Machine Learning with Weka: Classification and Evaluation

PU Tools, Ressourcen, Infrastruktur

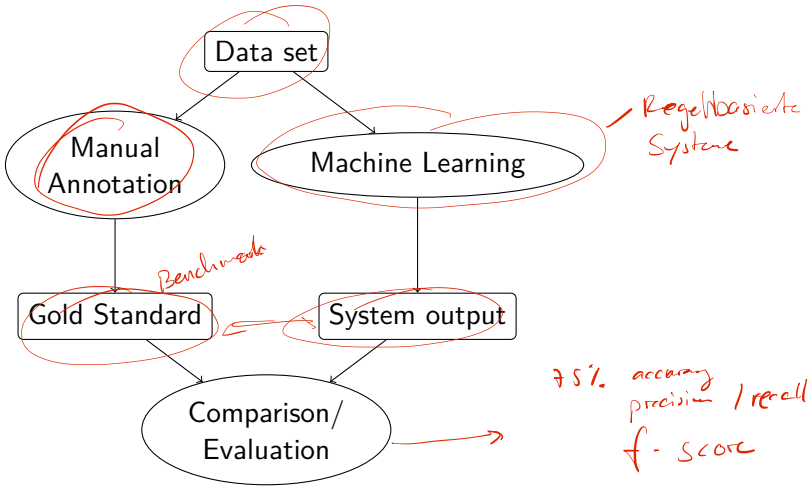
Nils Reiter,
nils.reiter@uni-koeln.de

December 3, 2020
(Winter term 2020/21)

Exercise 04

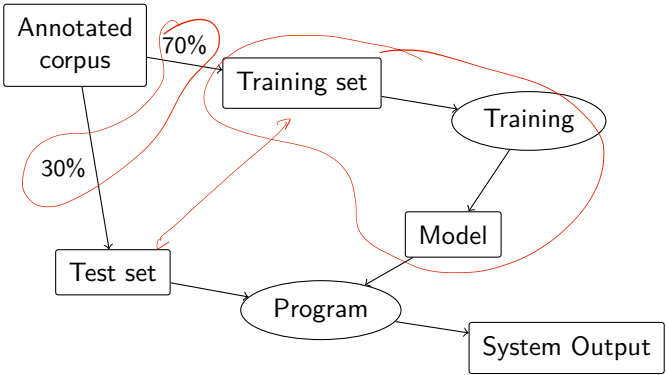
https://github.com/idh-cologne-tools-ressourcen-infra/exercise-04

Experiments



Evaluation

- Goal: Predict the quality on new data
- The program cannot have seen the data, so that it's a realistic test



Classification

- Assigning *classes* to *objects/instances/items*
- Many different algorithms available:
  - Decision trees
  - Support vector machines
  - Naïve bayes
  - Neural networks
  - Bayesian networks
  - ...

Classification

- Assigning *classes* to *objects/instances/items*
- Many different algorithms available:
  - Decision trees
  - Support vector machines
  - Naïve bayes
  - Neural networks
  - Bayesian networks
  - ...
- Libraries are available, not a technical challenge
  - Challenge: Use an appropriate one
  - Challenge: Use it correctly
  - Challenge: Interpret its results reasonably

What's an item?

- If you have the wrong idea of what an item is, you are lost
- Items are the things you want to classify or cluster
- In principle, this can be anything
- But you need to extract useful (and generalizable) features for your items!

Items and Classes

Examples

Task	Items	Classes
POS-tagging		
WSD		
Anaphora Resolution		

Items and Classes

Examples

Task	Items	Classes
POS-tagging	Tokens	POS-tags
WSD		
Anaphora Resolution		

Items and Classes

Examples

Task	Items	Classes
POS-tagging	Tokens	POS-tags
WSD	Content tokens	Synsets
Anaphora Resolution		

Items and Classes

Examples

Task	Items	Classes
POS-tagging	Tokens	POS-tags
WSD	Content tokens	Synsets
Anaphora Resolution	Pairs of anaphora and antecedent candidate	Yes / No

Features (a.k.a. attributes, properties)

- Decision is based on features (= properties)
- The prediction model only sees feature values!
  - What's not encoded in a feature doesn't play a role
  - It's our job to provide useful features
- Playground for being creative
  - It helps to understand the problem/task/phenomenon

Deep Learning

- Significant gains in recent years
- Neural networks expect the input to be encoded as vectors (which are used as features)
- Typical vectors in NLP: Word embeddings
  - Each word is represented as vector, derived from the words co-occurrences
  - »Distributional semantics«
- Mathematically reduced dimensions
- No manual feature engineering, but settings on input representation

Deep Learning

- Significant gains in recent years
- Neural networks expect the input to be encoded as vectors (which are used as features)
- Typical vectors in NLP: Word embeddings
  - Each word is represented as vector, derived from the words co-occurrences
  - »Distributional semantics«
- Mathematically reduced dimensions
- No manual feature engineering, but settings on input representation

Why does it work so well?

- Over decades, domain expertise was used to develop ›classical‹ ML systems
  - Tiny improvements were considered a success
- Larger data sets and more compute power: Sudden jumps in prediction performance
  - »Gold rush« in NLP
  - Limits become apparent

What do we need?

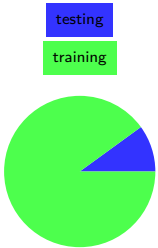
- Annotated data: Data with given classification
- Feature extraction: For each item, we extract feature values from the data
- Three use cases / Phases
  - Training: A classifier »learns probabilities«[1] by looking at the extracted features from the annotated data
  - Testing: The trained classifier is used on new, but annotated data. Then we can compare, how good the classifier performed
  - Application: Features are extracted from new, un-annotated data. The trained classifier then decides in which class an item belongs

[1] It's not learning and it's not always probabilities ...

Training and test set

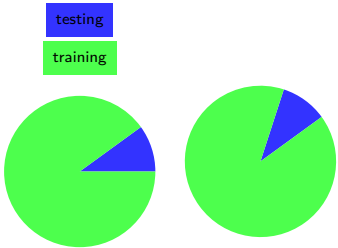
- In order to get fair and useful results, it is absolutely necessary not to train and test on the same data!
- Percentage Split: Randomly chosen 30% are used as test data (and omitted in the training)
- Cross Validation

Cross Validation



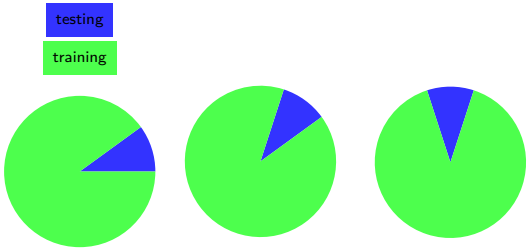
(a) Round 1

Cross Validation



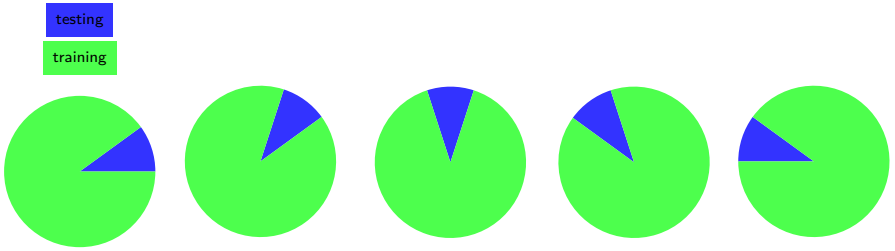
(a) Round 1 (b) Round 2

Cross Validation



(a) Round 1 (b) Round 2 (c) Round 3

Cross Validation



(a) Round 1 (b) Round 2 (c) Round 3 (d) Round 4 (e) Round 5

(f) Round 6 (g) Round 7 (h) Round 8 (i) Round 9 (j) Round 10

Figure: n-fold cross validation ($n = 10$)

Evaluation Metrics

Manning and Schütze (1999, pp. 267 ff.)

- Accuracy
  - Percentage of correctly classified instances

Evaluation Metrics

Manning and Schütze (1999, pp. 267 ff.)

- Accuracy
  - Percentage of correctly classified instances
- Precision
  - Percentage of correctly classified instances among those that have been classified as a particular class
- Recall
  - Percentage of correctly classified instances among those that belong to a class (according to the gold standard)

Evaluation Metrics

Manning and Schütze (1999, pp. 267 ff.)

- Accuracy
  - Percentage of correctly classified instances
- Precision
  - Percentage of correctly classified instances among those that have been classified as a particular class
- Recall
  - Percentage of correctly classified instances among those that belong to a class (according to the gold standard)
- F-measure
  - Harmonic mean between precision and recall

Evaluation Metrics

Manning and Schütze (1999, pp. 267 ff.)

- Accuracy
  - Percentage of correctly classified instances
- Precision
  - Percentage of correctly classified instances among those that have been classified as a particular class
- Recall
  - Percentage of correctly classified instances among those that belong to a class (according to the gold standard)
- F-measure
  - Harmonic mean between precision and recall

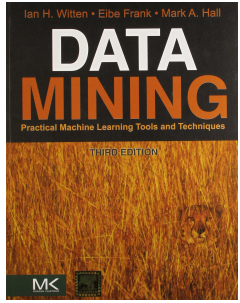
⚠ Important: P/R/F are measured *per class*, while accuracy is a single value for the entire classification

Section 3

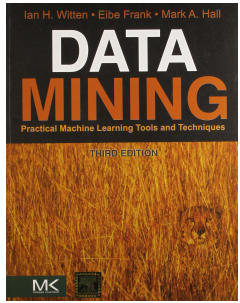
Weka

Introduction

Ian H. Witten et al. (2014). *Data Mining*. 3rd ed. Practical Machine Learning Tools and Techniques. Elsevier



Introduction



Ian H. Witten et al. (2014). *Data Mining*. 3rd ed. Practical Machine Learning Tools and Techniques. Elsevier

- Open source, Java: `https://www.cs.waikato.ac.nz/ml/weka/`
- Collection of machine learning algorithms
- Playground, GUI, well documented
- Technical limitation: Data sets have to fit in memory

Weka parts

- Preprocess: Remove attributes or instances, rebalance the data set, ...
- Classify: Train and test a classifier
- Cluster: Run a clustering algorithm
- Associate: Investigate associations between features[2]
- Select attributes: Rank attributes according to their importance for a class
- Visualize: Plotting

[2] Association != correlation

Data set: Wine quality prediction (Cortez et al., 2009)

- Given a number of physicochemical wine properties of wine, predict its quality
- Features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol Output variable (based on sensory data):
- Target class: Quality (score between 0 and 10)

demo

References I

- Cortez, P. et al. (2009). »Modeling wine preferences by data mining from physicochemical properties«. In: *Decision Support Systemts* 47, pp. 547–553.
- Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press.
- Witten, Ian H., Eibe Frank, and Mark A. Hall (2014). *Data Mining*. 3rd ed. Practical Machine Learning Tools and Techniques. Elsevier.