Statistics 202: Data Mining
Midterm review
Based in part on slides from textbook, slides of Susan Holmes

©Jonathan Taylor

October 29, 2012
Midterm review

Overview

- General goals of data mining.
- Datatypes.
- Preprocessing & dimension reduction.
- Distances.
- Multidimensional scaling.
- Multidimensional arrays.
- Decision trees.
- Performance measures for classifiers.
- Discriminant analysis.
Midterm review

General goals

- Definition: what is and isn’t data mining.
- Different types of problems:
  - Unsupervised problems.
  - Supervised problems.
  - Semi-supervised problems.
Midterm review

Datatypes

- Continuous, discrete, etc.
- How data is represented in R.
- Descriptive statistics for different datatypes.
- General characteristics:
  - Is it observational or experimental?
  - Is it very noisy?
  - Is there spatial or temporal structure?

Preprocessing

- General tasks: aggregation, transformation, discretization.
- Feature extraction: wavelet / FFT transform.
- Another type of feature extraction: dimension reduction.
Midterm review

**Dimension reduction**
- PCA as a dimension reduction tool.
- PCA in terms of the SVD.
- PCA loadings, scores.

**Graphical summaries**
- Various plots of univariate continuous data:
  - stem-leaf;
  - histogram;
  - density (using kernel estimate);
  - ECDF;
  - quantile.
  - Boxplot.
  - Pairs plot.
  - Correlation / similarity matrix (cases or features)
Midterm review

**Multidimensional scaling**
- Relation between distances and similarities.
- Graphical (Euclidean) representation of data that is “closest” to original dissimilarity.
- Relation to PCA when similarity is Euclidean.

**Multidimensional arrays**
- Data cubes.
- Standard types of operations on data cubes.
- A few examples in R.
Decision trees

- Definition of a classifier.
- Specific form of a decision tree classifier.
- Applying the decision tree model.
- Fitting a decision tree using Hunt’s algorithm.
- Various measures of impurity: Gini, entropy, Misclassification Rate.
- Pre and post-pruning to simplify tree.
Midterm review

Performance measures

- Sensitivity, specificity, true/false negatives, true/false positives.
- Confusion matrix.
- Using a cost matrix to evaluate a classifier.

Discriminant analysis

- Linear / Quadratic Discriminant Analysis;
- Logistic regression.
- ROC curve.