PGP IN AI/ML
Regression - Project - Part B

Submission Date: 2359hrs on07/07/2019

Total Marks: 24 marks

The objective of this part of the project is to develop a linear regression model for the given dataset using Gradient Descent methods. The regression model should be built using numpy, pandas, and sklearn.

Problem Description:

The problem statement is similar to the one in Part A of the assignment. Train the linear regression model using each of the following methods:

1) Stochastic Gradient Descent
2) Gradient Descent
3) Mini-batch Gradient Descent

Instructions:

These are some guidelines to help you get started with the code:

1. SGD can be implemented using **linear_model.SGDRegressor**(using **fit()** method)
2. Tune the following hyper-parameters for satisfactory results:
   - **a) max_iter**
   - **b) eta0**
   - **c) tol**

2) Every model will be evaluated using the following measures:
   - a) R^2 value
   - b) RMSE score
   - c) RSE - Residual Squared Error
3) Use t-test to confirm that **Son's Height** significantly depends on **Father's Height**
4) Plot the obtained line against the scatter plot of the data points.

Note:

1. Gradient Descent and Mini-Batch Gradient Descent methods do not have an in-built API in **sci-kit learn** library.

   To implement Mini-batch:
   - a. Create batches of the training dataset
   - b. Use **partial_fit** method of **SGDRegressor** class, to train the entire batch at once (batch size needs to be tuned to get satisfactory results)
   - c. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html

   To implement Gradient Descent by
   - a. Set the batch size as the size of the training dataset.

2. **t-Test Method**
    a. Analysis needs to be done on paper (not code)
    b. Refer to **Dependency of the dependent (target) variable on the independent (feature) variable** video in **MODULE 3**
    c. T-Tables can be referred from http://tyigit.bilkent.edu.tr/metrics/t-table.pdf
    d. Please refer to the example below for looking up the t-table.
    e. **Hint:** As we have 1000+ data points, consider the last row (infinite degrees of freedom) to get the t-statistic value. (Assume two-tail probabilities for selecting the column)

Submission Details:

Your submission should consist of:

1. SGD [3+1 marks]
    a. Code - **sgd.py**
    b. Visualization(Scatter Plot + Line) - **sgd_line.png**
2. Mini-Batch [3+1 marks]
    a. Code - **minibatch.py**
    b. Visualization(Scatter Plot + Line) - **minibatch_line.png**
3. Gradient Descent [3+1 marks]
    a. Code - **gradient_descent.py**
    b. Visualization(Scatter Plot + Line) - **gradient_descent_line.png**
4. Answers to the following question in docx file: (**question_answers_part_b.docx**)
    a. Values of intercept and coefficient obtained in Gradient Descent methods are slightly away from the values obtained from OLS method (implemented in Part-A). Which of the values are more reliable and why? [1 + 0.5 marks]
    b. The t-test for the three models [3 marks]
5. A separate document with the following information about implemented models (**model_output_part_b.docx**) [(5 values for each model) x 0.5 x 3 models = 7.5 marks]
    a. Coefficient obtained for SGD:
    b. Intercept obtained for SGD:
    c. RMSE, RSE and R^2 score for SGD:
    d. Coefficient obtained for mini batch:
    e. Intercept obtained for mini batch:
    f. RMSE, RSE and R^2 score for mini batch:
    g. Coefficient obtained for gradient descent:
    h. Intercept obtained for gradient descent:
    i. RMSE, RSE and R^2 score for gradient descent:

Make a folder with all these files and upload it in zip format with the name ‘<2018AIML_your_id>_regression_part_B.zip’.

Contacts:

- ✓ You should put up questions in discussion forum of the corresponding assignment folder only.
- ✓ Please put all your queries to the following TAs in Canvas but not to the instructor.
  - f20160688@hyderabad.bits-pilani.ac.in
  - f20160015@hyderabad.bits-pilani.ac.in
  - f20160484@hyderabad.bits-pilani.ac.in
  - himanshuchawla437@gmail.com
  - mohitrathore15@gmail.com

Example for t-Test:

We generally use 0.05 probability as the limit to prove the correctness of a mathematical hypothesis. Consider the following example:

1. 10 data points
2. There can be outliers towards both the tails in the distribution i.e, two-tail

Looking up in the t-table:

1. We have 10 data points, so the number of degrees of freedom is 10. So, that particular row needs to be selected.
2. Select the column corresponding to two-tail 0.05 i.e, 7th column from the left. So the t-statistic value is **2.228**
3. $t_{calculated} = (w_0 - 0)/\sigma$, if $t_{calculated} > t_{statistic\ from\ table}$, then w0 obtained from the model is significant, i.e. the relationship between independent and dependent features is significant.