Assignments on Logistic Regression #4 (Ref Video Lectures 15-20)
@Prof G C Nandi[1]

(Timely submission of assignments is essential. Copying/plagiarised submission from others will fetch fail (F) grade on this subject)

1. Our task is to build a Logistic Regression based classification model that estimates an applicant's probability of getting admission to an institution based on the scores from those two examinations whose data have been provided here (you may use 70% data for training and 30% for testing).

a) Design a Predictor with two basic features which are given using Batch Gradient Descent Algorithm, Stochastic Gradient Algorithm and mini batch Gradient Descent algorithms (determining minibatch size is your choice) with and without feature scaling and compare their performances in terms of % error in prediction. (only allowed to use NumPy library of Python, no other functions/libraries are allowed).

b) Inject more features from the data set in the model and repeat (a)

c) Add regularization term and repeat (b). Submit comparative analyses of your results.

2. After gaining experience of solving problem No 1) Design a classifier using logistic regression on Cleveland Medical data set for heart disease diagnosis. The processed dataset with some 13 features has been given with a label that a patient has a heart disease (1) or not (0). This design should have a professional touch within your ML knowledge in terms of data preprocessing, feature scaling, selecting appropriate features etc. The following link provides a description of the data set and it also contains other heart disease related data which you may also use, since the data provider has issued the following statutory warning: (As David says-"the file cleveland.data has been unfortunately messed up when we lost
node cip2 and loaded the file on node ics. The file processed.cleveland.data
seems to be in good shape and is usable (for the 14 attributes situation).
I'll clean up cleveland.data as soon as possible".

"Bad news: my original copy of the database appears to be corrupted.
I'll have to go back to the donor to get a new copy"-David Aha)

.
https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/ (the archive also contains other heart disease, which you may use)

Hints both for problem #1 and 2:

Evaluating logistic regression classifier:

1. After learning the parameters, you can use the model to predict whether a particular student will be admitted. For example a student with an Exam 1 score of 45 and an Exam 2 score of 85, you should expect to see an admission probability of 0.776. Another way to evaluate the quality of the parameters we have found is to see how well the learned model predicts on
our training set. Create a predictor function which will produce "1" or "0" predictions given a dataset and a learned parameter vector W.
2. Use Confusion matrix (it is a metric for testing the performance of a model) to evaluate the performance of your classifier. Can consult this blog for details: (https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62)

Deadline for submission: 7-11- 2022, midnight.

Full marks: 50+50=100

[1] Prof G C Nandi, IIIT-Allahabad