VU Machine Learning

WS 2020

Exercise 2

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

Exercise 2

- Groups of 3 students
- Implement two techniques for regression
- 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: 09.12.
- Presentations: 10.12., 11.12.

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 2018
- Must have different characteristics!
  - number of samples – small vs. large
  - number of dimensions – low vs. high dimensional
- Pre-process the data set if needed (scaling, missing values …)

Exercise 2 – Techniques

- Implement the gradient descent algorithm for linear regression
  - See the algorithm given in the lecture slides
  - Use the same cost function
  - Experiment with different learning rate
- Implement the k-nn algorithm for regression
  - Experiments with different k and distance functions
- You should implement these algorithms from scratch
- Please do not use any part of existing code
- You can use existing code/functions for general parts like
  - Partial derivatives, cost function, distance functions for k-nn
  - Code for reading the input and testing the algorithm (cross-validation, performance metrics for regression…)

Comparison

- Compare your implemented techniques (with best learning rate, k, …) with
  - The existing implementations of gradient descent and k-nn
  - Two other regression techniques (e.g., regression trees, random forest,…)
  - You can use the default parameters for the existing techniques
- Use at least two performance metrics for comparison
- Apply cross-validation
- Conclusions
  - How efficient are your algorithms
  - Performance of your algorithms regarding performance metrics for regression
  - Impact of learning rate, k, distance functions
  - Impact of pre-processing
  - Other findings

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**
- **Slides**
  - Around 25 - 40 slides, including tables & diagrams
  - No report needed
- Submission deadline: December 9, 18h
  - Late submission not allowed

Slides

- Details regarding the implementation (pseudocode…)
  - No source code in the slides
  - Lesson learned
- Characteristics of data sets & pre-processing (i.e. scaling etc.)
- Experiments, parameters tried 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 with parameters: 50%
  - The choice of data sets and pre-processing : 10%
  - Comparison to other techniques: 20%
  - Conclusions, lessons learned: 20%