ASSIGNMENT 1: BASIC ML

Full Marks : 100

In this assignment, we will explore two concepts been studied in class :

1. Linear Regression
2. Logistic Regression

It involves a 3-step work:

a. Dataset creation
b. Modelling
c. Reporting metrics

==========

1. Linear Regression

a. Dataset creation: (12)

i. Please generate a dataset (X,Y) of 1500 samples from a normal distribution N(0,1), each with 2 dimensional features.

If input = x, output(y) = $\beta x = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
Assign β to some values within the range 1-2 before generating y.

ii. Create a train-test split of 80-20.
iii. Visualize the data in a 2-D plane and plot them (Feature 1(x-axis), Feature-2(y-axis)).

b. Modelling (20)

i. Use scikit learn library to train a linear regression model and save the parameter values (named as say, P1).

ii. Design the linear regression algorithm using sample code available in https://www.dropbox.com/s/falgjz3h8ntf5ff/lr_algo.py?dl=0 . You've to complete the gradDes(...) algorithm which is essentially the update step of **stochastic gradient descent** as studied in class. Let's name the parameter values as P2.

iii. Plot the cost values obtained from Step ii (Iterations (x-axis) vs Cost values (y-axis)). Ideally, the values should decrease.

iv. Use scikit learn library to train a ridge regression model on the same data and save the parameters (say P3).

c. Metrics (18)

i. Use P1, P2, P3 to report Root Mean Square Error, Mean Absolute Error and Normalised Root Mean Square Error on the test set.

ii. Report a plot with x-axis as test set indices, y-axis as the predicted values using P1, P2, P3. Use 4 different colors and markers for the plot.

iii. Report a box-plot of errors across all test set points for the three different prediction modes.

Iv. Report the differences between the ground truth β and the obtained β values in P1, P2. P3.

****

2. Logistic Regression

a. Dataset creation: (20)

i. Please create two binary class datasets (X,Y) of 1500 samples, one with 2 dimensional features, and another with 3 dimensional features.

$X \sim N(0,1)$

$Y = 1$ if σ (ßx) > = 0.5

= 0 otherwise

Assign β to some values within the range 1-3 before generating Y.

ii. Create a train-test split of 80-20.
iii. Visualize the 1st dataset in 2-D plane and plot the same.
iv. Perform feature normalisation of the dataset with 3 dimensional features.

b. Modelling (20)

i. Use scikit learn library to train a Logistic regression model on both datasets.
ii. Vary the following parameters and save the parameters for both datasets:

1. Regularization - L1, Tolerance - 1e-2, Solver - liblinear (say parameter P11, P12) {The first index (1 in P12 for e.g.) denotes the combination number and second index (2 in P12) denotes dataset number}
2. Regularization - L1, * Tolerance - 1e-5, Solver - liblinear (say parameter P21, P22)
3. Regularization - L2, Tolerance - 1e-2, Solver - newton-cg (say parameter P31, P32)
4. Regularization - L2, Tolerance - 1e-5, Solver - newton-cg (say parameter P41, P42)

c. Metrics (10)

i. Use all the parameters (or models) [P*] to report mean accuracies and class-specific accuracies on the test sets for both the datasets in plot/table. Use the parameter names as distinguishing factors and clarify the notations.

ii. Plot a decision boundary using the 2-dimensional feature dataset using **any one of the four combinations** above (whichever gives the highest accuracy).

iii.Using 2 of the best features out of 3 in the 2nd dataset, plot a decision boundary in the same way as above.

==========

Submission Instructions:

1. Please submit your codes with filename as Assign<number>_Q<number>_<RollNumber>.py for e.g. Assign1_Q1_123456.py
2. Submit a write-up document with the desired results and plots and observations in pdf format, naming it as Assign<number>_<RollNumber>.pdf