Homework 1
ECS171

Atharva Chalke

October 12, 2018

1 Problem 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]

Answer.

Code: Part1+Part2.py
1. Anything below 21.5 falls in the low MPG rating.
2. Anything above 21.5 and below 33.5 falls in the mid MPG rating.
3. Anything above 33.5 falls in the high MPG rating.

2 Problem 2

Create a 2D scatter plot 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]

Answer.

Code: Part1+Part2.py
PLOTS: Plots2.pdf
The Weight-Displacement pair is the most informative.

3 Problem 3

Write a linear regression solver that can accommodate polynomial basis functions on a single variable for prediction of MPG. Your code should use the Ordinary Least Squares(OLS) estimator (i.e. the Maximum-likelihood estimator). Code this from scratch. Its recommended

to use a library (e.g. numpy) for basic linear algebra operations (addition,multiplication and inverse). [20p]

Answer.

Code :Part3.py

4 Problem 4

Split the dataset in the first 200 samples for training and the rest 192 samples for testing.Use your solver to regress for 0th to 3rd 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 4 lines in each plot, 7 plots total). Which polynomial order performs the best in the test set? Which is the most informative feature for mpg consumption in that case? [20pt]

Answer.

Code: Part4.py
Plots: Plots4.pdf

Note: The feature-output matrix is shuffled before extracting training-testing matrix's; therefore, running the code will change the error using MSE.

Training error on cylinders of degree 0 is 65.47457775
Testing error on cylinders of degree 0 is 56.30669360416664
Training error on cylinders of degree 1 is 25.077021879630582
Testing error on cylinders of degree 1 is 23.213773191558147
Training error on cylinders of degree 2 is 24.80550047999671
Testing error on cylinders of degree 2 is 23.163363729561762
Training error on displacement of degree 0 is 65.47457775
Testing error on displacement of degree 0 is 56.30669360416664
Training error on displacement of degree 1 is 22.294285357227846
Testing error on displacement of degree 1 is 20.74408302544121
Training error on displacement of degree 2 is 19.576601393878914
Testing error on displacement of degree 2 is 18.448556507309423
Training error on horsepower of degree 0 is 65.47457775
Testing error on horsepower of degree 0 is 56.30669360416664
Training error on horsepower of degree 1 is 24.95227346633671
Testing error on horsepower of degree 1 is 23.298468528941523
Training error on horsepower of degree 2 is 18.306753199229924
Testing error on horsepower of degree 2 is 20.23062983966611
Training error on weight of degree 0 is 65.47457775

Testing error on weight of degree 0 is 56.30669360416664
Training error on weight of degree 1 is 20.478396826809846
Testing error on weight of degree 1 is 16.856380940939886
Training error on weight of degree 2 is 41.06328476440424
Testing error on weight of degree 2 is 32.05292653798181
Training error on acceleration of degree 0 is 65.47457775
Testing error on acceleration of degree 0 is 56.30669360416664
Training error on acceleration of degree 1 is 52.81994598609364
Testing error on acceleration of degree 1 is 47.85742373705036
Training error on acceleration of degree 2 is 52.151873330664266
Testing error on acceleration of degree 2 is 46.391279412616086
Training error on model year of degree 0 is 65.47457775
Testing error on model year of degree 0 is 56.30669360416664
Training error on model year of degree 1 is 44.88238386042576
Testing error on model year of degree 1 is 36.05740175539402
Training error on model year of degree 2 is 43.01796197354468
Testing error on model year of degree 2 is 33.99250954916837
Training error on origin of degree 0 is 65.47457775
Testing error on origin of degree 0 is 56.30669360416664
Training error on origin of degree 1 is 42.47098791047973
Testing error on origin of degree 1 is 40.29400748687821
Training error on origin of degree 2 is 41.10284609780083
Testing error on origin of degree 2 is 40.44397969053755

Best Polynomial Degree for test-case: 1
Most informative feature for test-case: Weight

5 Problem 5

Modify your solver to be able to handle second order polynomials of all 7 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 (MSE). Use the same 200/192 split.[20pt]

Answer.

Code:Part5.py
Note: Data is shuffled again, so running the code might give you slightly different answers

Training error for degree 0 is 64.3502
Testing error for degree 0 is 58.345833333333395
Training error for degree 1 is 11.364934062955408
Testing error for degree 1 is 11.204873470842676
Training error for degree 2 is 8.376365333249028

Testing error for degree 2 is 9.095945246927668

6 Problem 6

Using logistic regression (1st order) for low/medium/high classification. Report the training/testing classification precision (you might want to look how precision is defined and ow it is calculated). You can use a library (e.g. scikit-learn) to perform logistic regression.[10pt]

Answer.

Code: Part6.py
Training precision for class 0 is 0.9444444444444444
Training precision for class 1 is 0.8705882352941177
Training precision for class 2 is 0.8
Testing precision for class 0 is 0.9090909090909091
Testing precision for class 1 is 0.797752808988764
Testing precision for class 2 is 0.6

7 Problem 7

If a USA manufacturer (origin 1) had considered to introduce a model in 1980 with the following characteristics: 6 cylinders, 350 cc displacement, 180 horsepower, 3700 lb weight,9m/sec2acceleration 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]

Answer.

Code: Part7and8.py,Part7support.py, Part7Supportlogistic.py

Using Weights from training second order term we get the value to be 23.55198552.
Expected Value = Medium Category (1).
Using logisitc regression, we get the class for this feature to be [1.]

8 Problem 8

Predict the mpg of the vehicle in the photo. Clearly state your assumptions. [3pt bonus]

Answer.

Code: Part7and8.py
Assumptions
Cylinder = 1, displacement = 80, horsepower = 1 , weight = 1800, acceleration = 3, model year = 4, origin = 1
Using Weights from training second order term we get the value of the horse cart to be 12.05258831.