VU Machine Learning

Exercise 0: Dataset description

Rudolf Mayer
(mayer@ifs.tuwien.ac.at)

Exercise “Dataset description”

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

  - Datasets should have different characteristics
    - number of samples – small vs. large
    - number of dimensions – low vs. high dimensional
    - missing values (i.e. some rows have no values for some attributes)
  - Choice of diverse data sets important for grading!

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 exercises
  - (you **may** change them, but then you will have to repeat the dataset description for that exercise)

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 lecture
  - Distribution/histograms of values in the input and target attributes
- Do not include code in written report
  - But include code & scripts in submission package (if you didn’t use just a GUI tool)

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

- Tools:
  - Python / scikitlearn
  - 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
  - Matlab
  - Rather not useful: GUIs (cannot easily reproduce / automate)
    - Rapid Miner
      - Very simplified GUI
    - Orange Data Mining: https://orange.biolab.si

Questions ?