What you need to know about mid-term

April 8, 2013

The exam will consist of 4 questions. Each question will have two parts. The four questions be on Map-Reduce, Clustering, Bayes Rule and Classification. You will need to bring a scientific calculator along. You are allowed to bring one A4 single-sided paper of notes.

1. (MR) Express computations in map-reduce framework. For example, dot product, database join, matrix multiplication [Read Chapter 2 from the Book]

2. (MR) Understand why redundancy is important in parallel settings. Give a probabilistic explanation.. $(1 - (1 - p)^n)$.

3. (MR) What is cluster computing?

4. (MR) How can k-means algorithm be expressed using map-reduce.

5. (Clustering) Take a simple data set and work through the k-means algorithm.

6. (Clustering) What is the difference between the k-means problem and the k-means algorithm. Does the k-means algorithm “solve” the k-means problem.

7. (Bayes Rule) You should be very familiar with Bayes Rule.

8. (Bayes Rule) How Bayes rule is the basis of classification problems.

9. (Bayes Rule) What is the Naive Bayes Algorithm. Why is it called Naive. What is the advantage and disadvantage of making the conditional independence assumption.

10. (Classification) How are classifiers evaluated.

11. (Classification) What is FPR, TPR, FNR, TNR.

12. (Classification) What is Precision and Recall.

13. (Classification) How to construct the ROC Curve.