Classification Methods

Computational Data Science,
Addis Ababa University

www.aau.edu.et

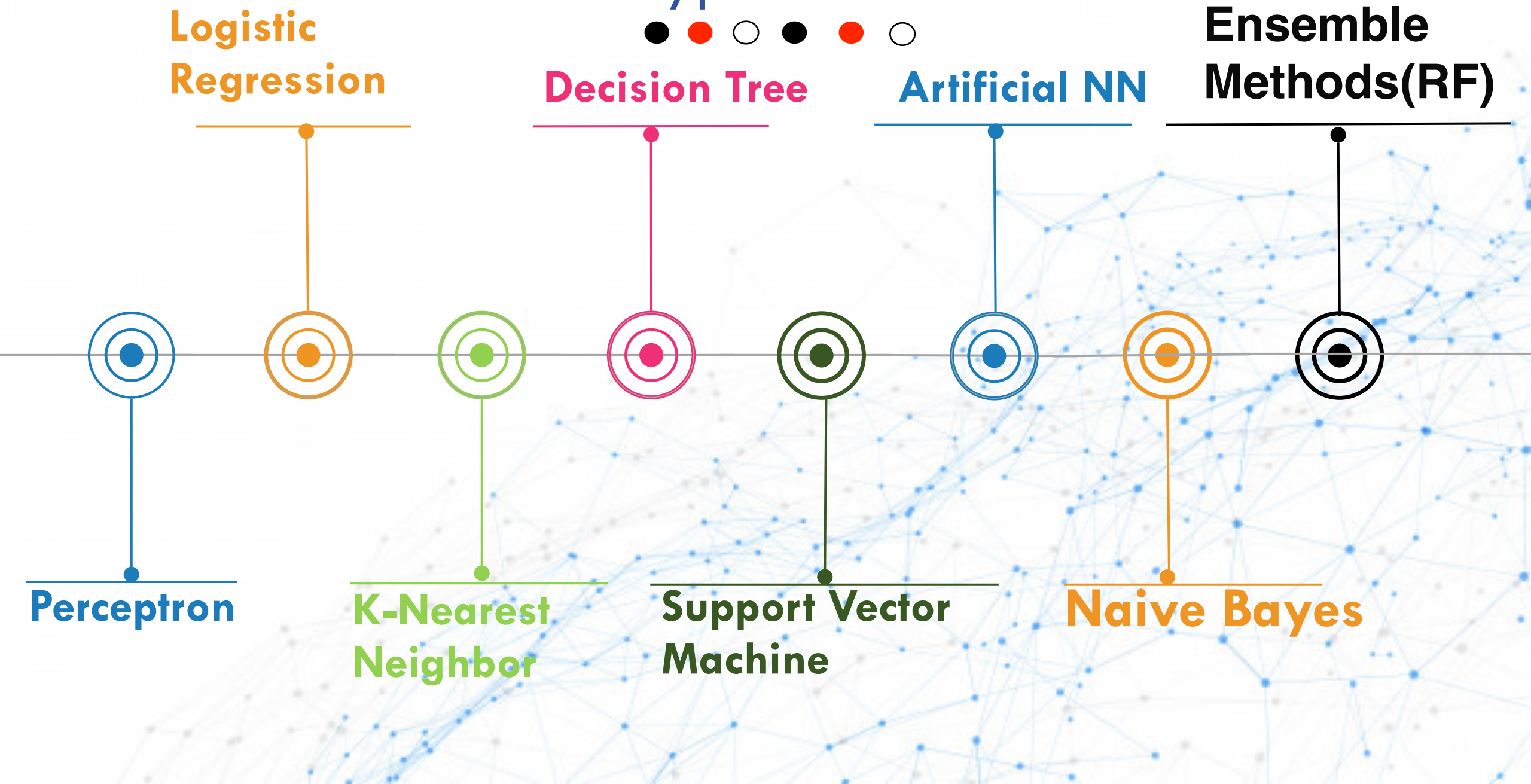
mesfin.diro@aau.eud.et

+251-912-086156

Classification Algorithms

- In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given features of input data.
- Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to features from the problem domain.

Different types of Classifiers

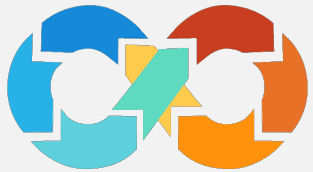


Main Types of Classification

Binary Classification

Supervised

Two Class Labels



Multi-Class Classification

Supervised

More than two class labels

Multi-Label Classification

Supervised

Two or more class labels where one or more class labels may be predicted for each examples

Imbalanced Classification

Supervised

the number of examples in each class is unequally distributed

Binary Classification

Logistic Regression

In this algorithm, the probability describing the possible outcomes of a single trial are modeled using a logistic function

K-Nearest Neighbors

Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data.

Decision Tree

Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

Naive Bayes

The Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features

Support Vector Machine

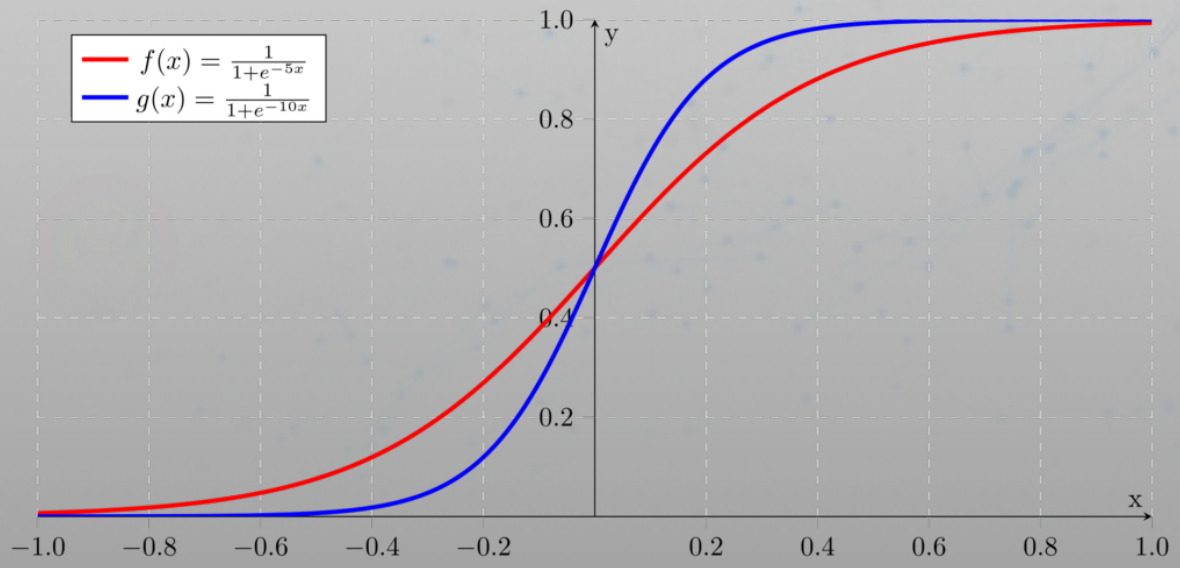
A representation of the training data as points in space separated into categories by a clear gap that is as wide as possible.

Logistic Regression

- Logistic Regression is an omnipresent and extensively used algorithm for classification.
- It is a classification model, very easy to use and its performance is superlative in linearly separable class.
- This is based on the probability for a sample to belong to a class that must be continuous and bounded between (0, 1).
- It is dependent on a threshold function to make a decision that is called Sigmoid or Logistic function.

Logistic Regression

- To understand the concept of Logistic Regression, it is important to understand the concept of Odd Ration (OR), Logit function, Sigmoid function or Logistic function, and Cross-entropy or Log Loss.



Logistic Regression: Odds Ratio(OR)

- Odds Ration (OR) is the odds in favor of a particular event. It is a measure of association between exposure and outcome.
- Lets X is the probability of subjects affected and Y is a probability of subjects not affected, then, odds = X/Y

$$odds\ Ratio = \frac{P}{1-p}$$

Where P is the probability of the positive events

- Let's say the probability of sucess(P) is 0.8, thus the probability of failure(Q) = 1 - P = 0.2
- odds(success) = P/Q = 4 , odds(Failure) = Q/P = 0.2/0.8= 0.25

Logistic Regression: Logit Function

- **Logit function is the logarithm of the Odd Ratio (log-odds). It takes input values in the range 0 to 1 and then transforms them to value over the entire real number range.**
- **Let's take P as probability, then P/(1-P) is the corresponding odds; the logit of the probability is the logarithm of the odd given below:**

$$logit(P) = \log\left(\frac{P}{1-p}\right)$$

Where P is the probability of the positive events

Logistic Regression: Logistic Function

- **Logistic function(Sigmod Function) is The inverse of the logit function is called the logistic function or Sigmoid function. It is named as Sigmoid function due to its characteristic shape.**
- **The equation of the Sigmoid function (from logit function):**

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

Classification Metrics: Confusion Matrix

- Confusion matrix is figure or a table that used to describe the performance to shows how wrongly or correctly an output is predicted by a classifier algorithm.
- There are four possible results of what the classification model do which defines the confusion matrix or contingency table:

	Actual Positive	Actual Negative	
Predicted Positive	True Positive(TP)	True Negative (TN)	
Predicted Negative	False Positive (FP)	False Negative(FN)	

Accuracy and Precision

- Accuracy is the first statistic measures of the ratio of the number of correct predictions over the total predictions of a classifier as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision measures how often a classifier is correct when it dares to say positive as:

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

Recall and F1-Score

- Recall measures how often we prove right on all positive instances of occurrences:

$$Recall = \frac{TP + TN}{TP + FN}$$

- It's a hard-wired to describe the performance of a system from a single measurement. Hence, for such system F1-score is a combination to return the harmonic mean of precision and recall

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

THANK YOU!