MA3832-Assessment 1

Weighting: 10% *Total marks: 12*

Due on Friday, 12^{nd} August 2022, 11:59pm AEST

Overview

This assessment aims to assess students' understanding on the topics covered in week 2. It addresses the following learning outcome(s):

- understanding the roles of linear algebra, and optimisation in the realm of machines learning;
- understanding algorithms underpinning various optimisation methods;
- applying and implementing concepts in linear algebra, and optimisation in Python.

Submission

You will need to submit the following:

- A PDF file clearly shows the assignment question, the associated answers, any relevant Python outputs, analyses and discussions.
- **Python/Jupyternotebook** script file to reproduce your work.
- The task cover sheet.

You have up to three attempts to submit your assessment, and only the last submission will be graded.

A word on plagiarism:

Plagiarism is the act of using someone else's words, work or ideas from any source as one's own. Plagiarism has no place in a University. Student work containing plagiarised material will be subject to formal university processes.

Question 1 **(12 marks)**

In this question, we consider the `marketing.csv` data which contains 200 observations and 4 variables. The response variable is sales, denoted as y. The explanatory variables—measured in thousands of dollars—are advertising budget spent on youtube, newspapers and facebook, respectively, which are denoted as X_1, X_2 and X_3, respectively.

To model impacts of the media strategies on logarithm of sales, a researcher uses the following multiple linear regression:

$$Y = \log(y) = \beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \beta_3 \log(X_3) + \epsilon, \tag{1}$$

$$= \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{2}$$

where $Y, \log(X_1), \log(X_2)$ and $\log(X_3)$ is a vector of $n \times 1$ (n is the number of observations in the dataset), $\mathbf{X} = (\mathbf{1}_n, \log(X_1), \log(X_2), \log(X_3))$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]'$.

The parameter $\boldsymbol{\beta}$ in Equation (2) can be estimated by minimising the following loss function - mean squared errors:

$$\mathcal{L} = \frac{1}{n}(Y - \mathbf{X}\boldsymbol{\beta})'(Y - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 \tag{3}$$

It is well-known that the optimal solution of $\boldsymbol{\beta}$ in Equation (3), denoted as $\hat{\boldsymbol{\beta}}$, has the following form:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \tag{4}$$

and its standard deviation is

$$s.e(\hat{\boldsymbol{\beta}}) = \sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}} \text{ where } s^2 = \frac{1}{n-4}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2, i = 1, 2, \ldots, n. \tag{5}$$

Your tasks are to:

(a) use equations (4) and (5) to estimate $\boldsymbol{\beta}$ and its standard deviation in `Python`. Comment on impacts of the media advertisement on sales.

(b) Write down a step-by-step procedure of Classical Gradient Descent to estimate $\boldsymbol{\beta}$ in Equation (3)

(c) Write a Python code to implement the Classical Gradient Descent procedure provided in (b).

(d) Discuss the results obtained from (c) and compare it with that obtained from (a).