LAB-5 K Nearest Neighbours

KNN (K Nearest neighbours)

Consider the following dataset for k=3 and test data (x, 35, 100) as (Person, Age, Salary) and predict the target.

Person	Age	Salary K	Distance	Rank	target
A	18	50	52.8		
B	23	55	46.6		
C	24	70	31.9	2	N
D	41	60	40.4	3	Y
E	43	70	31.1	1	Y
F	38	40	60.1		

Step 1: Distance (d) = $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

$x_2, y_2 = (35, 100)$

$d_1 = \sqrt{(35-18)^2 + (100-50)^2} = 52.8$

$d_2 = \sqrt{(35-23)^2 + (100-55)^2} = 46.6$

Step 2: Identify 3 nearest neighbours

1. E(31.1, Y)
2. C(31.9, N)
3. D(40.4, Y)

Step 3: Majority voting

Since 2 out of 3 belong to class 'Y'

∴ the predicted class for X(35,10) is 'Y'

For iris dataset
How to choose the k value? Demonstrate using accuracy rate and error rate

steps to choose k using accuracy rate & error rate

1) Split the dataset (training (70%), testing (30%))
2) Train KNN with different k value k={1,3,5...}
3) Calculate accuracy
4) Accuracy = $\frac{\text{Correct predictions}}{\text{Total predictions}} \times 100$
4) Calculate error rate = Error rate = 1 - accuracy
5) Plot accuracy vs k

For diabetes dataset
What is the purpose of feature scaling?
How to perform it?

→. Purpose of feature scaling
1) Diabetes dataset has features like glucose level, BMI, and age which have different ranges
2) Machine learning algorithms perform better when features are on similar scale
3) It improves convergence speed
4) Prevent dominant features from biasing the model

→ ~~Metho~~.

→ Methods to perform feature scaling
1) Min-Max scaling

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2) Standardization

$$x' = \frac{x - \mu}{\sigma}$$

09.04.2021