CS 696 Intro to Big Data
Fall Semester, 2019
Assignment 3

San Diego State University -- This page last updated 1/29/19

Due Mar 19 23:59

This a assignment uses datasets that can be downloaded at:

http://www.eli.sdsu.edu/courses/spring19/cs696/datasets/index.html

Problem 1. The dataset GPA-GRE contains 16 years of GPA, Verbal and Quantitative GRE scores for graduate students.

a. Use linear regression with the combined GRE score (Verbal + Quantitive) as the independent variable and the GPA as the dependent variable to create a model that predicts a students GPA given their combined GRE score.

b. What is the r2 score of the model?

c. Plot the residuals of the data from the model.

Problem 2. In the dataset dwell data there is a file multiple-site.tsv. The file contains two columns: a site number and dwell-time on that site. Each site contains multiple entires.

Using scikitlearn compute the mean and standard deviation of the dwell time per site.

Using Spark (not scikitlearn) compute the mean and standard deviation of the dwell time per site.

Instructions

You assignment 3 jupyter notebook should be self contained. All calculations and answers to the questions are to be in one notebook. This assignment requires you to use files, some of which are provided. You notebook needs to read the unmodified files, including names. Any needed modification to the files needs to be done in the notebook.

At the beginning of your notebook you should create variable that hold the path (plus name) of any input files that you use. It is likely that for grading purposes those paths will need to change. I should be able to run your notebook using my input files by just changing the path to files at the top of your notebook.

Grading

10 points per problem.

What to turn in

You are to turn in a Jupyter Python notebook containing the code and answers to the questions above. Since Jupyter notebooks can contain text and code, before each problem indicate which problem it is in text, not in code comment.

Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.