CSci 5523 – Assignment #1

1. The following questions from chapter 2 of the book: 3, 4, 7, 9, 15, 16, 18, 27.

2. Suppose that you had a set of arbitrary objects of different types representing different characteristics of widgets. A domain expert gave you the similarity value between every pair of objects. How would you convert these objects into a multidimensional data set for clustering?

3. Suppose that you had a set of discrete biological protein sequences that are annotated with text describing the properties of the protein. How would you create a multi-dimensional representation from this heterogeneous data set?

4. Consider the problem of dimensionality reduction in the context of an $n \times m$ image with grayscale values. The image can be represented as a matrix A that has n rows and m columns. If you compute a truncated SVD decomposition of this matrix in order to do dimensionality reduction, which are the objects whose dimensionality you are reducing and which ones are the original dimensions?

5. The following questions involve the "test set" described in https://pjreddie.com/projects/mnist-in-csv/. Please download that dataset and answer the following questions:

    a. For each image in the test set (*query* image) compute the closest other test image using the following approaches: Euclidean distance, cosine similarity, extended Jaccard similarity. Report the number of times the closest image is the same digit as that of the query image.

    b. Used truncated SVD to perform a dimensionality reduction using 5, 10, 20, and 40 dimensions. Represent the records using both the U and the $U\Sigma$ matrices. For each of the above dimensions and low-dimensional representations, perform the study that you did in part (a). You can use Matlab to compute the truncated SVD.

    c. Each image in the above data set is 28×28. Create a 7×7 image by averaging the values of each 4×4 patch of the image. Using this 49-dimensional representation perform the study that you did in part (a).