Data Mining and Machine Learning
Hyperparameter tuning and Assignment help

Gergely Horváth

November 3, 2021

Outline

1. Hyperparameter-tuning

2. Assignment help

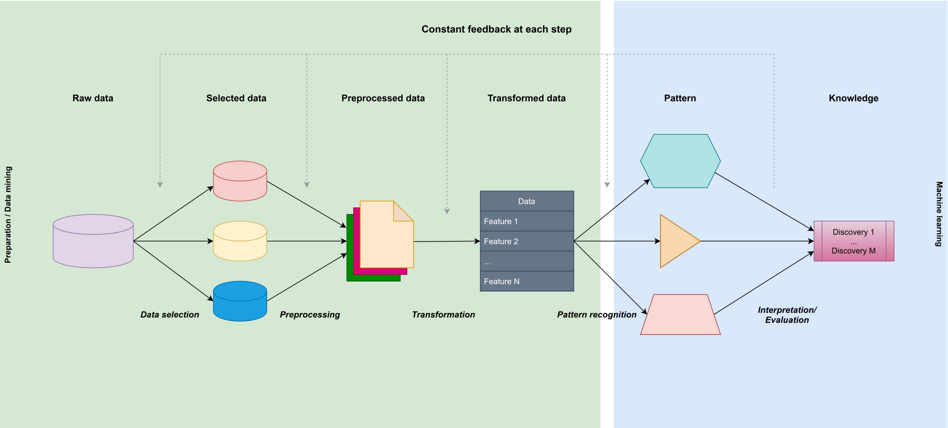
Hyperparameter-tuning

- Why?
- Lot of combinations even for the discrete case! Intractable to try each by hand!
- How?
  - Manual iteration
  - Grid-search (essentially an arbitrary subspace of an n-dimensional parameter space)
  - Random-search (random search in the n-dimensional parameter space)
  - Bayesian optimization (using acquisition function to have an approximate of the model performance w.r.t the parameter space)

Assignment details

- Kaggle competition closes at: 11th of December
- Report submission deadline: 13th of December, 11:59:59 am
- Maximum of 2 submissions per day on Kaggle
- Classification task into disease groups based on easily accessible data
- A data set of few million instances
- 43 features
- The data is not immediately usable for every algorithm, you have to apply transformation on the data-set
- You do not have to use all the data provided, sub-sampling is possible
- You do not have to use all the features provided
- Those models are appreciated, where you have a white-box solution, with the least possible features
- DO NOT leave this task to the last moment, it is advised to submit solutions almost each day!

Data science workflow*



*: From the OTSZ Hungarian Medical Journal

Assignment evaluation criteria

You do not have to be excellent at every point for maximum score!

- Kaggle position
- Public score
- Private score
- Number of uploads
- Dataset investigation
- Preprocessing
- Algorithms
- Tuning
- Performance evaluation
- Other aspects
- Code quality
- Documentation
- Delay

Assignments from the previous year

Student no. 1

Received 100% for a top Kaggle position, for trying several models (including a neural network), and by stacking relatively weaker classifiers

Student no. 2

Received 106% for extensive data investigation, feature engineering, not only performance measurements, but training and inference time measurements as well!

Student no. 3

Received 121% for preprocessing and testing with numeric evaluation, stacking and excellent model performance

sklearn.pipeline.Pipeline

- What is a pipeline? (in data science)
- Your whole data science workflow step-by-step in a coherent, well-tested framework, that lacks data and/or target leakage!
- Why do we bother?
- For prototyping it is not much of a help...
- BUT for large-scale model selection, validation and testing it is useful! As well as for preventing leakage!
- What tools do we have?
  - Manually constructed pipeline
  - *sklearn.pipeline.Pipeline*