Math 444 – Data Mining Spring 2020

Home Assignment 1

An essential part of the homework in this course consists 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 Canvas. Include snippets of your code in your report to make the point of what you are doing, and a complete list of your code in the appendix. I reserve the right to ask you to turn in your code, or/and to come by my office and go over your code with me. You are not supposed to use any of the built-in routines for the algorithms that can be found in Matlab, Phyton or R, since being in full control of the codes is part of the goals of the class. A code without a proper report is not acceptable, even if your code is correct and commented. A code that has been copied from others, or dowloaded from the internet will be treated as an honor code violation.

The grade for the assignment is based on the correctness and tquality your pdf report. To help you producing good quality report, a LaTeX template will be posted on Canvas under Announcements. You are strongly encouraged to use the template. The quality of the reports will be taken into account in the grading.

---

1. This is a warm-up exercise where you can visualize the data. Download from Canvas the file `DataAssignment1S2020.mat`. The file contains a data matrix $\mathsf{X} \in \mathbb{R}^{3\times 1500}$. Each column of X represents a data vector in three dimensional space.

   (a) Visualize the raw data in three dimensions. What do you see? Also, in preparation for when the dimensionality of the data is higher, using *scatter plots*, that is, select two components of the data at a time and plot them one against the other. This will generate three plots. Include them in your report and comment on them. Here is a suggestion of how to do that in Matlab.

   ```
   load DataAssignment1S2020
   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 do you conclude about the dimensionality of the data? Show the three scatter plots of the principal components. Make sure you use the same axes. Compare these three plots to the three scatter plots of the data and comment on what the principal components added to your understanding of the data. Do the plots suggest a presence of clusters in the data?

2. Now let's do the same for a data set that cannot be easily visualized. 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?

3. Download from Canvas 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 2, 4, 5 and 8. From each subgroup, select one representative, and approximate these sample specimen 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. Also, plot the residual $x^{(j)} - \mathsf{P}_k x^{(j)}$ with $k = 25$.

   Plot the norms of the errors as a function of k.

4. Download from Canvas the data file `IrisDataAnnotated.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, (1) *Iris setosa*, (2) *Iris versicolor*, and (3) *Iris virginica*, and the components of the data vectors have the following attributes:

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

   The vector I contains the annotation of each data vector, that is, $I(j) \in \{1, 2, 3\}$ tells which one of the three species $x^{(j)}$ represents.

   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. You can use different marker for each data point according to which species the data represents.