Data Mining Lab Mini Talk

- Preprocessing
  - Discretization
  - Feature Selection
- Credits to Stefan Kramer for slide material
Discretization

- Dealing with numeric attributes:
  - cannot be handled by the learning scheme
  - performance is improved => overfitting
  - the range of the feature is divided into a set of intervals

- Supervised vs. Unsupervised:
  - Supervised consider the relation of the attribute values to the class values
  - Unsupervised: only look at the distribution of values of the attribute
Unsupervised Discretization

1. Domain-dependent
   - Age: “baby” if in (0,3], “child” if in (3,6], “school child” if in (6,10], “teenager” if in (10,18]

2. Equal-width
   - divide value range into a number of intervals of equal width

3. Equal-frequency
   - divide value range into a number of intervals so that (approximately) the same number of data points are in each interval
Supervised Discretization

1. Entropy Split (Fayyad & Irani, 1993)
   - splitting (top-down): starts with single interval and successively splits the interval into sub-intervals
   - stops when a given number of intervals is reached or intervals becoming too small
   - entropy as splitting criterion

2. ChiMerge (Kerber, 1992)
   - merging (bottom-up): merging of adjacent intervals
   - use $X^2$–statistics to determine pair of interval to be merged

Data Mining Lab SS2015
Feature Selection

- Many features may be:
  - irrelevant
  - redundant
- Removing them can:
  - increase efficiency
  - improve accuracy
  - prevent overfitting
- Feature (subset) selection techniques try to determine appropriate features automatically
Unsupervised Feature Selection

• Using domain knowledge: some features may be known to be irrelevant or redundant, common sense

• Random sampling:
  • select a random sample from the features
  • may be appropriate in case of many weakly relevant features or in connected with so-called ensemble methods
Supervised Feature Selection

- Filter approaches:
  using some evaluation measure of attribute with respect to class

- Wrapper approaches:
  using learning algorithm as plug-in to evaluate feature set(s)
Feature Measures for Filters

- **Gini-Index**: describes how a given attribute supports the partition of a set of instances into two subsets with respect to the class label (0 no class separation at all, 1 perfectly separated classes)

- **Information Gain**: calculates the entropy reduction for the split in a given attribute

- **Relief**: determines attribute weights for best separation by distance to near Hit and nearMiss instance
Wrappers

- Search through the space of possible feature subsets
- Each subset encountered in search is tried with a learning algorithm
- Error rate in cross-validation as evaluation function
- Improve it by modifying the feature subset based on the result
Pro’s / Con’s

- Disadvantage:
  - very inefficient for certain learning schemes: many cycles necessary
  - higher risk of overfitting

- Advantage:
  - feature subset is tailor to the learning algorithm
  - can consider combination of features
  - can eliminate redundant features
Strategies

- **Forward Selection:**
  - start with trying a single feature
  - select and add feature with the best performance
  - new iteration to get the next best feature
  - terminates upon fixed number of features or plateau

- **Backward Selection:**
  - Starts with full feature set
  - search for attribute with the least loss of performance
  - new iteration to get next elimination candidate
Thank you for your attention