Problem Set 1

September 15, 2014

Due date: Wed, Oct 1 2014 at 1pm; before class.

Instructions: Return your answers to the teaching assistant, Natali Ruchansky by email `natalir@bu.edu`. You can either hand your paper in person or send it by email.

Discussing the problems with your colleagues is perfectly fine. However, you should write the answers by yourself. If you discuss with others, give proper acknowledgments. Looking for the answers in the internet is discouraged. You should at least make a serious effort to solve a problem by yourself before starting to look online. If you do, however, give proper acknowledgments.

Remember that you can be at most 3 days late; for every late date you lose 10% of your grade.

Typed solutions are encouraged, especially if your hand writing is messy. However, there will be no extra marks for typed answers.

Partial credit will be given for partial solutions, but not for long off-topic discussion that leads nowhere. Overall, think before you write, and try to give concise and crisp answers. Also, remember the 20% rule that states that: If you do not know the answer to a question and you write that you do not know, then you get 20% of the grade you would have taken had you answered the question correctly. You get 0 for a wrong answer.

Exercise 1 (20 points): You are given a set V consisting of n integers. The task is to report all n products of the n distinct $(n-1)$-cardinality subsets of V. Your algorithm should run in linear time and it should not use division.

Exercise 2 (20 points): Consider a set of n points $X = x_1, \ldots, x_n$ in some d-dimensional space, and distance function $d(x_i, x_j) = L_2^2(x_i, x_j)$. Let $\bar{x}$ be the d-dimensional vector that is the*mean* of all the vectors in X. Prove that $\bar{x}$ minimizes $\sum_{x_i \in X} d(\bar{x}, x_i)$, i.e., that the mean is the *representative* for distance function $d()$.

Exercise 3 (20 points): The Jaccard similarity between two sets X and Y is defined as:

$$\text{JSim}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}.$$

The Jaccard distance between sets X and Y is defined as:

$$\text{JDist}(X, Y) = 1 - \text{JSim}(X, Y).$$

Prove or disprove that the JDist function is a metric.

Exercise 4 (40 points):

1. **[10 points]**

   Consider the nearest neighbor problem: We are given a set of objects X and a distance function d. At query time we are given an object q and the goal is to find a point $x^* \in X$ that such that

   $$d(q, x^*) \leq d(q, x), \text{ for all } x \in X.$$

   The *linear-scan* algorithm has complexity $\mathcal{O}(nD)$, where n is the number of objects in X and D is the time required for one distance computation.

   In certain cases computing the distance function d is an expensive operation. In those cases it is desirable to have a *lower bound* on the distance d. A lower bound is a function d_L with the property $d(x, y) \geq d_L(x, y)$ for all $x, y \in X$. A lower bound d_L is useful when it is much faster to compute than the function d.

   Explain how a lower bound distance function can be used to speed up the linear-scan algorithm.

2. **[20 points]**

   Consider the d-dimensional Euclidean space $\mathbb{R}^d$. Let us define a *point cloud* x as a set $\{(\vec{x}_1, p_1), \ldots, (\vec{x}_n, p_n)\}$, such that $\vec{x}_1, \ldots, \vec{x}_n$ are vectors in $\mathbb{R}^d$ and $p_1, \ldots, p_n$ are positive numbers with $\sum_i p_i = 1$. In other words, a point cloud x can be seen as a *probability distribution* over points in $\mathbb{R}^d$.

   Given two point clouds $x = \{(\vec{x}_1, p_1), \ldots, (\vec{x}_n, p_n)\}$ and $y = \{(\vec{y}_1, q_1), \ldots, (\vec{y}_m, q_m)\}$, of n and m points, respectively, we define the *point-cloud shift* distance to be the *minimum amount of energy* required to shift cloud x to cloud y. In other words, if f_{ij} specifies how much of the i-th point of cloud x should be shifted to the j-th point of cloud y, we want to minimize

   $$\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} L_2(\vec{x}_i, \vec{y}_j),$$

   subject to $\sum_{j=1}^{m} f_{ij} = p_i$ and $\sum_{i=1}^{n} f_{ij} = q_j$, for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

   Prove the point-cloud shift distance is a metric.

   How can you compute the point-cloud shift distance? What is the complexity of your algorithm?

3. **[10 points]**

   Provide a lower bound for the point-cloud shift distance.

   A good lower bound should be as tight as possible and as fast to compute as possible.

   What is the complexity of computing your lower bound?