CSCE 633: Fall 2019 Homework 1

Assigned Mon, Sep 9, due on Mon, Sep 23, by 11:59 PM.

Submit PDF and zip file (of code + latex) on eCampus. You can also present your code and its explanation using Jupyter notebook similar to discussions. No late submissions accepted.

Name (please print):

Problem 1: Linear Regression. NOTE: In parts (1) and (2) there is no α_d, it is considered to be 1 to simplify the computations. You can use programming tools to help solve this problem. Please provide as much written detail of your solution as possible.

In this problem we learn that the approach taken in linear regression can be used when there are variables' powers of more than one in the function formula which is called *polynomial regression*. Consider the polynomial regression problem with four training data points $\{(0,1),(2,4),(3,9),(5,16)\}$ and two test data points $\{(1,3),(4,12)\}$ in 2-D space. The d-degree polynomial regression problem is to find a d-degree polynomial

$$\hat{f}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{d-1} x^{d-1} + x^d$$

that fit the data with optimal RSS (residual sum of squares).

(1) Solve the d-degree polynomial regression problems with $d = 0, 1, 2, 3, 4$. Show the data points and all your regressed curves in one plot. Make sure to represent the training points and test points in different color or markers.

(2) For each d, calculate the typical (squared) bias, variance, total error, training error and test error. Then, draw a plot of the above five curves as we go from simpler model to more complex model. The x-axis should be the order d of the polynomials, from 1 to 4, and the y-axis should be the values for each curve. Make sure to label each curve.

(3) Explain why each of the five curves has the shape displayed in part (2).

Hint: Polynomial regresion is no harder than linear regression. For example when $d = 2$, you can express the problem in the following form:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix}, \quad \omega = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ 1 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}.$$

The goal is to minimize $\|\mathbf{X}\omega - \mathbf{Y}\|_2^2$ with respect to ω.

(4) In this part we want to implement linear regression. Consider the *Smarket* dataset used in discussion week 1 available to download from here. In this assignment, we model *Today* based on *Lag1* and *Lag2*. We want to make a comparison between the RSS of the models trained using L^1-norm and L^2-norm in a 5-fold cross validation scheme. For each of the norms, please perform a 5-fold cross validation where in each fold the model is trained with a specific norm but the output of the trained model is assessed using L^2-norm. Please compare the mean values across all the 5 folds together. Justify your results.

Problem 2: Numerical Solution to Logistic Regression. Consider the logistic regression problem with two training points$\{(-3,1),(-1,0)\}$ and two test data points with features $\{-4,5\}$. We want to train our logistic regression model of the form:

$$p(y|\mathbf{x},\boldsymbol{\beta}) = \text{Ber}(y|\mu(\boldsymbol{\beta}^T\mathbf{x}))$$

Where here we have:

$$\boldsymbol{\beta} = (1,\beta_0)^T$$

and

$$\mathbf{x} = (x,1)^T$$

therefore, for each point we have:

$$\boldsymbol{\beta}^T\mathbf{x} = x + \beta_0$$

Note that we have assumed the slope to be 1.

(1) Please form the cross-entropy error similar to the slides and find the optimum β_0. For finding the optimum value, you can either use the derivative or you can plot the function using websites such as fooplot.

(2) What would be the model output for test data points?

Problem 3: Models for Heart Disease. NOTE: This is a programming assignment but please develop appropriate tables and figures for your write up. For this problem, please download the associated *hw1_input.csv* file. Heart disease is a major burden on the health care system and a leading cause of death in older adults. It is important to predict, from clinical data gathered on patients, who will have heart disease and who will not. In this problem, we will explore a dataset of such clinical data.

(1) **Data Exploration**: Please inspect the input features and provide scatter plots and histograms for the data, and explain the dataset (including difference between categorical and continuous features).

(2) **Logistic Regression Regularization Comparison with Bootstrapping**: Using 80% of the data as a training set and 20% as a testing set in each bootstrap repeated 1000 times each, please implement and compare the average and the standard deviation of coefficients obtained from Ridge regression and LASSO regularized logistic regression. By averaging the coefficients it is meant that all the different coefficient values for a specific variable is averaged. Coefficient average values then are to be compared by plotting them against each other. Are all input features necessary? Please describe how categorical features are handled.

(3) Please plot the ROC curve for both models for a single bootstrap data. What are the area under the curve measurements?

(4) What is the optimal decision threshold to maximize the f1 score?

(5) Please provide a mean and standard deviation for the AUROCs for each model .

(6) Please provide a mean and standard deviation for the f1 score for each model.