VU Machine Learning

Winter semester 2021

Exercise 2

Nysret Musliu (nysret.musliu@tuwien.ac.at)

Exercise 2

- Groups of 3 students
- Implement a simple AutoML algorithm for regression tasks
- Perform experiments and compare to existing/other techniques
- Submit the source code
- Prepare a slide presentation
  - Around 25-40 slides, including tables & diagrams
  - No report needed (only if you prefer to write a report)
- Submission: 13.12. , 18h
- Presentations: 14/16/17.12.

Exercise 2 – Data Sets

- Pick 3 regression data sets
  - 1 data set from the previous assignment
  - Two data sets from UCI ML Repository, Kaggle… that were published after 2019
- With different characteristics
  - number of samples – small vs. large
  - number of dimensions – low vs. high dimensional

Exercise 2

- Automated Machine Learning
  - Implementation of a simple search algorithm for automated selection/configuration of machine learning algorithms
  - Comparison with other state of the art approaches

Automated Machine Learning

- Implement a simple search algorithm that searches for the best machine learning technique (and best hyperparameters) for a particular regression data set
- Various search algorithms can be implemented
  - Simple local search (hill climbing, simulated annealing…)
    - You can start with an initial algorithm/hyperparameters and then improve the solution iteratively
      - by selecting the best solution in the neighborhood (solutions that are obtained by small changes of current solution) or
      - generating a random solution (change randomly one hyperparameter or the algorithm). The solution is then accepted based on certain probability that is based on the quality of generated solution (e.g. you can use the acceptance criteria of simulated annealing)
    - …
  - Exact techniques (e.g. techniques based on tree search, …)
  - Other techniques or your original algorithm …
  - The method should be more sophisticated than gird search or random search

Implementation of algorithm

- Search space:
  - At least three available machine learning regression algorithms
  - Most important hyperparameters that should be tuned for each of these algorithms. You can specify for each hyperparameter a reasonable range of possible values
  - The aim is to find for each data set a solution (the best algorithm/hyperparameters) in the search space that optimizes an evaluation score (e.g., RMSE)
- Please write me an email if you have any questions

Implementation of algorithm

- You should implement your AutoML algorithm from scratch
- You can use existing code/functions for general parts like
  - Existing regression algorithms
  - Code for reading the input and testing the algorithm (cross- validation, performance metrics for regression…)
  - …

Comparison with other approaches

- Compare you approach with two state of the art AutoML systems (e.g. auto-sklearn, TPOT…)
- Use for comparison the selected regression data sets
- Time limit: you should use at least 1h per data set and each AutoML system

Submission

A zip file with

- **Source code:**
  - You can use any programming language: Python, Mathlab, R…
  - Provide the information for the packages needed to run you code
- **Data sets (links to the data sets)**
- **Slides**
  - Around 30 - 40 slides, including tables & diagrams
  - No report needed
- Submission deadline: December 13, 18h

Slides

- Characteristics of data sets & pre-processing (i.e. scaling etc.)
- Details regarding the implementation (pseudocode…)
  - Main issues regarding the implementation (lessons learned)
  - Some parts of the source code can be given in the slides
  - Experiments and performance metrics used
- Comparison to other techniques
- Discussion of experimental results, comparison in regard of the different datasets & techniques (tables, figures)
- Conclusions/lessons learned

Presentations

- Length of presentations
  - 15 minutes (12 minutes 3 minutes Q&A)
- You can use the slides that you submitted and skip some of them during the presentation
- You may also get questions for your source code

Evaluation of assignment

- Total number of points: 16.5
  - Implementation of algorithms end experiments: 50%
  - The choice of data sets and pre-processing : 10%
  - Comparison to other techniques: 20%
  - Conclusions, lessons learned: 20%