Lab on Preprocessing

Introduce the datasets *sick*, *vote*, *mushroom* and *letter.*

1. Apply discretization:

- uderstand what discretization is
- load the *sick* dataset and look at the attributes
- classify using NB, evaluating with cross-validation
- apply the supervised discretization filter and look at the effect (in the Preprocess panel)
- apply unsupervised discretization with different numbers of bins and look at the effect
- use the FilteredClassifier with NB and supervised discretization, evaluating with cross-validation
- repeat using unsupervised discretization with different numbers of bins
- compare and interpret the results.

2. Apply feature selection using CfsSubsetEval:

- understand what feature selection is
- load the mushroom dataset and apply J48, IBk and NB, evaluating with cross-validation
- select attributes using CfsSubsetEval and GreedyStepwise search
- interpret the results
- use AttributeSelectedClassifier (with CfsSubsetEval and GreedyStepwise search) for classifiers J48, IBk and NB, evaluating with cross-validation
- interpret the results.

3. Apply feature selection using WrapperSubsetEval:

- load the vote dataset and apply J48, IBk and NB, evaluating with cross-validation
- select attributes using WrapperSubsetEval with InfoGainAttributeEval and RankSearch, with the J48 classifier
- interpret the results
- use AttributeSelectedClassifier (with WrapperSubsetEval, InfoGainAttributeEval and RankSearch) with classifiers J48, IBk and NB, evaluating with cross-validation
- interpret the results.

4. Sampling a dataset:

- load the *letter* dataset and examine a particular (numeric) attribute
- apply the Resample filter to select half the dataset
- examine the same attribute and comment on the results.