Assignment I

1. Image Data Analysis

Data: image_analysis.mat

Description: 64x64 facial images of 40 subjects, each subject has 10 images. The left figure is an example image. X is a cell array with each cell an 64x64 image matrix. label is a column vector with each dimension indicating the identity of the corresponding image. The values of the image matrices will be integers in [0,255], 0 is black, 255 is white.

Task: Stretching all images into column vectors and form the data matrix by concatenating all those image vectors, then do the following tasks under three conditions 1) raw data; 2) normalizing every column to unit norm; 3) standardize every column to zero mean and unit variance

• A) Perform PCA on the data matrix and plot the first 10 principle components as images. Those images are usually called “eigenfaces”
• B) Perform LDA on the data matrix and plot the first 10 discriminative components as images. Those images are usually called “Fisherfaces”
• C) Compare the eigenfaces and Fisherfaces and explain what you observe
• D) Do K-means clustering on the data matrix with number of clusters set to 5,10,20,30,40,50. Compare the clusters you get and the ground truth image labels (subject identity). Compute cluster purity and rand index.
• E) Repeat the process of D)100 times independently with different initializations, check the consistency of the results on different runs
• F) Repeat the procedure of D) on data matrix after PCA projection, set the number of principle components to 10%, 30%, 50%, 70%, 90% of the total number of principle components.

• G) Run hierarchical clustering on the data with and without PCA. Use standard Euclidean distance as the dissimilarity measure. Try max, min and average distance when merging clusters and compare the difference. Draw the final Dendrogram.

• H) After procedure G), set the number of clusters to 5, 10, 20, 30, 40, 50. Compute cluster purity and rand index.

• I) Run k-Nearest Neighbor classification on the data set with k=1,3,5. Use standard Euclidean distance before and after PCA. Report the average and standard deviation of classification accuracy over 10 and 5 fold cross validation.

• J) Construct image similarity network, where the nodes in the network are images, and the edges linking pairwise images reflecting their similarities. Compute the similarity with the following Gaussian function with standard Euclidean distance. Tune the width so that you can achieve the best 1,3,5-Nearest Neighbor classification accuracy averaged over 10-fold cross validation (you can also use the local scaling scheme as I introduced in the spectral clustering class). Plot the heat map of the resultant similarity matrix.

\[ s(x_i, x_j) = \exp \left( -\frac{d(x_i, x_j)^2}{2\sigma^2} \right) \]

2. Document Analysis

Data: document_analysis.mat
Term Dictionary: terms.txt

Description: 4852 documents from 5 classes. Bag-of-Words representation. 61188 distinct words in total. X is a sparse document
matrix, each column is a document, labels is the ground truth label vector for those documents.

Tasks: do the following tasks in three scenarios 1) raw data; 2) data after column normalization such that every column has unit norm; 3) data after TF-IDF normalization
• A) Perform PCA, plot the eigenvalues as a curve, choose the number of principle components you want to keep and explain why
• B) Perform LDA, check the generalized eigenvalues and eigenvectors, choose the number of discriminative components and explain why
• C) Do K-means clustering with number of clusters set to 5, compute clustering purity and rand index
• D) Perform A) 100 times independently with different initializations, check the consistency of the results
• E) Run hierarchical clustering on the data with and without PCA. Use standard Euclidean distance as the dissimilarity measure. Try max, min and average distance when merging clusters and compare the difference. Draw the final Dendrogram. Set the number of clusters to 5 and compute cluster purity and rand index
• F) Repeat E) with cosine similarity
• G) Construct document similarity network, where the nodes in the network are documents, and the edges linking pairwise documents reflecting their similarities. Compute the similarity with the Gaussian function with standard Euclidean distance and cosine distance. Tune the width so that you can achieve the best 1,3,5-Nearest Neighbor classification accuracy averaged over 10-fold cross validation (you can also use the local scaling scheme as I introduced in the spectral clustering class). Plot the heat map of the resultant similarity matrix.

Out-of-Memory: If you encounter out-of-memory error in any of these tasks, you are allowed choose k>=2 classes to perform the task, but you need to explain where this out-of-memory happens and which step in your code cost most memory. Also please report the configuration of your computer and the memory capacity you can use to perform the task.