UNIVERSITY OF CALIFORNIA, DAVIS
DEPARTMENT OF COMPUTER SCIENCE

ECS 171: Homework Set 1

Instructor: Ilias Tagkopoulos
TAs: Jason Youn, Ameen Eetemadi, and ChengEn Tan
{jyoun, eetemadi, cetan}@ucdavis.edu

September 30, 2019

General Instructions: The homework should be submitted electronically through Canvas. 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 python 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 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 it is 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

For this assignment, you will need to code your solution from scratch. Unless explicitly stated, it is fine to use open source code, for example sci-kit learn, to help you write your own implementation of the methods. Perform and report (code and results) the following:

1. Assume that we want to classify the cars into 4 categories: low, medium, high, and very high mpg. Find what the threshold for each category should be, so that all samples are divided into four 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 four mpg categories? [10pt]

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]

4. Split the dataset in the first 292 samples for training and the rest 100 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]

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 292/100 split. [15pt]

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

7. re-do the logistic regression training/testing, but now after you apply min-max normalization to the dataset. Do you see any difference in performance? [5pt]

8. If a USA manufacturer (origin 1) had considered to introduce a model in 1981 with the following characteristics: 4 cylinders, 400 cc displacement, 150 horsepower, 3500 lb weight, 8 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]