Problem Set 1

September 13, 2013

Due date: Mon, Sept 30 2013 at 4pm; before class.

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): 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
$$L_2^2(x,y) = \sum_{i=1}^{d} (x_i - y_i)^2$$
is a metric. (10 points)

Exercise 3 (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 *centroid* for distance function $d()$.

Exercise 4 (20 points): The Jaccard similarity between two sets X and Y is defined as:
$$\mathrm{JSim}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}.$$
The Jaccard distance between sets X and Y is defined as:
$$\mathrm{JDist}(X,Y) = 1 - \mathrm{JSim}(X,Y).$$
Prove or disprove that the JDist function is a metric.

Exercise 5 (20 points): In class we have defined the Edit Distance between two strings x and y, of length n and m respectively to be the minimum (weighted) number of insertions, deletions and substitutions that transform string x to string y. We also demonstrated that assuming different `deletion`, `insertion` and `substitution` costs for every letter (or pairs of letters), the following dynamic-programming recursion computes the edit distance between x and y:

$$D(x(1\ldots i), y(1\ldots j)) = \min \begin{cases} D(x(1\ldots i-1), y(1\ldots j)) + \texttt{delete}(x[i]), \\ D(x(1\ldots i), y(1\ldots j-1)) + \texttt{insert}(y[j]), \\ D(x(1\ldots i-1), y(1\ldots j-1)) + \texttt{substitute}(x[i], y[j]). \end{cases}$$

In the above equation $x(1\ldots i)$ (resp. $y(1\ldots j)$) is the substring of x (resp. of y) that consists of the first i (resp. j) symbols appearing in x (resp. y). Also, for symbol a, $\texttt{delete}(a)$, $\texttt{insert}(a)$ correspond to the cost of deleting or inserting a respectively. Finally, for symbols a, b, $\texttt{substitute}(a, b)$ corresponds to the cost of substituting symbol a with symbol b.

1. (**10 points:**) Prove or disprove that the edit distance function as defined above is a metric.

2. (**10 points:**) Find two instantiations of the edit-distance function that are metrics. An instantiation of the edit distance function is defined by a specific way of allocating costs to operations such as deletions, insertions and substitutions.