CS 562: Problem Set 3 (Part 2)

March 16, 2014

Instructions

These problems are to be started and worked on in class in groups of two (or individually). Please make sure that you have ready access to a computer - either the clamshell server or your own laptop running Linux-based sofware with all the required Python software.

Problems

1. PCA for Reduced Dimensionality in Clustering **[Dataset: segmentation_data.zip]**

   For this problem you will use an image segmentation data set for clustering. You will experiment with using PCA as an approach to reduce dimensionality and noise in the data. You will compare the results of clustering the data with and without PCA using the provided image class assignments as the ground truth. The data set is divided into three files. The file "segmentation_data.txt" contains data about images with each line corresponding to one image. Each image is represented by 19 features (these are the columns in the data and correspond to the feature names in the file "segmentation_names.txt". The file "segmentation_classes.txt" contains the class labels (the type of image) and a numeric class label for each of the corresponding images in the data file. After clustering the image data, you will use the class labels to measure completeness and homogeneity of the generated clusters. The data set used in this problem is based on the Image Segmentation data set at the UCI Machine Learning Repository. Your tasks in this problem are the following:

   (a) Load in the image data matrix (with rows as images and columns as features). Also load in the numeric class labels from the segmentation class file. Using your favorite method (e.g., sklearn's min-max scaler), perform min-max normalization on the data matrix so that each feature is scaled to [0,1] range.

   (b) Next, Perform K-means clustering on the image data (since there are a total 7 pre-assigned image classes, you should use K = 7 in your clustering). Use Euclidean dis-

tance as your distance measure for the clustering. Print the cluster centroids (use some formatting so that they are visually understandable). Compare your 7 clusters to the 7 pre-assigned classes by computing the Completeness and Homogeneity values of the generated clusters.

(c) Perform PCA on the normalized image data matrix. You may use the linear algebra package in Numpy or the Decomposition module in scikit-learn (the latter is much more efficient). Analyze the principal components to determine the number, **r**, of PCs needed to capture at least 95% of variance in the data. Then use these **r** components as features to transform the data into a reduced dimension space.

(d) Perform K-means again, but this time on the lower dimensional transformed data. Then, compute the completeness and homogeneity values of the new clusters. Discuss your observations based on the comparison of the two clustering results.

2. Exercise 11.3.1 from the MMDS textbook. Consider the matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix}$$

(a) Compute the matrices $M^T M$ and $M M^T$

(b) Find the eigenvalues for your matrices of part (a).

(c) Find the eigenvectors for the matrices of part (a).

(d) Find the SVD for the original matrix M from parts (b) and (c). Note that there are only two nonzero eigenvalues, so your matrix Σ should have only two singular values, while U and V have only two columns.

(e) Set your smaller singular value to 0 and compute the one-dimensional approximation to the matrix M .

(f) How much of the energy of the original singular values is retained by the one-dimensional approximation?