Solução Lista 01

Nome: Vinicius de Oliveira Bezerra
E-mail: v.bezerra@aluno.ufabc.edu.br
Nome: Deyved Kevyn Alves Lima
E-mail: deyved.lima@aluno.ufabc.edu.br

24 February, 2025

Importações

```r
library(tidyverse)
```

Solução Exercício 01

Classificação:

O problema de classificação ocorre quando o objetivo é atribuir uma entrada a uma de várias categorias predefinidas.

Exemplo prático:

- **Aplicação:** Identificação de objetos em imagens (por exemplo, identificar se há um "gato", "cachorro" ou "pessoa" em uma foto).
- **Vetores de características:** Pixels da imagem (ou características extraídas de técnicas como redes neurais convolucionais).

- **Rótulos/Respostas:** As categorias possíveis são "gato", "cachorro", "pessoa", etc.

Objetivo:

Classificar as imagens de forma que o modelo consiga identificar se elas pertencem a uma **categoria** específica (como "gato", "cachorro" ou "pessoa").

Regressão:

Em um problema de regressão, o objetivo é **prever um valor contínuo**. Ou seja, a saída não pertence a categorias discretas, mas sim a um intervalo de valores.

Exemplo prático:

- **Aplicação:** Previsão do preço de imóveis.
- **Vetores de características:** Tamanho do imóvel (em metros quadrados), número de quartos, localização, idade do imóvel, entre outros.
- **Rótulos/Respostas:** O preço de um imóvel em uma determinada moeda (valor númerico contínuo).

Objetivo:

O modelo tenta prever um valor contínuo (preço) com base nas características fornecidas do imóvel.

Agrupamento:

O probelma de agrupamento é **não supervisionado**, ou seja, não temoa rótulos para os dados. O objetivo é **dividir os dados em grupos** (clusters) com base na semelhança entre as características dos dados.

Exemplo prático:

- **Aplicação:** Segmentação de clientes em marketing.
- **Vetores de características:** Dados como idade, histórico de compras, comportamento da navegação, região geográfica, entre outros.
- **Rótulos/Respostas:** Não há rótulos específicos, pois o modelo irá dividir os dados em **grupos ou clusters**, como clientes "frequentes", "potenciais" ou "perdidos".

Objetivo:

O modelo tenta agrupar os clientes em diferentes segmentos com base nas semelhanças entre suas características, sem saber previamente quais segmentos existem.

Solução Exercício 02

Maldição da Dimensionalidade: A "maldição da dimensionalidade" acontece quando o número de variáveis ou características de um conjunto de dados cresce demais, tornando as coisas mais difíceis para os algoritmos. Com muitas dimensões, os dados ficam muito espalhados, e as distâncias entre eles quase se igualam, dificultando a tarefa dos modelos de encontrar padrões. Além disso, mais dimensões exigem mais dados para treinar corretamente e aumentam o custo computacional, deixando tudo mais lento e ineficiente.

Solução Exercício 03

```r
library(dplyr)
library(tibble)
library(purrr)

knn_classify <- function(k, x, D){

  # Calcular a distância Euclidiana quadrada para todos os pontos do dataset
  D2 <- D %>%
```

```r
    mutate(dist = (x[1] - x_1)^2 + (x[2] - x_2)^2) %>% # Distância euclidiana
    arrange(dist) %>% # Ordena pela menor distância
    head(k) %>%       # Seleciona os k vizinhos mais próximos
    count(y)          # Conta quantos vizinhos pertecem a cada classe

  # Retornando a classe mais frequente entre os k vizinhos
  return(D2$y[which.max(D2$n)])}

D <- tibble(
  x_1 = rnorm(100, 1, 1),
  x_2 = rnorm(100, 1, 2),
  y = factor(sample(c("one", "two", "three"), 100, replace = TRUE))
)

# Definindo um novo ponto para classificar e a quantidade de vizinhos
x <- c(1, 2)
k <- 10

# Chamando a função kNN
resultado <- knn_classify(k, x, D)
print(resultado)  # Exibe a classe mais provável
```

```
## [1] three
## Levels: one three two
```

Solução Exercicício 04

```r
library(tidyverse)
library(purrr)

data("iris") # Carrega o banco no ambiente global

iris <- as_tibble(iris) %>% # Converte para a dataframe tibble
  select(Petal.Length,Sepal.Length,Species) %>% # Seleciona colunas da dataframe
  rename( x_1 = Petal.Length, x_2 = Sepal.Length, y = Species) # Renomeia as colunas

knn_predict <- function(train_data, new_point, k){

  D2 <- train_data %>%
    mutate(dist = (new_point[1] - x_1)^2 + (new_point[2] - x_2)^2) %>% # Distância euclidiana
    arrange(dist) %>%
    head(k) %>%
    count(y)

  # Retornando a classe mais frequente entre os k vizinhos
  return(D2$y[which.max(D2$n)])
}

# Converte iris para uma lista (parausar pmap_lgl)
l_iris <- as.list(iris)
```

```r
# Testa o modelo com k = 10 e calcula quantos acertos temos
v_bool_k10 <- pmap_lgl(l_iris, function(x_1, x_2, y) {
  predicted_y <- knn_predict(iris, c(x_1, x_2, y), k = 10)
  return(predicted_y == y)
} )

# Testa o modelo com k = 1
v_bool_k1 <- pmap_lgl(l_iris, function(x_1, x_2, y){
  predicted_y <- knn_predict(iris, c(x_1, x_2, k), k = 1)
  return(predicted_y == y)
})

# Calcula a taxa de acerto
accuracy_k10 <- sum(v_bool_k10) / length(v_bool_k10)
accuracy_k1 <- sum(v_bool_k1) / length(v_bool_k1)

# Exibe os resultados
cat("Taxa de acerto para k = 10: ", accuracy_k10 * 100, "%\n")
```

```
## Taxa de acerto para k = 10:  95.33333 %
```

```r
cat("Taxa de acerto para k = 1: ", accuracy_k1 * 100, "%\n")
```

```
## Taxa de acerto para k = 1:  99.33333 %
```

Solução Exercício 05

Objetivo: Encontrar a função de regressão $f : \mathcal{X} \to \mathcal{Y}$ ótima que minimiza o risco esperado da função de perda do erro absoluto, tal que $\ell_2(y, y') := (y - y')^2$ Mostrar que essa função equivale à $f(x) :=$ Mediana$(Y|X = x)$

Demonstração

$$\mathcal{R}(f) = \mathbb{E}_{XY}[\ell(|Y - f(X)|)]$$

Condicionando em X

$$\mathcal{R}(f) = \mathbb{E}_X(\mathbb{E}_Y[\ell(|Y - f(X)| \mid X)])$$

Função de minimização

$$\min_z \mathbb{E}[|Y - z| \mid X = x]$$

Minimização da Esperança do Erro Absoluto

$$\mathbb{Q}(z) = \mathbb{E}[|Y - z||X = x] = \int_{-\infty}^{\infty} |Y - z| \; \mho_{Y|X} \; (y|x)dy$$

Encontrar a derivada

$$\mathbb{Q}'(z) = \frac{\partial}{\partial z}\left[\int_{-\infty}^{\infty} |Y - z| \ \mho_{Y|X} \ (y|x)dy\right]$$

$$\frac{\partial}{\partial z}|Y - z| = \begin{cases} -1 & \text{se } Y > z \\ 1 & \text{se } Y < z \end{cases}$$

Logo

$$\mathbb{Q}'(z) = \int_{-\infty}^{z} \mho_{Y|X} \ (y|x)dy - \int_{z}^{\infty} \mho_{Y|X} \ (y|x)dy$$

$$\mathbb{Q}'(z) = \int_{-\infty}^{f(x)} P(Y < z \mid X = x) - P(Y > z \mid X = x)$$

Para o ponto mínimo a derivada deve ser 0

$$\mathbb{Q}'(z) = 0 \Longrightarrow P(Y < z \mid X = x) = P(Y > z \mid X = x)$$

Com a soma das duas probabilidades deve ser 1

$$\begin{cases} P(Y < z \mid X = x) = P(Y > z \mid X = x) \\ P(Y < z \mid X = x) + P(Y > z \mid X = x) = 1 \end{cases} \Longrightarrow P(Y < z \mid X = x) = P(Y > z \mid X = x) = \frac{1}{2}$$

Esse valor divide igualmente a distribuição em duas partes iguais, com 50% de probabilidade para cada lado, que é por definição, a Mediana$(Y|X = x)$. Assim, a função de regressão que buscamos é

$$f(x) := \text{Mediana}(Y|X = x)$$

Solução Exercício 06

A probabilidade de um ponto estar a uma distância r da origem é proporcional à área da casca esférica de raio r em d dimensões. A função de distribuição acumulada (CDF) de R é:

$$F_R(r) = r^d$$

Dado que m: quantidade de pontos independentes e uniformemente distribuídos

$$F_R min(r) = 1 - (1 - r^d)^m$$

A mediana é o valor de r tal que a CDF é 0,5 Logo

$$1 - (1 - r^d)^m = 0,5(1 - r^d)^m = 0,51 - r^d = 0,5^{1/m} r^d = 1 - 0,5^{1/m}$$

Por fim

$$r = (1 - 0,5^{1/m})^{1/d} \ \square$$