ECS 171 HW1 - Solution

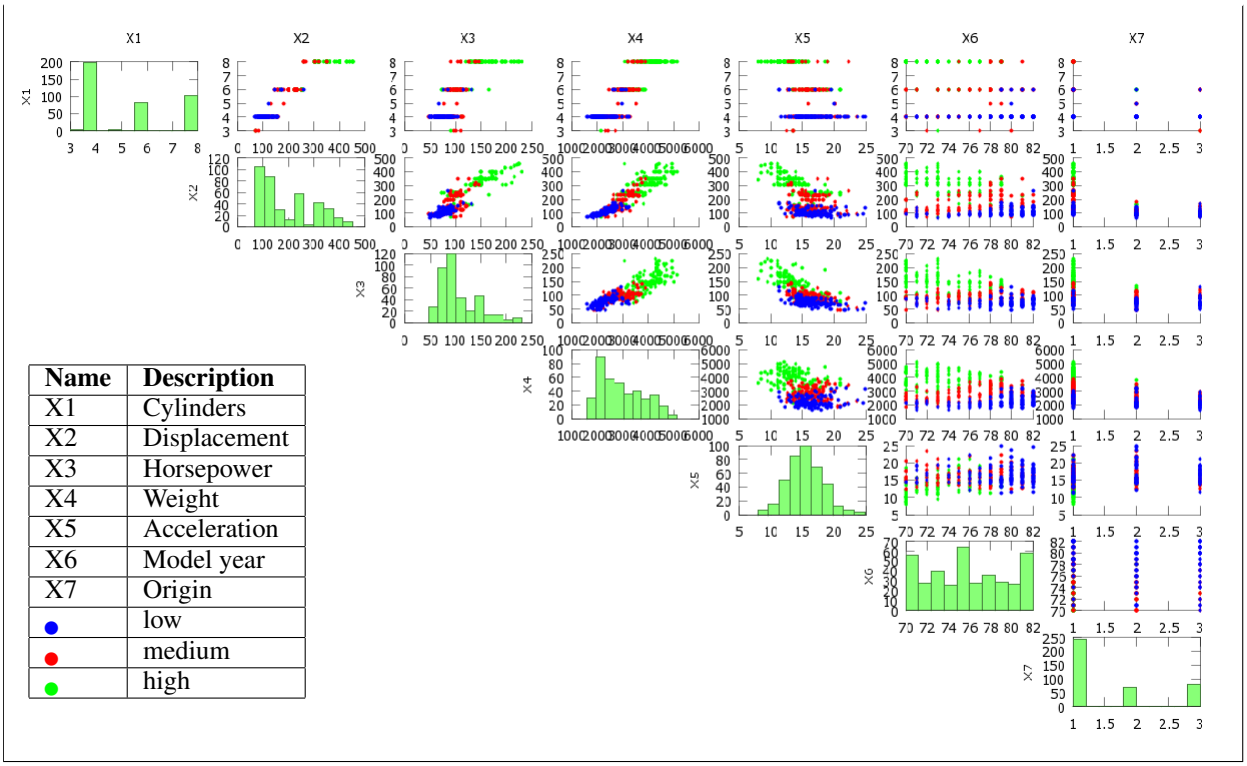
1 Find threshold for creating equally sized bins

There are 392 data points, hence 392 mpg values. Sort all the mpgs in an ascending order. Use the 131^{st} mpg value as the low threshold. Use the 261^{th} value as the medium threshold. Any value above the medium threshold would be considered high mpg. This divides the data into bins of size 130, 130 and 132.

- low threshold: 18.5
- medium threshold: 26.6

2 Scatterplot matrix visualization

Visual inspection of figure 2 indicates that "weight" and "horsepower" pair construct most informative pairs with regards to mpg. Having said that, there are other pairs which seem equally informative and hence acceptable answers.



Name	Description
X1	Cylinders
X2	Displacement
X3	Horsepower
X4	Weight
X5	Acceleration
X6	Model year
X7	Origin
●	low
●	medium
●	high

Figure 1: The upper right portion of the scatter-plot matrix

3 Linear regression solver

This would be solved using the OLS formula as in equation (3), where the X matrix is constructed using all the data loaded from the file except for the "mpg" column (and the last column which contains string only). Note that, a column of "1" should be added in order to account for the bias. The vector y consists of the corresponding "mpg" values.

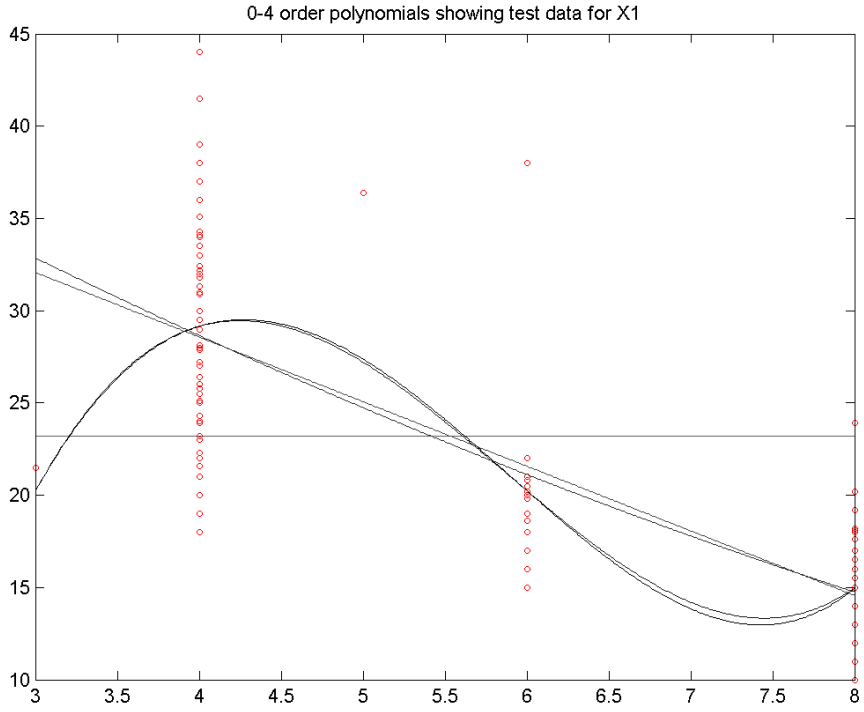
$$w = (X^T X)^{-1} X^T y \tag{1}$$

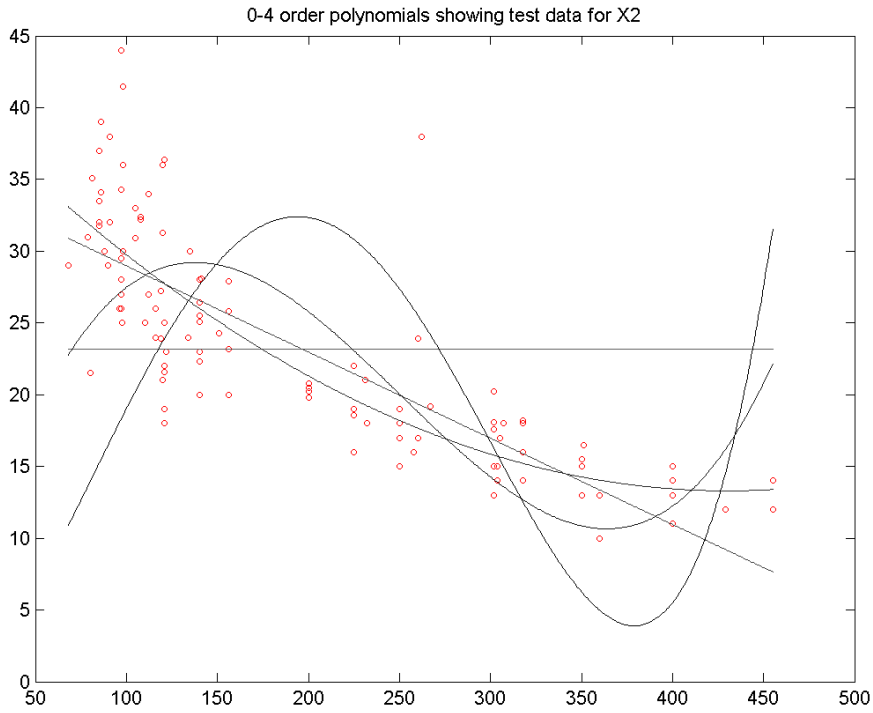
This would be solved using matrix operations support provided in matlab.

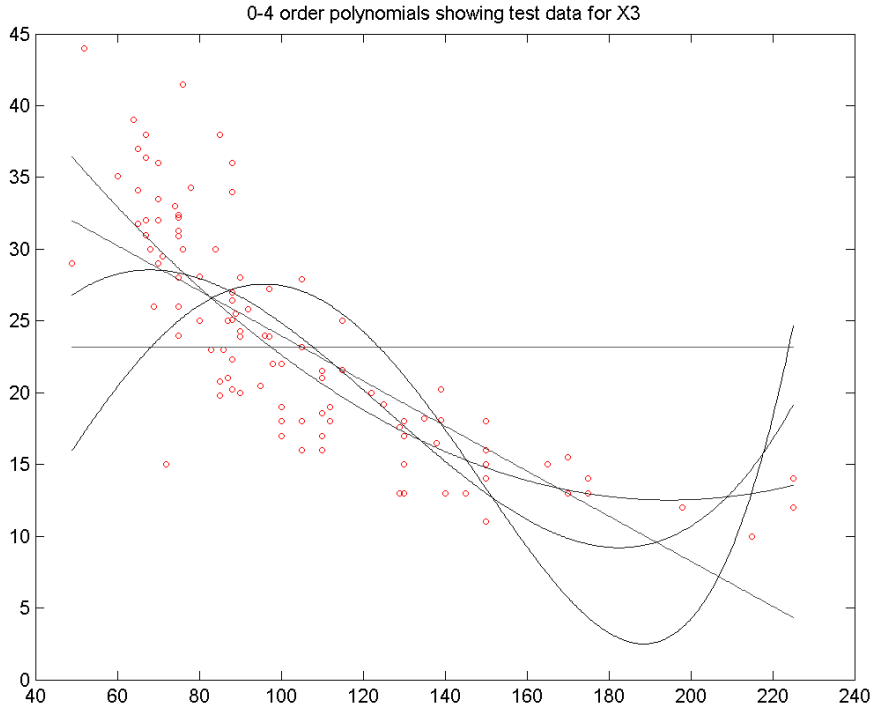
4 Single-variate polynomial regression

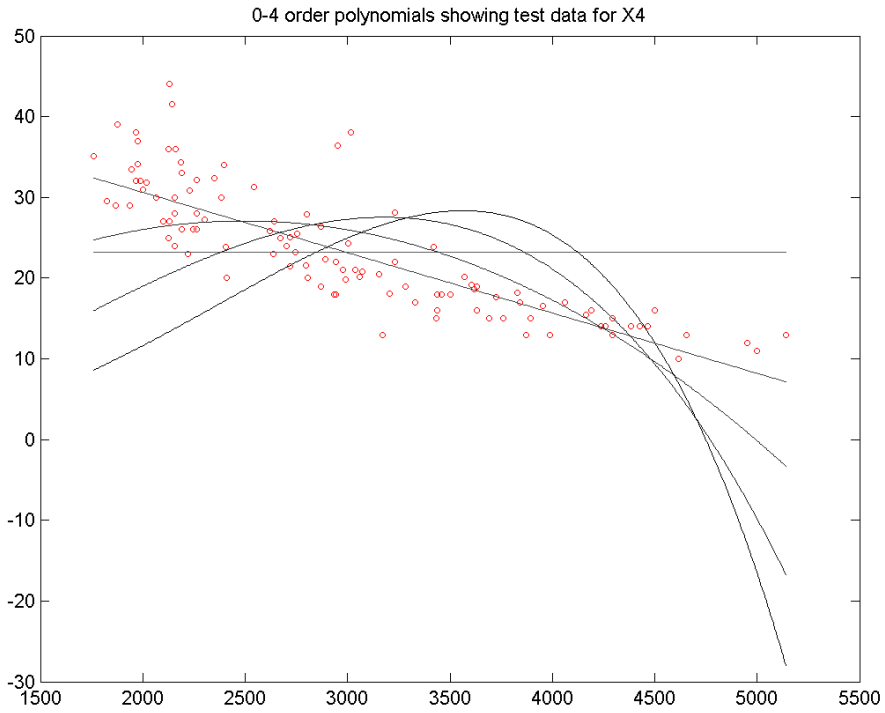
Table 4 shows the errors up to order 4. "tr" stands for training and "ts" stands for testing error.The second order polynomial performs the best for the test set. The most informative feature is displacement (X2).

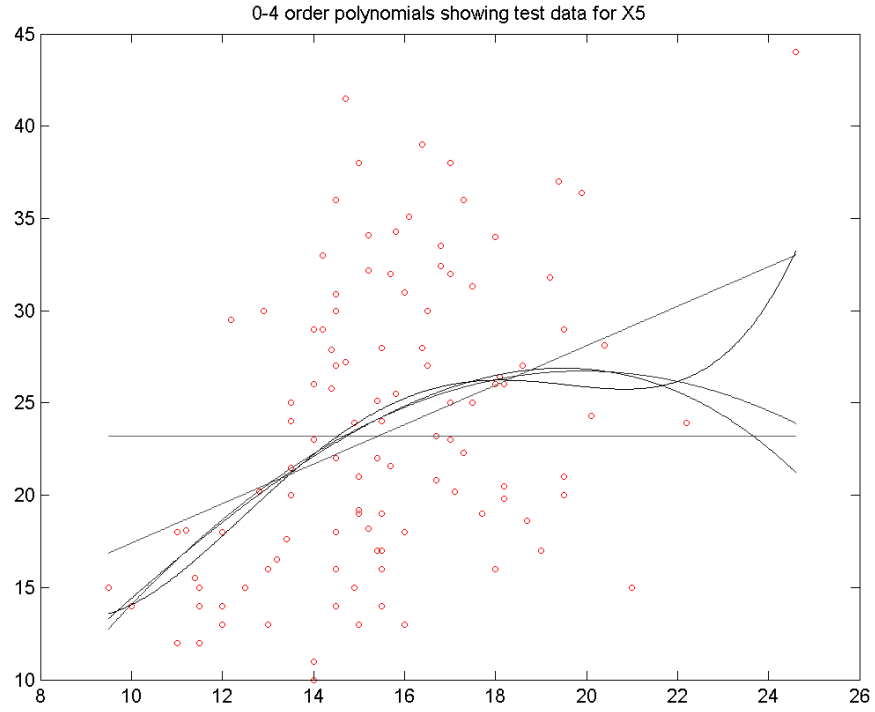
P	**0**		**1**		**2**		**3**		**4**	
F	ts	tr	ts	tr	ts	tr	ts	tr	ts	tr
1	28.8	29.9	10.8	12.4	10.8	12.3	9.7	11.5	9.7	11.5
2	28.8	29.9	9.3	10.4	8.3	9.4	14.8	16.9	53.6	54.8
3	28.8	29.9	11.5	10.9	9.4	8.2	13.1	12.3	28.8	28.8
4	28.8	29.9	8.7	8.7	17.4	18.7	43.4	48.9	86.7	96.9
5	28.8	29.9	24.6	24.5	23.8	24.4	23.7	24.8	23.5	22.7
6	28.8	29.9	19.9	19.1	19.2	17.4	19.2	17.4	19.2	17.4
7	28.8	29.9	20.2	20.4	19.6	19.6	19.6	19.6	19.6	19.6

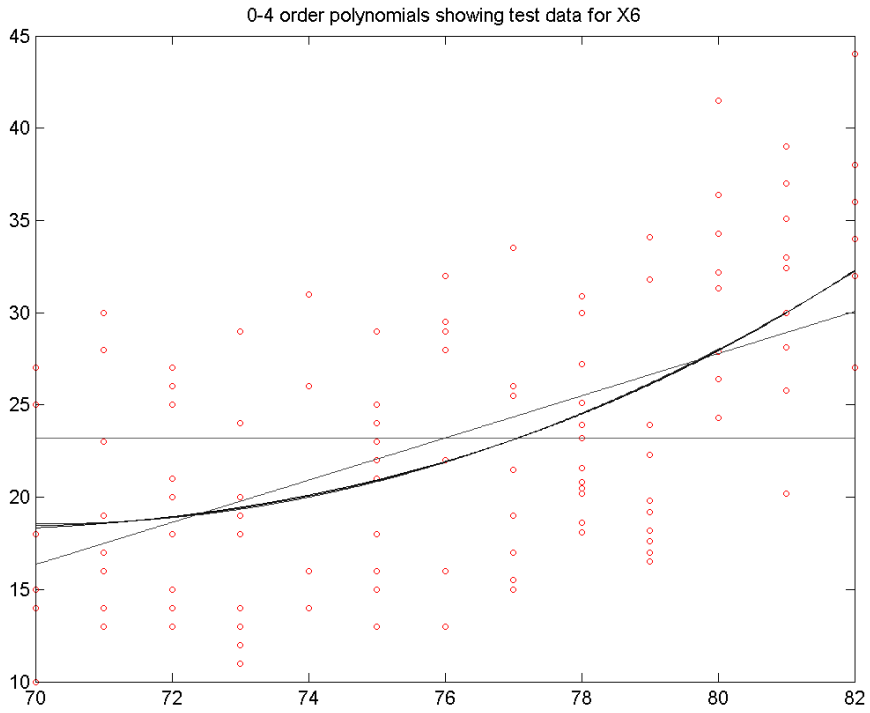


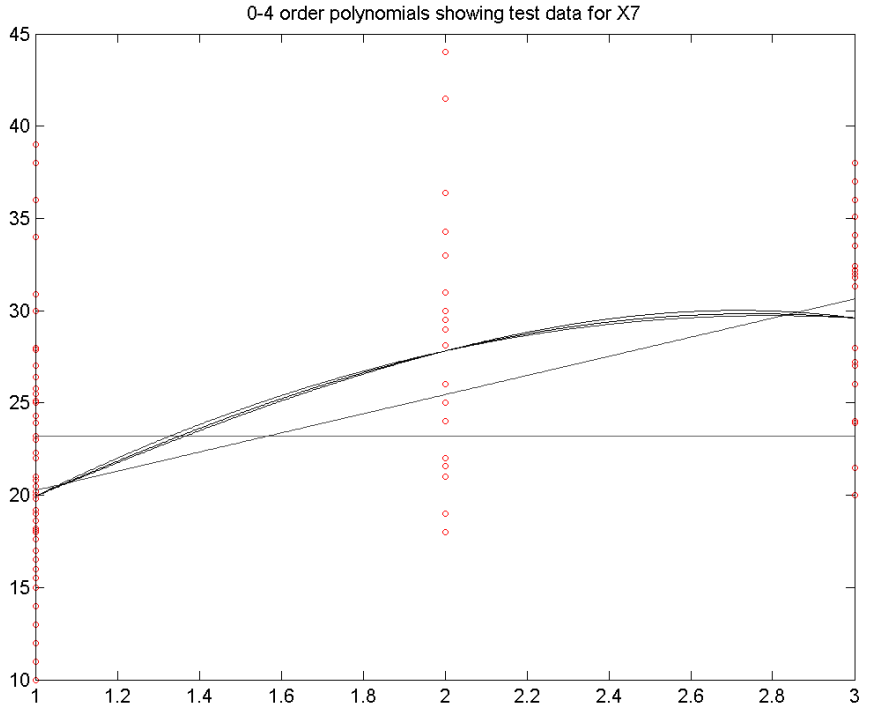

0-4 order polynomials showing test data for X2
45
40
35
30
25
20
15
10
5
0
50
100
150
200
250
300
350
400
450
500


0-4 order polynomials showing test data for X3
45
40
35
30
25
20
15
10
5
0
40
60
80
100
120
140
160
180
200
220
240


0-4 order polynomials showing test data for X4
50
40
30
20
10
0
-10
-20
-30
1500
2000
2500
3000
3500
4000
4500
5000
5500


0-4 order polynomials showing test data for X5
45
40
35
30
25
20
15
10
8
10
12
14
16
18
20
22
24
26


0-4 order polynomials showing test data for X6
45
40
35
30
25
20
15
10
70
72
74
76
78
80
82


0-4 order polynomials showing test data for X7
45
40
35
30
25
20
15
10
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3

5 multi-variate polynomial regression

The table 5 contains MSE for test and training set.

Order	Train	Test
0	28.77	29.90
1	5.23	5.35
2	3.94	4.29

Table 1: MSE for up to second order multi-variate polynomial

6 Logistic Regression

First, will normalize the input features (centralize and rescale). Then, will create three different binary classifiers, one for "low", one for "medium" and one for "high" mpg based on the labels found in question 1. For each classifier, will use the same formula as in the lecture to create the objective function. Then gradient descent is used to find the optimal set of weights.
The classification accuracy would be:
Training: 80.7%
Testing: 83.0%

Few points:

1. In order to make sure test and training sets are balanced, stratified folding was used.
2. Three binary classifiers are constructed a) low vs. medium or high, b) medium vs. low or high and c) high vs. low or medium
3. Any new record is given to all three binary classifiers and the final prediction is attributed to one with the highest score.

7 Prediction

Logistic regression prediction: low
Second-Order multivariate polynomial regression: 18.93

So the results of logistic regression and second-order polynomial regression disagree. Three binary classifiers were used. There was a tie between classifiers for medium and high classes. This confirms the result of the polynomial regression as the predicted value "18.93" is very close to the threshold between low and medium.

8 The donkey engine

The donkey consumes "Zero" gas per mile because it has zero cylinders. It's the most efficient. Humans have a long way to go, to create a car as efficient (and intelligent) as this.