Solução Lista 01

Nome: Guilherme Afonso Gigeck
E-mail: guilherme.gigeck@aluno.ufabc.edu.br
Nome: Vitor Moraes Bravo Alves
E-mail: vitor.bravo@aluno.ufabc.edu.br
(Não é preciso informar os RAs)

24 fevereiro, 2025

Exercício 01

a) Problema de classificação: É um tipo de problema de aprendizado supervisionado, onde o objetivo é prever a classe, categoria ou rótulo dos dados com base no vetor de característica. Um exemplo para esse problema é um sistema de detecção te spam em e-mails, em que o objetivo é reconhecer se um e-mail é spam ou não. O vetor de características pode ter elementos como a inclusão de palavras-chave específicas, frequência de links no corpo do e-mail, remetente, a estrutura do e-mail e análise linguística. Com isso é possível classificar um e-mail como "spam" ou "não spam".

b) Problema de regressão: É um problema de aprendizado supervisionado, onde o objetivo é prever um valor numérico contínuo com base no vetor de características, nesse problema não existem rótulos. Um exemplo é um modelo de regressão para prever o preço de um imóvel. Nesse caso o vetor de características pode conter elementos como a localização, o tamanho do imóvel em metros quadrados, número de quartos, salas e banheiros, idade do imóvel e proximidade do sistema de transporte público. Assim podemos estimar um valor de venda do imóvel.

c) Problema de agrupamento: É um tipo de problema de aprendizado não-supervisionado em que o objetivo é distribuir o conjunto de dados em grupos com características semelhantes, mas sem classificações ou rótulos definidos. Um exemplo pode ser um modelo para segmentar clientes em uma agência de marketing. O vetor de características pode conter elementos como a idade do cliente, localização, histórico de compras, frequência de compras, valor gasto. Dessa forma é possível direcionar as propagandas de um determinado produto à um público específico.

Exercício 02

A maldição da dimensionalidade é um fenômeno que ocorre quando trabalhamos com dados com muitas dimensões. Quanto maior for o número de variáveis ou características da base de dados, mais distantes estarão os vizinhos mais próximos do ponto desejado, ou seja, os dados estarão cada vez mais distantes uns dos outros. Isso dificulta muito na identificação de padrões e classificações, comprometendo a eficácia de métodos que usam a proximidade dos pontos, como o método kNN, e afeta o desempenho dos algoritmos.

Exercício 03

```
myKNN=function(k,x,D){
  D2 <- D %>%
        mutate(dist=(x[1]-x_1)^2+(x[2]-x_2)^2) %>%
        arrange(dist) %>% head(k) %>% count(y) %>% arrange(desc(n))
  return(D2)
}
```

```
#basic unit test
Dinstance = tibble(x_1=rnorm(100,1,1),
                   x_2=rnorm(100,-1,2),
                   y=factor(sample(c("one","two","three"),100,replace=T))
                   )
x = c(1.05,0.5)
ans = myKNN(6,x,Dinstance)
ans
```

```
## # A tibble: 3 x 2
##   y         n
##   <fct> <int>
## 1 one       3
## 2 two       2
## 3 three     1
```

Exercício 04

```
data("iris")
iris = as_tibble(iris) %>%
       select(Petal.Length,Sepal.Length,Species) %>%
       rename(x_1=Petal.Length,x_2=Sepal.Length,y=Species)

irislist = as.list(iris)

Kcheck = function(mylist,mytibble,k){
                  return(pmap_lgl(mylist,function(x_1,x_2,y){
                          ans=myKNN(k,c(x_1,x_2),mytibble)
                          return(as.character(ans$y[1])==y)
                          }
                  ))
                }
#k=1
IrisK1Check = Kcheck(irislist,iris,1)
K1accuracy = sum(IrisK1Check)
Irisk10Check = Kcheck(irislist,iris,10)
K10accuracy = sum(Irisk10Check)
result=t(list(total=length(iris$y),K1accuracy=K1accuracy, K10accuracy=K10accuracy))
result
```

```
##      total K1accuracy K10accuracy
## [1,] 150   149        143
```

A razão do único erro é, devido às características filtradas, que existem duas entradas de características idênticas e diferentes tipos e assim por conta do empate o KNN escolheu pseudo-aleatóriamente errado (provavelmente por conta da disposição inicial da lista e o algoritmo de ordenamento)

Exercício 05

Seja $D_m = ((x_1, y_1), ...(x_n, y_n))$ conjunto de dados amostrados, nos quais $x_k \in \mathcal{X}$ e $y_k \in \mathbb{R}$, e XeY variáveis aleatórias tais que D_m seja um exemplo de amostragem. Considere a função de predição $f\colon \mathcal{X} \to \mathbb{R}$ e a função perda $l_2\colon \mathbb{R} \times \mathbb{R} \to [0, \infty]; l(y, y') = |y - y'|$. Calcula-se então o risco esperado $\mathcal{R}(f)$ em relação as variáveis X e Y:

$$\mathcal{R}(f) = \mathbb{E}_{XY}[l(Y, f(X))] \tag{1}$$
$$= \mathbb{E}_X[\mathbb{E}_{Y|X}[l(Y, f(X))]] \tag{2}$$

Para minimizar o risco esperado basta minimizar $\mathbb{E}_{Y|X}[l(Y, f(X))]$ para todo espaço condicionado por X, logo busca-se o ponto crítico dessa expressão:

$$\frac{\mathrm{d}\mathbb{E}_Y[l(Y, f(X))|X = x]}{\mathrm{d}f} = \mathbb{E}_Y[\frac{\mathrm{d}l(Y, f(X))}{\mathrm{d}f}|X = x] \text{ utilizando o resultado fornecido} \tag{3}$$
$$= \int_{\mathbb{R}} \frac{\mathrm{d}l(y, f(X))}{\mathrm{d}f}\mathrm{d}y|X \tag{4}$$
$$= \int_{\mathbb{R}} \frac{\mathrm{d}|y - f(x)|}{\mathrm{d}f}\mathrm{d}y|X \text{ como estamos condicionando a x} \tag{5}$$
$$= (\int_{-\infty}^{f(x)} \frac{\mathrm{d}|y - f(x)|}{\mathrm{d}f}\mathrm{d}y|X) + (\int_{f(x)}^{\infty} \frac{\mathrm{d}|y - f(x)|}{\mathrm{d}f}\mathrm{d}y|X)) \tag{6}$$
$$= (\int_{-\infty}^{f(x)} -1y|X) + (\int_{f(x)}^{\infty} 1y|X)) \text{ probabilidade acumulativa} \tag{7}$$
$$= -\mathbb{P}(y < f(x)|X = x) + \mathbb{P}(y > f(x)|X = x) \tag{8}$$

Portanto para que $\frac{\mathrm{d}\mathbb{E}_Y[l(Y, f(X))|X = x]}{\mathrm{d}f} = 0$ temos que:

$$\mathbb{P}(y < f(x)|X = x) = \mathbb{P}(y > f(x)|X = x) \tag{9}$$

Ou seja, a função f que minimiza o risco é dada por $f(x) \colon = Mediana(Y|X = x)$ $\square$.

Exercício 06

Dada a descrição do exercício temos que a probabilidade de um subconjunto da hiperesfera é diretamente proporcional ao seu volume. O volume de uma hiperesfera de dimensão d de raio r é dado por $V_d(r) = C * r^d$ onde $C \in \mathbb{R}$ é constante conhecida cujo valor não é de interesse para esta análise. Calculando a mediana $M \leq 1$ do evento descrito:

$$\frac{1}{2} = \mathbb{P}(min\{X_1, \cdots, X_m\} > M) = \mathbb{P}(\bigcap_{i=1}^{m}(X_i > M)) \tag{10}$$
$$= \prod_{i=0}^{m} \mathbb{P}(X_i > M) \text{ por independencia dos eventos} \tag{11}$$
$$= \mathbb{P}(X > M)^m \text{ por semelhança dos eventos} \tag{12}$$
$$= (1 - V_d(M)/V_d(1))^m = (1 - M^d)^m \tag{13}$$

Isolando M obtemos $M = (1 - 0.5^{\frac{1}{m}})^{\frac{1}{d}}$ como desejado