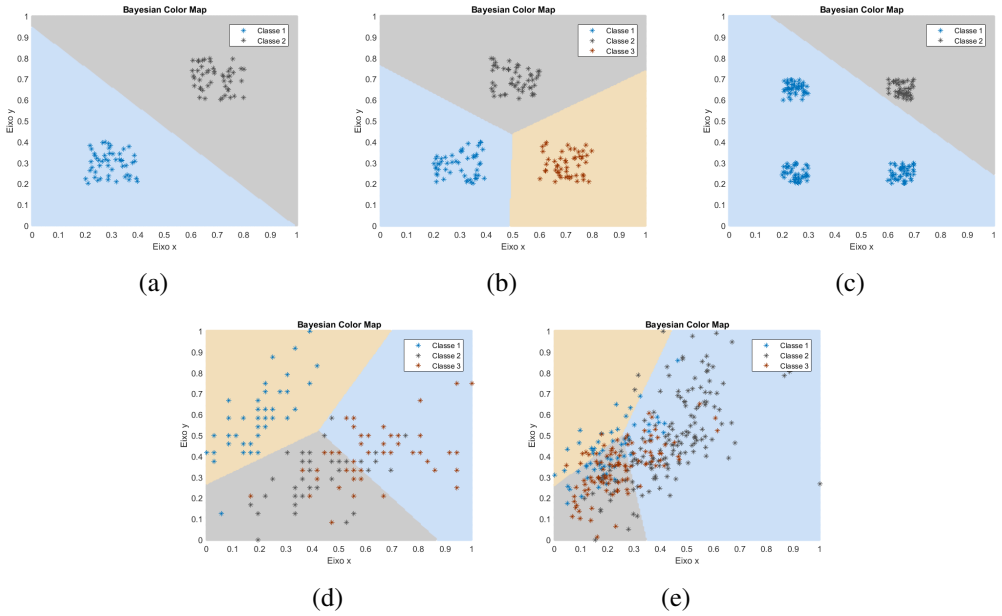

Figura 4. Superfície de decisão do classificador Bayesiano com discriminante linear. (a) Base artificial com 2 classes. (b) Base artificial com 3 classes. (c) Base artificial AND. (d) Íris com dois primeiros atributos. (e) Coluna com dois primeiros atributos.

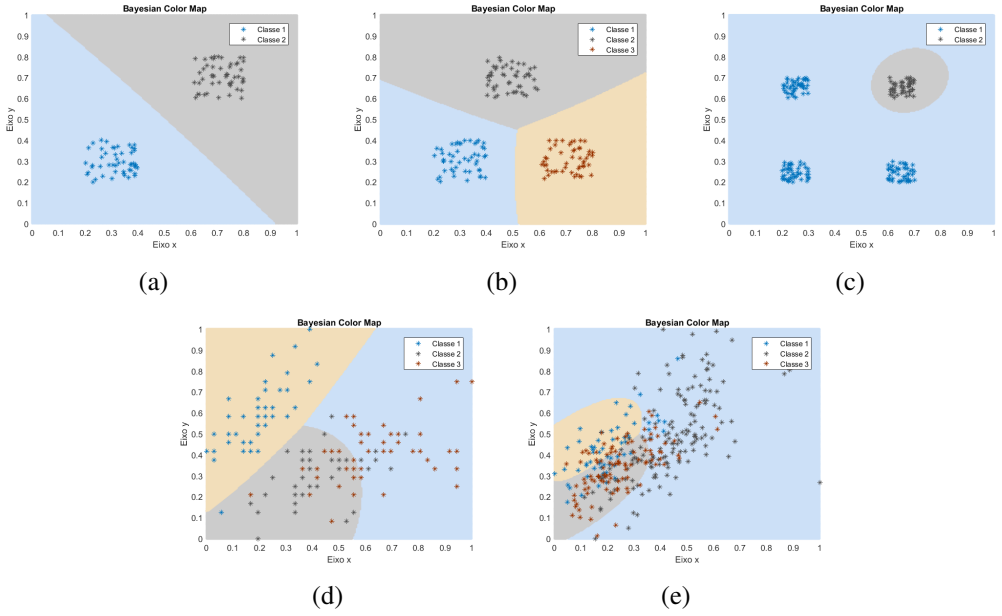

Figura 5. Superfície de decisão do classificador Bayesiano com discriminante quadrático. (a) Base artificial com 2 classes. (b) Base artificial com 3 classes. (c) Base artificial AND. (d) Íris com dois primeiros atributos. (e) Coluna com dois primeiros atributos.

5. Comparação com KNN e DMC

No classificador KNN foi utilizado o busca em grade com validação cruzada *k-fold* para encontrar o melhor k, o intervalo variou de 1 até 31, incrementando de 2 em 2. A Tabela 4 mostra os resultados do KNN em todas as bases de dados, levando em consideração as métricas já mencionadas.

	Bases de Dados					
Métricas	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	96,40	76,84	100,00	97,03	100,00	100,00
Taxa Mínima	86,67	67,29	100,00	94,59	100,00	100,00
Taxa Máxima	100,00	87,10	100,00	100,00	100,00	100,00
Desvio Padrão	03,42	04,34	00,00	01,87	00,00	00,00
Sensibilidade	96,57	73,38	100,00	96,63	100,00	100,00
Especificidade	98,19	84,41	100,00	99,43	100,00	100,00
Precisão	96,36	74,04	100,00	96,82	100,00	100,00
Tempo (s)	40,89	81,21	89,77	115,36	262,39	20,92
Intervalo de k	[3-11]	[13-27]	[1]	[3-13]	[1]	[1]

Tabela 4. Resultados do KNN.

A Tabela 5 mostra os resultados do DMC em todas as bases de dados, levando em consideração as métricas já mencionadas.

	Bases de Dados					
Métricas	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	94,00	75,19	100,00	96,49	99,03	100,00
Taxa Mínima	83,33	66,13	100,00	90,54	97,86	100,00
Taxa Máxima	100,00	88,71	100,00	100,00	100,00	100,00
Desvio Padrão	04,47	05,27	00,00	02,02	00,64	00,00
Sensibilidade	94,23	72,50	100,00	95,96	99,62	100,00
Especificidade	97,03	87,00	100,00	99,30	97,90	100,00
Precisão	94,19	74,07	100,00	96,34	98,92	100,00
Tempo (s)	00,24	00,29	00,20	00,36	00,21	00,15

Tabela 5. Resultados do DMC.

6. Resultados

Analisando os experimentos, pode-se visualizar que em geral, os dois classificadores se sai muito bem tanto na classificação de padrões, chegando a taxas de acerto de 100% quase em sua totalidade em algumas bases de dados. A base Iris tem as classes bastante separadas, tornando fácil encontrar retas que separe-as. Já em bases como a Coluna Vertebral e a Dermatologia, onde dados de diferente classes sobrepõem-se, pode-se ver que o desempenho da taxa de acerto decai. Na base do Câncer foi obtido ótimo desempenho. Já com a base de dados gerada artificialmente, o resultado já era esperado obter 100% de acerto, visto que as classes são separáveis uma das outras. O classificador Bayesiano com discriminante quadrático se saiu levemente melhor do que com o discriminante linear.

Analisando o tempo computacional, o classificador KNN demorou mais em relação aos outros, isso pode se dá porque foi utilizada a busca em grade para obter o melhor k, e também dependendo do valor de k, o custo poderá ser ainda maior. Já o tempo computacional do classificar Naive Bayes, Bayesiano e DMC são bastante similares.

Referências

Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`. Acesso em março de 2019.