AI Apprenticeship Program Technical Assessment Questions

Question 1: Data Ingestion and Preparation

Context: Real Real Estate Agency Pte. Ltd. is building a product that provides interested homebuyers with realistic house price forecasts. They would like to obtain a good database of **Singapore housing data** that is **consistently up to date**. The purpose of this dataset is to be used by data scientists, who are focused on statistical modelling and data exploration. As a data engineer, your role is to collect this data. Design the appropriate processes and pipelines to generate the present the data to the data scientist.

You are required to submit the code that you will write to prepare this data - if you intend to use multiple sources of a repetitive nature, one script on one target source is sufficient. In addition, submit a one-page summary of your thought process, code workflow and how this could be deployed. A good submission should demonstrate clean code with good engineering practices, as well as a summary that demonstrates what you identify to be critical to the success of the product in your role and how you address these concerns.

Question 2: Machine Learning

Business context: Real Real Estate Agency is building a tool that provides interested homebuyers with realistic house price forecasts.

In this section, you will be tested on the ability to build a basic model based on the dataset that has been provided. Your goal is to build a model that predicts the variable 'Y house price of unit area'. You are to submit the scripts or notebooks you've used in creating your model, for which both R/Python languages are accepted. Other than fulfilling the requirements of the tasks stated below, do ensure your code is of quality.

Tasks:

1. Data Cleaning: Show and explain your process on handling any noise that is available in the data.
2. Exploratory Data Analysis: Display your findings from the data with appropriate outputs and visualisations. Do elaborate and explain on the steps of EDA you have taken in leading up to the next task.
3. Model Performance: Provide an output of performance metrics for the model(s) you have created. Some metrics that you would like to use would include RMSE, RMAE, or R2. Do provide explanations for the metrics you have chosen.

Data Set Information:
The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan.

Attribute Information:
The inputs are as follows
X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
X2=the house age (unit: year)
X3=the distance to the nearest MRT station (unit: meter)
X4=the number of convenience stores in the living circle on foot (integer)
X5=the geographic coordinate, latitude. (unit: degree)
X6=the geographic coordinate, longitude. (unit: degree)

The output is as follows
Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Question 3: Quantitative Theory

Context: You are a machine learning engineer in the medieval times, where computers and sklearn do not exist. Your boss wants you to build a linear regressor to predict housing prices. He has 3 huts, which he wants you to conduct the problem on. The goal of this model is to predict the price, ***p*** in terms of the number of bedrooms ***b*** and toilets ***t***.

num_bedrooms, ***b***	num_toilets, ***t***	price, ***p***
3	1	30
6	1	55
3	3	70

A. Gradient Descent
For this question, no computing aids are allowed. You may choose to scan a copy of your worked solution, or write it in text or LaTeX within the notebook.

1. Derive the hypothesis for a linear regression model based on θ.
2. Derive the full (mean-squared error) MSE cost function of this problem.
3. Derive the gradient w.r.t. each variable.
4. Using a learning rate, α = 1, and initializing θ at [0, 0, 0], determine the new θ after one round of gradient descent.
5. Being computational in nature, this approach requires many steps, which is not human-friendly in nature. Using a more analytical approach, provide the optimal weights for θ.
6. How can we verify that this answer is correct? Please show your working.

B. Regularization

For this question, we would value answers that display both a strong theoretical understanding of the topic of regularisation, as well as the ability to apply this theoretical understanding to practical problems.

Where possible, you are also encouraged to use numerical examples as elaboration on theoretical concepts you introduce. In other words, blindly copying from Wikipedia or Stack Exchange is not encouraged.

1. In your own words of a few sentences, explain what regularisation means. Specifically, what is the problem that occurs naturally in unregularised cost functions, and how does regularisation fix it? Also, given a regularisation parameter λ, what is λ trying to account for? How does a λ slightly greater than the optimal value affect, and how does a slightly smaller λ affect it? How does $\lambda = 0$ and $\lambda = \infty$ affect it? How does regularisation affect the mathematical complexity and search space of the problem? (By complexity, we mean that a higher order polynomial would increase the complexity.)

2. Explain the differences in a few sentences between L1 and L2 regularisation. Specifically, which one is capable of eliminating coefficients, and why is it able to do so, while the other is unable to do so? In addition, are these regularisation methods differentiable, and if not, how do we implement gradient descent for it?