Problem Set 2

October 16, 2012

Due date:  Mon, October 29 2012 at 4pm.

Exercise 1:  (20 points) Assume two d-dimensional real vectors $x$ and $y$. And denote by $x_i$ ($y_i$) the value in the i-th coordinate of $x$ ($y$). Prove or disprove the following statements:

1. Distance function
   \[ L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i| \]
   is a metric.  (5 points)

2. Distance function
   \[ L_2(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \]
   is a metric.  (5 points)

3. Distance function
   \[ L_3^2(x, y) = \sum_{i=1}^{d} (x_i - y_i)^2 \]
   is a metric.  (10 points)

Exercise 2:  (30 points) The $k$-means clustering problem takes as input $n$ points $X$ in a $d$-dimensional space and asks for a partition of the points into $k$ parts $C_1, \ldots, C_k$. Each part $C_i$ is represented by a $d$-dimensional representative point $r_i$ such that

\[ \sum_{i=1}^{k} \sum_{x \in X, x \in C_i} d(x,r_i) \]

is minimized. In the $k$-means problem, the distance $d(x,r_i)$ from a point to its corresponding representative is $d(x,r_i) = L_2^2(x,r_i)$, and $r_i$ is the mean of the points in cluster $C_i$.

In class, we mentioned that the $k$-means clustering problem is NP-hard for $d \geq 2$. However, the $k$-means problem for $d = 1$ can be solved optimally in polynomial time. Design a polynomial-time algorithm for the $k$-means problem for $d = 1$. Compute the running time of your algorithm.
Exercise 3: (30 points) The k-center clustering problem takes as input n points X in a d-dimensional space and asks for a partition of the points into k parts C₁, . . . , Cₖ such that

$$\max_{i=1}^{k} \max_{x,y \in C_i} d(x, y)$$

is minimized. In the k-center problem, the distance d(x, y) between two points is measured using an any metric d.

In class, we mentioned that the k-center clustering problem is NP-hard for d ≥ 2. However, the k-center problem for d = 1 can be solved optimally in polynomial time. Design a polynomial-time algorithm for the k-center problem for d = 1. Compute the running time of your algorithm.

Exercise 4: (10 points) Consider a set of n points X = x₁, . . . , xₙ in some d-dimensional space, and distance function d(xᵢ, xⱼ) = L₂²(xᵢ, xⱼ). Let ¯x be the d-dimensional vector that is the mean of all the vectors in X. Prove that ¯x minimizes $\sum_{x_i \in X} d(\bar{x}, x_i)$.

Exercise 5: (10 points) Recall the problem of co-clustering (or biclustering) that we discussed in class. Think of cases where the regular clustering of the data points into k clusterings will give identical results as co-clustering.