Lab 6:

KNN (K Nearest Neighbours)

- Consider the following dataset, for K=3 and test data (X, 35, 100) as (Person, Age, Salary K) & predict the target.

Person	Age	Salary K	Distance(d)	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:

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

$x_2, y_2 = (35, 100)$

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

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

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 (target) for X (35, 100) is 'Y'.

For iris dataset:

How to choose the K value?

1) Accuracy rate approach:
- We train the model with diff K values and calculate the accuracy for each K.

2) Error Rate Approach:

Error Rate = 1 − Accuracy.

- A lower error rate indicates a better K value.

Demonstration of Accuracy Rate & Error Rate:
- Small K values may lead to overfitting.
- Large K values may lead to underfitting.

For Diabetes Dataset:

1) What is the purpose of Feature Scaling?

→ Feature Scaling is used to normalize the range of independent variables.

2) How to perform feature scaling?

1. Standardization $X_{scaled} = \frac{X - \mu}{\sigma}$

μ – mean

σ – standard deviation

Used when data follows a normal distribution.

09.04.