PROJECT 1. BINARY CLASSIFICATION USING LOGISTIC REGRESSION

Thana Shree Jeevanandam
University of Buffalo
thanashr@buffalo.edu

Abstract

Machine Learning,the art of making the computers learn by themselves is growing tremendously in all its aspects. It provides the machines to develop problem solving capabilities without human intervention. The topic under study in this project is to diagnose cancer cells from a group of cells using Binary Classification. The Logistic Regression classifier is used to handle the challenge.The logistic regression is coded from scratch on python.The training is given through logistic regression and gradient descent. The validation and testing are done to check the prediction accuracy. The different evaluation metrices are used to check the effectiveness of the solution. On the whole, the system is trained on this dataset to understand the relationship between the features and the relevance judgements so that when a new data is given, it could predict with the highest accuracy.

1.INTRODUCTION

An important challenge that is being grappled with in the medical industry is to classify a patient diagnosed with cancer, to their appropriate class. Class specific treatment can significantly reduce toxicity and increase the efficiency of the therapy. The goal is to increase the prediction accuracies of cancer cells. Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. In this project the 30 featured dataset is modeled to provide the diagnosis of cancer cells. Logistic Regression with Gradient Descent finds the way to solve the problem.

2. DATASET:

The wdbc (Wisconsin Diagnostic Breast Cancer) dataset used for the prediction task consists of 569 instances. The Dataset is used to populate the Training, Validation and Testing sets. The Dataset contains,

Classes	2
Samples per class	212(M),357(B)
Samples total	569

Dimensionality	30
Features	real, positive

Each instance in the dataset is associated with 32 attributes (ID, diagnosis (B/M), 30 real-valued input features). Ten real-valued features are computed for each of 3 cell nuclei, yielding a total of 30 descriptive features. The 10 features (in order) are:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1

The task at hand is to predict the Diagnosis for each instance based on the Logistic Regression of

3. PREPROCESSING OF DATASETS

The project starts with preprocessing of data. As a preprocessing measure, the process initiates reading the contents from the dataset in python. The original CSV data file is processed into a Pandas Dataframe. The steps in preprocessing of data includes,

1. Dropping the ID column.
2. Labelling the Diagnosis column.
3. Splitting the dataset.
4. Normalizing the data.

3.1 DROPPING THE ID COLUMN:

The ID column is dropped as it does not contribute to the analysis.

3.2 LABELLING THE DIAGNOSIS COLUMN:

The ID column is dropped as it does not contribute to the analysis. The categorical data, M and B in the Diagnosis column is converted to binary values, 1 and 0 respectively. This data is populated to variable 'y' in the project.

3.3 SPLITTING THE DATASET:

The next step of preprocessing is partitioning datasets into Training, Validation and Testing sets. The Training data comprises of 80% of the overall dataset; Validation set comprises of 10% and

Testing comprises of the last 10%. The partition should be such that there is no overlapping of data in any of the datasets. The model is then trained on the Training dataset and tested on the Validation and Testing sets.

3.4 NORMALIZING THE DATA:

The data has features varying high in their magnitudes and range. To bring all the features to the same level of magnitude/range for processing, we normalize the data. Hence,the normalization fits the data in the range 0-1.

4.PROCESSING OF DATASET:

The preprocessed dataset is all set to get into the process. Logistic Regression is one of the classifiers to fit models for categorical data, mainly a binary data. It comes under the generalized linear models class. It predicts the probabilities directly by preserving the marginality. The steps in the processing include,

4.1 INITIALIZING THE WEIGHTS:

The only set of parameters controlling how accurate our model is are the weights and the bias. The changes in the hyperparameter is reflected in the model's accuracy.They include:

1.Learning Rate

2.Epoch (Number of iterations)

As the first step, the weights are initialized with respect to the shape of x_train.

4.2 LOGISTIC REGRESSION:

In this project , it takes the 30 features as an input vector and gives the result as Malignant or Benign Cell.

The Logistic Regression undergoes two propagations namely,

1.Backward Propagation

2.Forward Propagation

4.2.1 FORWARD PROPAGATION:

The two operations as a part of the forward propagation process are,

1.Applying a linear transformation on the input features and

2.Applying a non linear transformation (sigmoid) on top of the previous output to give the final output.

4.2.2 BACKWARD PROPAGATION:

The Backward Propagation is done with gradient descent algorithm to stabilize the weights.

4.3 SIGMOID FUNCTION :

The activation function of Logistic Regression is the sigmoid function. It has the range between 0 and 1 which can be interpreted as uncertainties for the outputs of 0 and 1 in the model.

The sigmoid function applies a non linear transformation onto the initial value and the range of the sigmoid function is a set of real values between 0 and 1.

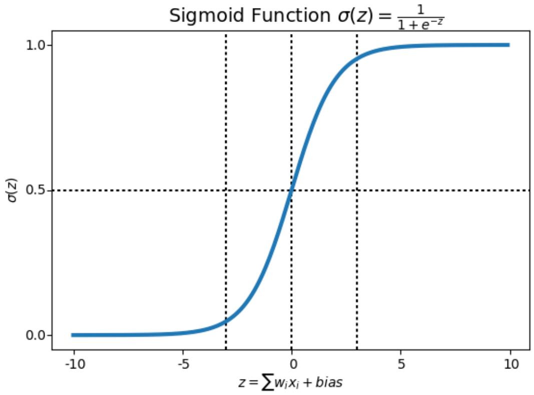

Fig 4.1: SIGMOID FUNCTION

4.4 GRADIENT DESCENT :

The cost must be decreased in order to increase the accuracy of the model. The Gradient Descent for logistic regression is used to train the model using a group of hyperparameters to bridge the model’s output and the true output together with a decrease in the loss/error/cost function.

$$W = W - \alpha \frac{\partial J}{\partial W}$$

$$b = b - \alpha \frac{\partial J}{\partial b}$$

The weights are stabilized by the Gradient Descent Algorithm. It is an optimization algorithm that finds the optimal weights (a,b) that reduces prediction error. It finds the least minimum between y_test (true) and output (predicted). Cost is the loss for the entire dataset.The Gradient Descent Algorithm is run repeatedly with updated weights to get the minimum cost.

5. ARCHITECTURE:

The Logistic Regression uses sigmoid function as its activation function.

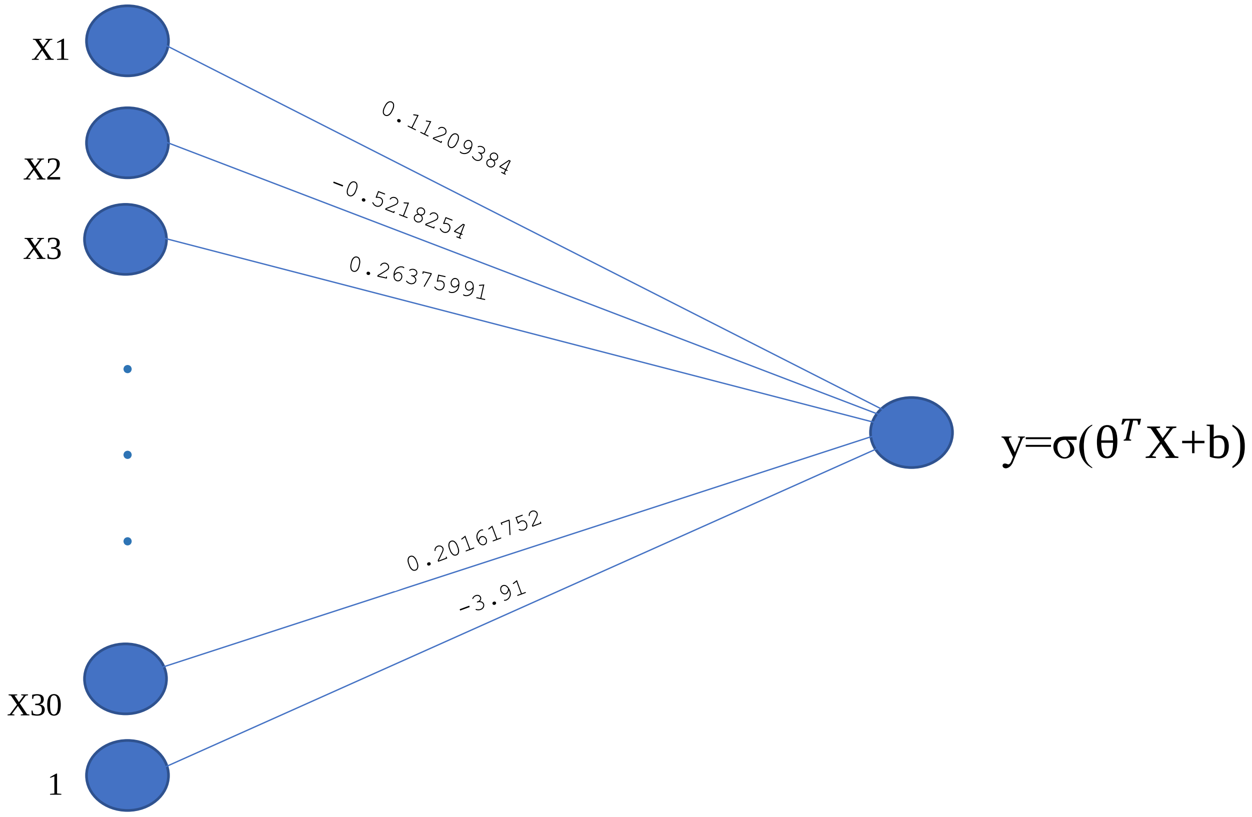

Fig 5.1: ARCHITECTURAL DIAGRAM

6.RESULT:

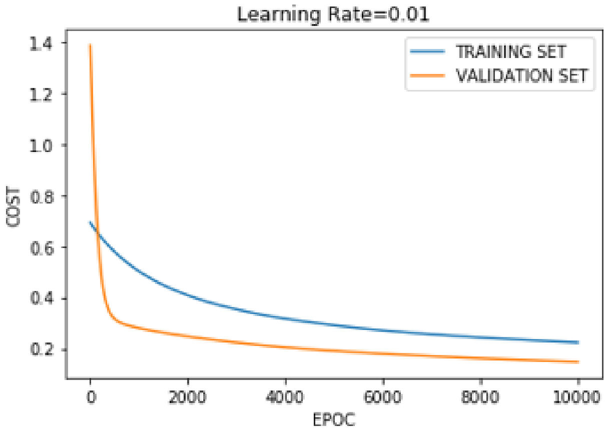

Fig 6.1: COST VS EPOCH (learning rate=0.01)

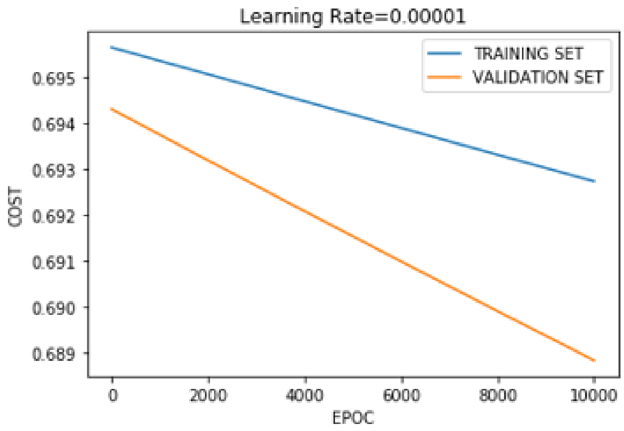

Fig 6.2: COST VS EPOCH (learning rate=0.00001)

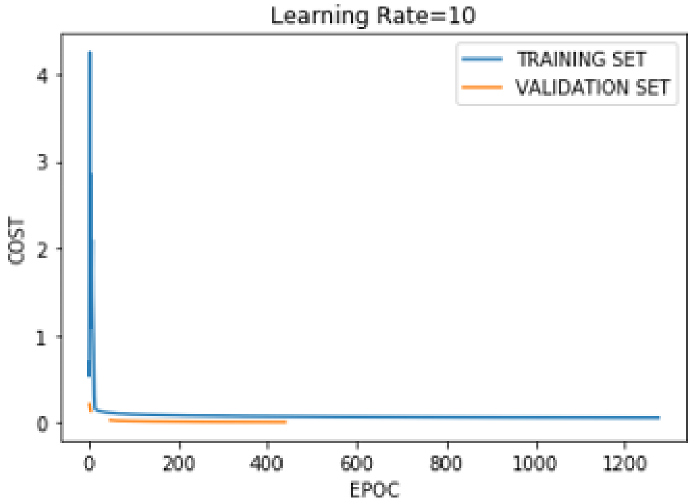

Fig 6.3: COST VS EPOCH (learning rate=10)

The above graphs represent the change in cost with respect to the change in epoch. The epoc is plotted along the x axis with its corresponding cost along the y axis. We can infer from the above graph that the cost of validation set is lower than the training set with a smooth curve on learning rate =0.01. In general, the ideal values for learning rate is 0.01. Very high values of learning rate causes overfitting. Smaller learning rate skes the graph which is not optimal.

LEARNING RATE	**Testing Accuracy**
0.01	0.9649122807017544
0.00001	0.5964912280701754
10	0.9473684210526315

Table 6.1: Tabulation of the Learning Rate and Testing Accuracy

4.CONCLUSION

This project helps get a clear idea of the working of Logistic Regression. It has also helped understand the functioning of the gradient descent. The updated weights by computing the cost function and minimizing it by each of the approaches. The final model is tested with testing set to check the accuracy which is 96%. The various evaluation models are used to evaluate the model. This project helped get a deeper insights of the impact of the hyper parameters on the performance of the system. Thus, we performed a logical regression to diagnose a cancer cell.

REFERENCES

[1] Li, Jun, José M. Bioucas-Dias, and Antonio Plaza. "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields." *IEEE Transactions on Geoscience and Remote Sensing* 50.3 (2011): 809-823.

[2] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12.Oct (2011): 2825-2830.

[3] Basis Functions–Introduction to Linear Basis Function Models - https://chemicalstatistician.wordpress.com/2014/03/10/machine-learning-lesson-of-the-day-introduction-to-linear-basis-function-models/

[4] Python functions - https://docs.scipy.org/