VU Machine Learning

Summer 2020

Exercise 2: Classification Experiments

Exercise “Classification”

- Groups of 3 students (exact)
  - Can (should ☺) use the same groups as last time
- Perform experiments in machine learning
- Write a report paper
  - 10-15 pages
    - Including tables & diagrams
    - And analysis
  - Prepare a presentation
  - Actually present (**either exercise 2, or exercise 3**)

Exercise “Classification” – details

- Use 4 data sets
  - 2 from our very own Kaggle competition
  - 1 you already chose – keep it, unless you encounter issues
  - 1 that you shall still use
    - *Must have different characteristics!!*
      - *number of samples – small vs. large*
      - *number of dimensions – low vs. high dimensional*
      - *number of classes – few vs. many classes*
      - *pre-processing needed…*
    - *Choice of diverse data sets important for grading !*
  - Need to register your chosen datasets in TUWEL
    - You can reuse the ones from Exercise 0
      - If they are classification data sets

- Chose 3 different classifiers, from at least 3 different types of learning algorithms
  - Argue & justify choice (part of grading…)
  - I.e. 4x3 combinations of dataset & classifier

Exercise “Classification” – details

- Experiment with the datasets and classifiers, by evaluating their performance
  - Chose a number of performance measures. Argue why you chose them, what they measure, and whether they are sufficient.
- Experiment with different parameter settings
  - And report on it - report not only one (best/random) result from a classifier on a specific dataset, but several results!
- Compare results among classifiers and datasets
  - Aggregated comparison, e.g. pick best settings for each combination
  - Significance testing against at least one baseline
- Evaluate effect of pre-processing (mostly scaling)
  - Compare results w/o pre-processing vs. applied pre-processing methods (be careful about built-in pre-processing in some implementation!)
- Compare holdout with cross-validation
- Record (approximate) runtimes of the classifiers
- **Summarise** your results - tables or figures

Exercise “Classification” – details

- Qualitative Analysis
  - Are there any patterns to be identified across the datasets and classifiers?
    - E.g. which methods worked generally good/bad, is there one outperforming?
    - How can you compare results on different datasets?
  - Analyse e.g. how sensitive an algorithm is to parameter settings
    - Are there any differences over the datasets?
  - How is the runtime behaviour changing with the dataset size (number of samples/features)
  - Does the pre-processing affect your results? Is there any trend?

Exercise “Classification”: Software

- Python / scikit
- WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
  - easy to use (GUI), also powerful API
- R (http://www.r-project.org/)
  - advanced & powerful software
  - if you know R already, or you want to learn it
- Rapid Miner
- Matlab

Exercise “Classification”: Written Report

- Report should be 10-15 pages
- Full report of your work
  - Experiments, parameters tried
  - Characteristics of data sets & pre-processing (i.e. handling of missing values, scaling etc.)
  - Characteristics of classifiers
  - Explanation of choice for data sets & classifiers
  - Discuss experimental results, compare them in regard of the different datasets & classifiers (tables, figures)
    - Do not include code in report, but include code & scripts in submission package
  - Analysis

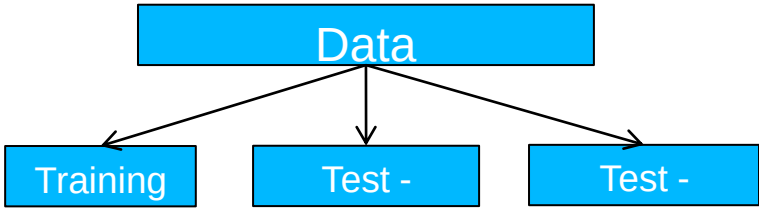
Exercise “Classification”: Presentation slides

- Presentation shall be <= 10 minutes
- Focus on the essentials of your findings
  - Only a short overview on datasets & algorithms
  - No code, ..

- Rather:
  - What was interesting, what worked, what didn’t work
  - What are your conclusions regarding the usefulness of the algorithms, on your specific datasets
  - How do the classifiers compare

Exercise “Classification”: competition

- Competition-style evaluation
  - We will use Kaggle in-class (https://inclass.kaggle.com) for a competition
  - Submission requires a simple CSV file
    - For each sample in the test set: <id>,<predicted class>
  - Pick two of the datasets provided in TUWEL
  - Number of uploads to Kaggle per day is limited ➔ start early!
    - Also so that you have an early feedback on your results compared to other groups!
    - First try locally what works
      → only then upload to Kaggle



- 15% Bonus points for the top 3 teams!
  - +5% Bonus if you also have your notebook running in Kaggle
    - See https://www.kaggle.com/notebooks
      - Keep it private, share it with mayer@ifs.tuwien.ac.at

Exercise “Classification”: step-by-step

- Get your data sets
  - Your existing ones & from Kaggle

- Import data file, scale/encode data, other preprocessing
- Run classifiers (with different parameters)
  - Select most interesting ones, ...
  - Document any problems/findings
  - Upload results from good algorithms to Kaggle
  - Not necessary to implement algorithms
    - Rely on libraries, modules etc.
    - Code just for loading data, pre-processing, running configurations, processing/aggregating results, ...
- Write your report

Questions ?