Assignment 1
Cluster Analysis and Dimension Reduction

Wenke Liu
wenke.liu@nyumc.org

Due on Oct 1st, 2018

1 Introduction

The MNIST (Mixed National Institute of Standards and Technology) database contains 60000 training examples and 10000 testing examples of handwritten digit images. Due to its popularity in performance testing, it has been recognized as "the Drosophila of machine learning". In this assignment, we ask you to apply some clustering and dimension reduction methods to a subset (n=1000) of the MNIST data.

The raw data for this assignment are stored in a tab-delimited plain text file, `dat.txt`. The file contains 1000 rows and 784 columns of integers ranging from 0 to 255. Each row represents intensity value of a 28×28 image, one number for one pixel. The succesive rows of the square pixel matrix is flattened into the row vector, from left to right, top to bottom, with the first element of the vector corresponds to the upper left pixel. You may have to rotate the matrix to visualize the digit correctly in your program. The `lab.txt` file contains a 1000-element vector, each element is the label (true class) of the corresponding row in `dat.txt`.

Please feel free to use your favorite programming language. Most (if not all) algorithms and procedures mentioned here have existing implementations in Python, R and Matlab.

2 Clustering

1. Subset the data and only keep examples whose true labels are 0 or 1 (there should be 211 of them in total). Run hierarchical clustering on this subset with Euclidean distance as dissimilarity measure. Plot the dendrogram for single, complete and average linkages.

2. For the three dengrograms obtained in question 1, cut each tree and obtain two-group assignments. Produce contingency tables for each assignment compared to the true labels. Which linkage performs the best?

3. **With the subset of 0s and 1s**, calculate the Gap statistic for K-means clustering with k ranging from 1 to 10. You can choose your own Monte Carlo sample numbers if you want to (in the R and Matlab functions this is set to 100 by default), and use uniform over principle conponents as the reference distribution. Which k value has the largest Gap statistic? Does this make sense to you? If not, could you make some speculations about why it is the case?

4. Run the K-means algorithm with $k = 10$ to cluster **all of the 1000 images** into 10 groups. You may want to run mutiple times with random initiation and pick a solution with the smallest within-cluster sum of squares. Visualize the cluster centers (centroids) as ten 28×28 images. Does each of the centers represent the digit 0-9?

5. Run K-medoids algorithm with $k = 10$ and Pearson correlation distance as the dissimilarity measure. Visualize the cluster centers (medoids) as ten 28×28 images.

3 dimension reduction

1. Run PCA on the **whole dataset** (n=1000) and generate the scree plot. Do you choose to standardize the columns? Why?

2. Represent all data points on a two-dimensional scatter plot using their first two principal component scores. Distinguish the points by color or symbol style according to their true label. Do examples of different digits appear to be well seperated?

3. Run t-SNE on the dataset to generate a two-dimensional representation. It may take a while for your computer to generate the results. Visualize all data points as in the previous question.

4. Subset the data and only keep examples whose true labels are 3 (there should be 107 of them). Run PCA on this 3-only dataset and only keep the first two principal components. Visualize the center of the data in the original feature space (i.e., construct an image of an 'average' 3). Loadings of the two principal component specify **directions** or **features** that the new representations are based on. Could you visualize them? How do you interpret them? (Hint: *The Elements of Statistical Learning* 14.5.1)

5. Visualize the original images (data points) with the maximum and minimum value of the two principal component **scores**.