CS 493: Algorithms for Massive Data Sets
Homework 3
Due: Thu, Apr 18

1. *(20 points)*
Consider the problem of learning a binary concept. Suppose we have a space of hypotheses H from which the target hypothesis is chosen uniformly and at random. We are also given labeled data D, consistent with the (unknown) target hypothesis.

   (a) What is the Bayes optimal classifier for this setting ?

   (b) Suppose we pick a random hypothesis h from the version space $VS_{H,D}$. Prove that the expected misclassification error for a new example is at most twice the optimal misclassification error of the Bayes optimal classifier.

2. *(20 points)*
Prove that the clusters generated by the Single-link algorithm are identical to the clusters obtained from the MST based graph theoretic approach. (If necessary, you can assume that all distances between pairs of points are unique).

3. *(20 points)*
Consider Gonzalez's furthest point heuristic for clustering: We pick a point x_1, then pick a point x_2 that is furthest away from x_1, then pick x_3 that maximizes $\min(d(x_1, x_3), d(x_2, x_3))$ and so on. Suppose we treat the first k points picked in this way as cluster centers and assign each to the closest center. Prove that the maximum diameter of the clusters produced is at most twice the optimal diameter, i.e. if there is an optimal solution that assigns the points to k clusters with maximum diameter D, then the diameter of the clustering produced by this algorithm is at most $2D$.

4. *(20 points)*
Consider the following (erroneous) scheme for estimating $F_3 = \sum_{i=1}^{n} m_i^3$, which is a generalization of the scheme for F_2: For each i, we pick X_i to be one of the cube roots of unity $\{1, \omega, \omega^2\}$ uniformly and at random. (Assume each X_i is picked independently). In one pass over the data, we compute $Z = \sum_{i=1}^{n} m_i X_i$. Our estimator is $Y = Z^3$.

   (a) Show that $\mathrm{E}[Y] = F_3$.

   (b) Explain the flaw in this scheme, i.e. why the estimator Y cannot be used to estimate F_3 using a small amount of space.

5. *(20 points)*
Given a collection of sets show how you can represent each set by a short sketch so as to estimate the following functions for any pair of sets A, B from their sketches:

(a) $|A \cup B|$

(b) $\frac{|A-B|}{|A \cup B|}$.