Assignment 3: CS 215

Due: 28th October before 11:55 pm

All members of the group should work on all parts of the assignment. Copying across groups or from other sources is not allowed. We will adopt a zero-tolerance policy against any violation.

Submission instructions:

1. You should type out a report containing all the answers to the written problems in Word (with the equation editor) or using Latex, or write it neatly on paper and scan it. In either case, prepare a single pdf file.
2. The report should contain names and roll numbers of all group members on the first page as a header.
3. Put the pdf file and the code for the programming parts all in one zip file. The pdf should contain the names and ID numbers of all students in the group within the header. The pdf file should also contain instructions for running your code. Name the zip file as follows: A2-IdNumberOfFirstStudent-IdNumberOfSecondStudent-IdNumberofThirdStudent.zip. (If you are doing the assignment alone, the name of the zip file is A2-IdNumber.zip, if there are two students it should be A2-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip).
4. Upload the file on moodle BEFORE 11:55 pm on the due date. We will nevertheless allow and not penalize any submission until 10:00 am on the following day (i.e. 29th October). No assignments will be accepted thereafter.
5. Note that only one student per group should upload their work on moodle, though all group members will receive grades.
6. Please preserve a copy of all your work until the end of the semester.

Questions:

1. If d-dimensional random variable $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{C}$ is a $d \times d$ covariance matrix, then derive the PDF of the random variable $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\nu}$ where $\boldsymbol{B}$ is a $d \times d$ matrix and $\boldsymbol{\nu} \in \mathbb{R}^d$. For simplicity, assume that both $\boldsymbol{B}$ and $\boldsymbol{C}$ are invertible. [15 points]
2. In some applications in machine learning, the dimensionality d of the samples is very high, so much so that we have $d > n$ where n is the number of samples. In all such applications, it is very hard to even store the $d \times d$ covariance matrix in memory, and computing its eigenvectors is even more challenging. We have seen that the sample covariance matrix can be computed as $\hat{\boldsymbol{C}} := \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ where $\bar{\boldsymbol{x}}$ is the mean of all the vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ with each $\boldsymbol{x}_i \in \mathbb{R}^d$. Show that $\hat{\boldsymbol{C}}$ can be expressed in the form $\boldsymbol{Z}\boldsymbol{Z}^\top$ where $\boldsymbol{Z} \in \mathbb{R}^{d \times n}$. What does $\boldsymbol{Z}$ contain here in terms of the samples $\{\boldsymbol{x}_i\}_{i=1}^n$? Write an expression for it.
   Now consider another matrix $\boldsymbol{L} = \boldsymbol{Z}^\top \boldsymbol{Z}$. This matrix has size $n \times n$ and it is much smaller in terms of size. Determine how to obtain the eigenvectors of $\hat{\boldsymbol{C}}$ from the eigenvectors of $\boldsymbol{L}$. Hint: Write down the eigenvector equations carefully and use the definitions of $\hat{\boldsymbol{C}}$ and $\boldsymbol{L}$. [15 points]
3. Using the `randn` function in MATLAB and our knowledge of transformation of random variables, write a MATLAB program to draw samples from the Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ in each of the following cases: [20 points]
   (a) $\boldsymbol{\mu} = (1; 1)^\top, \boldsymbol{C} = \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}$
   (b) $\boldsymbol{\mu} = (1; 2)^\top, \boldsymbol{C} = \begin{pmatrix} 9 & 1 \\ 1 & 4 \end{pmatrix}$

Plant	Before	After	$A - B$
1	30.5	23	-7.5
2	18.5	21	2.5
3	24.5	22	-2.5
4	32	28.5	-3.5
5	16	14.5	-1.5
6	15	15.5	.5
7	23.5	24.5	1
8	25.5	21	-4.5
9	28	23.5	-4.5
10	18	16.5	-1.5

Table 1: Caption

In both cases, draw $n = 100$ points from these Gaussians and plot separate scatter plots of the points. Write down the 2D directions of maximum and minimum variance of these points in both cases. Note that the variance of a 2D random variable $\boldsymbol{X}$ in the direction $\boldsymbol{v}$ (where $\boldsymbol{v}$ is a unit vector) is given by $E([\boldsymbol{v}^\top(\boldsymbol{X} - \boldsymbol{\mu})]^2)$ where $E(\boldsymbol{X}) = \boldsymbol{\mu}$.

4. We have seen the Kolmogorov-Smirnov test in class and we have shown that the distribution of $D := \max_x |F_n(x) - F(x)|$ is independent of $F(x)$. Here $F(x)$ is the true CDF and $F_n(x)$ is the empirical CDF with n samples. Now suppose that you have two empirical CDFs $F_n(x)$ and $G_m(x)$ (with n and m samples respectively) and you want to check whether they correspond to the same underlying CDF. Now, your test statistic is $D2 := \max_x |F_n(x) - G_m(x)|$. Show that under the null hypothesis that both empirical CDFs indeed emerged from the same underlying CDF $F(x)$, the distribution of $D2$ is independent of $F(x)$. [15 points]

5. An industrial safety program was recently instituted in the computer chip industry. The average weekly loss (averaged over 1 month) in labor-hours due to accidents in 10 similar plants both before and after the program are as follows: Calculate the test statistic and the p-value. Determine, at the 5 percent level of significance, whether the safety program has been proven to be effective. [15 points]

6. Use the MATLAB code `readMNIST` from the homework folder to read in images of digits (0 to 9) from the well known MNIST database via a function called

```
[I_train,labels,I_test,labels_test] = readMNIST();
```

Here

```
I_train
```

is a cell array in MATLAB. You can extract the ith image from this array using

```
A=I_train{i}; A=A';
```

where `A` is a 2D unsigned integer array of size 28×28. Note that each image has size 28×28. Vectorize each image upon reading to create a 784×1 vector. This has can be done in MATLAB using `c = A(:);` where A is the 2D image array and c is its vectorized format. For each digit (0 to 9), do the following: [20 points]

   (a) Compute the sample covariance matrix of the vectorized images of that digit.

   (b) Compute the eigenvectors and eigenvalues of that covariance matrix using the `eig` command in MATLAB.

   (c) Plot of a graph of the eigenvalues in decreasing order. What do you observe in the eigenvalues?

   (d) Take the eigenvectors corresponding to the five largest eigenvalues and reshape them into images of size 28×28. Rescale those images so that the minimum is 0 and the maximum is 1 and plot them using the `imagesc` or `imshow` commands. What do you observe in these eigenvectors?

Your report should include the eigenvalue plots and eigenvectors plots for the digits 0, 3, 6, 7.