

VU Machine Learning

Winter 2017/18

Exercise 1:

Classification Experiments

- Groups of 3 students
- Perform experiments in machine learning
- Write a report paper
 - 10-15 pages
 - Including tables & diagrams
 - And analysis
 - Prepare a presentation

Exercise “Classification” – details

- Pick 4 data sets (2 from UCI ML Repository, 2 from Kaggle)
 - Must have different characteristics!!
 - number of samples – small vs. large
 - number of dimensions – low vs. high dimensional
 - number of classes – few vs. many classes
 - missing values
 - pre-processing needed...
 - Choice of diverse data sets important for grading !
 - Need to register your chosen datasets in TUWEL
 - Limitation of groups working on the same datasets
- Chose 4 different classifiers, from at least 3 different types of learning algorithms
 - Argue & justify choice (part of grading...)
 - I.e. 4x4 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 (e.g. different strategies for missing values)
 - Compare results w/o pre-processing vs. applied pre-processing methods (be careful about built-in pre-processing in some implementation!)
- Record (approximate) runtimes of the classifiers

- 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?

- 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”: 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>
 - For a certain number of datasets (you need to chose 2 of those)
 - List will be provided in TUWEL

Exercise “Classification”: how-to

- Get your data sets from the UCI ML Repository:
<http://www.ics.uci.edu/~mlearn/MLSummary.html>
- Import data file, scale/encode data, other preprocessing
- Run classifiers, perform model selection, ...
 - Document any problems/findings
- Matlab/R/APIs: not necessary to implement algorithms – rely on libraries, modules etc.
 - Code just for loading data, pre-processing, running configurations, processing/aggregating results, ...

Questions ?