Exercício 8 - Classificador de Bayes aplicado a um problema Multivariado

Rodrigo Machado Fonseca - 2017002253

January 15, 2022

1 Introdução

Neste trabalho iremos implementar o algoritmo Classificador de Bayes 2. Em seguida, iremos utilizá-lo para classificar um conjunto de amostras.

2 Algoritmo de Bayes Multivariado

O algoritmo de Bayes não se altera, e será definido pelos seguintes passos:

- Calcular P[A] e P[B].
- Escolher um elemento para ser classificado.
- Calcular a Verossimilhança.
- Classificar o elemento.

Para o caso multivariado é necessário calcular a densidade conjunto da seguinte maneira:

$$p(x) = \frac{1}{\sqrt{((2 * \pi)^n |\Sigma|)}} \dot{e}xp(-0.5(x-\mu)^T \Sigma^{-1} * (x-\mu)), \tag{1}$$

onde Σ é a matriz de covariâncias, $|\Sigma|$ é o determinante da matriz de covariância e μ é o vetor de médias das distribuições marginais. Tal equação está expressa pelo seguinte linha de código:

```
> rm(list=ls())
> pdfnvar <- function(x, m , K, n){
+   first_term = (1/(sqrt((2*pi)^n*(det(K)))))
+   return(first_term*exp(-0.5*(t(x - m)%*% (solve(K)) %*% (x - m))))
+ }
```

O restante do algoritmo permanece inalterado como pode ser visto nos trechos a seguir:

```
> bayes_paramaters <- function(x_train){
+   c1 = x_train[x_train[,ncol(x_train)] == 1, ]
+   c2 = x_train[x_train[,ncol(x_train)] == 2, ]
+   meanc1 <- apply(c1[,1:ncol(x_train)-1], 2, mean)
+   covc1 <- cov(c1[,1:ncol(x_train)-1])
+   par_c1 <- list(meanc1, covc1)
+   meanc2 <- apply(c2[,1:ncol(x_train)-1], 2, mean)
+   covc2 <- cov(c2[,1:ncol(x_train)-1])
+   par_c2 <- list(meanc2, covc2)
+   return(list(par_c1, par_c2))
+ }

> bayes_classifier <- function(x_train, x_test){
+   pC1 = (nrow(x_train[x_train[ ,14] == 1, ])) / nrow(x_train)
+   pC2 = 1-pC1
+   parameters <- bayes_paramaters(x_train)
+   par_c1 <- parameters[[1]]
+   par_c2 <- parameters[[2]]
+   y_hat <- matrix(nrow = nrow(x_test), ncol = 1)
+   x_aux <- x_test[,1:ncol(x_test)-1]
+   for(i in sample(nrow(x_aux))){
+     p1 <- pdfnvar(x_aux[i, ],
+                   par_c1[[1]],
+                   par_c1[[2]],
+                   ncol(x_aux))
+     p2 <- pdfnvar(x_aux[i, ],
+                     par_c2[[1]],
+                     par_c2[[2]],
+                     ncol(x_aux))
```

```
+     if(p1*pC1/(p2*pC2) >= 1){
+       y_hat[i] <- 1
+     }
+     else{
+       y_hat[i] <- 2
+     }
+   }
+   return(y_hat)
+ }
```

3 Experimento

Neste experimento utilizaremos a base *Startlog (Heart)*[1].

```
> # Carregar a base de dados
> data <- as.matrix(read.table("heart.dat", sep = ' '))
```

A base apresenta os seguintes atributos:

- 1. idade
- 2. sexo
- 3. tipo de dor no peito (4 valores)
- 4. pressão arterial em repouso
- 5. colesterol sérico em mg / dl
- 6. açúcar no sangue em jejum> 120 mg / dl
- 7. resultados eletrocardiográficos de repouso (valores 0,1,2)
- 8. frequência cardíaca máxima alcançada
- 9. angina induzida por exercício
- 10. pico antigo = depressão de ST induzida por exercício em relação ao repouso
- 11. a inclinação do segmento ST de pico de exercício

- 12. número de vasos principais (0-3) coloridos por fluorosopia
- 13. tal: 3 = normal; 6 = defeito corrigido; 7 = defeito reversível

A predição é 1 quando não possui doença cardíaca, e 2 caso contrário. Há 150 amostras da classe 1 e 120 amostras da classe 2.

Neste experimento serão separados aleatoriamente 90% dos dados para treinamento e 10% para teste. Ao final, deveremos calcular a acurácia obtida com o classificador. Posteriormente, o mesmo procedimento será repetido com amostras de treinamento de 70% e 20%.

4 Resultados

A tabela a seguir estão expostos os resultados de cada experimento:

Porcentagem Treino	Acurácia
90%	0.962962962962963
70%	0.851851851851852
20%	0.689814814814815

Table 1: Atributos da base de dados *Startlog (Heart)*.

5 Conclusão

É possível notar que os valores de acurácia foram menores que aqueles encontrados no exercício anterior. Isso se deve ao fato de que este problema apresenta maior complexidade, principalmente por ser multivariado. Além disso, a proporção de dados não estava perfeitamente equilibrada como no exercício anterior. Nesse sentido, se trata de um problema de solução mais difícil e portanto realmente se espera acurácias mais baixas para o classificador.

Na tabela 1 podemos ver que quanto maior o número do conjunto de treino maior a acurácia. Isso ocorre porque o conjunto de treinamento é utilizado na construção do classificador, e portanto é importante se ter um conjunto grande para tornar o classificador o mais geral e eficiente possível.

É importante salientar que mesmo o conjunto com 20% de amostras obteve resultado bom levando em conta a limitação do conjunto de treinamento.

Com o experimento foi possível compreender melhor o funcionamento do classificador de Bayes e implementá-lo de forma satisfatória. Os resultados obtidos podem provar que o classificador de Bayes com um conjunto de amostras suficientemente grande possui um ótimo desempenho.

References

[1] `https://archive.ics.uci.edu/ml/datasets/statlog+(heart)`