JM2050 – Natural Language Processing

Machine learning basics

September 5th, 2024

Prof.dr.ir. U. Kaymak

Recap

- Three general approaches to NLP
  - Rule-based (rationalism)
  - Statistical (ML) models (empiricism)
  - Deep learning models (massive parallelism)
- Text is considered unstructured data
- Challenges with text data
  - Ambiguity
  - Variation
  - World knowledge
  - Context
- NLP tasks become easier by using limiting the domain, using knowledge resources and context information

[2]

Outline

- Types of learning
- Machine learning tasks
- Modeling methods
- Model evaluation
- Data mining pipeline

Machine learning basics

[4]

JADS Jheronimus Academy of Data Science

What is machine learning?

Oxford definition:

the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data

Solved through an optimization problem

[5]

Types of learning

Basic learning types

- Supervised
- Unsupervised
- Reinforcement learning

Other learning types

- Semi-supervised learning
- Transfer learning
- Active learning
- Etc.

[6]

Machine learning tasks

- Classification
- Regression
- Similarity matching
- Clustering
- Anomaly detection
- Co-occurrence grouping
- Profiling
- Link prediction
- Data reduction
- Causal modeling

[7]

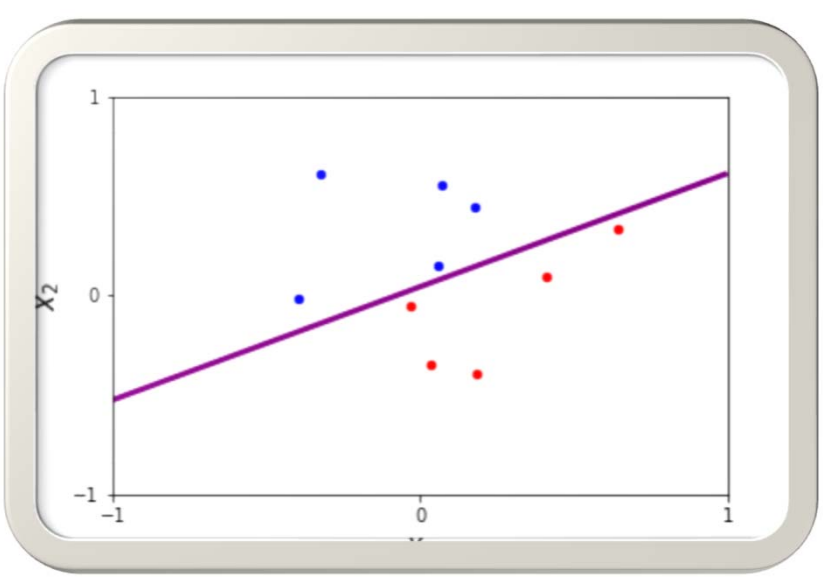
Machine learning / modeling methods

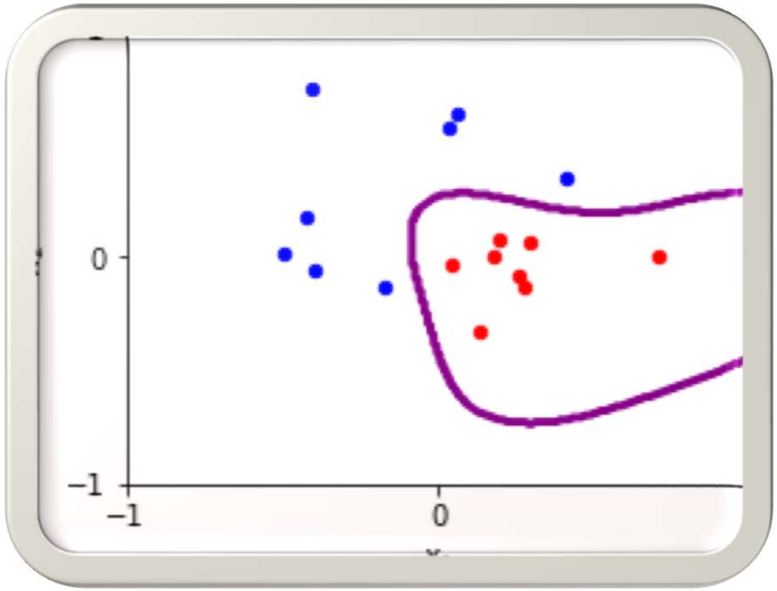
- Linear regression
- Logistic regression
- Decision trees
- k nearest neighbors classifier
- Naïve Bayes classifier
- Mixture models
- Support vector machines
- Neural networks
- Fuzzy inference systems
- Bayesian networks

(Nonlinear) input – output mappings

[8]

Linear vs. non-linear classifiers





[9]

Measuring classifier performance

Confusion matrix

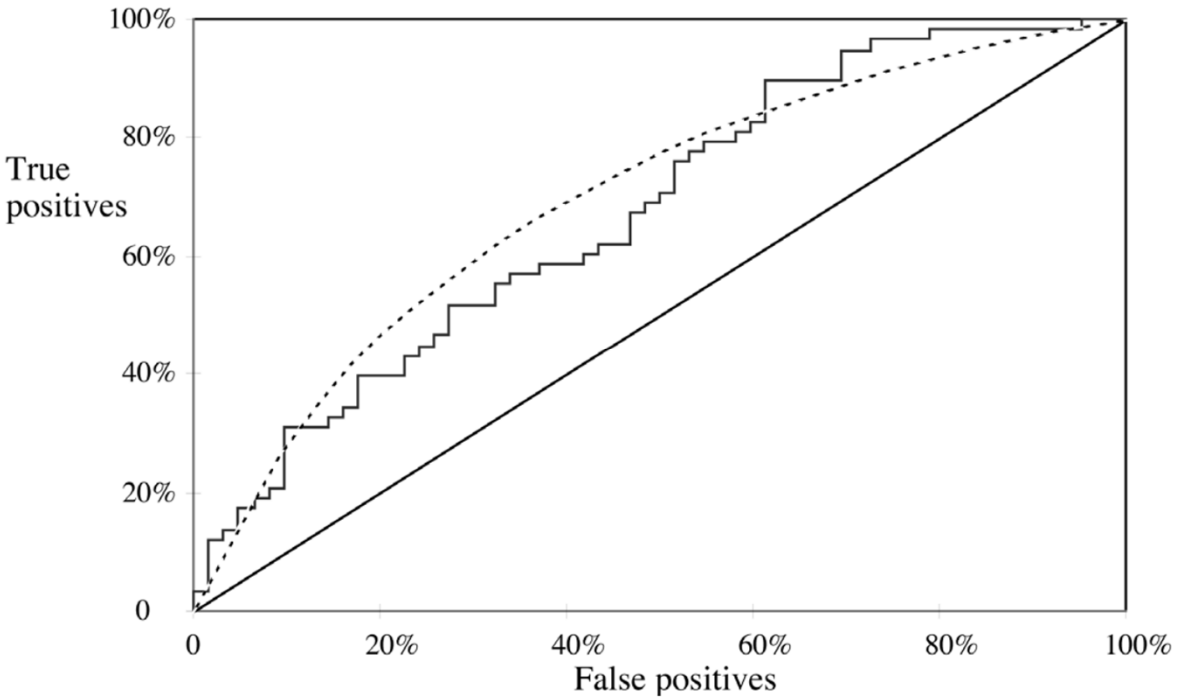
		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

wikipedia.org

- Accuracy = (TP+TN)/(TP+FN+FP+TN)
- Error = 1 – Accuracy
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1 = 2 (Precision · Recall) / (Precision+Recall)

[10]

Receiver Operating Characteristic (ROC) curves



Common metric:

Area under the ROC Curve (AUC)

[11]

Kappa statistic

$$\text{Kappa, } K = \frac{a - p}{1 - p}$$

a - accuracy

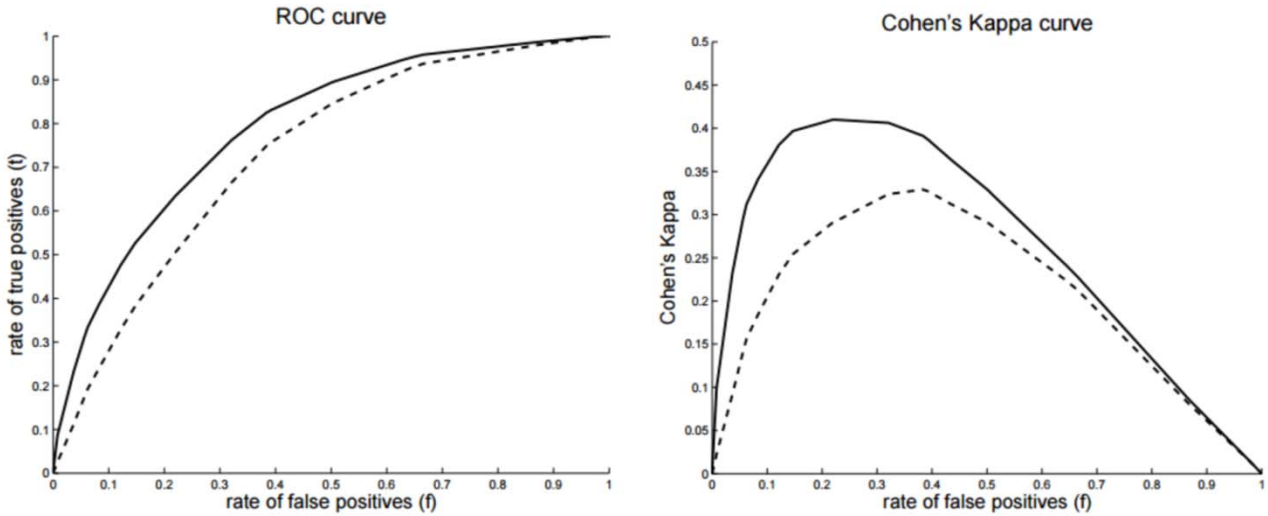
p - probability of predicting the correct class due to chance

Kappa = 1 => perfect model

Kappa ≈ 0 => no better than random guessing

[12]

Kappa curves

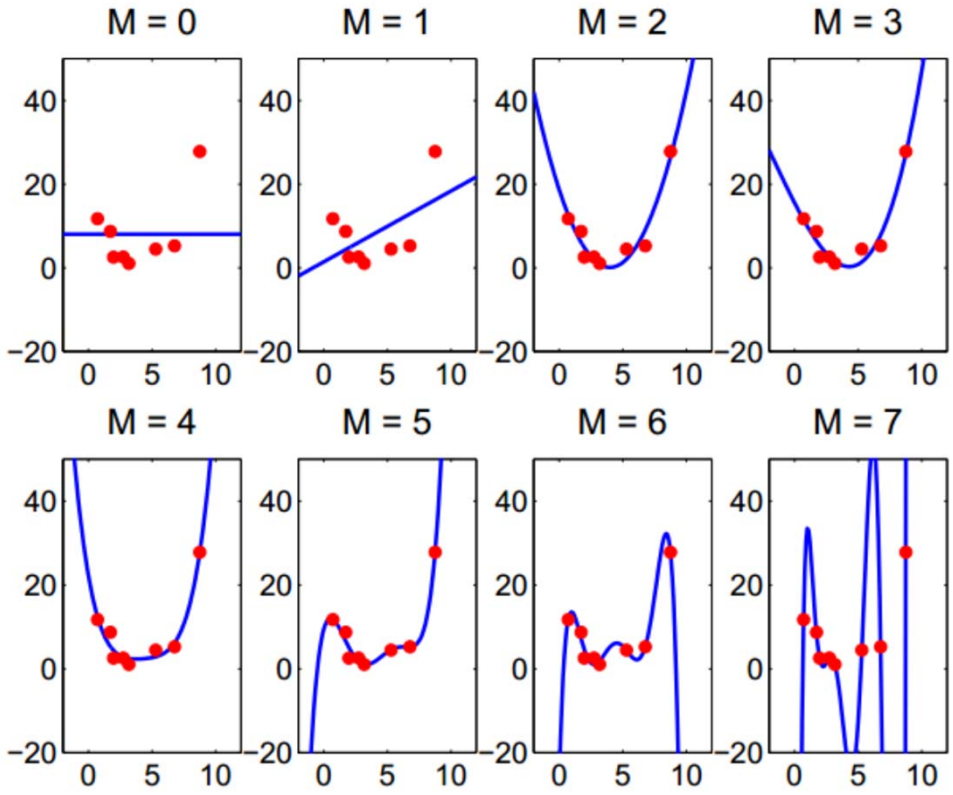


- Can be used to select optimal threshold
- AUK index - Area Under the Kappa curve

(U. Kaymak, A. Ben-David, R. Potharst, AUK: a simple alternative to the AUC, 2010)

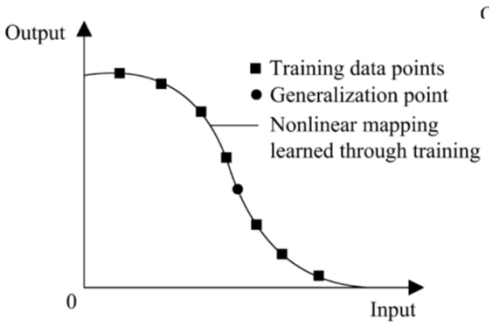
[13]

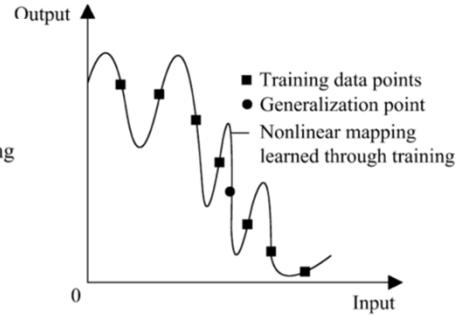
Overfitting

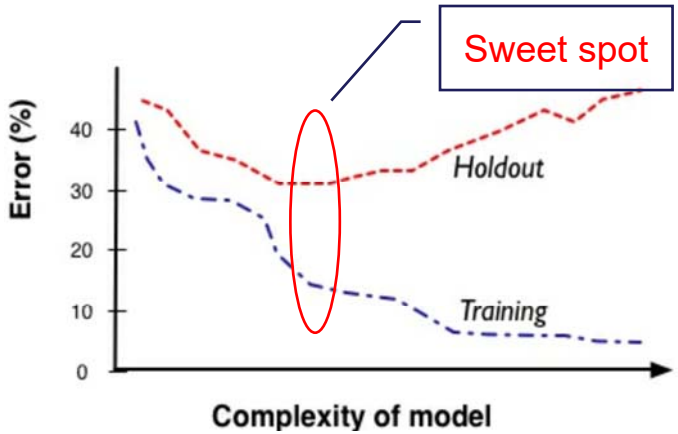


[14]

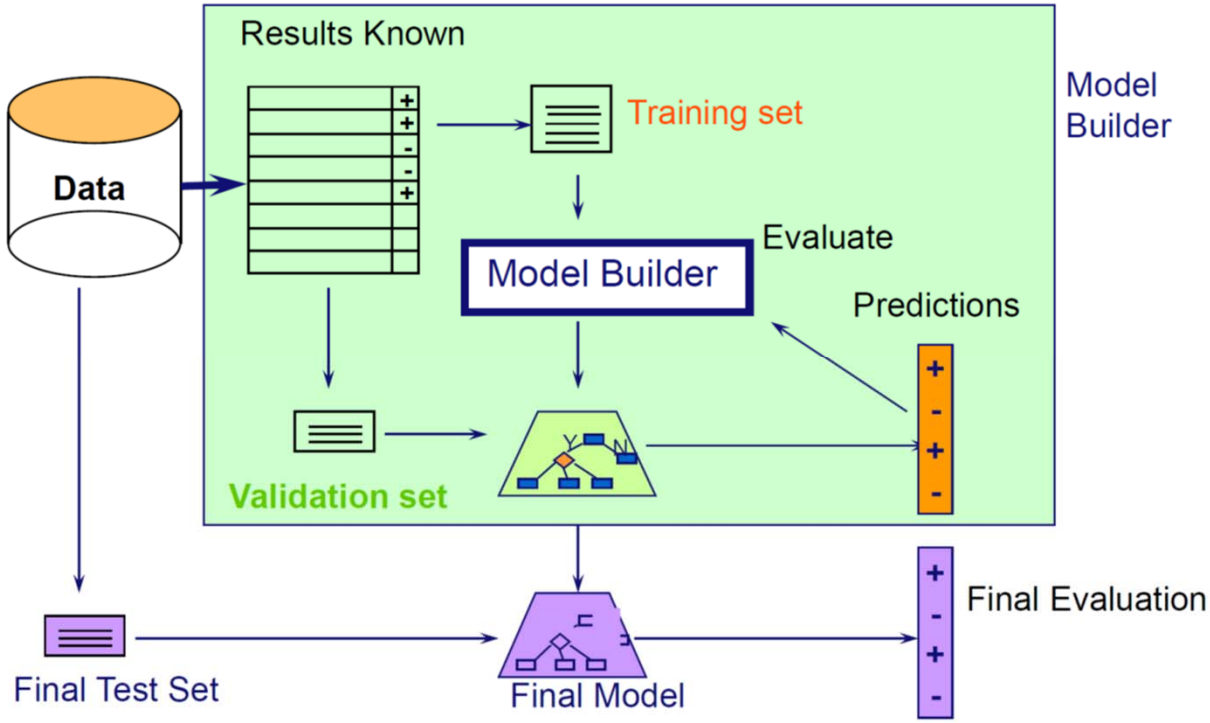
Generalization







Experiment design



[16]

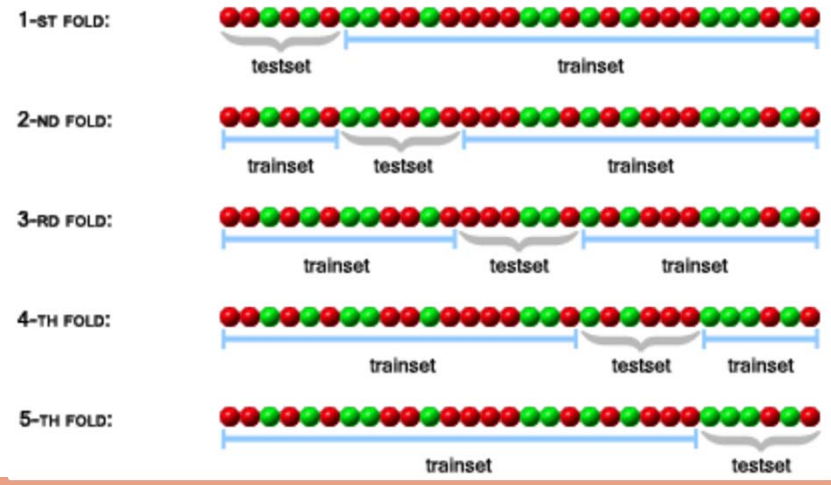
www.jads.nl

Jheronimus Academy of Data Science

Crossvalidation

- Split data into groups of same size

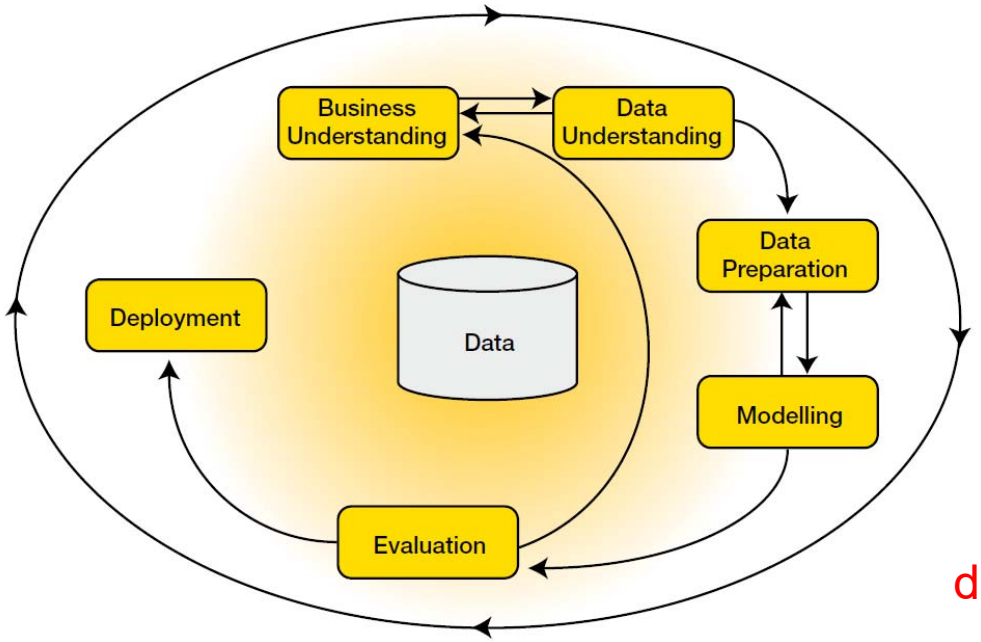
- Hold aside one group for testing and use the remainder for training



- Repeat for all groups

[17]

Crisp-DM Data Mining Framework



Industry-standard data mining process

www.jads.nl