ECS 171: Homework Set 1

Yilun Yang
912425744

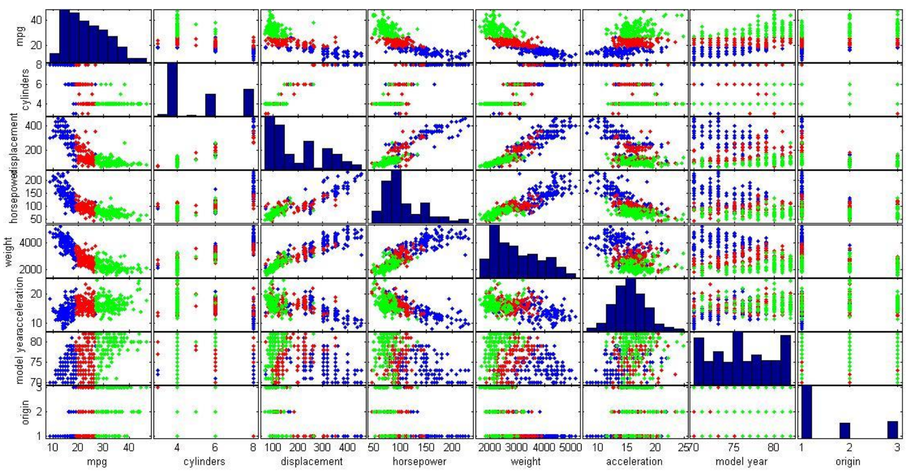
Instruction:
Problem1.m is for problem1, Problem2.m is for problem2, polyre.m is for problem3, Problem4.m is for problem4, polyremod.m and Problem5.m are for problem5 stoc_grad_desc_logi.m and Problem6.m are for problem6, Problem7.m is for problem7

1. *Assume that we want to classify the cars into 3 categories: low, medium and high mpg. Find what the threshold for each category should be, so that all samples are divided into three equally-sized bins. [10pt]*

   We have totally 398 observations in the dataset. Since this problem is only about mpg and it doesn't have missing value, we will not delete the 6 observations with missing value features. We first sort the mpg data, then identify the approximate Tertiles (which are the 133th and 265^{th} number) as 19 and 26.8.

2. *Create a 2D scatterplot matrix, similar to that of Figure 1.4 in the ML book (K. Murphy, page 6; also available on the lecture 1 slides - the figure with the flowers). You may use any published code to perform this. Which pair from all pair-wise feature combinations is the most informative regarding the three mpg categories? [10pt]*

   We applied the *gplotmatrix* function to draw the scatterplot matrix, the plot is as below:



   In this plot, blue represents low mpg, red is the median, green is high mpg. There are lots of plots very informative. If we look at the first row or column, we will find obvious linearity between mpg and cylinders, displacement, horsepower, weight, and model year. An older car (model year) with more cylinders, and higher displacement, horsepower, weight tends to have a low mpg (High fuel consumption), this is consistent with our common sense. This trend can also be found in other plots. For example, in the displacement vs weight plot. The upper right region (higher weight and displacement) has more blue points and the bottom left region has more green points even though those points are not totally separated. Besides, an obvious linearity between displacement and weight is also shown in this plot. Actually, linearity is very common among these features. Horsepower, weight, displacement have positive linear relationship while acceleration with displacement, horsepower have negative relationship. Even though so much information is given in the plots, I believe the displacement vs weight plot is most informative. Though there is still overlap on the three categories, they are separated the most compared with others.

3. *Write a linear regression solver that can accommodate polynomial basis functions on a single variable. Your code should use the Ordinary Least Squares (OLS) estimator which is also the Maximum-likelihood estimator for this problem (you will have to code it from scratch). [20p]*

I implement a function with various input and output. For input, the training data, polynomial order and optional test data can be given. For the output, coefficients are given by default. MSE and prediction are given only if the test data is also given as an argument. I also give unit test on some special situations where regression are not appropriate. (E.g. not enough sample, etc.)

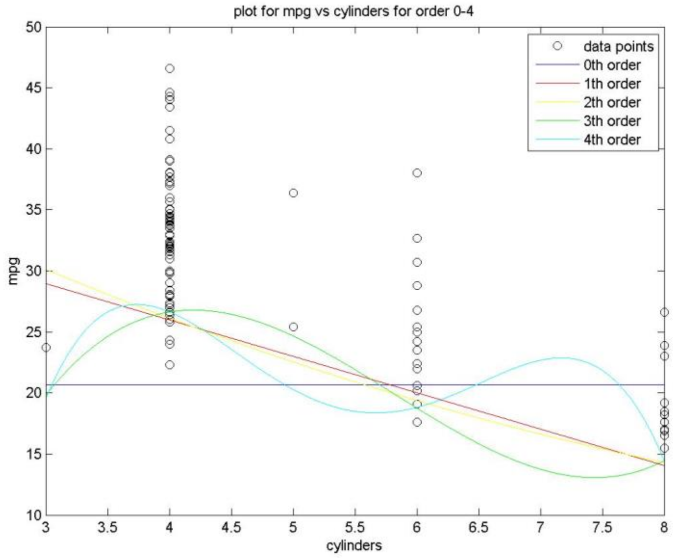
4. *Split the dataset in the first 280 samples for training and the rest 112 samples for testing. Use your solver to regress for 0th to 4th order polynomial on a single independent variable (feature) each time by using mpg as the dependent variable. Report (a) the training and (b) the testing mean squared errors for each variable individually (except the "car name" string variable, so a total of 7 features that are independent variables). Plot the lines and data for the testing set, one plot per variable (so 5 lines in each plot, 7 plots total). Which polynomial order performs the best in the test set? Which feature is the most informative regarding mpg consumption in that case? [20pt]*

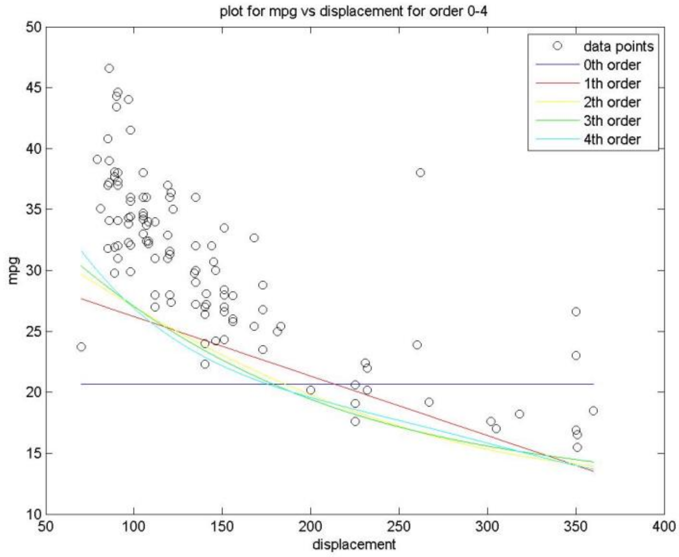
We should firstly notice that we should delete the 6 observations with missing value and use the remaining 392 observations to perform regression. For the first part to calculate MSE, I used a matrix to store mses(training and test) for different features and orders. The table is as below: (we use 1th tr, te to denote 1th order training and test mse)

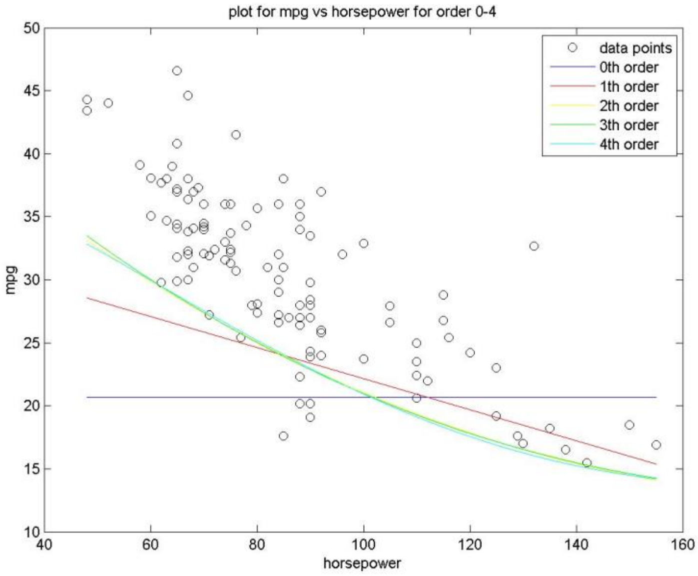
	cylinder	displace	horsepower	weight	accelerate	modelyear	origin
0^{th} tr	39.48	39.48	39.48	39.48	39.48	39.48	39.48
0^{th} te	141.03	141.03	141.03	141.03	141.03	141.03	141.03

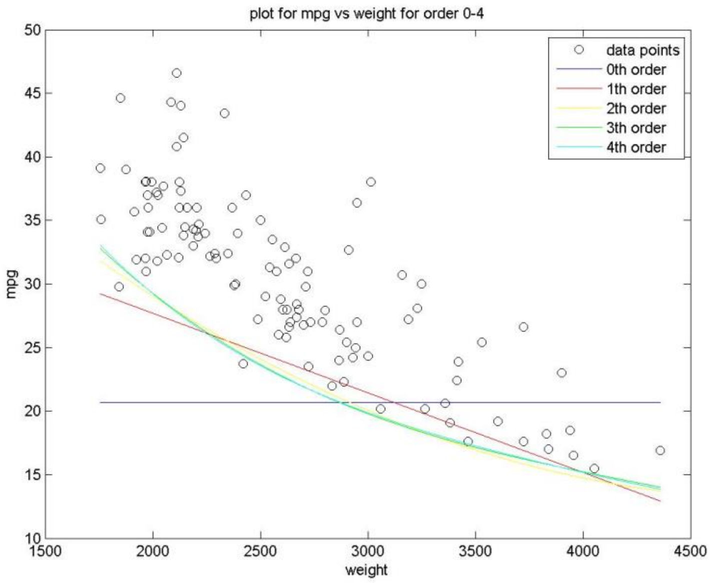
1st tr	12.47	10.93	14.18	8.43	30.50	36.16	24.91
1st te	68.19	64.11	66.28	61.23	118.76	71.81	101.81
2nd tr	12.33	9.00	10.58	6.65	29.71	35.95	23.93
2nd te	67.61	59.38	54.52	60.09	118.46	49.93	103.03
3rd tr	10.97	8.89	10.56	6.46	29.40	35.91	23.92
3rd te	62.90	59.72	54.36	61.59	123.75	71.17	103.03
4th tr	10.90	8.65	10.51	6.45	28.79	35.61	23.92
4th te	63.92	61.63	54.00	61.92	115.42	1230.30	103.03

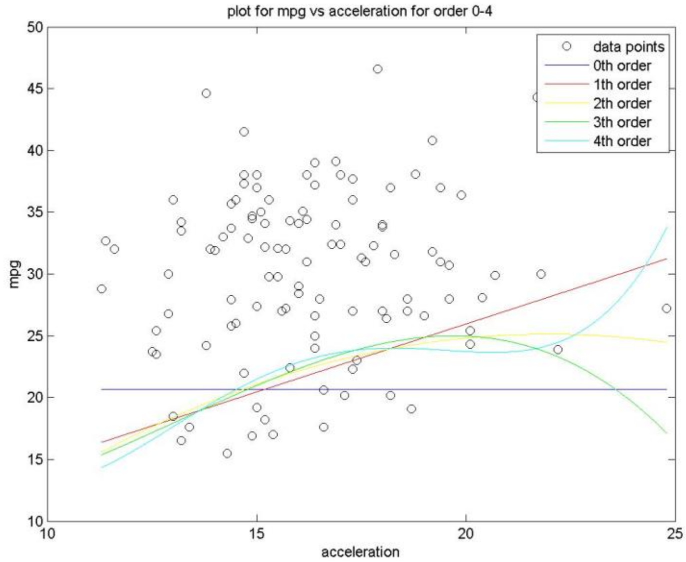
In the table, 4th order polynomial with horsepower has the smallest test MSE and 4th order polynomial with weight has smallest training MSE. Next, we will show the seven plots below:

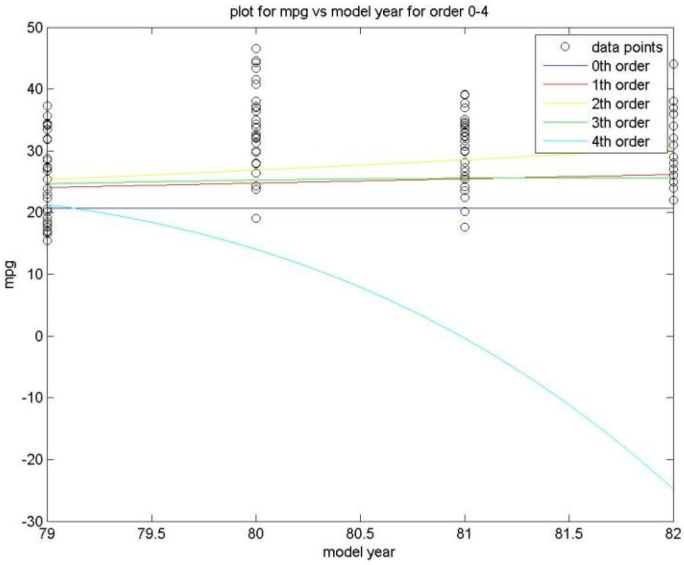


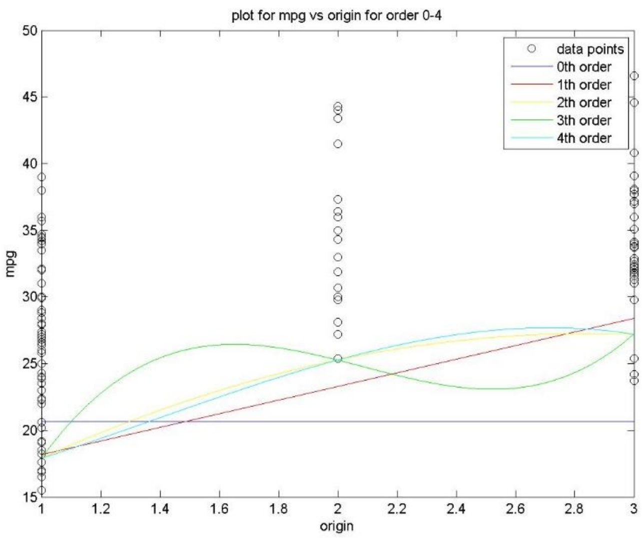



plot for mpg vs horsepower for order 0-4
data points
0th order
1th order
2th order
3th order
4th order
mpg
horsepower
50
45
40
35
30
25
20
15
10
40
60
80
100
120
140
160


plot for mpg vs weight for order 0-4
data points
0th order
1th order
2th order
3th order
4th order
mpg
weight
50
45
40
35
30
25
20
15
10
1500
2000
2500
3000
3500
4000
4500


plot for mpg vs acceleration for order 0-4
data points
0th order
1th order
2th order
3th order
4th order
mpg
acceleration
50
45
40
35
30
25
20
15
10
10
15
20
25


plot for mpg vs model year for order 0-4
data points
0th order
1th order
2th order
3th order
4th order
mpg
model year
50
40
30
20
10
0
-10
-20
-30
79
79.5
80
80.5
81
81.5
82


plot for mpg vs origin for order 0-4
data points
0th order
1th order
2th order
3th order
4th order
mpg
origin
50
45
40
35
30
25
20
15
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3

Based on the plots and table, we should conclude the 2nd order is the best. Obviously, test MSEs in 2nd order are smaller than 0th and 1st no matter in what feature thus 2nd it better than 0th and 1st order. Besides, although cylinder MSE would decrease a little in 3rd and 4th order compared with 2nd order, for other features, MSE would increase a lot more especially for model year. Thus 2nd order perform the best for us. In the second order, we should choose horsepower and model year as our most informative feature. Their test MSEs are both very low and they provide continuous and discrete data together thus are more informative.

5. *Modify your solver to be able to handle second order polynomials of all 8 independent variables simultaneously (i.e. 15 terms). Regress with 0th, 1st and 2nd order and report (a) the training and (b) the testing mean squared error. Use the same 280/112 split as before. [20pt]*

In this solver, I only made small modification to the one variable solver. The only problem here is how to expand original matrix into a bigger one with 2nd order terms interlaced. I used mod function to determine polynomial order for each column and round (i/2) + 1 to determine the correct column to get our data. Training MSE is 4.3237and test MSE is 17.5106.

6. *Modify your solver to allow for logistic regression (1st order) and report the training/testing mean squared error, as before. [10pt]*

Clearly, this is a multinomial logistic problem since we have low, median and high labels for mpg. From lecture notes, if we use π_1-π_3 to denote the probability of each mpg class, we would have the following properties:

$$\log(\pi_{i1}/\pi_{i3}) = \beta_{01} + \beta_{11}X_{i1} + \beta_{21}X_{i2} + \beta_{31}X_{i3} = \beta_1^T X_i$$
$$\log(\pi_{i2}/\pi_{i3}) = \beta_{02} + \beta_{12}X_{i1} + \beta_{22}X_{i2} + \beta_{32}X_{i3} = \beta_2^T X_i$$

We can get π_1-π_3 as

$$\pi_{i1} = \frac{\exp(\beta_1^T X_i)}{1 + \exp(\beta_1 X_i) + \exp(\beta_2^T X_i)},$$

$$\pi_{i2} = \frac{\exp(\beta_2^T X_i)}{1 + \exp(\beta_1 X_i) + \exp(\beta_2^T X_i)},$$

$$\pi_{i3} = \frac{1}{1 + \exp(\beta_1 X_i) + \exp(\beta_2^T X_i)}.$$

Also the maximum likelihood function for multinomial logistic regression is

$$L = \prod_{i=1}^{n} \left[\frac{\exp(Y_{i1}\beta_1^T X_i + Y_{i2}\beta_2^T X_i)}{1 + \exp(\beta_1^T X_i) + \exp(\beta_2^T X_i)} \right] = \frac{\exp\left(\sum Y_{i1}\beta_1^T X_i + \sum Y_{i2}\beta_2^T X_i\right)}{\prod_{i=1}^{n} \left[1 + \exp(\beta_1^T X_i) + \exp(\beta_2^T X_i)\right]}$$

Which we will use later to implement our stochastic gradient ascent.

In my function, I update the low mpg and median mpg coefficients concurrently and get the final beta. I also give the argument room for test data in order to get MSE. Note in this problem, class labels are what we get finally, thus the misclassification rate seems to be a good measurement of MSE. The training and test MSE are 0.1393 and 0.3125. In other word, we may predict 70% new observations correctly, it's a pretty good result in my opinion.

There are many things I want to point out in this problem. Firstly, notice we have standardized our training data in order to let stochastic gradient descent converge. Rescaling is necessary or the gradient would be infinite and never converge (It will give you NaN for all beta). We also transform from standard beta back to our original data for the convenience of future prediction. Secondly, I chose alpha as 0.1, initial beta as all 0, iteration time as 5000, this can be improved instead of determining these values arbitrarily. For alpha, we can use line search to determine. For iteration time, we may draw plot or use small epsilon to determine how many time to iterate thus to converge. Finally, I separated functions (true class, predclass) into sub functions to avoid repeat coding.

7. *If a USA manufacturer (origin 1) had considered to introduce a model in 1980 with the following characteristics: 6 cylinders, 300 cc displacement, 170 horsepower, 3600 lb weight, 9 m/sec2 acceleration, what is the MPG rating that we should have expected? In which mpg category (low,medium,high mpg) would it belong? Use second-order, multi-variate polynomial and logistic regression. [10pt]*

   We applied all three methods on the test data. For second order regression, we use horsepower and model year which were picked out as most informative features to predict. Predicted mpg are 13.285 and 26.8617 and should be categorized into low and median mpg group. For multi-variate polynomial regression, the prediction result is 20.1756 and should be categorized into median mpg. For logistic regression, probability of each class is 0.8538, 0.1462 and 1.1343e-6 thus the new observation should be classified into low mpg class.