Introduction to Classification Models in Machine Learning

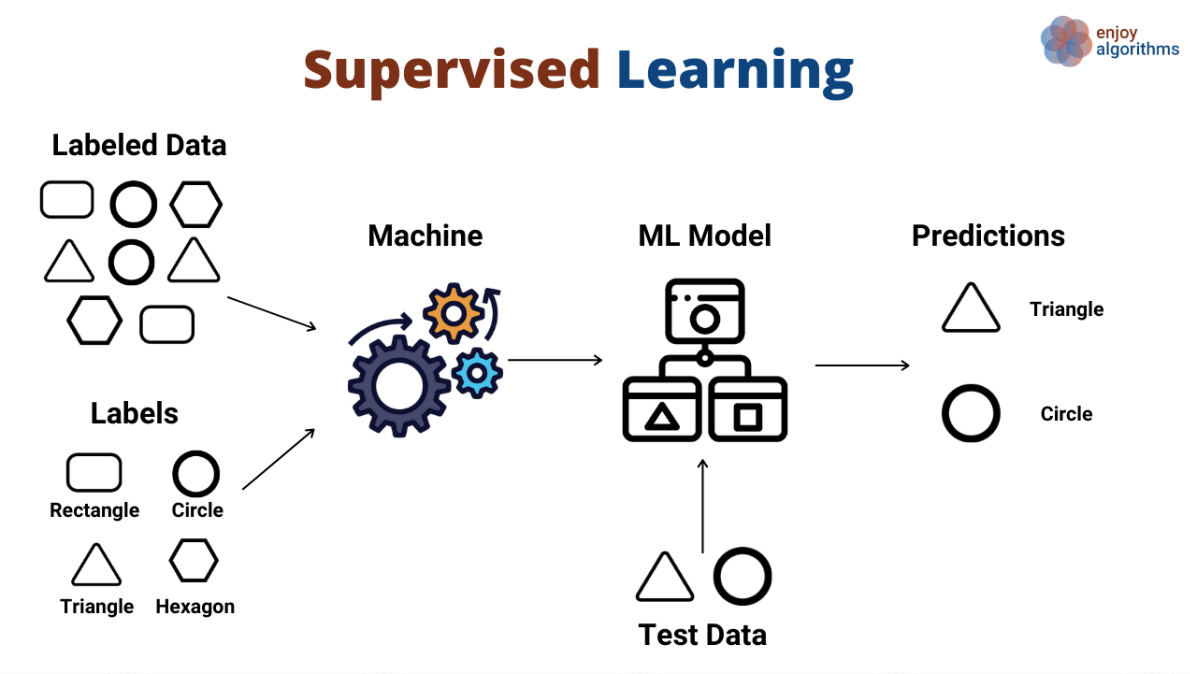
Alişan Köroğlu
2023900252
CSE5007 Machine Learning And Reasoning

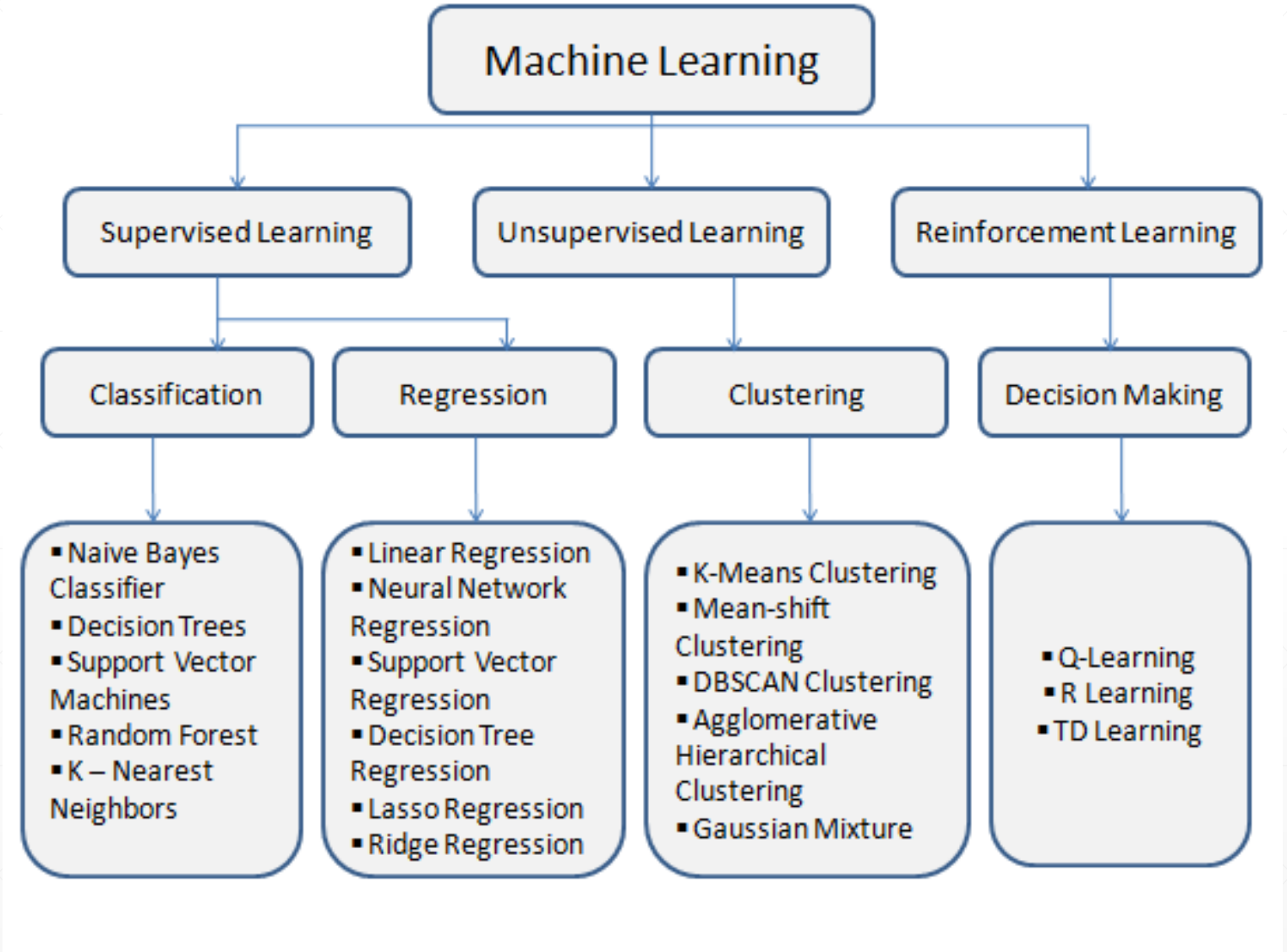
Agenda

- What is Classification?
- Importance of Classification in Machine Learning
- Classification Process
- Common Classification Models
- Performance Metrics

What is Classification?

- Classification is a type of supervised learning.
- Goal: To predict the category or class of a given data point
- Examples:
  - Is this mail spam or not? (Binary Classification)
  - What is the type of flower? (Multi-class Classification)




Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Classification
Regression
Clustering
Decision Making
▪ Naive Bayes Classifier
▪ Decision Trees
▪ Support Vector Machines
▪ Random Forest
▪ K – Nearest Neighbors
▪ Linear Regression
▪ Neural Network Regression
▪ Support Vector Regression
▪ Decision Tree Regression
▪ Lasso Regression
▪ Ridge Regression
▪ K-Means Clustering
▪ Mean-shift Clustering
▪ DBSCAN Clustering
▪ Agglomerative Hierarchical Clustering
▪ Gaussian Mixture
▪ Q-Learning
▪ R Learning
▪ TD Learning

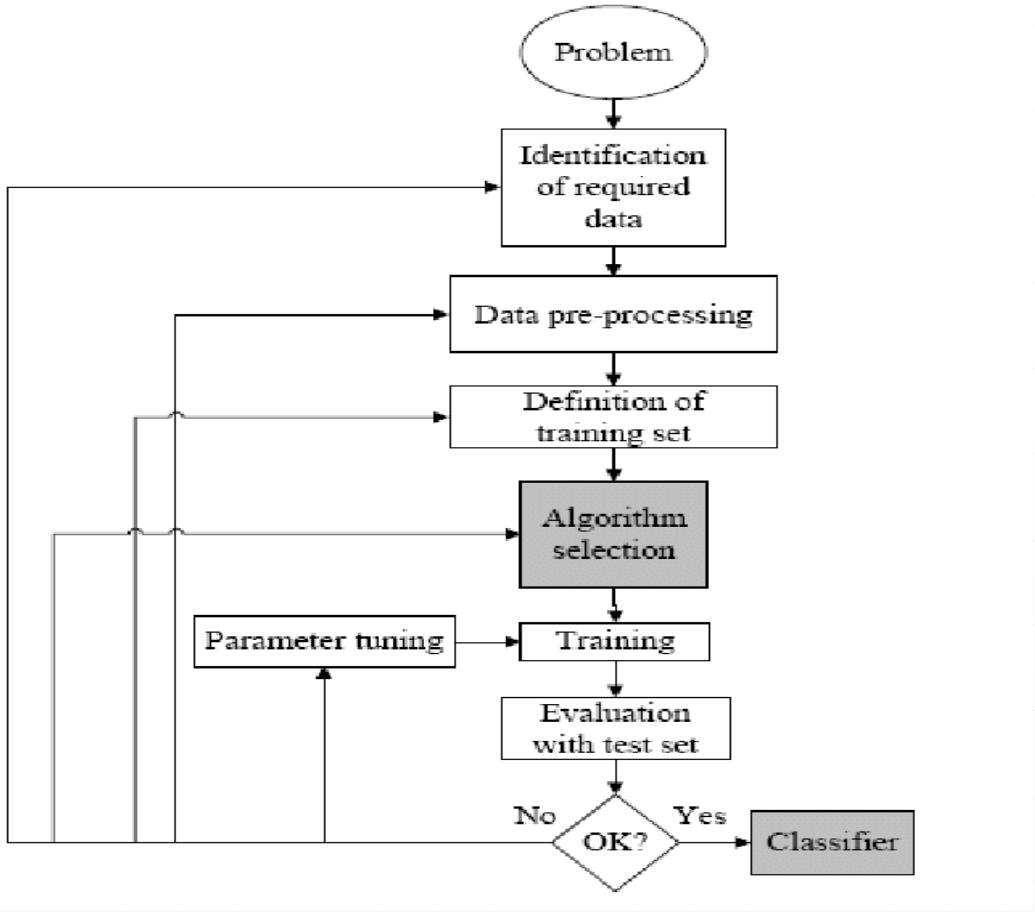
Classification Problems

- **Binary Classification**: It refers to the classification tasks having two class labels such as “true and false” or “yes and no” .
  - The presence (positive) or absence (negative) of a disease
- **Multi-Class Classification**: This refers to those classification tasks having more than two class labels.
  - Classifying an image of fruit as "apple", "banana", or "orange".
- **Multi-Label Classification**: Examples are associated with several classes or labels
  - Including an email in both the "important" and "business" categories.

Why is Classification Important?

- Classification is of critical importance in machine learning and data analytics because many real-world problems require the assignment of an object, event, or situation to specific categories or classes..
  - Healthcare.
  - Finance
  - Marketing
  - Technology
  - Safety and Security

Classification Process



Common Classification Models

- K-Nearest Neighbors (KNN)
- Support Vector Machine(SVM)
- Decision Trees
- Naive Bayes
- Neural Networks

Naive Bayes Algorithm

- The naïve Bayes algorithm is a family of probabilistic classification algorithms used for tasks like text classification, such as spam filtering and sentiment analysis. **It assumes that features are independent of each other, meaning the presence or absence of one feature doesn't impact the probability of another feature.** This assumption, though oversimplified, allows naïve Bayes classifiers to perform well in practice while being computationally efficient.

Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Thomas Bayes
1702 - 1761

Naive Bayes Algorithm

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

- **According to bayes theorem, P(A|B) = (P(B|A) * P(A)) / P(B)**
- Y= Play Tennis, N= Not Play Tennis
- B= {outlook = sunny, temperature = cool, humidity = normal, and windy = True}
- P(Y|B) = (P(B|Y) * P(Y)) / P(B),
  P(N|B) = (P(B|N) * P(N)) / P(B),
- P(B|Y) = 0 and P(B|N) =0

Naive Bayes Algorithm

- **P(sunny ∩ cool ∩ normal ∩ True | Yes) = P(sunny | Yes) * P(cool | Yes) * P(normal | Yes) * P(True | Yes)**
- **P(sunny ∩ cool ∩ normal ∩ True | No) = P(sunny | No) * P(cool | No) * P(normal | No) * P(True | No)**
- P(sunny | Yes) = 2/9, P(cool | Yes) = 3/9, P(Normal | Yes) = 6/9, P(True | Yes) = 3/9
- P(Y|B) = 2/9 * 3/9 * 6/9 * 3/9 * 9/14 = 972/91854
- P(sunny | No) = 3/5, P(cool | No) = 1/5, P(Normal | No) = 1/5, P(True | No) = 3/5
- P(Y|B) = 3/5 * 1/5 * 1/5* 3/5 * 5/14 = 45/8705

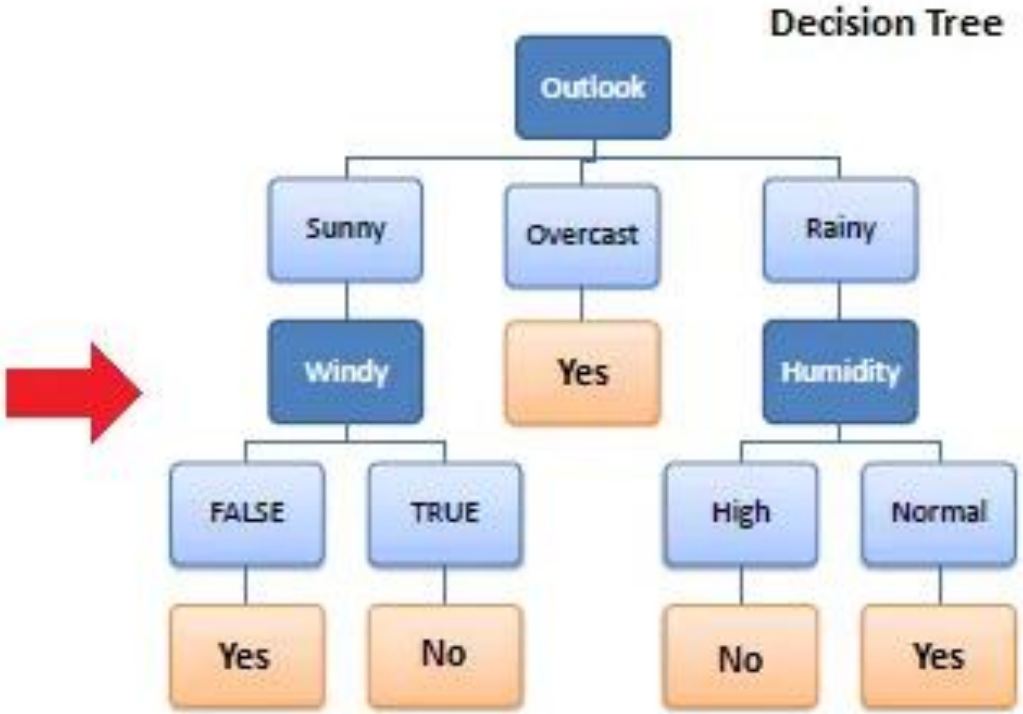
Naive Bayes Algorithm

- NB classification speed is fast, has great advantages when processing large data samples, and supports incremental operations, which can train new samples. But the fatal weakness of NB is the use of the assumption of independence of sample attributes. When the sample attributes are correlated, the classification performance will decrease rapidly.
- Types of naïve Bayes:
  - Bernoulli Naïve Bayes:
  - Multinomial Naïve Bayes:
  - Categorical Naïve Bayes:
  - Gaussian Naïve Bayes:

Decision Trees

Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Decision Trees

- The features that best divide the data are selected. This process is done using a specific criterion:
  - Gini Index: Measures which feature best divides the data.
  - Information Gain (Information Gain): Calculates the increase in information obtained with the data being available.
  - Entropy: Used to increase the change in the data.

Decision Trees

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

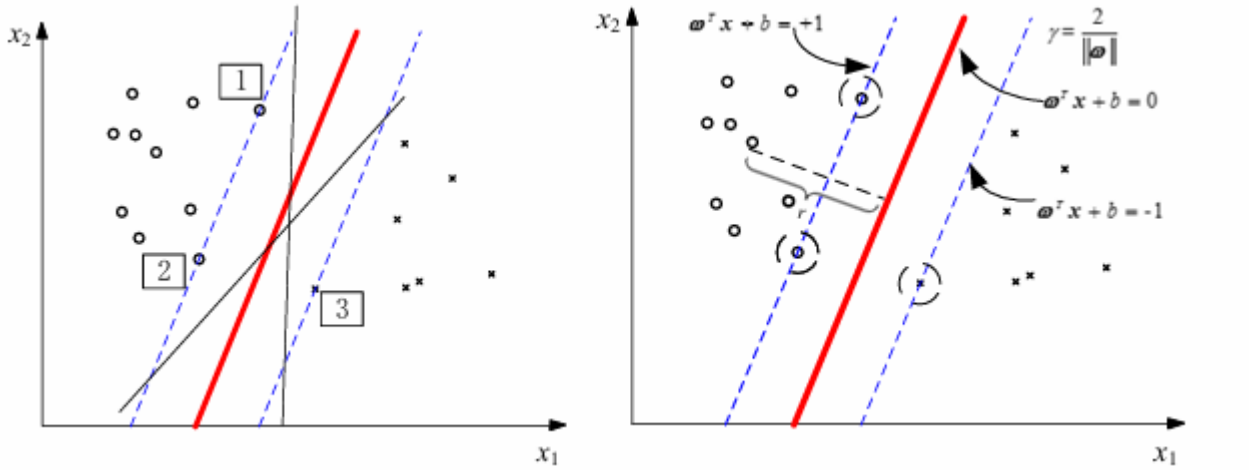
		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

Decision Trees

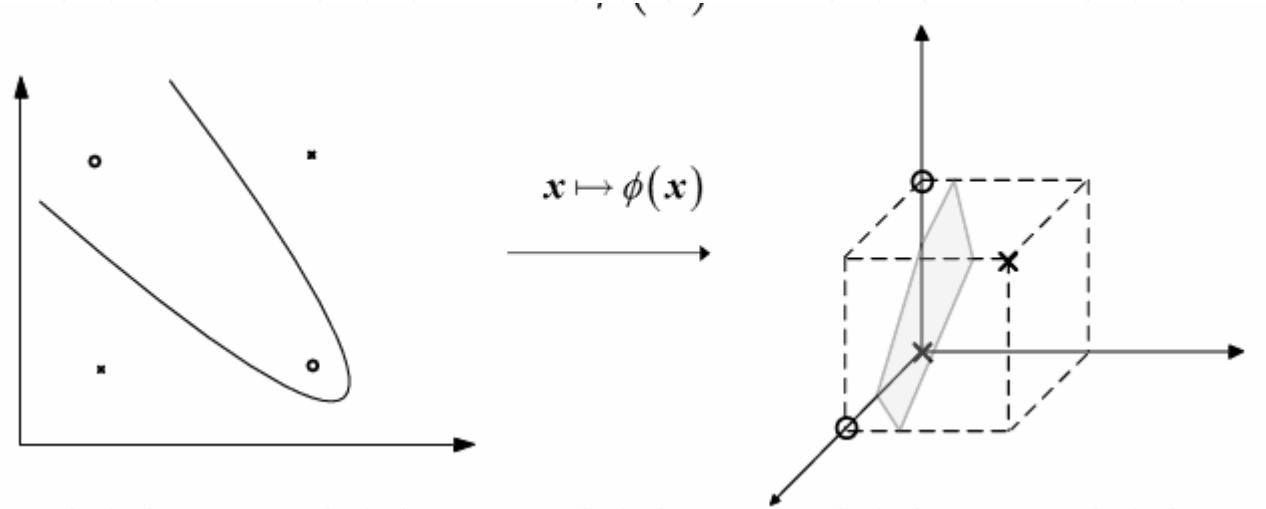
- DT has high classification accuracy and is easy to extract rules, allowing visual analysis. Its shortcomings are that it is not easy to process the actual data and it is prone to overfitting.
- **The Ensemble-based Decision Tree Techniques**
  - **Random Forest**: Combines multiple decision trees trained on random subsets of the data and features.
  - **Gradient Boosted Trees**: Sequentially builds decision trees where each tree corrects the errors of the previous one.
  - **AdaBoost (Adaptive Boosting):** Builds a sequence of weak decision trees (often decision stumps) and combines them to create a strong model.

Support Vector Machine



- SVM is to find a hyperplane that divides the sample. There may be multiple such hyperplanes, from which SVM is to select the relatively "optimal" hyperplane, as shown in the solid line in the Figure. Intuitively, the selection of hyperplane should try to meet the requirement of dividing two data sets to the maximum extent, that is, "interval maximization", so as to have a good tolerance for local "disturbances" existing in the data set. In other words, the classification results generated by such a hyperplane have good robustness.

Support Vector Machine

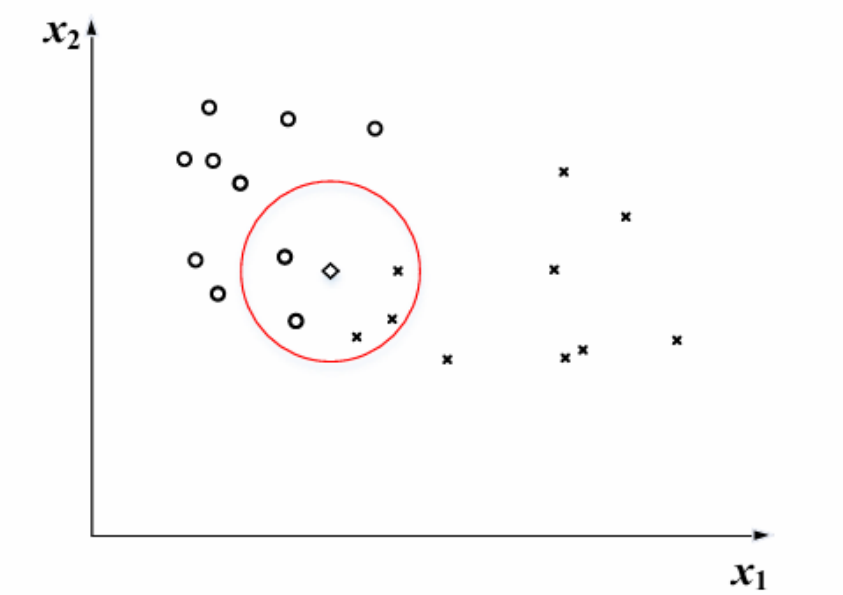


- In the nonlinear classification problem, the sample space of the current dimension can be mapped to the feature space of the higher dimension, so as to find the hyperplane for sample classification in the high dimensional space. This method is called the kernel method.

Support Vector Machine

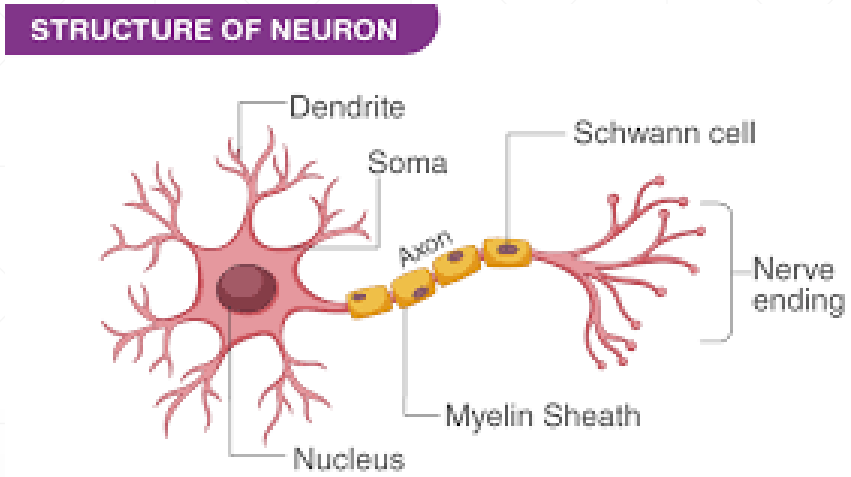
- Common kernels include:
  - **Linear Kernel**: Used for linearly separable data.
  - **Polynomial Kernel**: Maps data to a higher-degree polynomial space.
  - **RBF (Radial Basis Function) Kernel**: Uses Gaussian functions to create complex decision boundaries.
  - **Sigmoid Kernel**: Similar to a neural network activation function

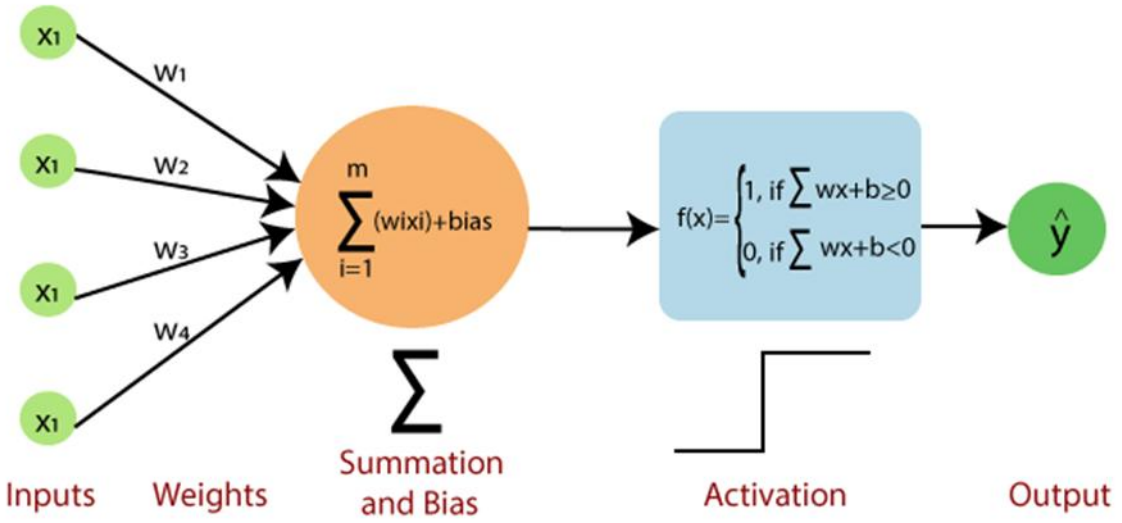
K-Nearest Neighbors (KNN)



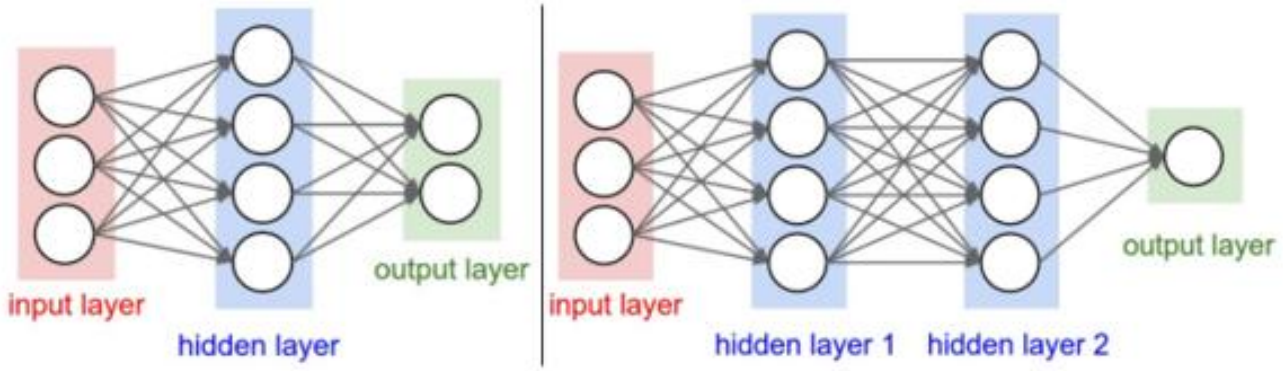
- The basic idea of the K-nearest neighbor algorithm is to the undetermined prediction samples, find out the nearest K samples and use the voting rule.
- The distance measurement methods commonly used by KNN include Euclidean distance, Manhattan distance and Minkowski distance.

Artificial Neural Networks (ANN)





Artificial Neural Networks (ANN)



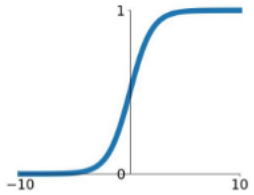
- Single-layer ANN model
  - 4 + 2 = 6 neurons (excluding the input layers)
  - [3x4] + [4x2] = 20 weights
  - 4 + 2 = 6 bias values
  - There are 26 parameters to learn
- Two hidden layer ANN model
  - 4 + 4 + 1 = 9 neurons
  - [3x4] + [4x4] + [4x1] = 12 + 16 + 4 = 32 weight
  - 4 + 4 + 1 = 9 bias
  - There are 41 parameters to learn

Artificial Neural Networks (ANN)

Activation Functions

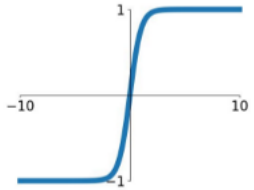
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



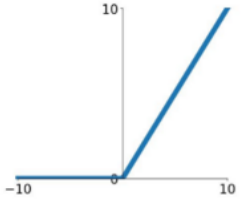
tanh

$$\tanh(x)$$



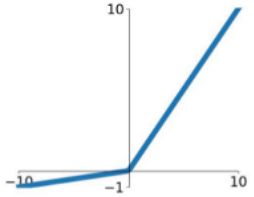
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

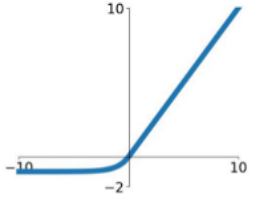


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Artificial Neural Networks (ANN)

- The ANN algorithm has high classification accuracy, strong learning ability, and strong robustness and tolerance to noisy data. In addition, ANN has the ability to associate, can approximate any non-linear relationship, and has good predictive and classification ability even on untrained data.
- However, the defects of ANN are also obvious. However, due to the large number of parameters, the training time of ANN is long, and the training process is a black box process, the intermediate results cannot be observed, and the interpretability is poor. ANN may also fall into a local minimum.
- ANN Models:
  - Feedforward Neural Networks (FNN)
  - Convolutional Neural Networks (CNN)
  - Recurrent Neural Networks (RNN)
  - Deep Neural Networks (DNNs)

Ensemble Classifier

- Ensemble classifier refers to a group of individual classifiers that are cooperatively trained on data set in a supervised classification problem.
  - Ensemble Classifier Generation by Manipulation of the Training Parameters
  - Ensemble Classifier Generation by Manipulation of the Error Function
  - Ensemble Classifier Generation by Manipulation of The Feature Space
  - Ensemble Classifier Generation by Manipulation of the Output Labels
  - Ensemble Classifier Generation by Clustering

Ensemble Classifier

- Advantages of Ensemble Learning

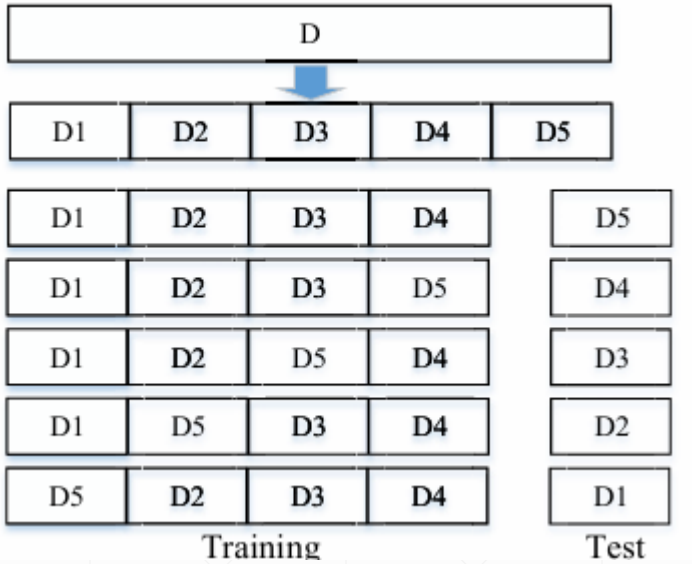
  1. Higher Accuracy: Combining multiple models allows you to get better results than the performance of an individual model.

  2. Generalization Ability: The risk of overfitting is reduced and the model performs better on different data sets.

  3. Flexibility: It offers the opportunity to combine the strengths of different algorithms.

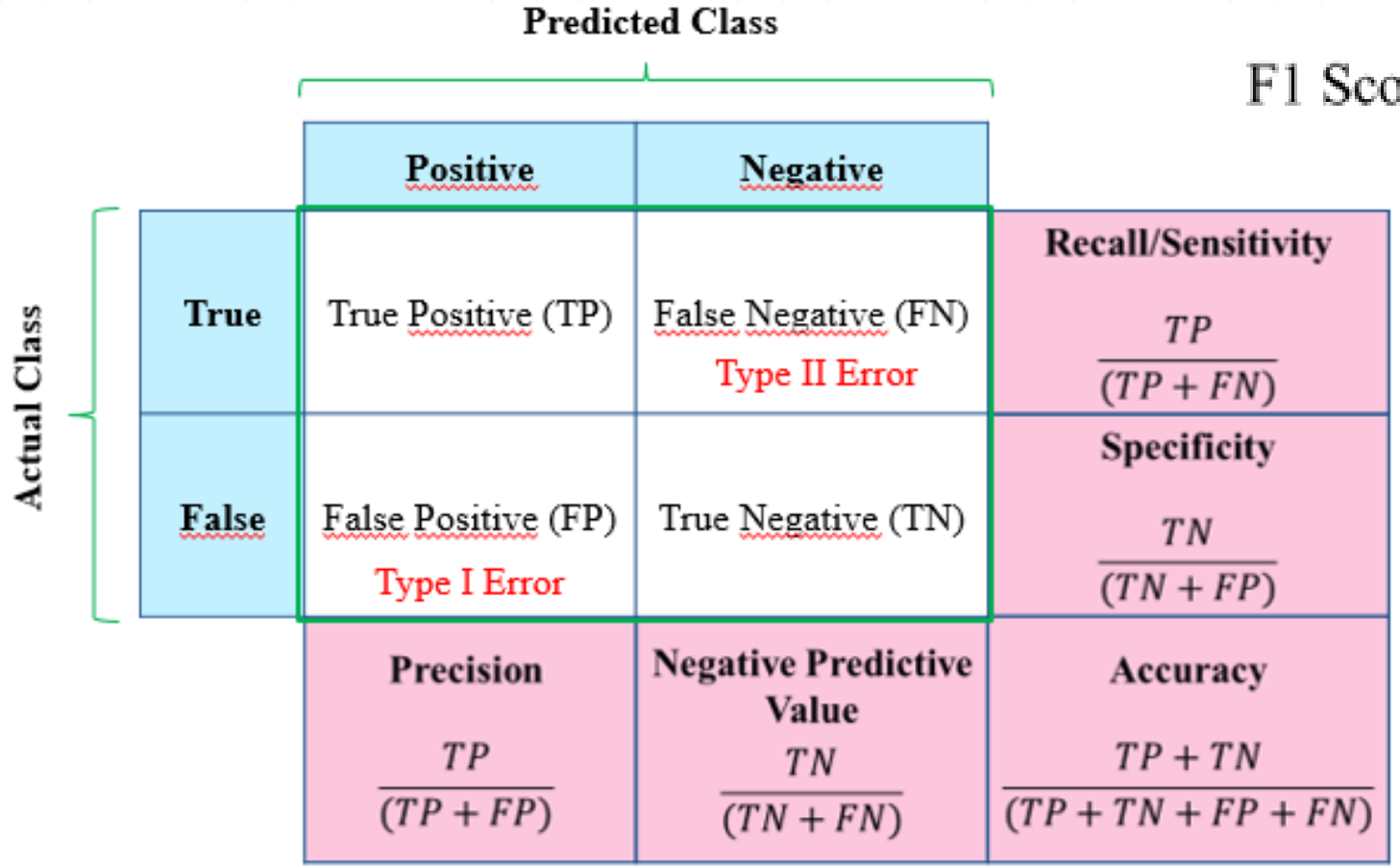
Model Evaluation Method

The cross validation method, is proposed on the basis of the idea of the set aside method also known as the K-fold cross validation method. As shown in the figure, through the stratified sampling method, the k-fold cross-validation method divides the data into k groups of mutually exclusive subsets with the same size. In the process of k training, each training selects k-1 group of samples as the training set, and the remaining group of samples as the test set, so k training results can be obtained,
and finally the final result is averaged.



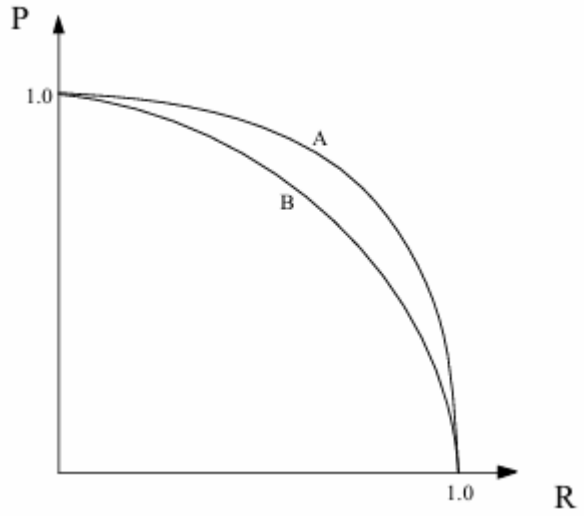
Confusion Matrix (Karmaşıklık Matrisi)

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



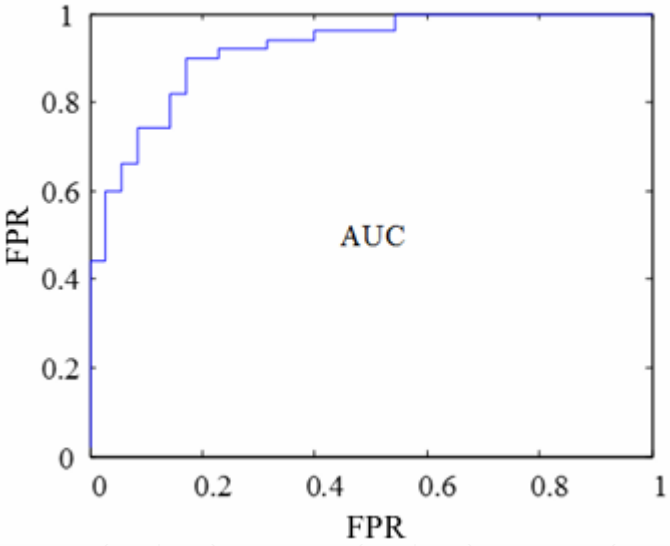
Performance index of classifier

- **PR Curve (Precision-Recall Curve)**
- PR Curve is a visualization of the relationship between a model's Precision and Recall metrics.
- Used when it is necessary to focus on the positive class, for example, detection of rare events.



Performance index of classifier

- **ROC Curve (Receiver Operating Characteristic Curve)**
- ROC Curve shows the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) of a model. ROC Curve is used to evaluate the classification performance of the model across all threshold values.
- It is suitable for evaluating the overall performance of the model. It is more commonly used in balanced data sets.



REFERENCES

- Ensemble Classifiers and Their Applications: A Review(Akhlaqur Rahman, 2022)
- An introduction to machine learning for classification and prediction(Jason E. Black, 2020)
- Review on Classification Algorithm and Evaluation System of Machine Learning(Zhaodong Wu, 2020)
- Machine Learning: Algorithms, Real-World Applications and Research Directions (Iqbal H. Sarker, 2021)
- Supervised Machine Learning Algorithm: AReview of Classification Techniques(Pankaj Saraswat, 2021)