ECE 6604: Machine Learning for Communication Systems – Fall 2019

Project 1: Linear and Logistic Regression
Due: Friday, October 18, 2019, 5:00pm

1. **Linear Regression.** The objective of this question is to help you visualize the working of linear regression. Let us first consider a simple linear regression problem with one input and one output. You are given three different training data sets: **P1-data1.txt**, **P1-data1a.txt** and **P1-data2.txt**. In each of these files, the first column represents the input variable while the second column represents the output variable. Note that you can import the training data sets from the txt files into MATLAB using the **load** command. For this setup, answer the following questions.

   (a) Visualize the training data set in file **P1-data1.txt** by plotting the output variable against the input variable.

   (b) Consider the linear regression model: $f(x) = \beta_0 + \beta_1 x$, where x is the input variable. Here, $f(x)$ is also called the hypothesis function. Assuming squared loss, plot the loss function for the following range of the learning parameters: $-10 \leq \beta_0 \leq 10$ and $-2 \leq \beta_1 \leq 4$. From the plot that you get, can you infer if the loss function has a global minimum? Justify your answer for full credit.

   (c) Fit the learning parameters β_0 and β_1 to the training data set in **P1-data1.txt** using the gradient descent algorithm. For the gradient decent algorithm, consider the learning rate to be 10^{-2} and the number of iterations to be 1500. Initialize both β_0 and β_1 to 0. Using the values of β_0 and β_1 obtained from the above procedure, plot the hypothesis function along with the training data set.

   (d) Repeat (c) for the learning rate values of 10^{-6} and 3×10^{-2}. Keep the same number of iterations.

   (e) From the plots that you obtained in (c) and (d), which hypothesis function would you use to fit the training data set? State clearly why the other hypothesis functions were not chosen.

   (f) Compare the values of the learning parameters obtained by gradient descent algorithm in (c) with the ones obtained directly from the normal equation discussed in the class.

   (g) Now assume that the training dataset has changed to **P1-data1a.txt**. Would you still choose a squared loss function? If not, which loss function is more appropriate for this? For this new loss function, train the linear regression model using the gradient descent method. Repeat the same for the squared loss function (using data set in **P1-data1a.txt**). Plot the two hypothesis functions in the same figure and explain your observations. Clearly state any assumptions that you make in this part.

   (h) Repeat (a) for the training data set given in **P1-data2.txt**.

   (i) Fit the learning parameters (β_0, β_1 and β_2) to the training data set contained in the file **P1-data2.txt** using the following hypothesis function:

   $$h(x) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

   where x is the input variable. For the gradient decent algorithm, initialize all the learning parameters to 0, and use proper values for both the learning rate and the total number of iterations based on your observations from parts (c)-(e). Plot the hypothesis function along with the training data set. Clearly state your choices for the learning rate and the number of iterations.

(j) Apply the k-NN algorithm to the datasets in **P1-data1.txt** and **P1-data2.txt** for different values of k ($1 \leq k \leq 5$). For each training data set, compare the plots obtained using k-NN algorithm with the hypothesis functions obtained using linear regression (i.e., the plot in (c) for **P1-data1.txt**, and the plot in (i) for **P1-data2.txt**).

2. **Logistic Regression.** In this problem, you will learn how to implement logistic regression. You are given: i) two different training data sets: **P2-data1.txt** and **P2-data2.txt**, and ii) two different validation data sets: **P2-valdata1.txt** and **P2-valdata2.txt**. As discussed in the class, the training datasets should be used to train your selected hypothesis classes while the validation datasets should be used to select the most appropriate hypothesis class. For each data set, there are two input variables, x_1 and x_2, and one output variable, y. In each file, the two inputs are represented by the first two columns, and the output variable is represented by the third column. Assume that the output variable is binary, i.e., $y \in \{0, 1\}$. For this setup, answer the following questions. *In parts (a) and (c) below, please limit your choices to the polynomial hypothesis classes of degree* 6 *or less.*

   (a) Visualize the training data set in file **P2-data1.txt** by plotting the labeled data points as a function of the two features. Based on this, select the candidate hypothesis classes (equivalently, models). Train these models using **P2-data1.txt**. While you used the gradient decent algorithm to obtain the optimal linear regression parameters in Problem 1, you are going to use the **fminunc** built-in function in MATLAB to train your models in this problem. The function **fminunc** is an optimization solver that finds the minimum of an unconstrained optimization problem (which is the case for regression problems). For the function **fminunc**, please use the maximum number of iterations as 400, and initialize all the learning parameters to 0. Plot the decision boundary along with the training data set.

   (b) Now use the validation dataset **P2-valdata1.txt** to select the most appropriate model from amongst the ones that you trained in part (a). You need to make sure that the validation error is at most 10% for your selected model. In other words, the percentage of prediction errors should not exceed 10% when your chosen model is applied to the validation dataset **P2-valdata1.txt**.

   (c) Repeat part (a) for the training data set in **P2-data2.txt** with the only difference being that you should use the **regularized** version of logistic regression in this part with a regularization parameter of 1. Similar to (b), consider maximum number of iterations to be 400 in the **fminunc** function and the initial values for all parameters to be zeros.

   (d) Similar to (b), use the validation dataset **P2-valdata2.txt** to select the most appropriate model from amongst the ones that you trained in part (c). Make sure that the validation error is at most 15% for your selected model.

   (e) Repeat (c) and (d) with a regularization parameter of 0. Explain your observations.

   (f) Repeat (c) and (d) with a regularization parameter of 100. Explain your observations.

Submission instructions: Please submit a zip file containing a pdf of your report and a subfolder containing your Matlab codes. Name your zip file as "LastnameFirstname.zip".

You should treat your project report as if your boss were going to read it. The report should not be a simple listing of the results and steps performed. It should be written as a standard technical report, where you should provide some background and context for each result. Your report should also contain a separate subsection titled "Matlab Code", where you should briefly describe the Matlab files that you submitted. Just a few lines description for each file is sufficient. This is just to ensure that we can easily locate the source file that was used to generate a specific result (in case we need to verify something).

Contact: Please direct any specific questions related to this project to GTA Mohamed Abd-Elmagid (maelaziz@vt.edu). Please keep the instructors cc'ed on your emails. Please feel free to contact the instructors directly for any conceptual questions related to the course material.