CS Ph.D. Qualifying Exam in CSE
Fall 2007
November 5, 2007, 9am-5pm

This exam has two sections: CSE algorithms and Data Analysis. Each sectino has five questions. You are supposed to answer six questions total (three from each section or four from one section and two from the other). If you answer more than six questions, then only the six lowest scored answers will count toward your total. This is an open-book/open-note exam but you are not allowed to discuss the exam with anyone. In case you need clarification, you may contact Profs. David Bader or Alex Gray. You need not type your answers if you need to save time but please write very clearly.

1 CSE Algorithms

1. A permutation of the integers from 1 to n is a sequence A with $a_1, a_2, \ldots, a_n$, where each integer appears exactly once. We define the *distance* of the permutation as the minimum number of reverals needed to transform the permutation to the identity $1, 2, \ldots, n$. A reversal (i, j) on the permutation A returns $a_1, a_2, \ldots, a_{i-1}, a_j, a_{j-1}, \ldots, a_i, a_{j+1}, a_{j+2}, \ldots, a_n$ What is the complexity to compute the distance? Make a strong case that justifies your answer.

2. A social network graph includes n vertices representing people and m edges, where an edge (i, j) connects v_i and v_j $(i \neq j)$ when person i and person j are friends. A clique is defined as a set of friends who all know each other (that is, there is an edge between each pair in the clique.)

   (a) How hard is it to find the largest clique in the social network graph?

   (b) Give an algorithm that returns the size of the largest clique (high-level description is fine).

   (c) Estimate the time this would take to run (in seconds) using a modern sequential computer and using the data extracted from Facebook.

3. Using the same social network graph described in Problem 2, give a high-level algorithm that finds people who are *central*, where *central* is defined as sitting on the highest number of shortest paths between all people in the graph. What is the algorithm's complexity? Estimate its running time (in seconds) on the Facebook network graph.

4. A *cut* in a graph is the set of edges that, when removed, separates a graph into two or more components. A *minimum edge cut* is the smallest number of edges that separates the graph into two components of nearly equal size (i.e., the number of vertices in each component).

(a) Give an algorithm that can give an approximate minimum edge cut partitioning of the graph. Analyze the cost of the algorithm.

(b) Give a qualitative analysis of how this method will perform on the following types of graphs:

- Erdős-Renyi random graphs
- 3-dimensional torus graph
- 2-dimensional finite-elements mesh
- A social network graph (e.g. the Facebook graph)

5. Given n points in 2-d space, give a cache-oblivious data structure that can return nearest neighbor queries.

2 Data Analysis

1. On generalization:

   (a) What is meant by 'generalization'?

   (b) How could we reasonably safely conclude that, for a given dataset, in practice, one classifier generalizes better than another?

   (c) How is learning theory related to the concept of generalization?

   (d) Does the concept of generalization apply only to classification? What about regression?

   (e) What about density estimation?

   (f) What about clustering?

2. In some emerging applications such as bioinformatics, typically we are trying to model complex phenomena, possibly with relatively few data points. Give examples of model selection methods which are most appropriate in this setting and examples of which are less appropriate in this setting, and explain why.

3. Why might someone want to be a Bayesian, in general? Why might someone want to be a frequentist, in general? Describe at least one setting in which Bayesian inference might have advantages, and at least one in which frequentist inference might have advantages.

4. Two key limitations of EM-based algorithm for learning Gaussian mixtures are:
   i. The number of mixture components has to be specified in advance.
   ii. The EM approach may find only a local minimum in the data likelihood.

   For both of these limitations answer the following questions.

(a) Describe approaches that can, in part at least, address the limitations.

(b) Discuss one significant strength of each of your approaches.

(c) Discuss one significant weakness of each of your approaches.

5. This problem is related to SVM binary classification and we only consider linearly separable case.

   (a) Show that a training set is linearly separable if and only if the convex hulls for the positive class samples (H+) and negative class samples (H-) do not intersect.

   (b) Why for a linear classifier large margin is desirable?

   (c) For the following training set with 2-dimensional feature vectors:
   one sample for the positive class: (0,0)
   two samples for the negative class: (0,1), (1,0)
   what is the maximum margin linear classifier?

   (d) Show that the SVM maximum margin linear classifier for the linearly separable case can also be found as follows: pick two points, one from H+ and one from H-, that are closest to each other, the hyperplane that bisects the line segment connecting the two points gives the maximum margin linear classifier. Convince yourself this is the case using the example in (c).