VU Machine Learning

WS 2024

Exercise 2

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

Exercise 2

- Groups of 3 students
- Implement a machine learning technique for predicting numeric values
- Compare to LLMs and existing techniques
- Submit the source code
- Prepare approximately 15-30 slides. No report is required for this assignment
- Individual discussions for each group (all members must be present)
- Submission: December 15th
- Discussions: December 16th– 20th (slots will be available in Tuwel)

Exercise 2 – Techniques

- Implement a **random forest algorithm (based on regression trees)** for predicting numeric values
  - Experiments with different number of trees and features for random forest
- You should implement the random forest algorithm, including regression trees, from scratch
- Please do not use any part of existing code for the implementation of your algorithm
- You can use existing code/functions for general parts like
  - Code for reading the input and testing the algorithm (cross-validation, performance metrics for regression…)

Comparison

- Apply an LLM tool to implement a random forest regressor. Discuss the differences between your custom implementation and the implementation provided by the LLM tool
- Compare the results of your implemented techniques with an existing random forest implementation and another regression technique (e.g., an existing regression tree implementation, k-NN, etc.)
  - You can use the default parameters for the existing techniques
- Use at least two performance metrics for comparison
- Conclusions
  - Performance of your algorithm regarding performance metrics for regression
  - How efficient is your algorithm
  - Comparison with an LMM tool
  - Impact of hyperparameters of random forest
  - Other findings

Exercise 2 – Data Sets

- Pick 2 regression datasets from UCI ML Repository, Kaggle…
- Should have different characteristics
  - number of samples – small vs. large
  - number of dimensions – low vs. high dimensional
- Pre-process the data set if needed (scaling, …)

Submission

A zip file with

- **Source code:**
  - You can use any programming language: Python, R, Matlab, …
  - Provide the information for the packages needed to run you code
- **Slides**
  - Around 15 - 30 slides
  - No report needed
- Submission deadline: December 15th, 23:00

Slides

- A discussion of implementations
- Comparison with LLMs implementation
- Comparision with the existing implementations/other algorithms
- Discussion of experimental results
- Conclusions/lessons learned

Individual discussions

- Length of discussion: 20 minutes
- You will have questions about
  - Implementation/Source code
  - Theoretical questions about techniques
  - Comparison with the LLM tool and the existing techniques
- All members of the group should be able to explain the code/experiments
  - Students in the same group can receive different numbers of points based on the discussion
- The evaluation will be based on your code, discussion, comparison, and conclusions/lessons learned