UNIVERSITY OF CALIFORNIA, DAVIS
DEPARTMENT OF COMPUTER SCIENCE

ECS 171: Homework Set 1

Instructor: Ilias Tagkopoulos
TAs: Minseung Kim and Ameen Eetemadi
{msgkim, eetemadi}@ucdavis.edu

September 25, 2015

General Instructions: The homework should be submitted electronically through Smartsite. Each submission should be a zip file that includes the following: (a) a report in pdf format ("report_HW1.pdf") that includes your answers to all questions, plots, figures and any instructions to run your code, (b) the matlab/octave code files. Please note: (a) do not include any other files, for instance files that we have provided such as datasets, (b) each function should be written in a seperate file, with the appropriate remarks in the code so it is generally understandable (what it does, how it does it), (c) do not use any toolbox unless is it explicitly allowed in the homework description. Shared/copied code from any source is not allowed, as it is considered plagiarism.

1 OF CARS AND MEN [100PT]

In this exercise, you will investigate the type of relationship that exists between the “miles per gallon” (mpg) rating of a car and several of its attributes. For this task, you will use the “Auto MPG” dataset (“auto-mpg.data” file; 398 cars, 9 features; remove the 6 records with missing values to end up with 392 samples) that is available in the UCI Machine Learning Repository:

https://archive.ics.uci.edu/ml/datasets/Auto+MPG

Perform and report (code and results) the following:

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]

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]

3. Write a linear regression solver that can accomodate 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]

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]

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]

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

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/sec^2 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]

8. Predict the mpg of the vehicle on the photo. Clearly state your assumptions and how your reached to that result. [3pt bonus]

GOOD LUCK!

THAN