Solução Lista 01

Nome: Guilherme Afonso Gigeck
E-mail: guilherme.gigeck@aluno.ufabc.edu.br
Nome: Fulano de Tal
E-mail: fdt@aluno.ufabc.edu.br
(Não é preciso informar os RAs)

22 February, 2025

Exercício 01

Exercício 02

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 three     3
## 2 two       2
## 3 one       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)
```

```r
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: $\mathcal{X} \to \mathbb{R}$ e a função perda l_2: $\mathbb{R} \times \mathbb{R} \to [0, \infty]$; $l(y, y') = |y - y'|$. Calcula-se então o risco esperado $\mathcal{R}(f)$ em relação as variáveis X e Y:

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

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:

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

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

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

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

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:

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

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