CSCI 4380/6380 Data Mining
Second Midterm Exam - Fall 2015
Open Notes

NAME:

Problem(1):

Problem(2):

Problem(3):

Total:

1. **[10 points] Short answers please**

   (a) Give two methods for choosing a good value for K (number of clusters) when using K-means clustering.

   (b) Why is it customary to initialize the weights of a back-propagation neural network with small random values?

   (c) Why is the re-substitution (training set) error usually unreliable for performance evaluation?

   (d) **For 6380 students only** When would the correlation coefficient be better than relative, absolute and squared errors for evaluating numeric prediction?

2. **[10 points] Short answers please**

   (a) Why is over-fitting less likely to happen using support vector machines than many other methods?

   (b) Give two ways to speed up back-propagation learning of neural networks.

   (c) Give one advantage for using model trees over regression trees for continuous prediction.

   (d) Give one advantage for using regression trees over model trees for continuous prediction.

3. **[10 points] Short answers please**

   (a) Give two major differences between bagging and boosting.

   (b) Why is it customary to initialize the weights of a back-propagation neural network with small random values?

   (c) Give one advantage for using ensemble learning instead of single classifier learning.

   (d) **For 6380 students only** Which method do you think should be used as meta learner for stacking in classification problems? which method should be used for regression problems? Briefly justify your answers.