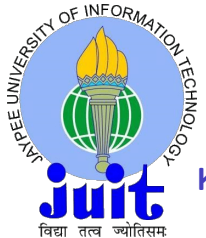

Machine Learning

(Course Code: 18B1WCI634)

KNN: Instance Based Learning

Mr. Sandeep Kumar Patel

Assistant Professor

Jaypee University of Information Technology

Waknaghat, Solan, HP-173234

Instance-Based Learning

Definition

Instance-Based Learning (IBL) is a learning approach in which the model does not build an explicit general model during training.

Instead, it:

- Stores training examples (instances)
- Compares new examples with stored instances
- Predicts output based on similarity

Key Idea

"Similar inputs produce similar outputs"

Instance-Based Learning

Characteristics

- Lazy learning approach
- No explicit training phase
- Uses distance/similarity measures
- Learns locally instead of globally

Examples:

- k-Nearest Neighbor (k-NN)
- Case-Based Reasoning

k-Nearest Neighbor (k-NN) Algorithm

k-NN is the most popular instance-based learning algorithm.

Working Principle

1. Store all training examples
2. Choose value of k
3. Compute distance between test sample and all training samples
4. Select k nearest neighbors
5. Predict class using majority voting

k-Nearest Neighbor (k-NN) Algorithm

Distance Function

Most common: Euclidean Distance

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Prediction Rule

$$\hat{y} = \arg\max_{c_j} \sum_{x_i \in N_k} I(y_i = c_j)$$

where:

- N_k = set of k nearest neighbors
- $I(\cdot)$ = indicator function

Example of k-NN Classification

Suppose:

$$k = 3$$

Training samples belong to two classes:

- Red circles = Class A
- Blue squares = Class B

For a new test point:

1. Compute distances to all points
2. Select 3 nearest neighbors
3. Count majority class

Example of k-NN Classification

Example:

Nearest neighbors: A, A, B

Majority class:

$$\boxed{\text{Predicted Class} = \text{A}}$$

Effect of k

- Small k: sensitive to noise
- Large k: smoother decision boundary

Instance-Based Learning vs Other Learning Models

Feature	Instance-Based Learning	Model-Based Learning
Learning Style	Lazy Learning	Eager Learning
Training Phase	Stores examples only	Builds explicit model
Prediction Time	Slow	Fast
Memory Usage	High	Lower
Generalization	Local	Global
Examples	k-NN, Case-Based Learning	Decision Tree, SVM, Neural Network

Important Difference

IBL compares with stored examples instead of learning explicit rules

Advantages and Disadvantages of k-NN

Advantages

- Simple and easy to implement
- No training phase
- Effective for small datasets
- Naturally handles multi-class problems

Disadvantages

- Slow prediction for large datasets
- Requires large memory
- Sensitive to irrelevant features
- Performance depends on choice of k

Advantages and Disadvantages of k-NN

Applications

- Recommendation systems
- Pattern recognition
- Medical diagnosis
- Image classification

Numerical Example of k-NN Classification

Training dataset:

Point	Coordinates	Class
A	(1,1)	Red
B	(2,2)	Red
C	(3,3)	Blue
D	(6,5)	Blue
E	(7,7)	Blue

Test sample:

$$X = (3, 2)$$

Choose:

$$k = 3$$

Goal:

Predict class of X

Step 1: Compute Distances

Use Euclidean Distance:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Distances from test point $X = (3, 2)$:

$$d(X, A) = \sqrt{(3-1)^2 + (2-1)^2}$$

$$= \sqrt{4+1} = \sqrt{5} = 2.23$$

Step 1: Compute Distances

$$d(X, B) = \sqrt{(3-2)^2 + (2-2)^2}$$

$$= \sqrt{1} = 1$$

$$d(X, C) = \sqrt{(3-3)^2 + (2-3)^2}$$

$$= \sqrt{1} = 1$$

Step 2: Remaining Distances

$$d(X, D) = \sqrt{(3-6)^2 + (2-5)^2}$$

$$= \sqrt{9+9} = \sqrt{18} = 4.24$$

$$d(X, E) = \sqrt{(3-7)^2 + (2-7)^2}$$

$$= \sqrt{16+25} = \sqrt{41} = 6.40$$

Step 2: Remaining Distances

Distance Summary

Point	Distance	Class
A	2.23	Red
B	1.00	Red
C	1.00	Blue
D	4.24	Blue
E	6.40	Blue

Step 3: Select k Nearest Neighbors

For:

$$k = 3$$

Nearest neighbors are:

Point	Class
B	Red
C	Blue
A	Red

Step 3: Select k Nearest Neighbors

Count classes:

$$\text{Red} = 2$$

$$\text{Blue} = 1$$

Majority class:

$$\boxed{\text{Predicted Class} = \text{Red}}$$

Observations from Numerical Example

Important Points

- k-NN does not create an explicit model
- Entire training dataset is stored
- Prediction depends on nearby examples

Effect of Different k

- $k = 1$: very sensitive to noise
- Small k: complex decision boundary
- Large k: smoother classification

Key Insight

k-NN classifies based on similarity to nearest stored instances