University of Washington
Department of Computer Science and Engineering
CSEP 521, Winter 2021

February 23, 2021

Homework 8, Due Thursday, March 4, 2021

Problem 1 (10 points):
The $L^{1/2}$ norm can be defined as

$$\|(x,y)\|_{1/2} = \left(\sqrt{|x|} + \sqrt{|y|}\right)^2 .$$

Is the $L^{1/2}$ norm a proper distance function (a metric)? Prove or disprove.

Problem 2 (10 points):
Suppose that $|U| = n$ and you select random subsets, $A \subseteq U$ and $B \subseteq U$ with $|A| = m$ and $|B| = m$. What is the expected size of $A \cap B$?

Using this estimate of the size of the intersection, what is the value of the Jaccard similarity of A and B?

Give an expression for the value of the Jaccard similarity of A and B if $m = \frac{n}{k}$. (The purpose of this exercise is to get an idea of what the Jaccard similarity should be on uncorrelated data.)

Problem 3 (10 points):
This problem is to work out the details of a bucketing approach to the nearest neighbors problem in 1-D. You *do not* need to implement your algorithm.

Let S be a set of n points from $[0, 1)$ with minimum separation $\delta \geq 2^{-k}$. Think of the line segment as being divided into overlapping buckets $B = \{B_j^i \mid 0 \leq j \leq k \text{ and } 0 \leq i < 2^j\}$ where B_j^i corresponds to the interval $[i2^{-j}, (i+1)2^{-j})$. Describe a nearest neighbors algorithm that relies on looking for the query point in appropriate buckets, and uses hashing to avoid storing unnecessary buckets. You should describe the run time of your algorithm in terms of the expected number of buckets accesses per query point. (You can assume that hashing is an $O(1)$ time operation.)

Programming Problem 4 (20 points):
You will use a well-known data set of articles from various newsgroups. Each article is represented by a "bag of words," which is a vector indexed by words with each component indicating the number of times the word occurs in a given article.

The data is stored in three files: data50.csv, label.csv, and groups.csv.

The data50.csv file is a sparse representation of each "bag of words." Every row contains three fields: articleId, wordId, and count. To find out which group an article belongs to, use the file label.csv. If articleId is n, then line n of label.csv contains the corresponding groupId. Line m of groups.csv contains the name of the m-th group.

You will examine three measures of similarity for bag-of-words vectors X and Y; higher numbers mean the vectors are more similar:

Jaccard similarity:

$$\text{Jaccard}(X, Y) = \frac{\sum_j \min(x_j, y_j)}{\sum_j \max(x_j, y_j)}$$

Cosine similarity:

$$\text{CS}(X, Y) = \frac{\sum_j x_j y_j}{\|X\|_2 \|Y\|_2}$$

L^2 **similarity**:

$$\text{H}(X, Y) = -\|X - Y\|_2 = -\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

a) Import the data sets. Remember the total number of words is large, so you probably want to use a sparse representation.

b) Implement the three similarity measures. For each metric, prepare the following plot: Rows and columns are indexed by newsgroups (in the same order). For each entry (A, B) of this 20×20 matrix, compute the average similarity over all ways of pairing one article from A with one article from B. After computing these 400 numbers (for each of the three similarity measures), plot your results using a heatmap. Label your axes with the group names. Starter code for heat maps in python is linked from the homework page. (Don't spend too much time on the visualization if your environment doesn't provide convenient tools.)

c) Based on the heatmaps you generated, which of the measures seems the most reasonable? Are there any pairs of newsgroups that are very similar? Would have you expected this? Explain.

d) Now compute another set of 20×20 matrices as follows. For each article a_1, find the article a_2 from a different newsgroup that has the largest Jaccard similarity with a_1. (You can do this with brute-force search.) Now for each pair of groups A, B, count how many articles in group A have their nearest-neighbor in group B. Plot these results in a heatmap.

e) Your plot for part (b) was symmetric, but for part (d) was asymmetric. Explain.

f) Which groups seem similar? Compare the plots from parts (b) and (d). Which method seems more suited to comparing newsgroups?