Classificador Bayesiano - Gaussiana no R^n

ENG121 - Reconhecimento de Padrões

Autor: Matheus Castro Silva

RESUMO

O trabalho a seguir trata-se de uma desafio que utilizará apenas o classificador bayesiano e o modelo gaussiano multivariáveis como uma ferramenta para treinamento e classificação das amostras. Foi disponibilizado a base de dados *Iris* no R. Utilizando a base de dados e a função de densidade de probabilidade normal fornecida.

Palavras Chaves **- algoritmos, Naive Bayes, classificação, reconhecimento de padrões e Iris.**

I. INTRODUÇÃO

Nesta atividade, foi utilizado uma base de dado na plataforma R, conhecida como Iris. As características já haviam sido extraídos a priori. A base de dados apresenta 3 classes; Setosa, Versicolor e Virginica. Na qual foram unificadas em dois grupos, sendo o primeiro somente com dados de Setosa e o outro com os dados de Versicolor e Virginica. Por fim foi aplicado o classificador de Naive-Bayes para identificação de cada uma das classes.

II. DEFINIÇÃO DO PROBLEMA

Dados o conjunto de dados apresentados na base de dados *Iris*, que apresenta 3 classes (Setosa, Versicolor e Virginica). Dividiu-se os dados em dois grupos sendos estes grupos compostos por:

- 1º Grupo: Dados de Setosa
- 2º Grupo: Dados de Versicolor e Virginica

As características presentes em cada grupo são:

- Comprimento da Sépala
- Largura da Sépala
- Comprimento da Pétala
- Largura da Pétala

É válido ressaltar que o 1º grupo consiste em 50 amostra e o segundo grupo com um número de 100 amostras. Pois como as classes estão desbalanceadas, apresentará uma alteração no cálculo da probabilidade de cada umas destas classes.

III. DIVISÃO DOS DADOS

Para cada grupo, foi pedido que houvesse uma separação, apresentando dados de testes e dados treinamento para o modelo.

A separação escolhida foi da seguinte forma:

- Treinamento: 70%
- Testes: 30%

Estes valores são bem coerentes para o problema vigente, pois queremos ajustar melhor os parâmetros do modelo. Desta maneira é necessário um maior número de amostras no treinamento do que na fase de testes.

IV. TREINAMENTO DO MODELO E FUNÇÃO DE DENSIDADE

O modelo foi treinado utilizando dois parâmetros conhecidos, sendos estes: média e matriz de covariância para cada classe. Os parâmetros obtidos foram utilizados na função de densidade de probabilidade normal, que pode ser observada abaixo.

$$\mathrm{p(x)} = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

em que μ e Σ são, respectivamente, média e matriz de covariância para a classe em questão.

Este problema consiste em 4 características para cada amostra, ou seja, um problema de 4 dimensões. Com isto, não será possível mostrar a imagens de resultado referente a cada um dos grupos.

V. RESULTADOS

Utilizou-se o conjunto de treinamento, estimar qual a classe de cada uma das amostras baseado nos valores de densidade de probabilidade encontrados no item anterior. Também foi apresentado a matriz de confusão. utilizando o comando em R abaixo.

$$table(y, \widehat{y})$$

Para o conjunto de teste, calculou-se a densidade de probabilidade de cada amostra para cada uma das classes, estimou-se a qual classe cada uma das amostras pertence baseado nos valores de densidade de

probabilidade encontrados e apresentou-se a matriz de confusão resultante.

Por fim, foi realizado um laço *for* para repetir os processos anteriores e calcular a média e desvio padrão dos acertos do grupo de teste após 30 repetições.

Os resultados obtidos foram:

- Média de Acerto: 100%
- Desvio-Padrão: 0

O resultado da matriz de confusão está presente na imagem 1 abaixo.

```
Confusion Matrix and Statistics

                    Reference
Prediction           setosa versicolorVirginica
  setosa                 15                  0
  versicolorVirginica     0                 30

               Accuracy : 1
                 95% CI : (0.9213, 1)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 1.191e-08

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.3333
         Detection Rate : 0.3333
   Detection Prevalence : 0.3333
      Balanced Accuracy : 1.0000

       'Positive' Class : setosa
```

Imagem 1: Resultado da matriz de confusão das amostras da base de dados *Iris*.

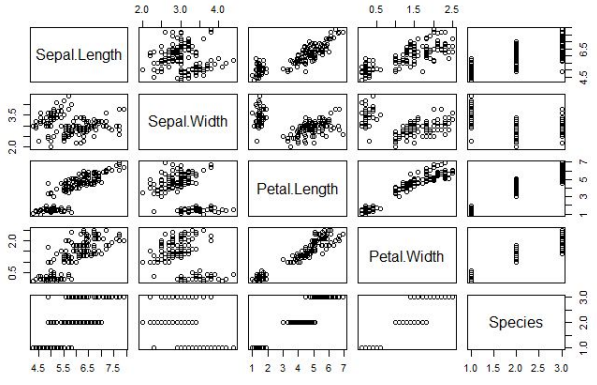

Imagem 2: Ilustração das amostras da base de dados *Iris*.

Apartir da imagem 2 é de fácil percepção que os dados estão bastante separados entre si, desta maneira é possível comprovar a taxa de 100% de acerto encontrado no teste de 30 repetições calculando a média e desvio padrão feito anteriormente.

VI. CONCLUSÕES

Os classificadores Bayes fazem parte de uma família simples de classificadores probabilísticos com base na aplicação do teorema de Bayes com fortes pressupostos de independência entre os recursos.

Apartir dos resultados de média e desvio-padrão, é de fácil visualização que as classes estão distantes umas das outras e com isso o classificador conseguiu gerar uma curva de separação bastante interessante para o problema proposto. Separando os grupos em 100% dos testes gerados.

Utilizando este classificador e do conjunto de amostras referente a cada classe, foi possível ter ótimos resultados e bastante eficiência perante o classificador de Bayes.

VII. REFERÊNCIAS

[1] R tutorial, http://www.r-tutor.com/r-introduction, acessado 30/08/2017.

[2] The Database of Faces - http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html, acessado 30/08/2017.