Lab Assignment 03

Total marks: 15

Instructions:

1. For question 1 and 4, compile into a single PDF (Figures.pdf) the relevant code snippets along with the displayed portion of data and submit in the gradescope.

1 Logistic Regression - Linear Classifier [10 Marks]

Apply the logistic regression (linear classifier) algorithm discussed in the lab session to predict next-day rain based on the 10 years of daily weather observations from many locations within a country. The dataset contains many factors taken into consideration to specify whether it rained or not on that particular day. The training and testing dataset is provided in the files titled 'weather_train.csv' and 'weather_test.csv', respectively. Carry out the following tasks as assignment problems:

1. Inspect and plot some portion of the training data using `pandas`. Segregate the training and testing data into two separate variables consisting of 'feature values' and corresponding 'predictions' (the prediction column is titled '`RainTomorrow`' in the dataset). To simplify the problem a bit, clean the whole data by carrying out the following sub-tasks:

   (a) Convert the predictions in the binary format by using '`1`' for '`YES`' and '`0`' for '`NO`'.

   (b) Identify and drop the feature columns having datatype 'object'.

   (c) Identify cells having '`NaN`' or '`NA`' values and replace them with mean values of their respective columns.

   (d) Normalize all the feature values by scaling them between `0` and `1`. The values in feature matrix '$\mathbf{X}$' can be normalized as:

   $$\mathbf{X}_{norm} = \frac{\mathbf{X} - min(\mathbf{X})}{max(\mathbf{X}) - min(\mathbf{X})}$$

   Execute the above sub-tasks and display some portion of the data and its head after each data cleaning step. [1 Marks]

2. Calculate the mean of the data column titled as '`MaxTemp`' from the normalized training dataset. [1 Marks]

3. Classify the cleaned dataset using the regularized binary classification algorithm discussed in the class and calculate the optimized weights and training set accuracy for the model (take regularization parameter $\lambda = 0.1$ and use Truncated Newton's Method for optimization). [4 Marks]

4. Plot the cost history (J) vs. the number of iterations. [Hint: You can make use of 'callback function' in `Optimize.minimize` to store the cost history] [2 Marks]

5. Apply the trained model on the cleaned test dataset to predict the testing accuracy of the model. [2 Marks]

2 Logistic Regression - Non-linear classifier [5 Marks]

Instructions: For question 1, compile into the same PDF (Figures.pdf) the code snippet, relevant figures along with your interpretations and submit in the gradescope.

1. Implement the logistic regression as a non-linear classifier on the microchip problem discussed in the lab session and briefly discuss the effect of using different values of regularization parameter with the help of decision boundary. [5 Marks]