Course Project: Data Mining (Fall 2016)

Name: _____ Student ID: _____ Grade: _____

Notes:

- **Print this sheet on A4 paper and append your project report as an appendix.**
- **You may complete the project in any programming language you prefer.**
- **The questions marked by * are optional. Answering these questions may get bonus points.**
- **Course project due date: November 14, 2016.**

1. Implement decision tree induction and deduction algorithms and evaluate decision tree classifiers.
   (a) Implement a decision tree induction algorithm that selects the best splitting attribute according to the information gain. Download three data sets from the UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/`). Train and evaluate decision trees on the data sets using the holdout method, the 10-fold cross-validation method, and the Bootstrap method.
   (b) Implement a decision tree induction algorithm that selects the best splitting attribute according to the Gini index. Train and evaluate decision trees on the data sets given in part (a).
   (c) Implement a decision tree induction algorithm that selects the best splitting attribute according to the misclassification error. Train and evaluate decision trees on the data sets given in part (a).
   (d) Based on the decision tree induction algorithm in part (a), implement a decision tree induction algorithm with pre-pruning and a decision tree induction algorithm with post-pruning. Compare the decision trees inducted by the three algorithms on the data sets given in part (a).
   (e) *Based on the decision tree induction algorithm in part (a), implement a decision tree induction algorithm that can handle missing values. Randomly generate data sets with missing values based on the data sets given in part (a). Train and evaluate decision trees on these incomplete data sets. Explain the effect of the fraction of missing values on the error rate of a decision tree.
   (f) Implement the AdaBoost algorithm. Use the decision tree induction algorithm in part (a) to train the base classifiers. Evaluate the ensemble classifier on the data sets given in part (a). Compare the evaluation results with the ones obtained in part (a).
2. Implement and evaluate the k-means algorithm.
   (a) Download three data sets from the UCI Machine Learning Repository. Evaluate the k-means algorithm for different k values. Explain the effect of k on the sum of squared errors (SSE) of the clusters found by the k-means algorithm. Show the best k value for each data set.
   (b) For a specific value of k, evaluate the k-means algorithm with respect to different initial centroids on the data sets given in part (a). Explain the effects of initial centroids on the quality of clusters.
   (c) *Implement the bisecting k-Means algorithm. Compare it with the k-means algorithm on the data sets given in part (a) in terms of (but not limited to) time efficiency and clustering quality.