# Problem Set 2

October 4, 2013

**Due date:** Mon, Oct 21, 2013 at 4pm; before class.

**Exercise 1 (20 points):** A *locality-sensitive hashing* function is a distribution on a family $\mathcal{F}$ of hash functions operating on a collection of objects, such that for two objects $x, y$,

$$\mathbf{Pr}\left[h(x) = h(y)\right] \quad = \quad sim(x, y). \tag{1}$$

The term $sim(x, y) \in [0, 1]$ is some similarity function defined on the collection of objects and $h()$ is a hash function from $\mathcal{F}$.

Show that for any similarity function *sim* that admits a locality sensitive hash function family as defined in Equation (1), the distance function $1 - sim(x, y)$ satisfies the triangle inequality.

**Exercise 2 (30 points):** Consider the following 1-dimensional clustering problem: given a set of $n$ 1-dimensional data points $X = \{x_1, \ldots, x_n\}$ and an integer $n$ partition these points into $k$ clusters such that

$$\sum_{i=1}^{k} \sum_{x_j \in C_i} (x_j - \mu_i)^2$$

is minimized. In the above equation, $C_i$ is the $i$-th cluster and $\mu_i$ is the representative of this cluster – in this case the mean of the data points in the cluster.

Design a polynomial-time algorithm for finding the optimal clustering $\mathcal{C}$.
[**Bonus points:**] If your algorithm runs in time linear in $n^2k$, then you get extra 10 points.

**Exercise 3 (25 points):** Consider a set of $n$ data points $x_1, \ldots, x_n$.

- Assume a random number generator $R()$ that generates values in the interval $[0, 1]$. Let distance function $d_R$ between two points $x_i$ and $x_j$ be $d_R(x_i, x_j) = R()$. Prove or disprove that $d_R$ is a metric. (10 points)
- Construct a graph $G = (V, E)$, where a point $x_i$ is represented by node $v_i \in V$. For every pair of nodes $v_i, v_j$, there exists a directed edge in $G$ with weight $w_{ij} = d_R(x_i, x_j)$. We define the distance function between two nodes $v_i$ and $v_j$, denoted by $d_G(v_i, v_j)$, be the *weight of the shortest path between* $v_i$ *and* $v_j$ in graph $G$. Prove or disprove that $d_G$ is a metric. (15 points)

**Exercise 4 (25 points):** Consider the edit distance between two labeled graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ with the same set of nodes to be the number of edges in $E_1$ that are not in $E_2$ plus the number of edges in $E_2$ that are not in $E_1$. That is,

$$\Delta(G_1, G_2) = |E_1 \setminus E_2| + |E_2 \setminus E_1|.$$

- Prove that $\Delta()$ is a metric. (10 points)
- Given a set of graphs $G_1, G_2, \ldots, G_n$ consisting of $n$ graphs all sharing the same set of labeled nodes design an algorithm for finding the centroid of the set of clusters, when distance $\Delta$ is used as a distance function between graphs. (15 points)