Math 444 – Data Mining Spring 2017

Exercise 1

An essential part of the homework in this course consist of presentation of results in a graphical form. Therefore, it is important that your homework is of good quality, typeset by Word, Latex or alike, Latex being the preferred choice. Save your homework as *one single pdf file*, and turn it in using Blackboard. Submit your Matlab-code, too, using Blackboard. The grading is based on your pdf report, so separate Matlab-plots or pieces of m-files will not be accepted as part of your report. To help you producing good quality report, a template will be posted on Blackboard under Assignments. We strongly encourage you to use the template. The quality of the reports will be taken into account in the grading.

1. Download from Blackboard the file `ModelReductionData.mat`. The file contains a data matrix $\mathsf{X} \in \mathbb{R}^{6\times 4000}$. Each column of X represents a data vector in the six dimensional space.

   (a) Visualize the raw data using *scatter plots*, that is, select two components of the data at a time and plot them one against the other. You have $\binom{6}{2}$ different two-dimensional projections. Here is a suggestion of how to do that in Matlab. The snippet below plots the first component against the second, so modify it as you think is necessary:

   ```
   load ModelReductionData
   figure(1)
   plot(X(1,:),X(2,:),'k.','MarkerSize',7)
   axis('equal')
   set(gca,'FontSize',20)
   ```

   (b) Center the data and compute the SVD. Plot the singular values. What can you say about the dimensionality of the data? Show the scatter plots of the first few principal components. Do the plots suggest a presence of clusters in the data?

2. Download from Blackboard the data file `HandwrittedDigits.mat`, containing the data matrix X of size $256 \times 1\,707$ containing pixel images of the handwritten digits, and the label vector I of length $1\,707$ containing numbers from 0 to 9, indicating the digits that the corresponding images represent.

   Extract from X the images that correspond to numbers 0, 1, 3 and 7. From each subgroup, select 5 samples, and approximate these samples by the linear combination of the first k feature vectors, $k = 5, 10, 15, 20, 25$, that is,

   $$x^{(j)} \approx \mathsf{P}_k x^{(j)} = \sum_{\ell=1}^{k} z_\ell^{(j)} u^{(\ell)},$$

   where $z_\ell^{(j)}$ are the principal components of $x^{(j)}$. Plot the approximation as images, as well as the residual, $x^{(j)} - \mathsf{P}_k x^{(j)}$. Plot the norms of the errors as a function of k.

3. Download from Blackboard the data file `IrisData.mat`. The matrix X consists of 150 vectors, each one having four components. The data correspond to measurements of certain dimensions in three species of flowers, *Iris setosa*, *Iris versicolor*, and *Iris virginica*, and the components of the data vectors have the following attributes:

   $$x = \begin{bmatrix} \text{sepal length in cm} \\ \text{sepal width in cm} \\ \text{petal length in cm} \\ \text{petal width in cm} \end{bmatrix}$$

By using the PCA, investigate if the data set suggests the presence of clusters that would make it possible to separate the three species from each other. (Later on, the flower species corresponding to each data point will be made available.)