Data Mining Assignment 1

Vivek Vaidyanathan - vaidy083@umn.edu - 5416749

October 2017

1 Question 1

Convert the similarity values as "distances" between heterogeneous objects, by taking the inverse of similarity. Let δ_{ij} denote the distance between each pair of heterogeneous objects i and j. Multidimensional Scaling is an ideal technique for this problem. We convert the given objects into a distance graph with $\binom{n}{2}$ edges, with the weight of each edge representing the distance between the corresponding nodes on the edge.
Once we have a graph, we have to solve the equations to get the vectors in the k dimensional space.

2 Question 2

Since we already have the data, all we need to is label the data by region. We also might have to time stamp each of the values. We might also have to use averaging techniques to take the average of all 10*10 squares for taking the data of a particular region as a whole. Time, coordinates, and other weather realated details for each region should be their own separate dimensions.

3 Question 3

The annotated text will describe the data point in question. In this case, it is the protein sequence. Apart from the inherent dimensions like amino acids or their order that determines the property of proteins, we should also consider the unique words in their description as separate dimensions. (Each different word and order of amino acids are different coordinates)

4 Question 4

We are reducing the ppi on the width of the image by doing a dimensional reduction directly on the image. We will be operating on greyscale values, which would be represented as a strength of white pixel (0,255). SVD cases the m*n image data to be stored as $m*k+k$ (U and Σ) matrices. This would case the compression of image, and the amount of detail retained in the image is determined by how big the value of k is.

5 Question 5

5.1 a

Raw data: J:9601 C:9614 E:9558

5.2 b

SVD U - 10,000 * 5 J:6813 C:6666 E:6799
SVD U - 10,000 * 10 J:8939 C:8901 E:8853
SVD U - 10,000 * 20 J:9536 C:9531 E:9501
SVD U - 10,000 * 40 J:9562 C:9559 E:9507
SVD $U\Sigma$ - 10,000 * 5 J:6846 C:6687 E:6848
SVD $U\Sigma$ - 10,000 * 10 J:8899 C:9022 E:8906

SVD $U\Sigma$ - 10,000 * 20 J:9556 C:9596 E:9562
SVD $U\Sigma$ - 10,000 * 20 J:9647 C:9690 E:9629

5.3 c

Average of 4x4 patches produced a 10,000*49 matrix J:9401 C:9439 E:9378