HMS analytical software

MACHINE LEARNING FOR HUMANS

Katja Glass
Eike Nicklas

Hands-on Workshop

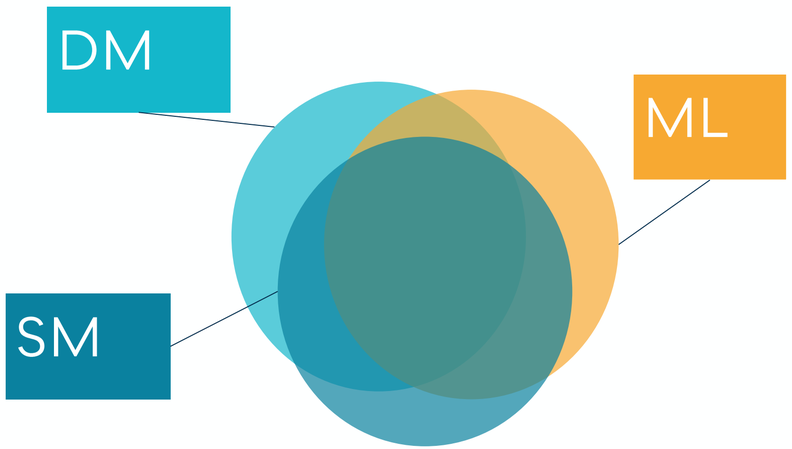
PhUSE SDE Heidelberg 2019

10.04.2019

Agenda

- Introduction to Machine Learning
- Hands-on: Breast Cancer Sample Analysis
- Summary and Wrap-Up

Statistical Modeling, Data Mining, Machine Learning



- Statistical Modeling:
  Deals more with „small data“ (compared to DM&ML). Apply statistical tests to derive as much information as possible.

- Data Mining:
  Focus more on explorative analysis (compared to SM&ML). Tries to find new information in the data.

- Machine Learning:
  Focus more on big data (compared to SM&DM). Tries to find information by „brute-force analysis“ of big data („the data speaks for itself“) -> often „black-box“ models, i.e. hard to interpret.

Machine Learning Areas

SUPERVISED LEARNING

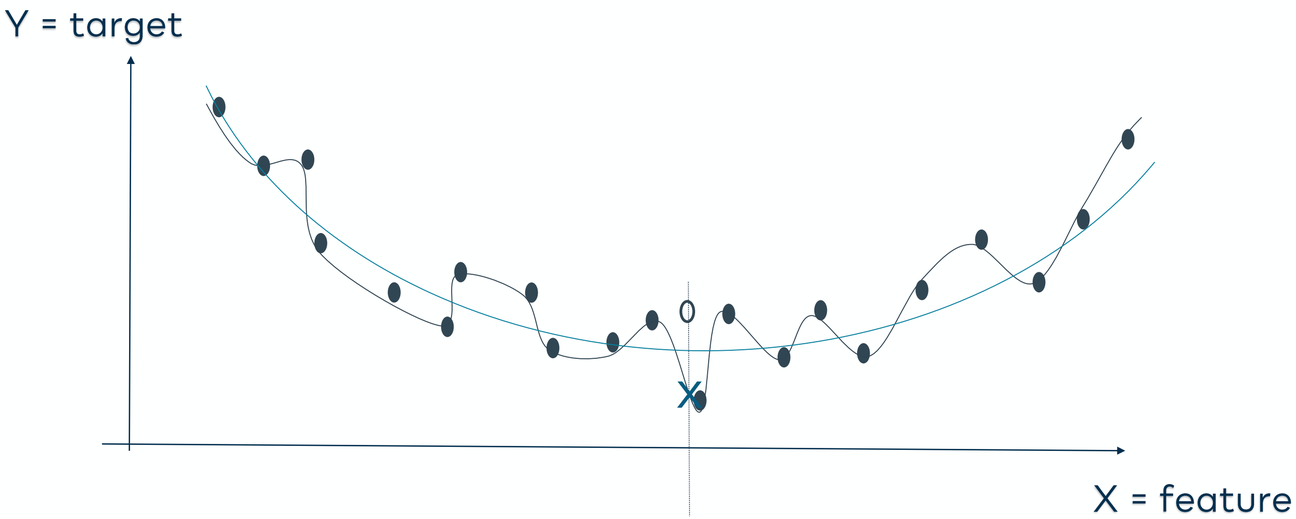
- There is a label (in my historic data)
- → I want to predict this label (in my future data)
- Example: Predict a disease based on the symptoms
- Most known algorithms: Linear Regression, Decision Trees, Random Forest, Boosted Trees, Neural Nets (Deep Learning)

UNSUPERVISED LEARNING

- There is no label
- → I want to „derive a label“
- Example: Find cluster of similar patients (in order to group them for a trial)
- Most known algorithmic approaches: Clustering, Association Analysis, Dimension Reduction

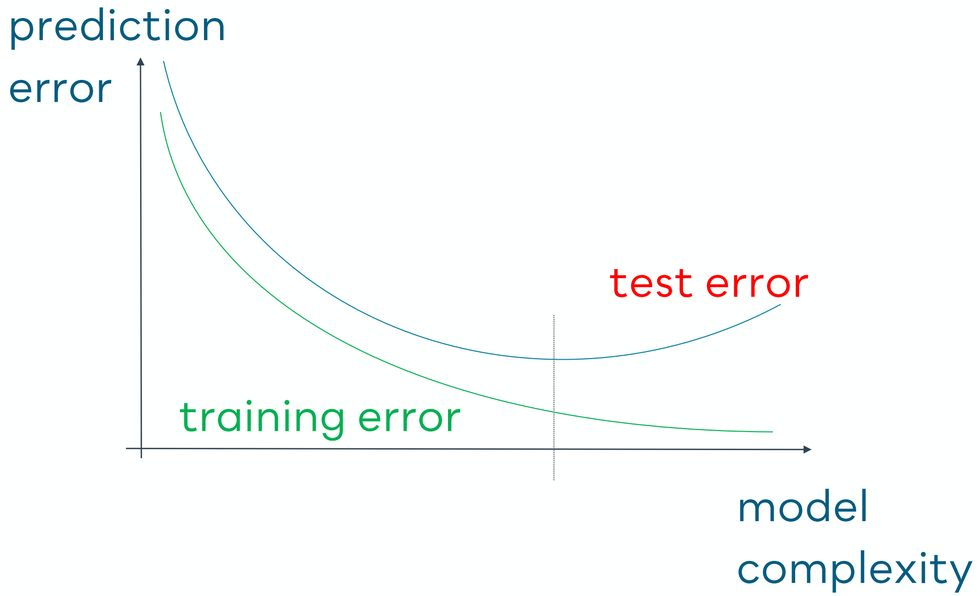
Supervised Learning

- Goal: Approximate a function that describes the relationship between features and target (=label)
  target = f(feature)
- Fit/train a model/algorithm on your data: Estimate model parameters (e.g. coefficients in a linear regression)
- Naïve approach: Minimize deviation from training data → Overfitting!
- Train/Test cycle: Fit your model on training data (by maximizing predictive accuracy) and test the „honest“ predictive accuracy on test data



Supervised Learning

- Tuning: Try several values of your tuning parameter and select the best one (regarding predictive accuracy on test data)
- Feature Engineering: Derive additional predictor variables (by applying domain knowledge)
- Imbalance: One group of the population is overrepresented



Hands-on: Breast Cancer Sample Analysis

- Activate WLAN access on your laptop
- Connect to hands-on working environment (Chrome or Firefox)
  - https://phusemlworkshop2019.westeurope.cloudapp.azure.com:8000
  - Ignore the warning about a potentially unsafe connection and proceed
  - Log-in with the credentials provided (user<xy>, PhUSE2019)
- Open the notebook **ML_Hands_On_Workshop_Phuse2019.ipynb**
- Work in teams, discuss, explore
- Raise your hand, if you run into any issues or have questions

Summary 1/2

- Predicted whether tumor is benign or malignant (Supervised Learning)
- Algorithms used: Decision Tree, Random Forest, Boosted Trees, Neural Nets
  - How to choose an algorithm?
- Concepts used:
  - Training / Test Cycle
  - Cross Validation
  - Model quality evaluation – What is the best quality metric for the task?

Summary (2/2)

Quality assurance of Machine Learning models

- What are the consequences if a prediction is wrong?
- How can I understand the black box? Or at least describe it?

Golden rules of Data Science (in the real world)
It is simple to get started, but....

- If results are to good or too bad to be true, your have made an easy fault.
- “WITT”: What is the target? I.e. the target variable as well as the overall objective of the model
- Your data is always bad: Missings, Outliers, False content, Unknown content, ...)
- KNOW YOUR DATA!
  -> Each variable: What, when, where, who,
  -> Explore, explore, explore = plot, plot, plot
- Always try to evaluate the robustness/variance/variability of your results.
- DO NOT SWITCH OFF YOUR MIND! ... because the algorithm cannot think.

HMS analytical software

THANK YOU FOR YOUR ATTENTION

Katja Glass

katja.glass@glacon.eu

Eike Nicklas

eike.nicklas@analytical-software.de