PA1-skeleton-Ashis-Biswas

February 17, 2019

1 Programming Assignment 1

- CSCI-4930/5930 ML Spring 2019 (Be sure to discard which section you are not enrolled)
- Author: Ashis Biswas (Replace my name with yours)

1.1 Tasks for everyone (Tasks 1-15)

1.1.1 TASK 1: Import all the necessary packages here

In [ ]:

1.1.2 TASK 2: Load the dataset into memory so that you can play with it here

In [ ]:

1.1.3 TASK 3: Compute mean, stdev, min, max, 25% percentile, median and 75% percentile of the dataset (BWEIGHT variable)

In [2]:

1.1.4 TASK 4: Also, draw the histogram plot for the BWEIGHT variable

In [ ]:

1.1.5 TASK 5: Present the skewness and kurtosis of the BWEIGHT target variable

In [ ]:

1.1.6 TASK 6: Do variable selection from the pool of 36 variables based on correlation score with the target variable BWEIGHT

1.1.7 Please report all the variables you kept for training.

In [ ]:

1.1.8 TASK 7: Check for missing data, and apply a "good" strategy to tackle it

In [ ]:

1.1.9 TASK 8: Tackle the dummy categorical variables by introducing dummy variables

In [ ]:

1.1.10 TASK 9.1: Randomly split the dataset into training, Tr (80%) and testing, Te (20%)

In [ ]:

1.1.11 TASK 9.2: On the training dataset, apply a normalization technique

In [ ]:

1.1.12 TASK 9.3: Apply the training data statistics to normalize the testing data as well.

In [ ]:

1.1.13 TASK 10: Find the linear regression function describing the training dataset using a technique you recently learned in class. CLOSED-FORM vs. Gradient Descent (batch or stochastic or mini-batch).

1.1.14 PLEASE DO NOT CALL ANY LIBRARY FUNCTION THAT MIGHT DO THE TASK FOR YOU. If you do, you are most likely get a ZERO for this assignment.

In [ ]:

1.1.15 Task 11: Predict BWEIGHT target variable for each of the testing dataset using the regression line you learned in Task 10, and report RMSE(testing) (Root Mean Squared Error)

In [ ]:

1.1.16 Repeat TASK 10 additional four times : Run linear regression training again

1.1.17 After each run, Report RMSE(testing)

In [ ]:

1.1.18 Task 12: Finally, Report RMSE(testing) = Average(RMSE_test) $\pm$ Stdev(RMSE_test)

1.1.19 Here Average(RMSE_test) = average of all the 5 RMSE(testing) scores you got above.

1.1.20 And, stdev(RMSE_test) = standard deviation of all the 5 RMSE(testing) scores above.

In [ ]:

1.1.21 Task 13: Run linear regression one last time on the whole dataset (i.e, training+testing which is preprocessed by you above).

In [ ]:

1.1.22 Task 14: Preprocess the judge-without-label.csv file according o the strategy you applied above on the whole dataset (task 13)

In [ ]:

1.1.23 Task 15: Predict BWEIGHT for each of the samples from the judge-without-label.csv file, and save the results in judge-submission-run-1.csv in the format below. Please change the run number and report what changes you have made in a corresponding file run-1.txt.

In [ ]:

2 Tasks only for CSCI-5930 (Grad) students

2.0.1 Task 16: Repeat tasks 9-12 three times, and report the ultimate RMSE_test average $\pm$ ultimate RMSE_test stdev

In [ ]:

2.0.2 Task 17: Make an entry in the Kaggle challenge below:

- [https://www.kaggle.com/c/csci-ml-s19-pa1/]
- Please oin the challenge and submit a judge-submission-run1.csv file, and please report your Kaggle handle here too. ### There is limit of 5 entries each day untile the deadline. ### For each of the runs you submit, please report here the RMSE you got (as reported by the Kaggle platform).

In [ ]: