Assignment 1: Dataset exploration and Regression

for AIDI1009-25W: Introduction to Neural Networks

—

Due Date: **12th June 2025 @ 11:55 pm**

1 Goals

The main goal of this assignment is to:

1. Learn basics of Regression.
2. Understand how to evaluate regression using different metrics.
3. Learn how to use Scikit Learn Package.
4. Learn how to manipulate datasets.

2 Introduction

Regression analysis is used when you want to predict a continuous dependent variable from a number of independent variables. If the dependent variable is dichotomous, then logistic regression should be used. Simple linear regression is when you want to predict values of one variable, given values of another variable. For example, you might want to predict a person's height (in inches) from his/her weight (in pounds). Standard multiple regression is the same idea as simple linear regression, except now you have several independent variables predicting the dependent variable. To continue with the previous example, imagine that you now wanted to predict a person's height from the gender of the person and from the weight. You would use standard multiple regression in which gender and weight were the independent variables and height was the dependent variable.
For more information on linear regression following resources are a good start.

1. Introduction to Regression Analysis
2. Linear Regression Models
3. Simple Linear Regression
4. Regression Analysis: Step by Step

3 Dataset

The dataset named "Engineering_graduate_salary.csv" is placed at Blackboard. This dataset is originally downloaded from Kaggle and we would like to thank "Aspiring Minds Research" for making this dataset available publicly.

3.1 Datset fields description.

1. **ID**: A unique ID to identify a candidate
2. **Salary**: Annual CTC offered to the candidate (in INR)
3. **Gender**: Candidate's gender
4. **DOB**: Date of birth of the candidate
5. **10percentage**: Overall marks obtained in grade 10 examinations
6. **10board**: The school board whose curriculum the candidate followed in grade 10
7. **12graduation**: Year of graduation - senior year high school

8. **12percentage**: Overall marks obtained in grade 12 examinations
9. **12board**: The school board whose curriculum the candidate followed
10. **CollegeID**: Unique ID identifying the university/college which the candidate attended for her/his undergraduate
11. **CollegeTier**: Each college has been annotated as 1 or 2. The annotations have been computed from the average AMCAT scores obtained by the students in the college/university. Colleges with an average score above a threshold are tagged as 1 and others as 2.
12. **Degree**: Degree obtained/pursued by the candidate
13. **Specialization**: Specialization pursued by the candidate
14. **CollegeGPA**: Aggregate GPA at graduation
15. **CollegeCityID**: A unique ID to identify the city in which the college is located in.
16. **CollegeCityTier**: The tier of the city in which the college is located in. This is annotated based on the population of the cities.
17. **CollegeState**: Name of the state in which the college is located
18. **GraduationYear**: Year of graduation (Bachelor's degree)
19. **English**: Scores in AMCAT English section
20. **Logical**: Score in AMCAT Logical ability section
21. **Quant**: Score in AMCAT's Quantitative ability section
22. **Domain**: Scores in AMCAT's domain module
23. **ComputerProgramming**: Score in AMCAT's Computer programming section
24. **ElectronicsAndSemicon**: Score in AMCAT's Electronics & Semiconductor Engineering section
25. **ComputerScience**: Score in AMCAT's Computer Science section
26. **MechanicalEngg**: Score in AMCAT's Mechanical Engineering section
27. **ElectricalEngg**: Score in AMCAT's Electrical Engineering section
28. **TelecomEngg**: Score in AMCAT's Telecommunication Engineering section
29. **CivilEngg**: Score in AMCAT's Civil Engineering section
30. **conscientiousness**: Scores in one of the sections of AMCAT's personality test
31. **agreeableness**: Scores in one of the sections of AMCAT's personality test
32. **extraversion**: Scores in one of the sections of AMCAT's personality test
33. **nueroticism**: Scores in one of the sections of AMCAT's personality test
34. **openess_to_experience**: Scores in one of the sections of AMCAT's personality test

4 Exercise

This exercise will test your ability to train, develop and deploy a Machine Learning regression model. The dataset (Engineering_graduate_salary Dataset), consists of 33 features and a label (Salary). Notice that several feature values are either missing or duplicated.

1. Download the "Engineering_graduate_salary.csv" dataset from D2L.
2. Perform Data Exploration:
   (a) Display the first few records in the dataset.
   (b) Display the number of rows and columns of the dataset.
   (c) Display the dataset statistics (min, max, ...)
   (d) Display the Null values of each feature.
   (e) Plot some graphs of the data to assist you in data exploration.
3. Perform initial cleaning of the data:
   (a) Delete columns which mainly contain Null values.
   (b) Remove duplicate columns (obvious redundant information).
   (c) Fill missing values in numeric columns if necessary.
   (d) Display the number of rows and columns of the data after cleaning.
   (e) Display the features left after cleaning.
   (f) Plot the distribution (histogram) of each of the following features: " DOB", "12percentage".
4. Analyze the pair-wise relationship between the features of the Data set.
5. Plot the correlation heatmap from the pairwise plots.
6. Perform necessary Data Preprocessing (Transformation) for final data preparation.
   (a) Scale values in numeric columns to a (0,1) range if needed.
   (b) Encode categorical data into one-hot vectors.
   (c) Split the dataset into training, validation and testing
7. Choose the Simple Linear Regression Model from Scikit Learn and use it to perform regression.
8. Use only the features after cleaning in the dataset and obtain the following measures to compare the performance of the different models.
   (a) MSE.
   (b) RMSE.
9. Perform Feature Selection (after cleaning the dataset) and repeat the previous task (step 8).
10. Perform Hyper-parameter Tuning on the models and again compare the different models.

5 Report Format

Name your report AIDI1009-24F-10827: Assignment#1 LastNameFirstName.pdf. Below is the general format of the report required:

1. The front page (i.e. title page) should contain only the following:
   - Course #, Course Name and Date
   - Your Name and ID
   - Assignment # and title of the assignment.
2. ntroduce the problem to be solved:

- Problem Statement
  (a) Briefly describe the problem solved in the assignment.
  (b) Assumptions and Constraints
  (c) Constraints could be for example using certain libraries, datasets, or a specific programming language.

3. Answer all questions posed in Section 4. Append the following to your answers:
   (a) Plots and Graphs.
   (b) Tables.
   (c) Figures.
4. Python Code
   (a) Python code for the problem
   (b) Document your code.