C S 487/519 Applied Machine Learning I
Fall 2018
Project 4: Compare regression methods

1 Objective

In this *individual* project, you are required to understand and compare several regression algorithms.

2 Requirements

- (60 points) Write code to conduct regression by
  - (30 points) utilizing several regression functions: (i) `LinearRegression`, (ii) `RANSACRegressor`, (iii) `Ridge`, and (iv) `Lasso`. These functions are provided by the Python scikit-learn library (`http://scikit-learn.org/stable/`).
  - (10 points) implementing the normal equation solution, and
  - (20 points) using one approach to conduct non-linear regression.
- (20 points) Each regressor (except the normal equation method) needs to be tested using two datasets: (1) the `housing` dataset, which is on Canvas, and (2) the California Renewable Production 2010-2018 dataset at `https://www.kaggle.com/cheedcheed/california-wind-power-generation-forecasting/data`. For the Normal equation solution, you just need to test it using the housing dataset.
- (15 points) Properly analyze the regressors' behavior by applying the knowledge that we discussed in class. Such analysis should include at least Mean squared error (MSE) (or R^2 score, or residual plots) and running time.
- (5 points) Write a readme file `readme.txt` with the commands to run your code.
- (10 points; **Bonus**) Design a new approach to improve the existing regressor. You need to describe your method in detail in the readme file. Your analysis should also include this new method. If your new approach outperforms the existing methods in a reasonable way (e.g., largely decrease MSE while using similar amount of running time, or generate similar MSE but reduce the running time dramatically), you will get the bonus points. If your new method does not outperform the existing method, you will only get two bonus points.
- Your Python code should be written for Python version 3.5.2 or higher.
- Please properly organize your Python code.

3 Submission instructions

- Compress your python code to a zip file named `proj4.zip` and upload it to Canvas. Note, **this instruction is changed!**

4 Grading criteria

(1) The score allocation has already been put beside the questions.

(2) Please make sure that you test your code **thoroughly** by considering all possible test cases. **For this project, your code will not be tested using more datasets. Thus, it does not need to be flexible to accept different datasets as input**

(3) At least 5 points will be deducted if submitted files (including files types, file names, etc.) do not follow the instructions.