Machine Learning with Weka

PU Machine learning mit Java, Weka und UIMA

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

November 17, 2020
Winter term 2020/21

Take-Home message: The Machine-Learning Recipe

1. Your problem
   - What are its structural properties?
   - What do you believe/know is relevant information to make a prediction?
   - To what extent can humans solve it?
   - Do you have non-formalisable requirements?

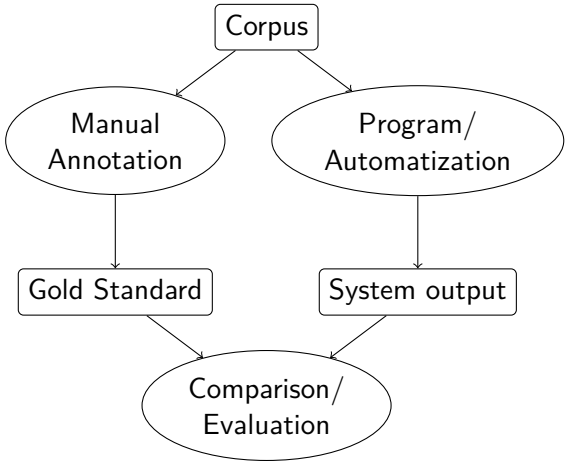
Take-Home message: The Machine-Learning Recipe

1. Your problem
   - What are its structural properties?
   - What do you believe/know is relevant information to make a prediction?
   - To what extent can humans solve it?
   - Do you have non-formalisable requirements?
2. Pick an appropriate algorithm
   - There are many out there
   - Appropriate: It should fit to your problem!
   - Gather training/testing data

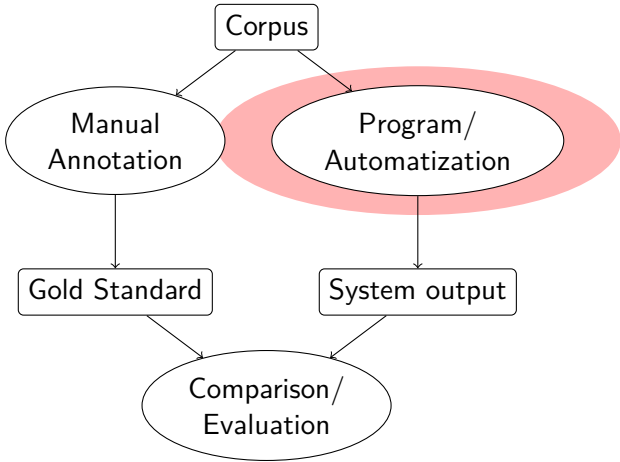
Take-Home message: The Machine-Learning Recipe

1. Your problem
   - What are its structural properties?
   - What do you believe/know is relevant information to make a prediction?
   - To what extent can humans solve it?
   - Do you have non-formalisable requirements?
2. Pick an appropriate algorithm
   - There are many out there
   - Appropriate: It should fit to your problem!
   - Gather training/testing data
3. Evaluate fairly
   - What's the performance of a realistic and optimised baseline?
   - What's an appropriate evaluation metric?
   - Report evaluation results including all settings and constraints

Experiments

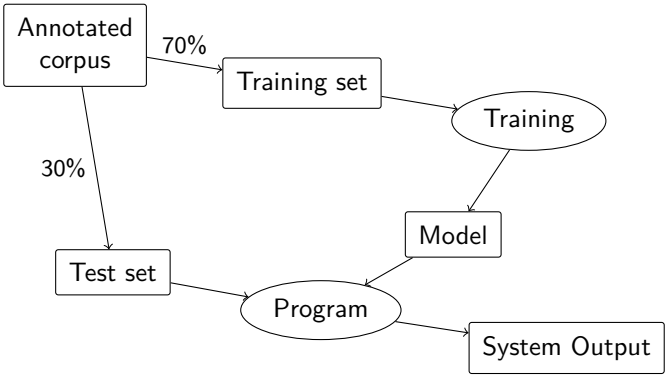


Experiments



Evaluation

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



Section 2

Machine Learning Basics

Introduction

- What is machine learning?
  - Method to find patterns, hidden structures and undetected relations in data

Introduction

- What is machine learning?
  - Method to find patterns, hidden structures and undetected relations in data
- It's all around us
  - Stock market transactions
  - Search engines
  - Surveillance
  - Data-driven research & science
  - ...

Introduction

- What is machine learning?
  - Method to find patterns, hidden structures and undetected relations in data
- It's all around us
  - Stock market transactions
  - Search engines
  - Surveillance
  - Data-driven research & science
  - ...
- Why is it interesting for text analysis?
  - Big data analyses
    - Automatic prediction of phenomena
    - Canonisation, Euro-centrism
    - Statements about 1000 texts more convincing than abt 10
  - Insights into data
    - By inspecting features and making error analysis

Machine Learning

Classification

- Assigning *classes* to *objects/instances/items*
  - Words → parts of speech
  - Photo portraits → gender of the depicted person
  - Texts → genres
- Many different algorithms available:
  - Decision trees
  - Support vector machines
  - Naïve bayes
  - Neural networks
  - Bayesian networks
  - ...

Machine Learning

Classification

- Assigning *classes* to *objects/instances/items*
  - Words → parts of speech
  - Photo portraits → gender of the depicted person
  - Texts → genres
- 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

Machine Learning

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

Machine Learning

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

- 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

Section 3

Weka

Introduction

Ian H. Witten and Eibe Frank (2005). *Data Mining*. 2nd ed. Practical Machine Learning Tools and Techniques. Elsevier

Introduction

Ian H. Witten and Eibe Frank (2005). *Data Mining*. 2nd 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
  - = Doesn't work for *really* large data sets

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!

What are classes?

- Items are classified into classes
- In the end, each item belongs into one class[1]
- Sometimes, the classes Yes and No are used

[1] We will not discuss multi-label classification here

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

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
  - Training: A classifier "learns probabilities"[2] 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

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

Features

Major part of work:

- Decide, which features should be used
- Implement code to extract them from the data

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

File formats

CSV (Comma-separated values)

- One record per line
- Feature values separated by comma (or semicolon, or tab)

Example

```
Darth, upper, ""
Vader, upper, Darth
war, lower, Vader
ein, lower, war
Lord, upper, ein
der, lower, Lord
Sith, upper, der
...
```