A Comparative Study of Various Machine Learning Classification Algorithms

Paraschopoulos Kyriakos

Supervisor: Konstantinos E. Psannis

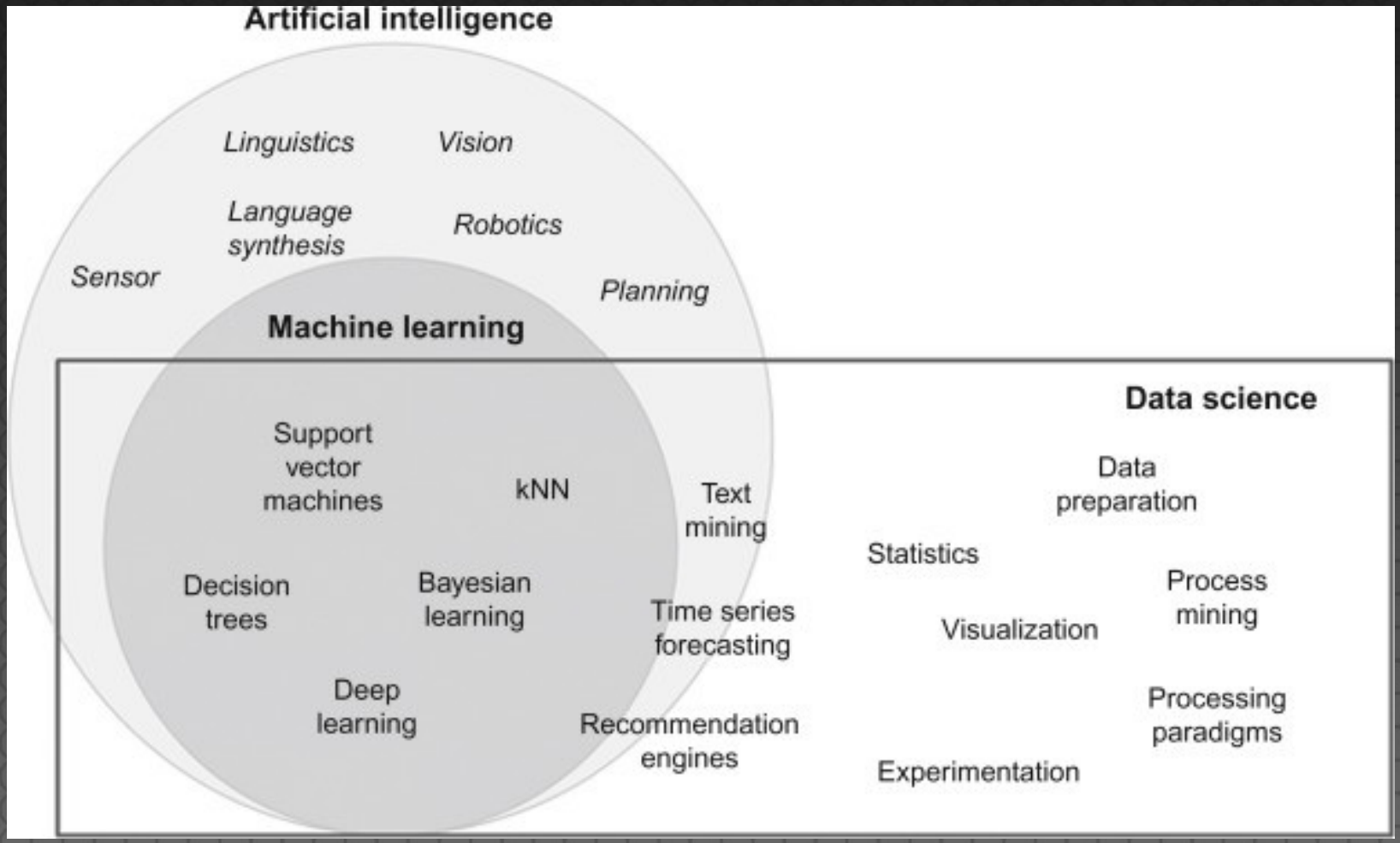
M.Sc Program

Department of Applied Informatics

University of Macedonia

Data Science, AI & Machine Learning

Data Science, AI & Machine Learning



Data Science

- Data science refers to the process of extraction of useful insights from data. It merges different techniques from various fields of computer science, mathematics and statistical models in order to extract insights in automated ways.

Artificial intelligence

- Artificial intelligence (AI) refers to the process of making machines able to simulate the human brain function, to understand data, learn from the data, and make decisions based on patterns hidden in the data. AI is defined as a collection of mathematical algorithms that leads to computers' understanding of relationships between different types and pieces of data.

Machine Learning

Definitions

“Field of study that gives computers the ability to learn without being explicitly programmed”, 1959 Arthur Samuel

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”, Tom M. Mitchell

Types of Machine Learning

- Supervised Learning
  - Data has known labels or output
- Unsupervised Learning
  - Unknown labels or output
  - Focus on finding patterns and gaining insights from the data
- Reinforcement Learning
  - Focus on making decisions based on previous experience

ML: Supervised Learning

Develop predictive model based on both input and output data.

- Classification

Classification is a technique used for predicting a discrete class label.
In classification problems data are classified into one of two classes (binary classification) or more classes (multi-class classification)

- *Regression*

Regression is a technique for predicting a continuous quantity.
Regression models are used to predict numerical or continuous variables based on previous observed data from the trained dataset.

Dataset

- A dataset represents a collection of data points
- It makes reference to a database table, where each column of the table represents a specific variable and each row corresponds to the observation of each member

Dataset splitting

- Training set

  The training set is the actual collection of data used to train the machine learning model by matching the input with the expected output.

  The training set represents the 60% of the original data set.

- Validation set

  The validation set is a set of examples used to tune the hyperparameters of the classifier to train the machine learning model with the optimal process. It is a way to evaluate how well the model has been trained.

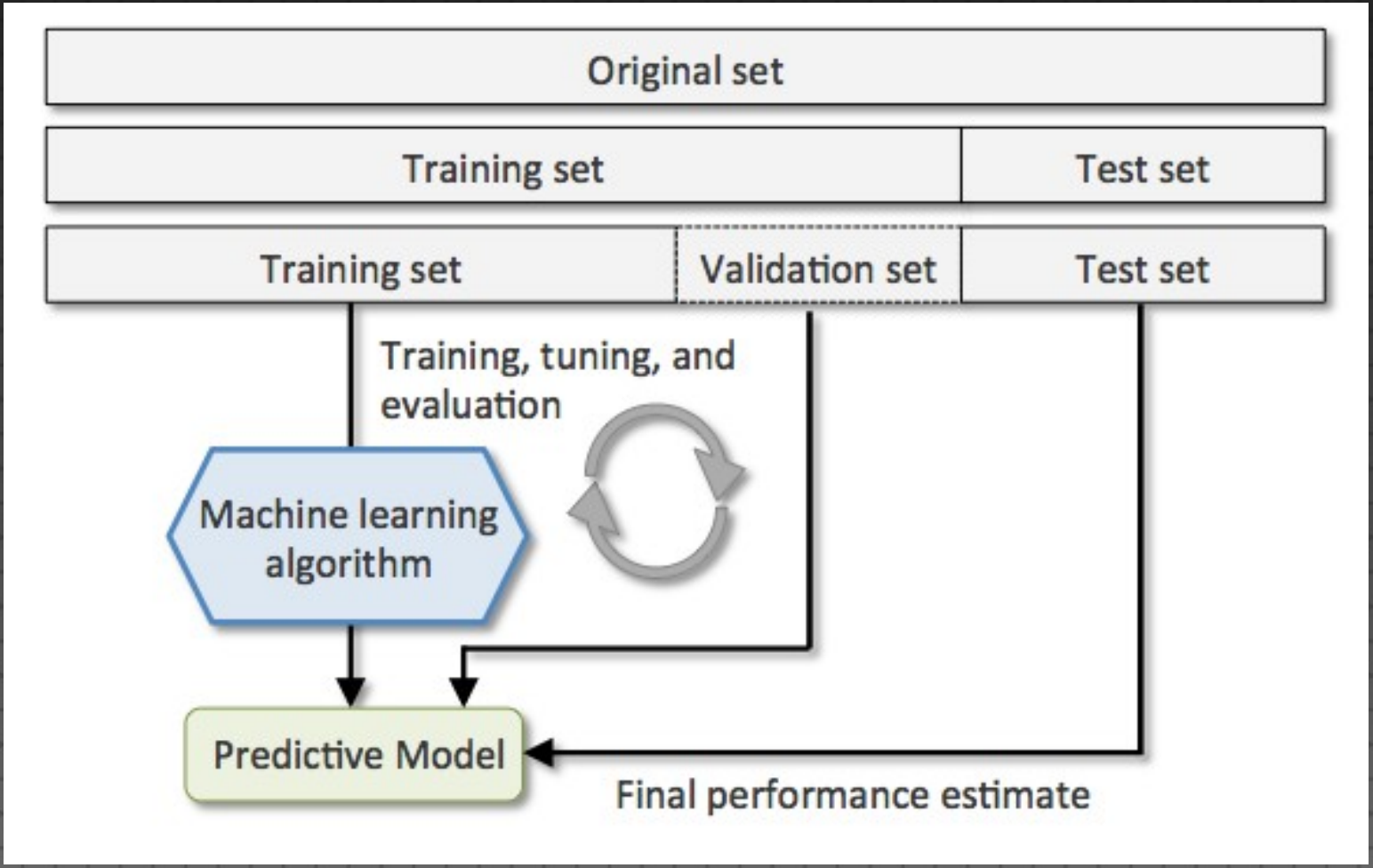
  The validation set represents the 20% of the original data set.

- Test set

  The test set is a set of examples that is applied at the final step when the model is fully trained and it is used to evaluate the performance of the classifier.

  The test set represents the 20% of the original data set.

Dataset splitting



Wine dataset

- The Wine dataset from UCI Machine Learning Repository will be used to perform the experiments of the classification algorithms

Evaluation Metrics & Classification Algorithms

EM - Confusion Matrix

- Table which visualizes the performance of an algorithm

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	True Positive (tp)	False Positive (fp)
Predicted Negative Class	False Negative (fn)	True Negative (tn)

Evaluation Metrics

- Accuracy - The number of correct predictions made by the model over the total number of instances evaluated.

$$\text{Accuracy (acc)} = \frac{tp + tn}{tp + fp + tn + fn}$$

- Recall - Proportion of positives that are correctly classified and the number of correct positives divided by the number of all samples that have been identified as positive.

$$\text{Error Rate (err)} = \frac{fp + fn}{tp + fp + tn + fn}$$

Evaluation Metrics

- Precision - What proportion of predicted positives is actual positive or correctly predicted as positive at the trained model.

$$\text{Precision (p)} = \frac{tp}{tp + fp}$$

- F1-score - Measure how precise a classifier is, by finding the balance between precision and recall.

$$\text{F-Measure (FM)} = \frac{2 * p * r}{p + r}$$

Logistic Regression

- The model finds the correct decision boundary for one of the two categories in the data set.
- Sigmoid function

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

  - The sigmoid function can take any real-valued number between 0 and 1.

$$f(x) \geq 0.5, class = 1$$

$$f(x) < 0.5, class = 0$$

LR - Applications

- Spam Detection
- Credit Card Fraud
- Tumour Prediction
- Marketing

Naïve Bayes

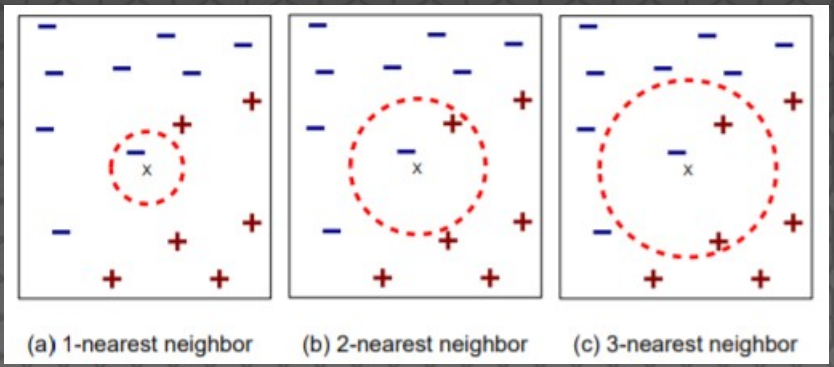
- Predict class labels by best approximating the probabilistic relationship between the independence attributes and the class label.

NB - Applications

- Text classification
- Recommendation System
- Real-time Prediction
- Multi-class Prediction

K-Nearest Neighbors

- Stores all available cases and classifies new cases based on a similarity measure (distance function).
- The value of *k* stands for number of dataset items that are considered for the classification.
- One of the top data mining algorithms used today.



K-NN: Pros & Cons

+ Simple and easy to implement. Only two parameters required, the value of *k* and the distance function.

+ No training step. It stores the training dataset

- Limitations on large datasets

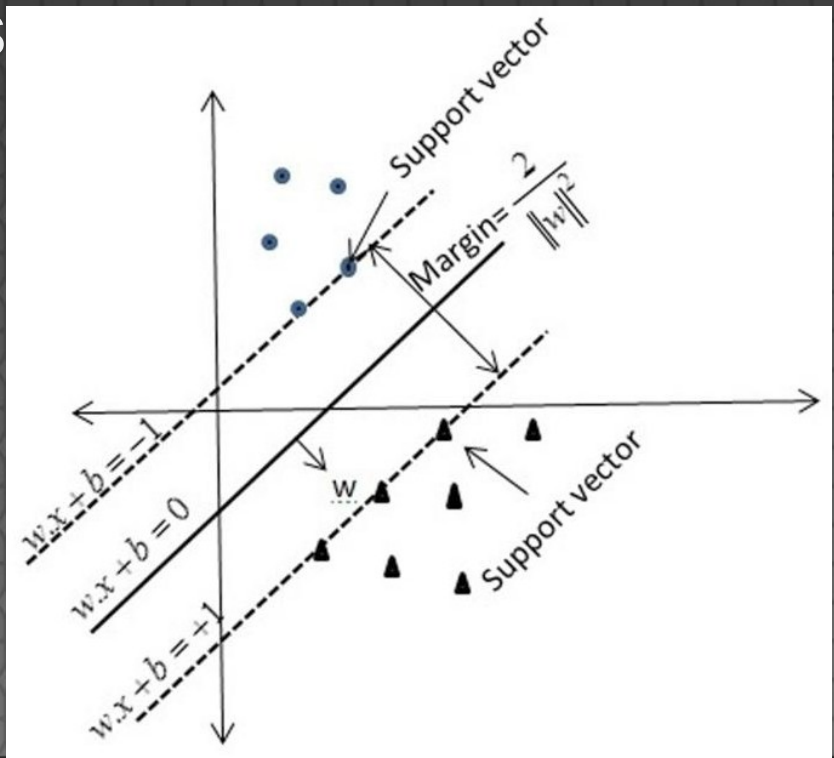
-

K-NN applicationss

- Text classification
- Finance
  - Market trends
  - Planning investment strategies

Support Vector Machine

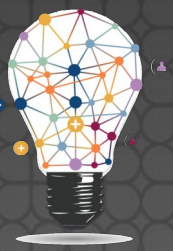
- Suitable for binary classification tasks.
- Constructs a hyperplane to find the maximum margin linear classifier.
- (w.x + b) > +1 for positive cases
  (w.x + b) < - 1 for negative cases



SVM Applications

- Text categorization
- Handwriting recognition
- Image classification

Thank you!



Paraschopoulos Kyriakos
kparaschopoulos@gmail.com