The University of Queensland
Australia
CREATE CHANGE

Machine Leaning

COMP4702/COMP7703

Prac 3

Different types of Learning

- **Supervised Learning** :

(y_i is available for all x_i)

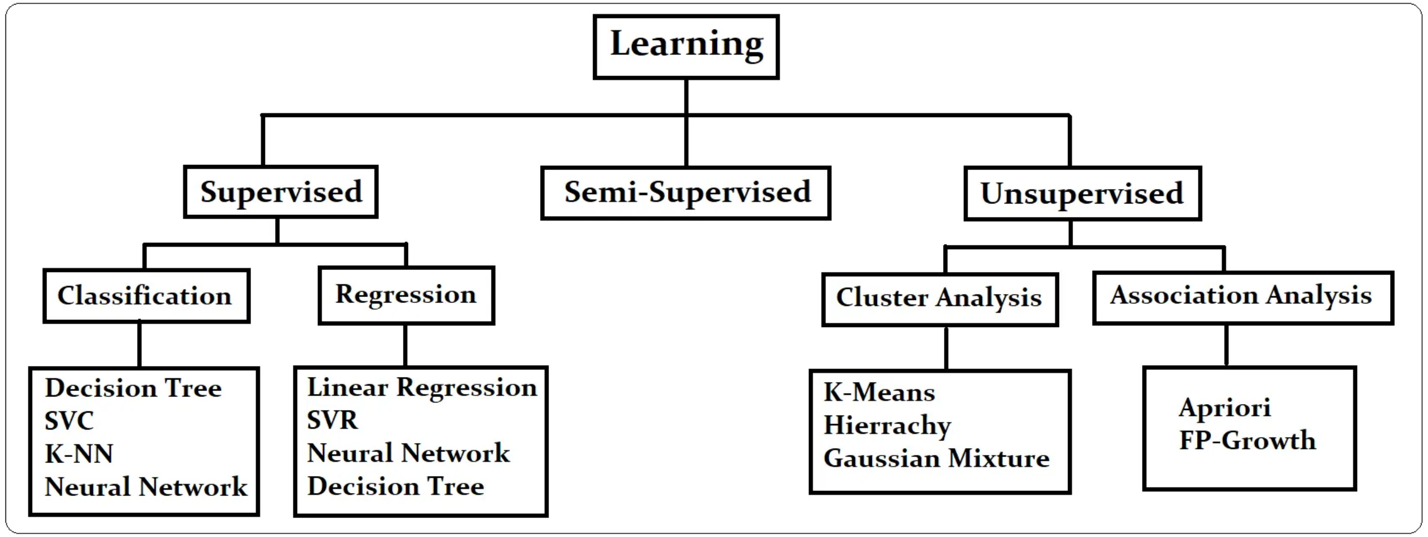
- classification: quantitative y_i
- regression: categorical y_i

- **Unsupervised Learning**:

(y_i is unavailable for all x_i)

- **Semi-Supervised Learning**:

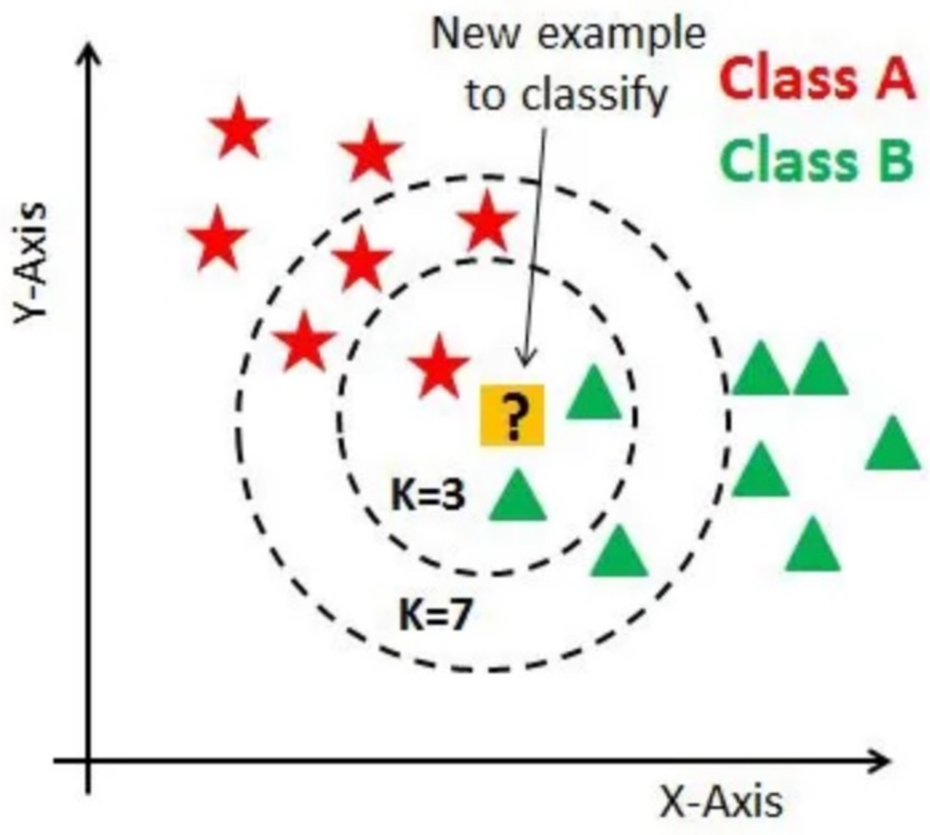
(y_i is unavailable for some x_i)



K-Nearest Neighbour (k-NN)

How does k-nn work:

1. Calculate distances
2. Find neighbours
3. Majority Vote / Averaging



K-Nearest Neighbour (k-NN)

Choosing a k value

- **Small k value** – wriggled decision boundary – Overfitting – Sensitivity to Noise
- **Large k value** – smooth decision boundary – Underfitting – Robust to Noise

K-Nearest Neighbour (k-NN)

Calculating Distance

- **quantitative** features:

  Euclidean Distance, Manhattan Distance, Mahalanobis Distance, …

- **Categorical** features :

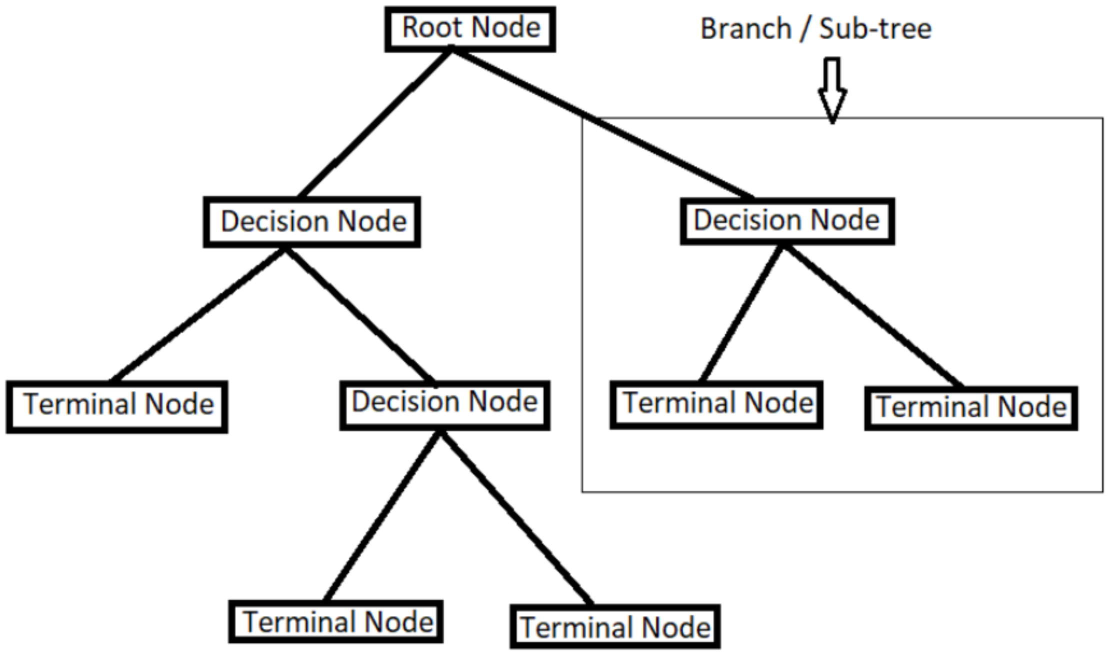
  Hamming Distance, Jaccard Similarity, ...

- **Normalisation** or **standardisation** is advised.

Decision Tree

Structure:

- **Root Node**: Represents the entire dataset. It is from this node that the initial splitting starts.
- **Decision/Internal Nodes**: Nodes that occur between the root node and the leaf nodes. Each represents a “if-the-else” statement.
- **Leaf/Terminal Node**: Nodes that do not split further, representing the outcome or decision.



Decision Tree

Algorithm: Recursive Binary Partitioning

1. All observations in a single set
2. Sort values on first variable
3. Compute split criteria for all possible splits into two sets
4. Choose the best split on this variable
5. Repeat 2-4 for all other variables
6. Choose the best split among all variables. Your data is now in two sets.
7. Repeat 1-6 on each subset.
8. Stop when stopping rule is achieved.

Decision Tree

Split Criteria:

- Classification

  - The Gini index measures total variance across the K classes, for subset m:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

  - Entropy is defined as

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk})$$

  - If all $\hat{p}_{mk}$'s close to zero or one, G and D are small. Lower is better!

- Regression

  - Split the data where combining MSE for left bucket (MSE_L) and right bucket (MSE_R), makes the biggest reduction from the overall MSE:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Decision Tree

Stopping Rules:

- **max_depth**: The maximum depth of the tree.
- **min_split**: The minimum number of samples required to split an internal node.
- **min_samples_leaf**: The minimum number of samples required to be at a leaf node.
- **max_leaf_nodes**: The maximum number of leaf nodes a tree can have.
- **min_impurity_decrease**: A node will be split if this split induces a decrease of the impurity greater than or equal to this value.