Introduction to Machine Learning and Big Data

Exercise 02

Volker G. Göhler

TU Bergakademie Freiberg

April 18, 2024

Machine Learning

Checklist

1. Frame the problem and look at the big picture
2. Get the data
3. Explore the data to gain insights
4. Prepare the data to better expose the underlying data patterns to machine learning algorithms
5. Explore many different models and shortlist the best ones
6. Fine-tune your models and combine them into a great solution
7. Present your solution
8. Launch, monitor, and maintain your system

Problem

- Get your Research Question right!
  - current solution?
  - requirements?
  - similar problems and solutions?
  - how does the manual solution look?
  - what are your assumptions?
- For the examples: it gets provided

Data

- openml.org
- kaggle.com/datasets
- paperswithcode.com/datasets
- archive.ics.uci.edu/ml
- registry.opendata.aws
- tensorflow.org/datasets

Meta Portals

- dataportals.org
- opendatamonitor.eu

Group Work Exercise

- you get a problem statement provided
- search for datasets (they could be walled by paywall/registration)
- you should get around three candidate datasets
- describe size and composition of the datasets and what they could provide to answering your question
- do a short presentation describing your problem and the datasets, what is your critique on the found datasets?
- Time: 30 min (20 min preparations, rest presentation/discussion)
- Groups: 2-4

Work with the Code

-> jupyter.lab

Home Work Assignment

- take one of the datasets out of the group work exercise (not too big, should be a table (csv))
- explore the data like with the californian housing prices
- present your findings next time

Questions?

?

Reference I