Programming Questions Report :

5.

(i) The logistic model $P(\hat{y} = 1 / x_1, x_2) = \frac{1}{1 + e^{-(-1+1.5x_1+0.5x_2)}}$

if $A = \sigma(W^T x + b)$, then the cross entropy error function (or) the cost is given as,

$$\boxed{\text{Cost} = \frac{-1}{m} \sum_{i=1}^{m} [y * \log(A) + (1-y) * \log(1-A)]}$$

(ii) The values of θ_0, θ_1 and θ_2 after one iteration are

$\theta_0 = -1.00316$

$\theta_1 = 1.49035$

$\theta_2 = 0.4969$

∴ The updated logistic regression model is

$$\sigma(f_\theta(x_1, x_2)) = \frac{1}{1 + e^{-(-1 + 1.49x_1 + \cancel{0.497x_2} \downarrow 0.497x_2)}}$$

(iii) With learning rate = 0.1 ~~and~~, 10,000 iterations and regularizer constant $\lambda = 0.01$,

Accuracy = 0.8333

Porecision = 0.75

Recall = 1.0

6.

The top 2 scores of my model is obtained by Random Forest and XG-Boost

Model	Score
Random Forest	3.50296
XG - Boost	3.44991

Random Forest

Random Forests (RF) are a combination of tree predictors. It ~~is~~ belongs to one of the types of Ensemble Classifiers known as Bagging. In random forest, several trees are buiet using the boot strapped data (i.e sampling with replacement) and at each node, only a subset of ~~nar~~ feantures are selected for mode splitting. Hence we train several trees using bootstrapped datasets and the majority classification among all trees is selected

XG Boost

XG Boost stands for "Extreme Gradient Boosting". It is a decision-tree based ensemble machine learning algorithm that uses a gradient boosting framework. It performs Optimized Gradient boosting algorithm through parallel processing, tree pruning, handling missing values and regularization to avoid overfitting / bias .

why random forest performed better compared to other methods ?

- It ran efficiently as the data set is large due to bootstrapping

- It does not overfit by design. This Taxi-data particularly, contained many outliers which random forest handled effectively.

- They are easily adapted to distributed computing

why XG boost performed better than others and even random forest ?

It basically improves upon the base Gradient Boosting framework through system optimization and algorithmic enhancements .

System Optimization include Parallelization, Tree Pruning and Hardware optimization

Algorithmic Enhancements include Regularization, Sparsity Awareness, Weighted Quantile Sketch and Cross validation .

Advantages of XG boost when compared to random forest

1.) XG boost straight away prunes the tree with a score called "similarity score" before entering into the actual modeling purposes

2) XG-boost is a good option for unbalanced datasets but we cannot trust random forest in unbalanced datasets

3) One of the most important differences between a XG boost and Random forest is that the XG boost always gives more importance to functional space when reducing the cost of a model while random forest tries to give more preference to hyper parameters to optimize the model.