# DATA 240, Spring 2024

## Assignment #3

**Release on April 9, 2024**
**Due 11:59pm on April 21, 2024**

### Notes

*This assignment should be submitted in Canvas as a format of ipython notebook (assignment3_yourFirstName_LastnName.ipynb).*
No late assignments will be accepted.
You may collaborate on homework but must write independent code/solutions. Copying and other forms of cheating will not be tolerated and will result in a zero score for the homework (minimal penalty) or a failing grade for the course. Your work will be graded in terms of correctness, completeness, and clarity, not just the answer. Thus, correct answers with no or poorly written supporting steps may receive very little credit.

*NOTE: Please do not use any package/library including scikit-learn library except NumPy, Pandas, and Matplotlib.*

Please download housing.csv file. This is a real-life dataset consisting of housing sales prices in the city of Windsor, Ontario, Canada. You can find a description of the variables on housing.txt file.
Our target variable is 'price'.

### 1. (4 pts) Linear regression with multiple variables

*Instruction:* Linear regression using gradient descent method from scratch
This is the task of Linear regression with multi variables.

$$\hat{y} = W_0 + W_1 x_1 + W_2 x_2 + \ldots + W_n x_n$$

*NOTE: Please do not use any open-source algorithm for gradient decent method. Instead, you need to write gradient descent method from scratch.*

You need to find the optimum Weights using gradient decent method from scratch.
Before applying gradient descent method, you might need to normalize variables, which is called feature scaling or normalization.
In gradient descent algorithm, weight need to be updated every iteration.

1-1. (2pts) Please print out the Weight values and the Root Mean Squared Error (RMSE) after optimization.
1-2. (1pt) Please fit the data using the Linear regression model with the optimum Weight
To simply the problem, please plot 'price' vs 'lotsize' as the below.

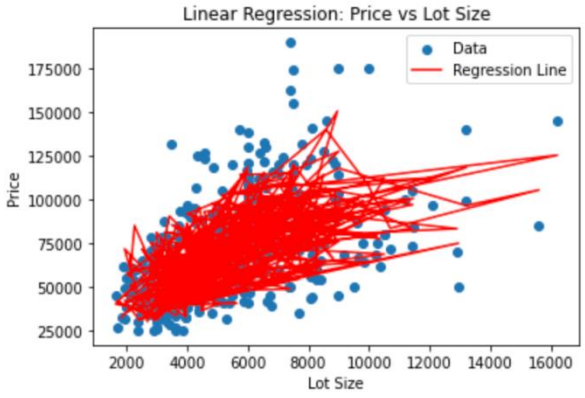


Fig. Bad example

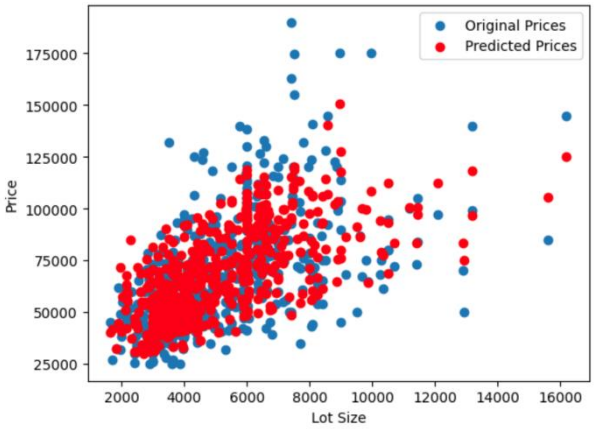


Fig. Good example

1-3. (1pt) Please plot 'true-price'(y) vs 'predicted-price'($\hat{y}$) and display $R^2$. The y and $\hat{y}$ should be original scale instead of normalized scale.

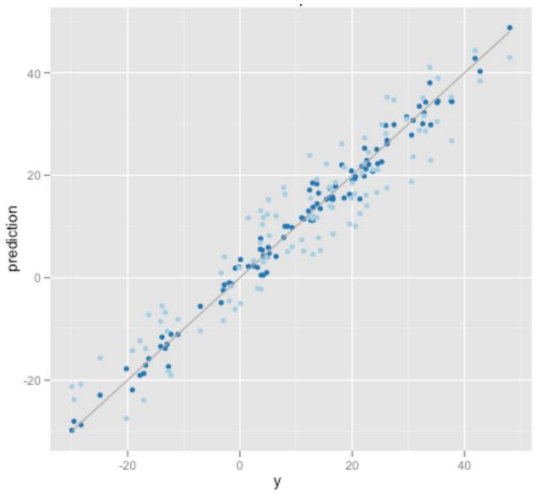


Fig. Plot of y vs $\hat{y}$

**2. (6 pts) Linear regression with Regularization**

*Instruction:* This is the task of polynomial regression.

$$\hat{y} = W_0 + W_1 x_1 + W_2 x_1^2 + W_3 x_1^3 + \cdots$$

To simply the problem, please consider '*lotsize*' as only the input variable.

You need to split the data into training/testing data set with the ratio of 70% / 30% to check overfitting.
*NOTE: you can use scikit-learn library package for modeling and splitting the dataset. Please use random state=123 for splitting the data.*

**2-1 (3 pts). Fit the training data using 5th order polynomial regression model and Ridge (L2 penalty) regularization.**

You need to try many different L2 penalty (for example, λ = 0.1, 0.5, 1, 5, 10, …, etc).
Search optimum L2 penalty, λ based on Root Mean Squared Error (RMSE) of testing data.
Print out λ vs. RMSE of testing data.
Choose the optimum L2 penalty value λ.
(0.5pt) Print out RMSE for training/ test data based on the optimum, λ.
(0.5pt) Print out the optimized weight values based on the optimum, λ.

(1pt) Plot weight coefficients with the different λ as the below.

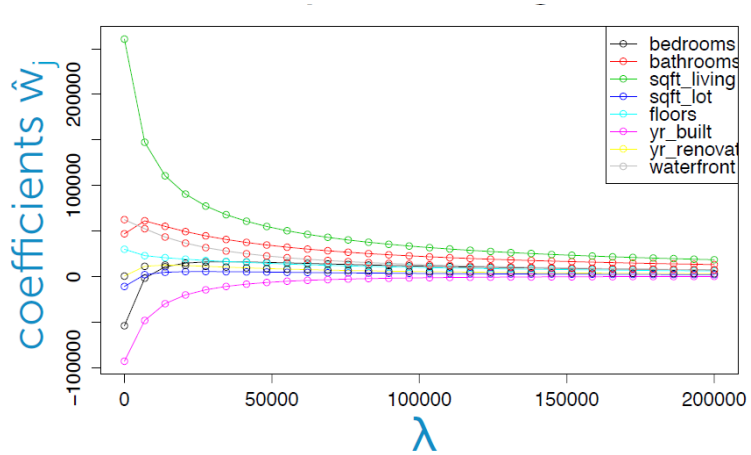


Fig. Plot W vs. λ. Here $W_j$ is [$W_0$, $W_1$, ..,$W_5$]

(1pt) Please fit the train/test data using the Linear regression model with and without the optimum L2 regularization *(Please plot 'price' vs 'lotsize')*.

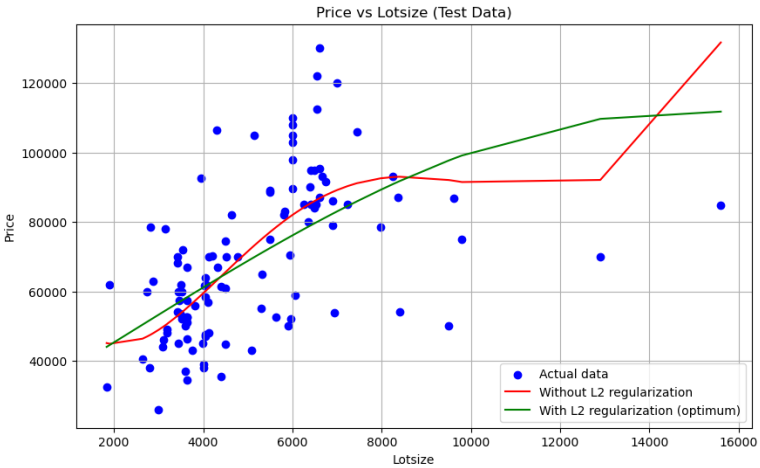


Fig. Plot of price(y, $\hat{y}$) vs. lotsize

**2-2 (3pts). Fit the training data using 5th order polynomial regression model and Lasso (L1 penalty) regularization.**

You need to try at least 5 different L1 penalty (for example, λ = 0.1, 0.5, 1, 5, 10, …, etc).
Search optimum L1 penalty, λ based on Root Mean Squared Error (RMSE) of testing data.
Print out λ vs. RMSE of testing data.
Choose the optimum L2 penalty value λ.
(0.5pt) Print out RMSE for training/ test data based on the optimum, λ.
(0.5pt) Print out the optimized weight values based on the optimum, λ.

(1pt) Plot weight coefficients with the different λ as the below.

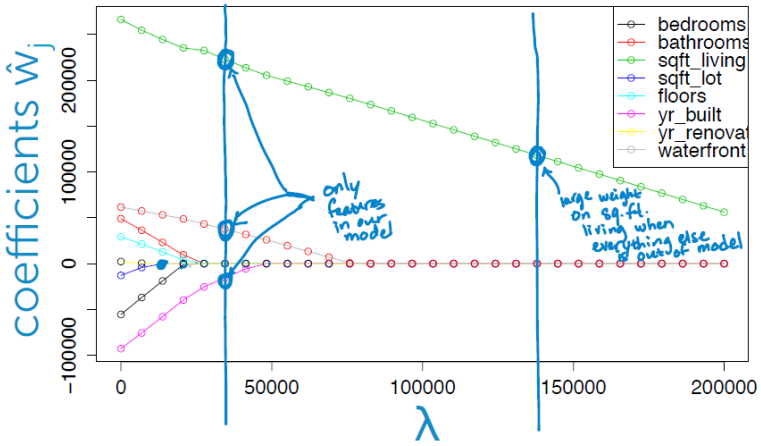


Fig. Plot W vs. λ. Here $W_j$ is [$W_0$, $W_1$, ..,$W_5$]

(1pt) Please fit the train/test data using the Linear regression model with and without the optimum L1 regularization *(Please plot 'price' vs 'lotsize' as the below example)*.

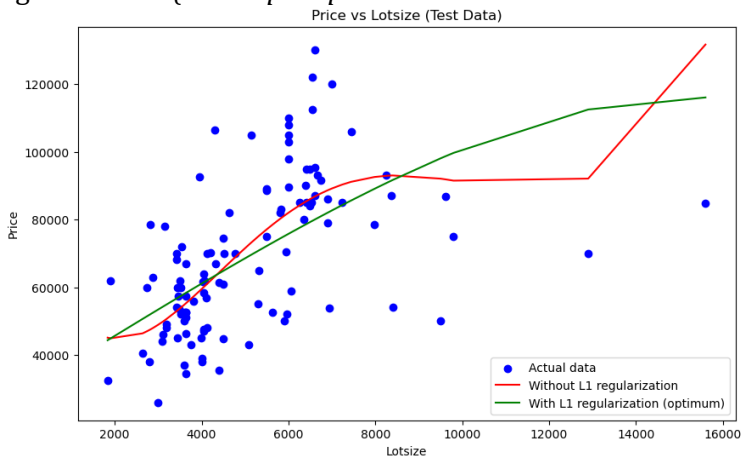


Fig. Plot of price(y, $\hat{y}$) vs. lotsize