Handed out: April 6 (Tue)

Hints for solution: Exercise class on April 8 (Thu)

Hand in: April 14 (Wed), by email to doris.entner@helsinki.fi. Please submit your report as a single PDF containing figures, relevant program output and **discussion**. *Submit the source code as separate files.*

Solving the exercises below gives you points, which will, at the end, determine your grade for the computer project. Each exercise in all the assignments gives you an equal amount of points. In total, there will be approximately 15-20 exercises.

Ex. 1 — PCA and linear regression

This exercise illustrates the use of PCA in linear regression, i.e. the use of unsupervised machine learning techniques in supervised machine learning tasks.

In linear regression, it is assumed that $y_k = \mathbf{x}_k^T \beta + \epsilon_k$, for $k = 1, \ldots, n$ where the ϵ_k are iid Gaussian with mean zero and variance σ^2, and $\mathbf{x}_k \in \mathbb{R}^p$ (cf math ex 4).

The files `data1_train.txt` and `data1_test.txt` contain training set and test set, respectively. The first $p = 6$ lines contain 100 observations of the random vector $\mathbf{x} \in \mathbb{R}^p$. The seventh line contains the observations of the random vector $y \in \mathbb{R}$.

1. Load the training set and make a plot which shows the proportion of variance explained by the first m, $m = 1, \ldots p$ principal components (cf. p26 in the lecture hand-out). How much variance in the data is explained by the first two principal components? (10 %)
2. Project each random variable in the random vector $\mathbf{x}$ on the first two principal components. That is, create a figure similar to the one on p27 in the lecture hand-out (but don't show the projections of the observations, there are too many of them!) (20 %)
3. Calculate the correlation matrix of $\mathbf{x}$, and explain the figure you just made in terms of the correlation structure of $\mathbf{x}$. (20 %)
4. Estimate β with standard linear regression and PCA-regression. For standard linear regression the estimate is

$$\hat{\beta}_{lr} = (XX^T)^{-1} X\mathbf{y}, \tag{1}$$

while for PCA-regression, it is

$$\hat{\beta}_{pc}(m) = U_m D_m^{-1} U_m^T \frac{1}{n} X\mathbf{y}. \tag{2}$$

(see equation (4) and equation (9) of math exercise 4.) For PCA-regression, calculate $\hat{\beta}_{pc}(m)$ for $m = 1, \ldots p$. Compare the obtained estimates, and explain the differences between them. (20 %)

5. Load the test set, and calculate from $\mathbf{x}$ a prediction $\hat{y}$ of y,

$$\hat{y} = \mathbf{x}^T \hat{\beta}. \tag{3}$$

   Make a plot which shows the average prediction error in function of m, the number of principal components used in the regression. Show in the same plot also the prediction error for the training set. (25 %)
6. How many principal components should you best use in the regression? How much of the variance in the data is thus "relevant" for the prediction? Why is the performance on the test set so different from the performance on the training set? What is the connection between the difference of the performance and the covariance structure of the data? (5 %)

Ex. 2 — Image compression

This exercise illustrates how PCA can be used to compress an image. The basic idea is to represent an image not in the normal pixel basis, but in the basis given by the first couple of principle component directions.

1. Load the file `image.txt` and visualize the image in gray scale. (5 %)
2. Cut the image into non-overlapping blocks of size $10px \times 10px$. Each block will be called an image patch. (15 %)
3. Consider an image patch as a 100 dimensional random vector $\mathbf{x}$ (parse each image patch column-, or row-wise). Every block is then an observation of $\mathbf{x}$. Perform PCA on that image data. Make a plot which shows the first 49 principal component directions $\mathbf{u}_i$. How do they look like? (Note: every $\mathbf{u}_i$ is a 100 dimensional vector and corresponds thus to a $10px \times 10px$ image. Show us these $10px \times 10px$ images.) (30 %)
4. How much of the variance of the image data is retained by the first $m = 50$, $m = 25$, and $m = 3$ principal components ? (5 %)
5. Represent each image patch in the basis given by the first m principal component directions. Use $m = 50$, $m = 25$, and $m = 3$ as before. Note that these are approximations to the true image patches. (20 %)
   "Glue" the approximations together to obtain an image: invert the cutting process in question 2. Show us these images. Comment on their quality, and relate the quality to the appearance of the principal component directions $\mathbf{u}_i$ of question 3. (15 %)
6. Assume that the receiver knows the principal component directions $\mathbf{u}_i$. By how many percentages can you compress your image when

you transmit in the basis given by $\mathbf{u}_i$ as compared to transmission in the original pixel basis? Give your answer for $m = 50$, $m = 25$, and $m = 3$. (10 %)

Ex. 3 — Statistical structure of numbers

Sometimes it is hard to distinguish handwritten digits: The numbers 8 and 9, or 1 and 7 may look rather the same, depending on the writing. Here, we use PCA to characterize numbers.

The file `digits.txt` contains 1000 digits of size $28px \times 28px$ in vectorized form. The first 100 columns contain the number 0, the next 100 columns the number 1, and so on.

1. Load the file and, as preprocessing step, remove from each digit the average value, and normalize each digit to norm 1. Show an example of each number (10 %)
2. Calculate the "mean digit". Show how the mean digit looks like. To which number is it most similar to? (5 %)
3. Do PCA on the digits. Show the first 9 principal component directions. Comment on their structure. (20 %)
4. Calculate the first three principal components for all digits. (15 %)
5. Make a plot which shows where the numbers 0,8,9,1 are projected to when we project them onto the first two principal component directions (cf. the figure on p27 in the lecture hand-out). Comment on the pattern of the projections. (20 %)
6. Do the same but now, project the numbers onto the second and third principal component direction. How does the pattern of the projections change? (10 %)
7. Add now the projection of the number 4 to the previous plots. To which number is it most similar according to the plots? (10 %)
8. Same as previous question but for a number of your choice. (10 %)