CSCI 4900/6900 Data Mining

Assignment Number 5: Due 4/17/2014 **(in class)**

1. **[20 points] Short answers please**

   (a) Give a major difference between lift charts and ROC curves.

   (b) Give one advantage to using the mean absolute error instead of the root mean squared error for performance evaluation.

   (c) Give one advantage to using the relative absolute error instead of the mean absolute error for performance evaluation.

   (d) Give an example of a data mining domain in which recall is more important than accuracy.

2. **[20 points] Short answers please**

   (a) Why is FP-growth usually much faster than Apriori for finding large item sets?

   (b) Give one advantage to using ball trees instead of kD trees for finding nearest neighbors.

   (c) Give one advantage to using post-pruning instead of pre-pruning for decision tree learning.

   (d) Give one advantage to using sub-tree replacement instead of sub-tree raising for decision tree pruning.

3. **[60 points]**

   Choose four of the research papers presented (or to be presented) in class and provide a brief review for each of them (no more than half a page each). Your review for each paper should include the paper title and may also include:

   - Which data mining technique(s) was/were used in the paper
   - What was the best thing you liked about the paper
   - What was the worst thing about the paper
   - How can you continue/extend the research described in the paper

   However, you may follow a different structure or include other items instead of or in addition to the suggested items above.