Lab Assignment 4 – CSE 564, Spring 2011

Due: Friday, April 15, 11:59pm

This lab will introduce you to two basic data analyses tasks: clustering and dimension reduction. You can use Matlab (the statistics toolbox) for analysis and visualize the result with the visual tools you developed in lab3. It is sufficient to just use one of the lab 3 datasets (say, the housing dataset) for all your experiments. You can read much information about the statistics toolbox on the Matlab website: http://www.mathworks.com/help/toolbox/stats/

Upon completion of the assignment please submit the following via blackboard:

- the matlab code
- a comprehensive report that illustrates with screenshots, narrative text and code snippets of all aspects of your work
- a link to your webpage so the work can be tested for grading (the data should be there as well)

1. Clustering with k-means

Matlab has a function KMEANS that takes as input the cluster number K and the data matrix. Use the default Euclidian distance. Try different K and for each color the data points according to their cluster ID and display them in the scatterplots, scatterplot matrix, and the parallel coordinate visualization interfaces. Finally, determine the optimal number of K using the elbow finding technique mentioned in the lecture notes and visualize these cluster as well.

2. Clustering with expectation maximization

Matlab has a function GMDISTRIBUTION.FIT that uses Gaussian Mixture models with EM-optimization to determine the clusters. Compare the results with those obtained with k-means, both for different K and the optimal number of K.

3. Dimension reduction with Principal Component Analysis (PCA)

Compute the covariance matrix of your data and then run Matlab's PRINCOMP function. If you have D dimensions, COEFF will contain the D Eigenvectors and 'latent' will contain the D Eigenvalues. You could visualize these with the parallel coordinate interface (with D+1) axes.

Sort the Eigenvectors by their corresponding Eigenvalues. Pick the set of vectors that explain, say 90% of the data variance (when (sum of Eigenvalues) / (total sum of Eigenvalues) >0.9). Project the data into them and visualize with scatterplot matrix and parallel coordinates (using the Eigenvectors as axes).

4. Dimension ordering for parallel coordinates

Compute Pearson's correlation coefficient for each pair of dimensions. The objective is to order the dimensions in such a way that the sum of correlation coefficients is maximized (you could just use the absolute values). Then we have an ordering in which each axes pair shows good correlation patterns. To obtain such an optimal ordering you need to solve the traveling salesman problem (TSP). Matlab Central http://www.mathworks.com/matlabcentral has several (approximate) TSP solutions. Show the ordering you find. You could also try to only use positive correlations for the TSP.