Final Review Topics  
CSCI 406 – Data Mining  
Stephanie Schwartz

1. Decision Trees in R (only the rpart library)  
   a. Interpreting R output  
   b. Classification vs Regression trees  
   c. Pruning  
   d. Error rates  
2. Cross Validation and Sampling  
   a. Training, testing, validation sets  
   b. Cross validation, k-fold and leave-one-out  
   c. Simple random sampling  
   d. Weighted sampling  
   e. Sampling with and without replacement  
   f. Stratified sampling  
   g. Determining sample size  
3. Ensemble Methods  
   a. Rationale  
   b. Construction methods, manipulating  
      i. Training methods – bagging and boosting  
      ii. Input features – random forest  
      iii. Class labels (general concept only)  
      iv. Learning algorithms (general concept only)  
4. Feature Subsets  
   a. The curse of dimensionality  
   b. Systematic approaches for choosing features (embedded, filter, wrapper)  
   c. Adjusted $R^2$ statistic  
5. Class Imbalance  
   a. Confusion Matrix Counts: True Positive, False Negative, False Positive, True Negative  
   b. Confusion Matrix Rates (TPR, FNR, FPR, TNR)  
   c. Precision and recall  
   d. $F_1$ measure  
   e. ROC curve  
   f. Multiclass problems  
6. Support Vector Machines  
   a. Maximal margin classifier (definition, when is it applicable…)  
   b. Support vector classifier (definition, when applicable, tuning parameter $C$…)  
   c. Support vector machine (definition, when applicable)  
   d. Output in R, review provided examples  
7. Clustering  
   a. Unsupervised vs supervised learning  
   b. Applications
c. Partitional vs Hierarchical, Complete vs Partial, Exclusive vs Non-Exclusive
d. Types of clusters (well-separated, center-based, contiguous)
e. K-means clustering
   i. Choosing initial centroids (and problems with this)
   ii. How to measure distance/assign points to clusters
   iii. Evaluations
   iv. Empty clusters, outliers
f. Hierarchical clustering
   i. Agglomerative vs divisive
   ii. Algorithm
   iii. Defining proximity
g. Output in R, review provided examples

8. Nearest Neighbors
   a. Similarity vs dissimilarity
   b. Euclidean vs Manhattan distance
   c. Distance matrix
   d. Weights and standardization
   e. K-nearest neighbors algorithm
   f. Lazy vs eager learning algorithms
   g. Output in R, review provided examples