Elective Course
Course Code: CS4103
Autumn 2025-26

Lecture #34

# Artificial Intelligence for Data Science

**Week-9:**

**MACHINE LEARNING (Part II)**

**Classification using K-Nearest Neighbor (KNN) Algorithm**

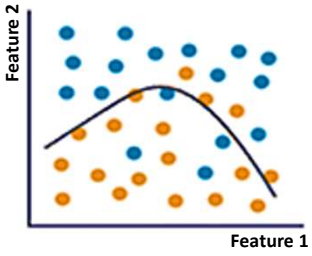
**Course Instructor:**

**Dr. Monidipa Das**
**Assistant Professor**
**Department of Computational and Data Sciences**
**Indian Institute of Science Education and Research Kolkata, India 741246**

## Classification

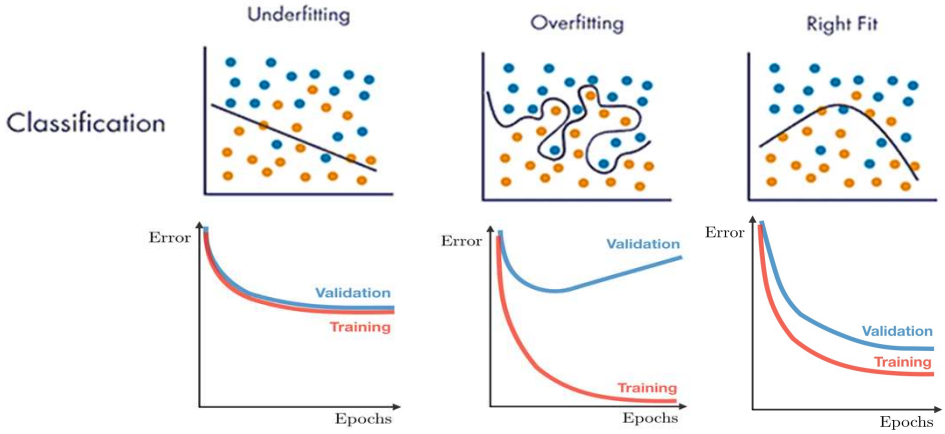
2

- Computer program asked to specify which of $k$ categories some input belongs to
  - Learning algorithm is asked to produce a function $f: \mathbb{R}^n \rightarrow \{1, .., k\}$, where $n$ =no of input variables
    - When $y = f(\boldsymbol{x})$ model assigns input vector $\boldsymbol{x}$ to a category identified by a numeric code $y$
- Other variants of classification task:
  - $f$ outputs a probability distribution over classes



# Generalization, Underfitting and Overfitting

In ML, **generalization** is the ability to perform well on previously unobserved inputs

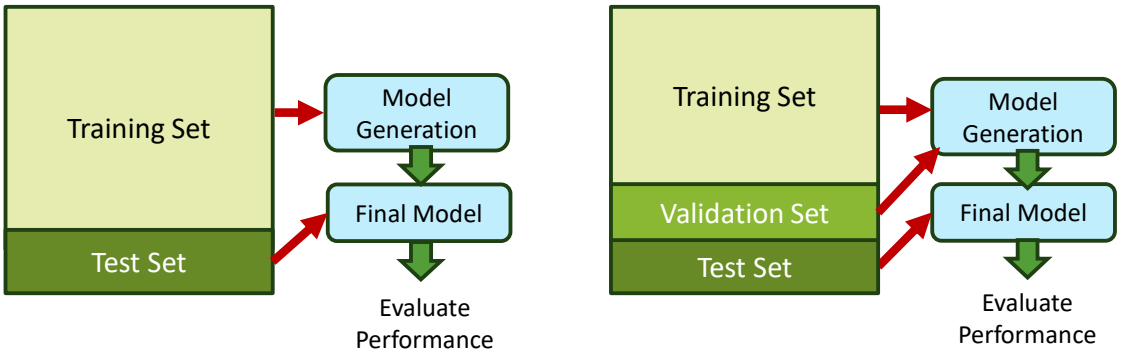



**Underfitting:** Inability to obtain low enough error rate on the training set
**Overfitting:** Gap between training error and testing error is too large



# Data Splitting

- **Training Set**: For training of the model.
  **Validation Set**: For unbiased evaluation of the model.
  **Test Set**: For final evaluation of the model.



# How does classification work?


- Classification determines the class label for an unlabeled data point (test case)

| Age | Gend | Cp | Trestbps | Chol | Fbs | Thalach | Oldpeak | Slope | Ca | Hdisease |
|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 1 | 145 | 233 | 1 | 150 | 2.3 | 3 | 0 | 0 |
| 67 | 1 | 4 | 160 | 286 | 0 | 108 | 1.5 | 2 | 3 | 1 |
| 67 | 1 | 4 | 120 | 229 | 0 | 129 | 2.6 | 2 | 2 | 1 |
| 37 | 1 | 3 | 130 | 250 | 0 | 187 | 3.5 | 3 | 0 | 0 |
| 41 | 0 | 2 | 130 | 204 | 0 | 172 | 1.4 | 1 | 0 | 0 |
| 56 | 1 | 2 | 120 | 236 | 0 | 178 | 0.8 | 1 | 0 | 0 |

Modeling

Classifier

Input feature

| Age | Gend | Cp | Trestbps | Chol | Fbs | Thalach | Oldpeak | Slope | Ca | Hdisease |
|---|---|---|---|---|---|---|---|---|---|---|
| 62 | 0 | 4 | 140 | 268 | 0 | 160 | 3.6 | 3 | 2 | ? |

Predicted Label/class



# Multi-class Classification: Example


| Age | Gend | BP | Chol | Nal | K | Drug |
|---|---|---|---|---|---|---|
| 23 | F | HIGH | HIGH | 0.793 | 0.031 | drugY |
| 47 | M | LOW | HIGH | 0.739 | 0.056 | drugC |
| 47 | M | LOW | HIGH | 0.597 | 0.069 | drugC |
| 28 | F | NORM | HIGH | 0.564 | 0.072 | drugX |
| 61 | F | LOW | HIGH | 0.559 | 0.031 | drugY |
| 22 | F | NORMAL | HIGH | 0.677 | 0.079 | drugY |

Modeling

Classifier

Input feature

| Age | Gend | BP | Chol | Nal | K | Drug |
|---|---|---|---|---|---|---|
| 36 | F | LOW | HIGH | 0.697 | 0.069 | ? |

Predicted Label/class

# Evaluation Metrics for Classification

| Age | Gend | Cp | Trestbps | Chol | Fbs | Thalach | Oldpeak | Slope | Ca | Hdisease |
|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 1 | 145 | 233 | 1 | 150 | 2.3 | 3 | 0 | 0 |
| 67 | 1 | 4 | 160 | 286 | 0 | 108 | 1.5 | 2 | 3 | 1 |
| 67 | 1 | 4 | 120 | 229 | 0 | 129 | 2.6 | 2 | 2 | 1 |
| 37 | 1 | 3 | 130 | 250 | 0 | 187 | 3.5 | 3 | 0 | 0 |
| 43 | 1 | 4 | 150 | 247 | 0 | 171 | 1.5 | 1 | 0 | 0 |
| 56 | 1 | 2 | 120 | 236 | 0 | 178 | 0.8 | 1 | 0 | 0 |
| 40 | 1 | 4 | 110 | 167 | 0 | 114 | 2 | 2 | 0 | 1 |
| 69 | 0 | 1 | 140 | 239 | 0 | 151 | 1.8 | 1 | 2 | 0 |

Training Set

Modeling

Classifier

Predicted Label

Input feature

| Age | Gend | Cp | Trestbps | Chol | Fbs | Thalach | Oldpeak | Slope | Ca | Hdisease |
|---|---|---|---|---|---|---|---|---|---|---|
| 62 | 0 | 4 | 140 | 268 | 0 | 160 | 3.6 | 3 | 2 | 1 |
| 40 | 1 | 4 | 110 | 167 | 0 | 114 | 2 | 2 | 0 | 1 |
| 60 | 1 | 4 | 130 | 206 | 0 | 132 | 2.4 | 2 | 2 | 1 |

Test Set

Actual Label ($y$)

| Hdisease |
|---|
| 1 |
| 0 |
| 1 |

Predicted Label ($\hat{y}$)



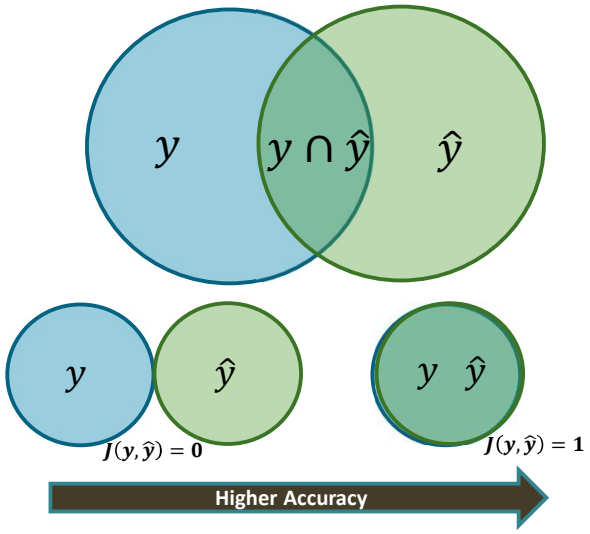
# Evaluation Metrics for Classification: Jaccard Index

- $y$: Actual labels
- $\hat{y}$: Predicted labels

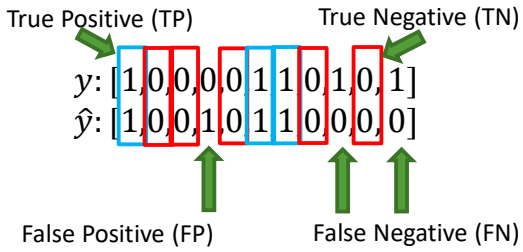
$$J(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

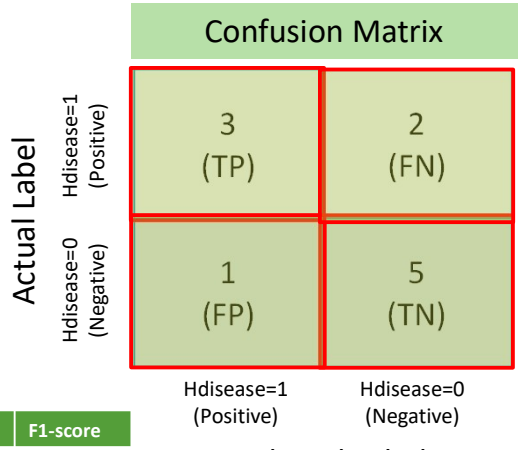
$y$: [1,0,0,0,0,1,1,0,1,0,1]
$\hat{y}$: [1,0,0,1,0,1,1,0,0,0,0]

$$J(y,\hat{y}) = \frac{8}{11 + 11 - 8} = 0.57$$



# Evaluation Metrics for Classification: F1-Score


$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{(Precision + Recall)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

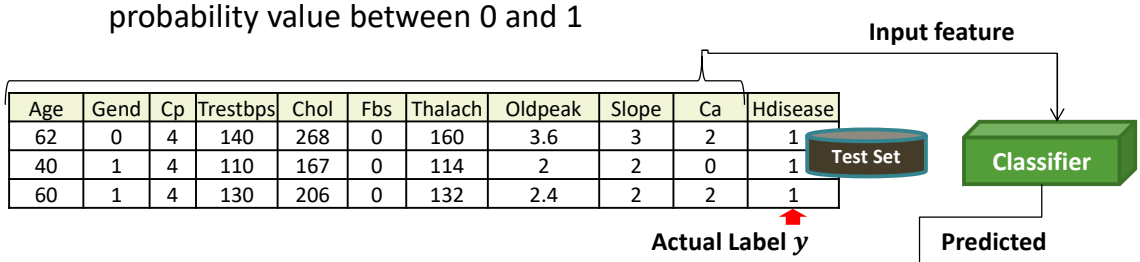
| | Precision | Recall | F1-score |
|---|---|---|---|
| Hdisease=1 | 0.75 | 0.60 | 0.67 |
| Hdisease=0 | 0.71 | 0.83 | 0.77 |

Average → 0.72

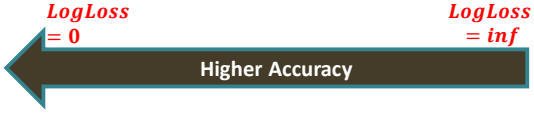


# Evaluation Metrics for Classification: Log Loss

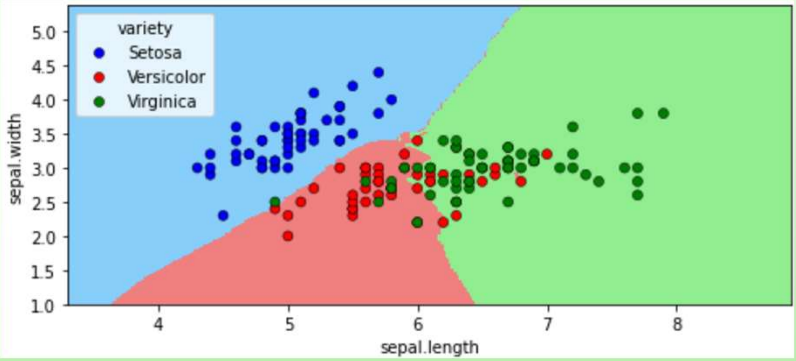

- Performance of a classifier where the predicted output is a probability value between 0 and 1




$$LogLoss = -\frac{1}{n}\sum(y \times log(\hat{y}) + (1 - y) \times log(1 - \hat{y}))$$



# K-Nearest Neighbors (KNN)







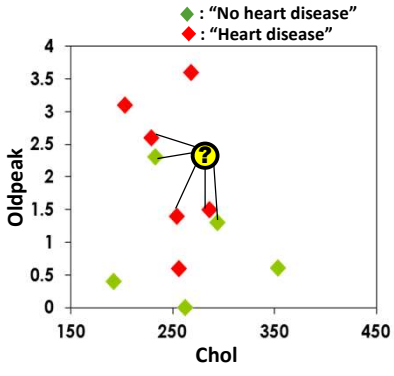

# Determining class using KNN

| Age | Gend | Cp | Trestbps | Chol | Fbs | Thalach | Oldpeak | Slope | Ca | Hdisease |
|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 1 | 145 | 233 | 1 | 150 | 2.3 | 3 | 0 | 0 |
| 67 | 1 | 4 | 160 | 286 | 0 | 108 | 1.5 | 2 | 3 | 1 |
| 67 | 1 | 4 | 120 | 229 | 0 | 129 | 2.6 | 2 | 2 | 1 |
| 37 | 1 | 3 | 130 | 250 | 0 | 187 | 3.5 | 3 | 0 | 0 |
| 41 | 0 | 2 | 130 | 204 | 0 | 172 | 1.4 | 1 | 0 | 0 |
| 56 | 1 | 2 | 120 | 236 | 0 | 178 | 0.8 | 1 | 0 | 0 |

| Age | Gend | Cp | Trestbps | Chol | Fbs | Thalach | Oldpeak | Slope | Ca | Hdisease |
|---|---|---|---|---|---|---|---|---|---|---|
| 54 | 1 | 4 | 124 | 266 | 0 | 109 | 2.2 | 2 | 1 | ? |

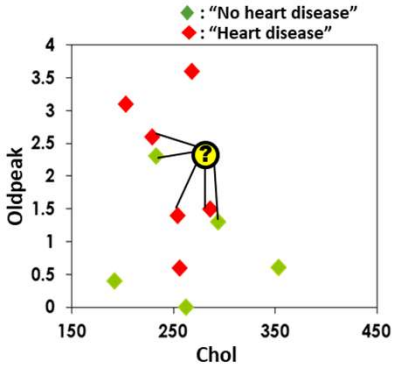
**0** → "No heart disease" ⟸ **1-NN**

**1** → "Heart disease" ⟸ **5-NN**



# K-Nearest Neighbors (KNN)


- A method for **classifying** cases/data-points based on their similarity to others cases/data-points
- Data-points that are near to each other are said to be "**neighbors**"
- Assume data-points with the same class labels are near to each other.





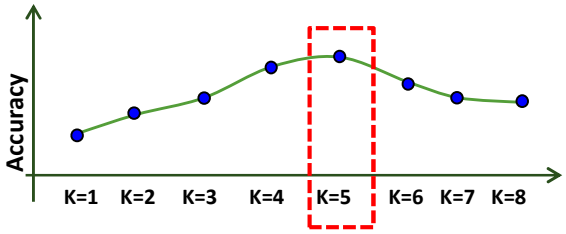

# KNN Algorithm (basic steps)

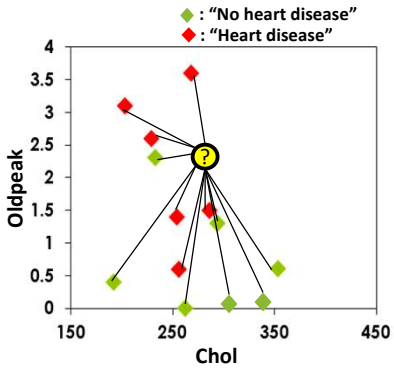

- Pick a value for K
- Calculate the distance of unknown data-points from all the known data-points
- Select K known data-points that are "nearest" to the unknown data-points
- Get the labels of the K selected data-points
- Return the mode of the K labels

# Finding the best value for K

15

- K=1 Class label 0
- K=7 Class label 1
- K=13 ?




**Better to avoid even number** as a value for K in the KNN algorithm **to prevent ties**



# Calculating Distance

16

- Most common way--- Euclidean Distance

$x_1 = (x_{11}, x_{12}, \dots, x_{1m})$

$x_2 = (x_{21}, x_{22}, \dots, x_{2m})$

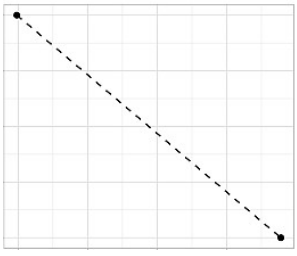
$$EDist(x_1, x_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \cdots + (x_{1m} - x_{2m})^2}$$

$$= \sqrt{\sum_{i=1}^{m} (x_{1i} - x_{2i})^2}$$

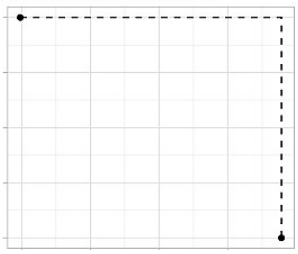
$$MDist(x_1, x_2) = \sum_{i=1}^{m} |x_{1i} - x_{2i}|$$

Minkowski distance: $Dist(x_1, x_2) = \left[\sum_{i=1}^{m} |x_{1i} - x_{2i}|^p\right]^{1/p}$

(A) Euclidean distance



(B) Manhattan distance



# Calculating Distance

17

Patient-1 ($x_1$)

Patient-2 ($x_2$)

| Age | Chol | Oldpeak |
|---|---|---|
| 54 | 236 | 0.8 |
| 67 | 284 | 1.5 |

$$EDist(x_1, x_2) = \sqrt{(54-67)^2+(236-284)^2+(0.8-1.5)^2} = 49.7$$

$$MDist(x_1, x_2) = |54-67| + |236-284| + |0.8-1.5| = 61.7$$



# Measuring Distance for Categorical Features

18

- **Hamming Distance**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Data point 1 | **0** | **0** | **0** | **1** | **1** | **0** | **0** |
| Data point 2 | **1** | **1** | **0** | **1** | **1** | **1** | **0** |

Hamming distance is: 3

**When all the attributes/features are categorical:**

| | Color | Shape | Size |
|---|---|---|---|
| *Data point 1* | Red | Circle | Small |
| *Data point 2* | Green | Rectangle | Small |

mismatch mismatch match

Hamming distance is: 1 + 1 + 0 = 2

**Using One-Hot Encoding for Categorical Attributes**

| | Color | Weight | Length |
|---|---|---|---|
| *Data point 1* | Red | 40.5 | 5 |
| *Data point 2* | Green | 53 | 3 |
| *Data point 3* | Blue | 78.5 | 15 |

[0,0,1] Red
[0,1,0] Green
[1,0,0] Blue

Hamming distance is: 2

# Feature Scaling

19

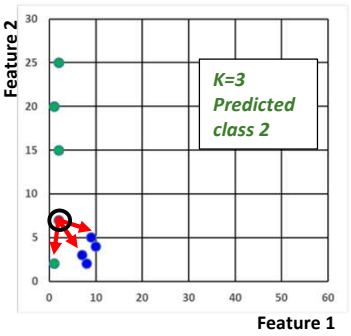
**Dataset**

| Feature1 | Feature2 | Class |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 15 | 1 |
| 1 | 20 | 1 |
| 2 | 25 | 1 |
| 10 | 4 | 2 |
| 9 | 5 | 2 |
| 7 | 3 | 2 |
| 8 | 2 | 2 |
| 2 | 7 | ? |

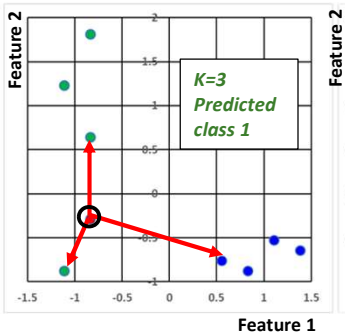
$$x' = \frac{x - \bar{x}}{\sigma} \qquad x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

*class 1*
*class 2*

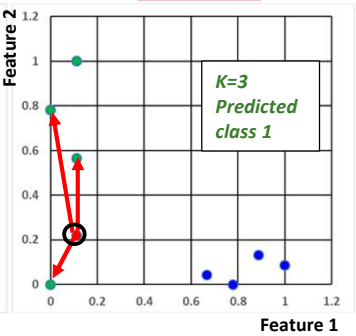
**Original Features**




**Features after Standardization (Z-score Normalization)**




**Features after Min-Max Normalization**






20

# Questions?