Mineração de Dados

Trabalho 6

Arthur do Prado Labaki - 11821BCC017

22-04, 2023

GBC212

Estudo do Conceito

O algoritmo de mineração de dados k-vizinhos mais próximos (kNN) é um método simples e efetivo de classificação e regressão em aprendizado de máquina. O algoritmo funciona encontrando os k exemplos de treinamento mais próximos do exemplo de teste e usando a classe majoritária ou a média desses k exemplos para prever a classe ou valor do exemplo de teste.

O algoritmo kNN pode ser resumido nos seguintes passos:

Calcular a distância entre o exemplo de teste e todos os exemplos de treinamento usando uma métrica de distância, como a distância euclidiana ou a distância de Manhattan.

Selecionar os k exemplos de treinamento mais próximos do exemplo de teste com base nas distâncias calculadas.

Usar a classe majoritária dos k exemplos selecionados para prever a classe do exemplo de teste em um problema de classificação ou usar a média dos k exemplos selecionados para prever o valor do exemplo de teste em um problema de regressão.

O valor de k pode ser escolhido pelo usuário e é um hiper parâmetro importante que afeta o desempenho do modelo. Um valor maior de k pode levar a uma melhor generalização, mas pode perder detalhes locais importantes, enquanto um valor menor de k pode levar a uma maior sensibilidade ao ruído e a overfitting.

O algoritmo kNN é fácil de entender e implementar, mas pode ser computacionalmente caro para grandes conjuntos de dados e pode ser sensível à escala dos dados e à escolha da métrica de distância. No entanto, é uma técnica muito útil para classificação e regressão em problemas de aprendizado de máquina.

Resolução do Exercício 1)

Resolver o problema abaixo usando o classificador k-vizinhos mais próximos:

Com a base de dados, escolhemos um valor adequado para k, que no nosso caso será 3. O primeiro passo é calcular a distância entre cada ponto desconhecido e cada ponto no conjunto

sepal		petal		class
length	width	length	width	
6.3	2.3	4.4	1.3	versicolor
6.2	3.4	5.4	2.3	virginica
5.2	3.4	1.4	0.2	setosa
6.9	3.1	5.4	2.1	virginica
5.7	4.4	1.5	0.4	setosa
5.4	3.7	1.5	0.2	setosa
5	3.3	1.4	0.2	setosa
6.4	2.8	5.6	2.1	virginica
6	3	4.8	1.8	virginica
5.5	2.5	4	1.3	versicolor

sepal		petal		class
length	width	length	width	
7.3	2.9	6.3	1.8	?
6.1	2.9	4.7	1.4	?
4.6	3.4	1.4	3.0	?

Figura 1: Base de dados

de treinamento. Usaremos a distância Euclidiana. Sua formula se baseia em que x e y para o tamanho e largura da pétala e z e w para a largura da sépala e 1 e 2 são duas classes distintas. Com isso temos:

$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2 + (z2 - z1)^2 + (w2 - w1)^2}$$

Então, será necessário realizar o calculo da distancia euclidiana entre cada um dos pontos testes e todos os pontos da base de dados. Com isso montamos a tabela abaixo que demonstra esse calculo. Nela o D1 é a distancia euclidiana entre o ponto atual (p1) e o ponto da primeira linha na base de dados (Versicolor) e assim segue até D10 sendo o ultimo ponto. Com isso criamos a tabela (contas no final do trabalho):

ID	x	y	z	w	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
1	1.8	6.3	2.9	7.3	2.284	1.589	5.588	1.048	5.459	5.464	5.657	1.183	1.987	2.988
2	1.4	4.7	2.9	6.1	0.707	1.248	3.659	1.288	3.694	3.579	3.701	1.183	0.435	1.009
3	3.0	1.4	3.4	4.6	3.998	4.364	2.863	4.710	2.996	2.929	2.830	4.695	3.888	3.357

Em seguida, selecionamos os k pontos mais próximos de cada ponto desconhecido com base nas distâncias calculadas. No nosso caso, k = 3, então escolhemos os três pontos mais próximos para cada ponto desconhecido. Os resultados são mostrados na tabela abaixo:

ID	Ponto 1	Ponto 2	Ponto 3
1	D4 = Virginica	D8 = Virginica	D2 = Virginica
2	D9 = Virginica	D1 = Versicolor	D10 = Versicolor
3	D7 = Setosa	D3 = Setosa	D6 = Setosa

Finalmente, contamos as classes dos pontos selecionados para cada ponto desconhecido e escolhemos a classe mais frequente para atribuir ao ponto desconhecido. Com isso temos que o ponto desconhecido 1 pode ser atribuído para a classe Virginica (3 de 3 ocorrências). Já o ponto desconhecido 2 provavelmente deve ser da classe Versicolor (2 de 3 ocorrências). E por fim, o ponto desconhecido 3 é atribuído para a classe Setosa (3 de 3 ocorrências).

Contas das distâncias

Primeiro ponto desconhecido: [1.8 6.3 2.9 7.3]

D1: [1.3 4.4 2.3 6.3]

$$d = \sqrt{(1.3 - 1.8)^2 + (4.4 - 6.3)^2 + (2.3 - 2.9)^2 + (6.3 - 7.3)^2}$$

$$d = \sqrt{(-0.5)^2 + (-1.9)^2 + (-0.6)^2 + (-1)^2}$$

$$d = \sqrt{0.25 + 3.61 + 0.36 + 1}$$

$$d = \sqrt{5.22}$$

$$d = 2.284$$

D2: [2.3 5.4 3.4 6.2]

$$d = \sqrt{(2.3 - 1.8)^2 + (5.4 - 6.3)^2 + (3.4 - 2.9)^2 + (6.2 - 7.3)^2}$$

$$d = \sqrt{2.52}$$

$$d = 1.589$$

D3: [0.2 1.4 3.4 5.2]

$$d = \sqrt{(0.2 - 1.8)^2 + (1.4 - 6.3)^2 + (3.4 - 2.9)^2 + (5.2 - 7.3)^2}$$

$$d = \sqrt{31.23}$$

$$d = 5.588$$

D4: [2.1 5.4 3.1 6.9]

$$d = \sqrt{(2.1 - 1.8)^2 + (5.4 - 6.3)^2 + (3.1 - 2.9)^2 + (6.9 - 7.3)^2}$$

$$d = \sqrt{1.1}$$

$$d = 1.048$$

D5: [0.4 1.5 4.4 5.7]

$$d = \sqrt{(0.4 - 1.8)^2 + (1.5 - 6.3)^2 + (4.4 - 2.9)^2 + (5.7 - 7.3)^2}$$

$$d = \sqrt{29.81}$$

$$d = 5.459$$

D6: [0.2 1.5 3.7 5.4]

$$d = \sqrt{(0.2 - 1.8)^2 + (1.5 - 6.3)^2 + (3.7 - 2.9)^2 + (5.4 - 7.3)^2}$$

$$d = \sqrt{29.85}$$

$$d = 5.464$$

D7: [0.2 1.4 3.3 5.0]

$$d = \sqrt{(0.2 - 1.8)^2 + (1.4 - 6.3)^2 + (3.3 - 2.9)^2 + (5.0 - 7.3)^2}$$

$$d = \sqrt{32.02}$$

$$d = 5.657$$

D8: [2.1 5.6 2.8 6.4]

$$d = \sqrt{(2.1 - 1.8)^2 + (5.6 - 6.3)^2 + (2.8 - 2.9)^2 + (6.4 - 7.3)^2}$$

$$d = \sqrt{1.4}$$

$$d = 1.183$$

D9: [1.8 4.8 3.0 6.0]

$$d = \sqrt{(1.8 - 1.8)^2 + (4.8 - 6.3)^2 + (3.0 - 2.9)^2 + (6.0 - 7.3)^2}$$

$$d = \sqrt{3.95}$$

$$d = 1.987$$

D10: [1.3 4.0 2.5 5.5]

$$d = \sqrt{(1.3 - 1.8)^2 + (4.0 - 6.3)^2 + (2.5 - 2.9)^2 + (5.5 - 7.3)^2}$$

$$d = \sqrt{8.94}$$

$$d = 2.988$$

Segundo ponto desconhecido: [1.4 4.7 2.9 6.1]

D1: [1.3 4.4 2.3 6.3]

$$d = \sqrt{(1.3 - 1.4)^2 + (4.4 - 4.7)^2 + (2.3 - 2.9)^2 + (6.3 - 6.1)^2}$$

$$d = \sqrt{0.5}$$

$$d = 0.707$$

D2: [2.3 5.4 3.4 6.2]

$$d = \sqrt{(2.3 - 1.4)^2 + (5.4 - 4.7)^2 + (3.4 - 2.9)^2 + (6.2 - 6.1)^2}$$

$$d = \sqrt{1.56}$$

$$d = 1.248$$

D3: [0.2 1.4 3.4 5.2]

$$d = \sqrt{(0.2 - 1.4)^2 + (1.4 - 4.7)^2 + (3.4 - 2.9)^2 + (5.2 - 6.1)^2}$$

$$d = \sqrt{13.39}$$

$$d = 3.659$$

D4: [2.1 5.4 3.1 6.9]

$$d = \sqrt{(2.1 - 1.4)^2 + (5.4 - 4.7)^2 + (3.1 - 2.9)^2 + (6.9 - 6.1)^2}$$

$$d = \sqrt{1.66}$$

$$d = 1.288$$

D5: [0.4 1.5 4.4 5.7]

$$d = \sqrt{(0.4 - 1.4)^2 + (1.5 - 4.7)^2 + (4.4 - 2.9)^2 + (5.7 - 6.1)^2}$$

$$d = \sqrt{13.65}$$

$$d = 3.694$$

D6: [0.2 1.5 3.7 5.4]

$$d = \sqrt{(0.2 - 1.4)^2 + (1.5 - 4.7)^2 + (3.7 - 2.9)^2 + (5.4 - 6.1)^2}$$

$$d = \sqrt{12.81}$$

$$d = 3.579$$

D7: [0.2 1.4 3.3 5.0]

$$d = \sqrt{(0.2 - 1.4)^2 + (1.4 - 4.7)^2 + (3.3 - 2.9)^2 + (5.0 - 6.1)^2}$$

$$d = \sqrt{13.7}$$

$$d = 3.701$$

D8: [2.1 5.6 2.8 6.4]

$$d = \sqrt{(2.1 - 1.4)^2 + (5.6 - 4.7)^2 + (2.8 - 2.9)^2 + (6.4 - 6.1)^2}$$

$$d = \sqrt{1.4}$$

$$d = 1.183$$

D9: [1.8 4.8 3.0 6.0]

$$d = \sqrt{(1.8 - 1.4)^2 + (4.8 - 4.7)^2 + (3.0 - 2.9)^2 + (6.0 - 6.1)^2}$$

$$d = \sqrt{0.19}$$

$$d = 0.435$$

D10: [1.3 4.0 2.5 5.5]

$$d = \sqrt{(1.3 - 1.4)^2 + (4.0 - 4.7)^2 + (2.5 - 2.9)^2 + (5.5 - 6.1)^2}$$

$$d = \sqrt{1.02}$$

$$d = 1.009$$

Terceiro ponto desconhecido: [3.0 1.4 3.4 4.6]

D1: [1.3 4.4 2.3 6.3]

$$d = \sqrt{(1.3 - 3.0)^2 + (4.4 - 1.4)^2 + (2.3 - 3.4)^2 + (6.3 - 4.6)^2}$$

$$d = \sqrt{15.99}$$

$$d = 3.998$$

D2: [2.3 5.4 3.4 6.2]

$$d = \sqrt{(2.3 - 3.0)^2 + (5.4 - 1.4)^2 + (3.4 - 3.4)^2 + (6.2 - 4.6)^2}$$

$$d = \sqrt{19.05}$$

$$d = 4.364$$

D3: [0.2 1.4 3.4 5.2]

$$d = \sqrt{(0.2 - 3.0)^2 + (1.4 - 1.4)^2 + (3.4 - 3.4)^2 + (5.2 - 4.6)^2}$$

$$d = \sqrt{8.2}$$

$$d = 2.863$$

D4: [2.1 5.4 3.1 6.9]

$$d = \sqrt{(2.1 - 3.0)^2 + (5.4 - 1.4)^2 + (3.1 - 3.4)^2 + (6.9 - 4.6)^2}$$

$$d = \sqrt{22.19}$$

$$d = 4.710$$

D5: [0.4 1.5 4.4 5.7]

$$d = \sqrt{(0.4 - 3.0)^2 + (1.5 - 1.4)^2 + (4.4 - 3.4)^2 + (5.7 - 4.6)^2}$$

$$d = \sqrt{8.98}$$

$$d = 2.996$$

D6: [0.2 1.5 3.7 5.4]

$$d = \sqrt{(0.2 - 3.0)^2 + (1.5 - 1.4)^2 + (3.7 - 3.4)^2 + (5.4 - 4.6)^2}$$

$$d = \sqrt{8.58}$$

$$d = 2.929$$

D7: [0.2 1.4 3.3 5.0]

$$d = \sqrt{(0.2 - 3.0)^2 + (1.4 - 1.4)^2 + (3.3 - 3.4)^2 + (5.0 - 4.6)^2}$$

$$d = \sqrt{8.01}$$

$$d = 2.830$$

D8: [2.1 5.6 2.8 6.4]

$$d = \sqrt{(2.1 - 3.0)^2 + (5.6 - 1.4)^2 + (2.8 - 3.4)^2 + (6.4 - 4.6)^2}$$

$$d = \sqrt{22.05}$$

$$d = 4.695$$

D9: [1.8 4.8 3.0 6.0]

$$d = \sqrt{(1.8 - 3.0)^2 + (4.8 - 1.4)^2 + (3.0 - 3.4)^2 + (6.0 - 4.6)^2}$$

$$d = \sqrt{15.12}$$

$$d = 3.888$$

D10: [1.3 4.0 2.5 5.5]

$$d = \sqrt{(1.3 - 3.0)^2 + (4.0 - 1.4)^2 + (2.5 - 3.4)^2 + (5.5 - 4.6)^2}$$

$$d = \sqrt{11.27}$$

$$d = 3.357$$