VU Machine Learning

Exercise 0:
Dataset description

Rudolf Mayer

(mayer@ifs.tuwien.ac.at)

Exercise “Dataset description”

- Select two **classification** datasets sets, from either
  - UCI ML Repository (http://www.ics.uci.edu/~mlearn/), or
  - Open ML (https://www.openml.org/search?type=data)

  - The Datasets should have different characteristics, e.g. differ in
    - Number of samples – small vs. large, or
    - Number of dimensions – low vs. high dimensional, or
    - Types of attributes (numeric vs. categorical),

  - And require some form of pre-processing, e.g.
    - Missing values (i.e. some rows have no values for some attributes), or
    - Scaling of attributes
    - …

Exercise “Dataset description”

- Groups of 3 students (exact)
  - Register for a group on TUWEL
- Need to register your chosen datasets in TUWEL
  - Limitation of # of groups working on the same datasets
- You will re-use these datasets for the next exercise
  - (You **may** change them if you do run into issues with them)

Exercise “Dataset description”: Written Report

- Report should be ~2 pages
  - Make sure that the document contains information on the group members that contributed

- Explanation of choice for data sets

- Characteristics of data set
  - How many samples, how many attributes
  - What types of attributes (nominal, ordinal, interval, ...)
    - See slides of first lectures
  - Distribution/histograms of values in selected input and target attributes
    - Don’t need to show all attributes, but the interesting ones

- Do not include code in written report
  - But include code / scripts in submission package
    - All plots etc. should be re-creatable from the code/scripts

Written Report: pointers on what to describe

- Target attribute
  - Distribution/range of values
    - Why is this important?
- Numeric values
  - Description on value ranges
  - Whether you need to treat these attributes in a pre-processing step
- Categorical data: which types? nominal, ordinal, ...
  - Why is that important?
- Other important aspects

Exercise “Dataset description”: Software

- Rely on libraries, modules to load data, plot, visualise, etc.
  - You need to develop just the boilerplate code/scripts
  - Do not use a GUI-only tool, as that generally does not allow to reproduce / automate the analysis

- Tools:
  - Python (using e.g. numpy, scikitlearn, matplotlib, …)
  - R (http://www.r-project.org/)
    - Recommended: use R only if you know already how to program it - it is likely too much to learn it along the exercises…
  - Matlab (or Octave)
    - Again, best if you know it already!
  - WEKA (http://www.cs.waikato.ac.nz/ml/weka/) (use the API!)
  - …

Questions ?