31/12/2024

Fall 2024-2025
AIE-207 Scientific Computation Techniques
PROJECT

DUE DATE: 15/01/2025

You will implement linear regression to predict students' performance using six variables:

Description:

The Student Performance Dataset is a dataset designed to examine the factors influencing academic student performance. The dataset consists of 10,000 student records, with each record containing information about various predictors and a performance index.

Variables:

- **Hours Studied**: The total number of hours spent studying by each student.
- **Previous Scores**: The scores obtained by students in previous tests.
- **Extracurricular Activities**: Whether the student participates in extracurricular activities (Yes or No).
- **Sleep Hours**: The average number of hours of sleep the student had per day.
- **Sample Question Papers Practiced**: The number of sample question papers the student practiced.

Target Variable:

- **Performance Index**: A measure of the overall performance of each student. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

Implementation Steps:

1. Upload the provided data (Train_Data.csv and Test_Data.csv), write your own codes for regression algorithm for predicting performance index.
2. Convert the Extracurricular Activities column outputs to numeric values; set "Yes" = 1 and "No" = 0.
3. After all value types in the dataset are numeric, apply min-max normalization to all variables except the target variable:

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

4. The steps for Linear Regression are as follows:

a) Determining the Hypothesis Function

The hypothesis function for a linear regression model is defined as follows: $h_\theta(x) = \theta_0 + \theta_1 x$. Here, θ_0 and θ_1 are the parameters of the model, and x represents the independent variable.

Hint: For every feature (excluding the target variable), a theta parameter (θ) needs to be assigned.

b) Defining the Cost Function

The cost function is used to measure how well the model performs. This function measures how different the model's predictions are from the actual values. The cost function is defined as follows:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Here, m represents the number of samples, and $x^{(i)}$ and $y^{(i)}$ represent the independent and dependent variable values of the i^{th} sample, respectively.

c) Applying the Gradient Descent Algorithm

The gradient descent algorithm is used to minimize the cost function. This algorithm calculates the gradient of the cost function and iteratively updates the parameters θ_0 and θ_1 The update equations are as follows:

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$$

Here, α (alpha) is the learning rate, which determines how quickly the learning will occur.

d) Iterating the Algorithm

The gradient descent algorithm is run iteratively over a certain number of iterations or until the change in the cost function falls below a certain threshold. Here, the iteration limit is determined by the iteration variable. Here, the goal is to test the theta parameters (θ) updated during training on the test data to determine the error rate.

5. Adjust the alpha and iteration values in your regression model to generate a cost graph, include this graph in your report, and provide an analysis. Try alpha values of 1, 0.1, 0.01, 0.001 and iteration counts of 100, 1000, 2000, 4000.

6. The performance of the model can be evaluated using the cost function calculated on the test dataset. Evaluate the performance of your model on the test data. Use R^2 (coefficient of determination) as error metric to measure the prediction accuracy of your model. Prepare an Excel file containing your model's prediction values and these error metrics and present this file in your report.

Important Notes:

- You can find the explanation and formulas related to linear regression in the attached PDF files "***Lecture2-Regression***" and "***Lecture3-Regression***."
- Your work must be original; you should create your own code and report. Using Chat GPT or similar will result in a direct score of 0.
- The use of built-in functions is strictly prohibited, except for (min, max, sum, dot, iloc, read_csv, randint, shape, etc.).
- Make sure to implement the formulas in the project within the functions you create.

Project Submission Format:

After completing your project, you will be required to submit your working code (configured to work with the dataset) and your report in PDF format. Upload the functioning code and the report in the format ***"name_surname_number.zip"*** to the class environment.

Within the report, you should include:

- Explain the preparation of the dataset and the purpose of each step used.
- Include the outputs requested in the project in the report and write detailed explanations for each.