Lab2

February 7, 2023

Lab 2 has two parts (A) and (B) and you are required to implement it in Python. **A good coding style should be followed, consider PEP 8 [1] or Google Python coding style [2].**

Part A

Continuous Distribution, Normal Distribution, Central Limit Theorem and Explore the data: Describe your results and graphs in the report for all tasks in part A.

1. Continuous Distribution

Implement in Python the following exercise in Ref [3]. Answer all questions.

2. Normal Distribution

Implement in Python the following exercise in Ref [4]. Lap times (see table C1) can be found here [5]. **Answer all questions and do not sketch with ruler and pencil as suggested :-).**

3. Central Limit Theorem

Implement in Python the following exercise in Ref [6]. Answer all questions.

4. Explore the Melbourne Real-Estate Dataset [7]

Perform any three data anaylsis tasks other than simple mean and standard deviations. For example, use the dataset to see if there is any correlation between distance from CBD and prices of properties. Can you find/analyze something about house sizes and prices. How does it look for different suburbs in terms of number of properties, which have the most homes and least homes. Which councils/suburbs have homes built in which years? How does it look for different sellers when selling homes? How are the different regions in terms of price point and house sizes? Think creatively!

Part B

Gradient Descent

Gradient Descent is widely used to optimize the parameters of machine learning models iteratively. In this lab exercise, you will explore the gradient descent algorithm (GDA) and implement it in Python.

Case study: Consider a simple linear model and find the best model parameters usually written in the form: $y = mx+c$ where m is the gradient or slope, and c is the intercept. With the Melbourne Real-Estate Dataset provided previously, explore:

- The relationships between the number of bedrooms with property prices in a particular suburb for example Caulfield North. In doing so, represent this relationship using a plot;
- try to find the best model parameters (m and the c) using GDA:
- explore different learning rates and its effect on m and c;
- explore the cost function; and
- plot your graph with the best model parameters.

Work Environment

This lab project will be created using a virtual environment in Anaconda to show application level isolation.

Lab Report

The report should be written in the lab report format template file, can be found here [8] (Do not edit this template file, make a copy in your own Overleaf account to edit!) using LaTeX in Overleaf. Download the pdf file after you finish writing and submit the pdf along with your zipped code files to Canvas. **You must document your code properly so its readable.

References

[1] "PEP 8 — Style Guide for Python Code." https://www.python.org/dev/peps/pep-0008/e. Accessed: 2021-02-11.

[2] "Google Python Style Guide." https://google.github.io/styleguide/pyguide.html. Accessed: 2021-02-11.

[3] "Continuous Distribution." https://openstax.org/books/introductory-statistics/pages/ 5-4-continuous-distribution. Accessed: 2021-01-28.

[4] "Normal Distribution." https://openstax.org/books/introductory-statistics/pages/ 6-3-normal-distribution-lap-times. Accessed: 2021-01-28.

[5] "Normal Distribution Lap Times Table." https://openstax.org/books/introductory-statistics/pages/ c-data-sets. Accessed: 2021-01-28.

[6] "Central Limit Theorem." https://openstax.org/books/introductory-statistics/pages/ 7-5-central-limit-theorem-cookie-recipes. Accessed: 2021-01-28.

[7] "Melbourne Realestate Dataset." https://www.kaggle.com/dansbecker/melbourne-housing-snapshot. Accessed: 2021-01-28.