Mathematics and Statistic for Data Analysis

Coursework 2

Winter term 2023, School of Computing, University of Buckingham

---

This coursework is made up of 4 questions and worth 100 marks. To score full marks, your answers must be correct, well-referenced and well-explained.

This coursework is due on Monday 3rd of April at 10am.

You must submit your answers in three parts:

- The first part is a single **pdf** file answering all the questions in Section A and required items in Section B.
- In the second part you need to provide your code in python.
- For third part a ZIP file providing the **csv** files required only for Section B.

You must submit your files in Moodle, using the file names:

1. **studentnumber_coursework_2.pdf** for the main file answering all the questions;
2. **studentnumber_coursework_2-code.py** for your code;
3. **studentnumber_coursework_2-Q4.zip** that include the *.csv* files in Question 4 (name the separate files in the same manner).

Section A

Question 1. (20 marks).

Let

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 2 & 0 \\ 0 & 4 & 3 \end{bmatrix}.$$

(a) Compute the rank of $\mathbf{A}$;

(b) Provide the characteristic polynomial for $\mathbf{A}$ and compute the eigenvalues of $\mathbf{A}$.

(c) Let $\mathbf{u_i}$ be the i-th column of $\mathbf{A}$ for $i = 1, 2, 3$. Use the Gram-Schmidt process to generate an orthonormal basis from $\{\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3}\}$.

Question 2. (10 marks). The multidimensional random variables $\begin{bmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{bmatrix}$ follows $\mathcal{N}_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 7 & 3 & Y & 2 \\ 3 & 6 & 0 & 4 \\ Y & 0 & 5 & -2 \\ 2 & 4 & -2 & 4 \end{bmatrix};$$

where for $\boldsymbol{\Sigma}$ you need to substitute the last digit of your student ID for Y. Answer the followings after the substitution of Y, as mentioned, in $\boldsymbol{\Sigma}$:

(a) Determine all the independent univariate random variables.

(b) What is the probability density function for $x_1 + 2x_2$?

(c) Let $\mathbf{x} := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $\mathbf{y} := \begin{bmatrix} y_1 \end{bmatrix}$. Determine the conditional distribution density function $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{y} = 1)$.

Question 3. (10 marks). Consider the following function

$$f(x_1, x_2, x_3) = -2x_1^2 + 4x_1 + x_2x_1 - x_2 + x_3x_2 - x_3^2.$$

.

(a) Compute ∇f;

(b) Find the solutions for $\nabla f(\mathbf{x}) = 0$;

(c) Starting at a value $\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$, and consider learning rates $\{1, \frac{1}{2}\}$. Execute two steps of the gradient descent algorithm for each learning rate separately. Discuss the outcomes and also compare the result with item (b) (whether they are distancing from points in (b) or getting closer)?

Section B - practical

Section B contains only one question which has several parts (referred to here as items). In this section you need to provide the required answers clearly, as mentioned in each item. For some items you need to provide your final answer in the **pdf** file of your answer script (together with the answers from previous questions), and for some items you need to generate a **csv** file. For the items where you have been asked to generate a **csv** file you need to ZIP them altogether and submit it in Moodle as a single file in the required submission point. Furthermore, you need to also submit your final **code** that contains clear explanation (in the code) for each part.

Question 4. (25 + 35 marks). Consider the data set "fish_data.csv" that contains information about two types of fish. Rows associate with each fish and in columns their type, the age of the fish, water temperature in degrees Celsius, and the length of the fish is recorded.

(a) Let the length variable (length of the fish) in the data set be dependent on the age and water temperature.

i. Consider the first 60 rows and find parameters for the linear regression where we consider the loss function to be the square of residuals. Provide the regression function with the computed parameters in the **pdf** file. Compute the empirical risk $\mathbf{R_{empf}}$ for the linear regression and provide the final value in the **pdf** file.

ii. Similarly, consider the first 60 rows and find parameters for the following non-linear regression with the described feature map, where we consider the loss function to be the square of residuals

$$\phi(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}.$$

Compute the empirical risk. Provide the parameters and the empirical risk in the **pdf** file.

iii. Use the remaining 10 rows (i.e. rows 61–70) as your test data and discuss your models in items i. and ii. Which one of the models do you recommend? Provide your argument in the **pdf** file.

iv. Assume there is noise in the prediction function that follows the normal distribution $\mathcal{N}(0, \sigma^2)$. Consider the empirical risk for the regression in i. and set it as σ^2. Now, consider the training data to be the last 60 rows and set the prior distribution for the parameters to follow the normal distribution $(\boldsymbol{\theta}_\mathbf{i}, 4I)$ where $\boldsymbol{\theta}_\mathbf{i}$ are the parameters computed in part **i**. Compute $\boldsymbol{\theta}_\mathbf{MAP}$ parameters and provide the final answer in **pdf** file. By comparing the average error on the whole data set compare the models in **i**, **ii** and $\boldsymbol{\theta}_\mathbf{MAP}$.

(b) The two different types of fish are recorded with numerical values 1 and -1. During the measurement process, due to some error, some types of fish were not properly recorded and in the table they are identified with 0 (in the first column). We want to create a model based on the first 60 rows that predicts the type of the fish accurately:

i. The first model is based on only the Mahalanobis distances of the data points where we ignore the first column, in following steps:

A. Consider only the second, third and the fourth column (i.e. age, temperature and length) and find two data points $\mathbf{x_a}$ and $\mathbf{x_b}$ with the maximum Euclidean distances from each other. Provide the row numbers of $\mathbf{x_a}$ and $\mathbf{x_b}$ in your **pdf** file and generate the **csv** file of all the Euclidean distances.

B. Classify the remaining data points (within the first 60 rows) according to their Euclidean distance from $\mathbf{x_a}$ and $\mathbf{x_b}$: Create two lists $\mathbf{L_a}$ and $\mathbf{L_b}$ where the Euclidean distance of each data point in list $\mathbf{L_a}$ from $\mathbf{x_a}$ is smaller than its distances from $\mathbf{x_b}$. Similarly, the distance of each data point in $\mathbf{L}_b$ to $\mathbf{x_b}$ is smaller than its distance to $\mathbf{x_a}$. Provide the list $\mathbf{L_a}$ and $\mathbf{L_b}$ in your **pdf** file.

C. Consider only the second, third and the fourth columns (i.e. age, temperature and length) for each element in the

lists $\mathbf{L_a}$ and $\mathbf{L_b}$. Compute the empirical mean and empirical covariance; respectively (for list $\mathbf{L_a}$ the empirical mean $\boldsymbol{\mu_a}$ and empirical covariance matrix $\boldsymbol{\Sigma_a}$ and for list $\mathbf{L_b}$ similarly $\boldsymbol{\mu_b}$ and $\boldsymbol{\Sigma_b}$). Provide your final solutions in the **pdf** file.

D. Verify whether the empirical covariances $\boldsymbol{\Sigma_a}$ and $\boldsymbol{\Sigma_b}$ computed in the previous item are positive definite. Provide your reasoning in the **pdf** file.

E. Create two Mahalanobis distances based on two empirical covariances $\boldsymbol{\Sigma_a}$ and $\boldsymbol{\Sigma_b}$. Generate two separate **csv** files that store all the Mahalanobis distances elements in the lists $\mathbf{L_a}$ and $\mathbf{L_b}$. In other words you generate two matrices that store the corresponding Mahalanabis distance to $\boldsymbol{\Sigma_i}$ in list $\mathbf{L_i}$ where $i \in \{a, b\}$.

F. The suggested model M_1 is the following: Given a new data $\mathbf{x}_*$ which contain only three values (for age, temperature and length), first the corresponding Mahalanobis distances of $\mathbf{x}_*$ from $\mathbf{x_a}$ and $\mathbf{x_b}$ is computed. The model M_1 labels $\mathbf{x}_*$ according to the label of the smaller value of the corresponding Mahalonobis distances from $\mathbf{x_a}$ and $\mathbf{x_b}$. Using model M_1, label for $\begin{bmatrix}55 & 27 & 460\end{bmatrix}$.

ii. The second model M_2 is a linear regression model where the type of the fish is dependent on the age, temperature and length of the fish. Provide the function and the parameters for the regression function in the **pdf** file and explain carefully how your approach. Find the label for $\begin{bmatrix}55 & 27 & 460\end{bmatrix}$.

iii. For each model M_1 and M_2 compute the relevant error measures. Furthermore, use the remaining data (rows 61–70) as your test data to discuss the two models. Which one do you recommend? Provide your answer in **pdf** file.

Zaniar Ghadernezhad, March 2023.