

2/14/25

Lab-5KNNClassification

Qn1 Consider the following dataset, for $K=3$ & test data $(X, 35, 100)$ as (Person, Age, Salary K) solve using Knn Classifier & predict the target.

Person	Age	Salary K	Target	Distance	Rank
A	18	50	N	52.8	
B	23	55	N	46.6	
C	24	70	N	31.9	2
D	41	60	Y	40.4	3
E	43	70	Y	31.1	1
F	38	40	Y	60.1	
X	35	100	?		

→ Euclidean distance: $d = \sqrt{(Age_2 - Age_1)^2 + (Salary_2 - Salary_1)^2}$

Step-1

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

$$B \rightarrow 46.6$$

$$C \rightarrow 31.9$$

$$D \rightarrow 40.4$$

$$E \rightarrow 31.1$$

$$F \rightarrow 60.1$$

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:

→ Yes

→ $X(35, 100)$ is 'Y'

Qn: For iris dataset

1. How to choose the k value? Demonstrate using accuracy rate & error rate.

Qn: 2. For diabetes dataset.

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

→ 1. Train KNN with different k values (eg. 1 to 20)

- * Measure accuracy (higher is better) & error rate (lower is better)
- * The best k is where accuracy is highest & error rate is lowest (typically 5-10)

→ 2. KNN relies on distance, so large scaling features can dominate the smaller ones.

- * Standardization ensures all features contribute equally.
- * Scaling improves accuracy & prevents biased predictions.