VU Machine Learning

Summer semester 2021

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: 25.05.
- Presentations: 26.05. – 28.05.

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
- With different characteristics
  - number of samples – small vs. large
  - number of dimensions – low vs. high dimensional

Exercise 2 – Literature

- Implement a regression tree algorithm and a model tree algorithm for predicting numeric values
  - You can find various implementations for these algorithms. However, you can also apply your own ideas for splitting of instances
  - 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
  - Code for reading the input and testing the algorithm (cross- validation, performance metrics for regression…)

Literature:

https://www.cs.waikato.ac.nz/ml/weka/slides/Chapter7.pptx (Pages 49 -60)

Wang, Y., & Witten, I. H. (1997). Induction of model trees for predicting continuous classes. In M. van Someren, & G. Widmer (Eds.), Proceedings of the of the Poster Papers of the European Conference on Machine Learning (pp. 128–137).

Quinlan, J.R. (1992). “Learning with continuous classes,”. Proceedings Australian Joint Conference on ArtificialIntelligence(pp. 343–348). World Scientific, Singapore.

https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.8?casa_token=381f3hWpHysAAAAA:N7Kq-iuJIINPcOT5_yiB28-F3sbi-iyPtGI27q8srf2prsiskoOh8mySrH_c_59b8eg9GA_CP05sWQU

https://www.nature.com/articles/nmeth.4370.pdf

Comparison

- Compare your implemented techniques with the existing implementations of regression/model trees and two other regression techniques (e.g., linear regression, 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 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 (links to the data sets)**
- **Slides**
  - Around 30 - 40 slides, including tables & diagrams
  - No report needed
- Submission deadline: May 25, 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%