**CSE 881: Data Mining (Fall 2016 Homework 3)**
Due date: Sept 27, 2016 (before midnight).

A soft copy of your homework must be submitted via handin. All submitted homework must be your own work.

1. Proximity Measure

   Let $\mathbf{x}$ and $\mathbf{y}$ be a pair of non-negative vectors, i.e., $\forall i : x_i \geq 0, y_i \geq 0$, where $x_i$ and $y_i$ are elements of the vectors. Consider the following distance measure:

   $$d(\mathbf{x}, \mathbf{y}) = 1 - c(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}, \tag{1}$$

   where $c(\mathbf{x}, \mathbf{y})$ is the cosine similarity between two non-negative vectors.

   (a) What are the maximum and minimum possible values for the distance measure?

   (b) Does the measure satisfy the positivity property?

   (c) Does the measure satisfy the symmetry property?

   (d) Does the measure satisfy the triangle inequality property?

2. Nearest-neighbor Search

   Suppose you are given a set of points $S$ in Euclidean space, as well as the distances between every point $\mathbf{x} \in S$ to a reference point $\mathbf{y}$. Assume the distance measure you use is a metric. Note: It does not matter if the reference point $\mathbf{y} \in S$.

   (a) If the goal is to find all the points in $S$ within a specified distance $\epsilon$ of a query point $\mathbf{q} \neq \mathbf{y}$, explain how you can use the triangle inequality and the already calculated distances to $\mathbf{y}$ to potentially reduce the number of distance calculations necessary. Hint: calculate the upper and lower bounds of the distance between the query point $\mathbf{q}$ and a data point $\mathbf{x} \in S$ using triangle inequality.

   (b) Suppose that you can find a small subset of points, $\Psi$, from the original data set $S$, such that every point $\mathbf{x} \in S$ is within a specified distance $\epsilon$ of at least one of the points in $\Psi$, and that you also have the pairwise distance matrix between all pairs of points in $\Psi$. Assume you can easily identify the corresponding point $\mathbf{x}^* \in \Psi$ for each point $\mathbf{x} \in S$, where $d(\mathbf{x}, \mathbf{x}^*) \leq \epsilon$. Describe a technique that uses this information to compute, with a minimum number of distance calculations, all pairs of points in $S$ that are within a distance of $\beta$ of each other.

3. KD-tree

   Consider the 2-dimensional data set shown in the table below:

| Data point | $x$ | $y$ |
|---|---|---|
| 1 | 0.1190 | 0.6991 |
| 2 | 0.2238 | 0.1493 |
| 3 | 0.2551 | 0.8407 |
| 4 | 0.3404 | 0.5472 |
| 5 | 0.4984 | 0.8909 |
| 6 | 0.5060 | 0.2543 |
| 7 | 0.5853 | 0.1386 |
| 8 | 0.7513 | 0.2575 |
| 9 | 0.9597 | 0.9593 |

   (a) Draw a KD-tree for the data set. If there are even number of points, e.g., 2 points, define median as the smaller of the two midpoints after sorting the data on the given dimension.

   (b) Draw a 2-dimensional plot of the data. Partition the space into rectangular regions based on the KD-tree.

   (c) Suppose you need to find the nearest-neighbors of a query point $[0.6, 0.5]$ within the following bounding box: $0.5 \le x \le 0.7$ and $0.4 \le y \le 0.6$. List all the nodes (starting from the root) that are visited to search for the nearest neighbors.

4. Hashing

   Consider the binary data shown below.

| Document | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ | $w_8$ | $w_9$ | $w_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $d_1$ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| $d_2$ | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| $d_3$ | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| $d_4$ | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |

   (a) Compute the pair-wise Jaccard similarity between the documents by filling out the following similarity matrix:

$$
\begin{array}{c c}
 & \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \end{array} \\
\begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \end{array} & \left( \begin{array}{cccc} \phantom{d_1} & \phantom{d_2} & \phantom{d_3} & \phantom{d_4} \\ \\ \\ \\ \end{array} \right)
\end{array}
$$

(b) Construct a MinHash signature of length 6 for each document by applying the following permutation matrices $\pi$ with column ordering shown below:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\pi_1$ | 10 | 3 | 8 | 9 | 6 | 5 | 4 | 2 | 1 | 7 |
| $\pi_2$ | 3 | 9 | 7 | 5 | 4 | 2 | 6 | 8 | 1 | 10 |
| $\pi_3$ | 4 | 2 | 3 | 1 | 7 | 5 | 6 | 9 | 8 | 10 |
| $\pi_4$ | 6 | 4 | 3 | 10 | 7 | 8 | 5 | 9 | 1 | 2 |
| $\pi_5$ | 9 | 6 | 8 | 5 | 1 | 4 | 7 | 2 | 3 | 10 |
| $\pi_6$ | 10 | 1 | 4 | 6 | 9 | 2 | 7 | 3 | 8 | 5 |

(c) Based on the matrix you found in part (b), compute the probabilities $p[h(d_i) = h(d_j)]$ between every pair of documents (where $h$ is the minHash function) and enter them into the following $4 \times 4$ matrix:

$$\begin{array}{c c} & \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \end{array} \\ \begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \end{array} & \left( \begin{array}{cccc} \quad & \quad & \quad & \quad \\ & & & \\ & & & \\ & & & \end{array} \right) \end{array}$$

(d) Compare the pair of documents with highest Jaccard similarity (in part (a)) against the pair of documents with highest probability (in part (c)). Are the results consistent with each other?

(e) Compare the pair of documents with lowest Jaccard similarity (in part (a)) against the pair of documents with lowest probability (in part (c)). Are the results consistent with each other?