Breast Cancer Classification

Danish Hafeez
dept. of E & CE
Air University
Islamabad
171458@students.au.edu.pk

***Abstract*—This Project is about early detection of breast cancer in which we divide the data into two based on the features i.e binary classification. To do so we used logistics regression algorithm and manipulate data using K-Folds Cross validation technique.**

I. Introduction

Nowadays, Breast cancer is one of the most common cancers in women. The healthy cells in the breast grow and change into a tumor. A tumor can be cancerous or benign. In 2017, the American Cancer Society has announced that, over 2, 50,000 new cases of invasive cancer were diagnosed each year in women. The American Cancer Society recommended that women with an average risk of breast cancer should undergo regular screening mammography, starting at the age of 45 to 54 years

Breast cancer can be categorized into two, which are malignant breast cancer and benign breast cancer. The classification of breast cancer as either malignant or benign is possible by scientifically studying the features of breast tumors, lumps, or any abnormalities found in the breast. At the benign stage the cancer has less risk and is not life threatening while cancer that is categorized as malignant is life-threatening

We modeled Logistics Regression to form a binary classifier for this problem of detection of cancer

II. DataSet

Data mining is most important aspect in machine learning. We used the data set which was extracted from mammograms. This dataset consists of 699 instances having given features

Features Information:

Attribute	Domain
1. Sample code number	id number
2. Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 - 10
11. Class:	0-Benign 1-Malignanat

I first Read the CSV file than preprocessed the data to add any missing points and Mean normalized the dataset.

III. Methodology

I used Logistics regression to perform my Classification weather it is benign or malignante

A. *Cross validation*

I split the K-Folds Cross validation to compute my gradient it showed appropriate prediction with mean of 97.43.my cross fold validation K=10 i.e it divides our data set in 10 different data sets which is a long theory

B. *Hypothesis*

The Gradient descent of logistics regression is calculated in different Steps First we have to calculate the hypothesis on our initialized Thetas used the Sigmoid Function to form an intuition about in which class our data set fits

$$h_\theta(x) = g(\theta^T x)$$

$$z = \theta^T x$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

Sigmoid plot is like that we build an intution if hypothesis is les than 0.5 than its in class 0 i.e benign or if hypothesis is greater than 0.5 its in class 1 i.e malignant Sigmoid Funtion gives reslut between 0 and 1 and its one of the best working binary classifier

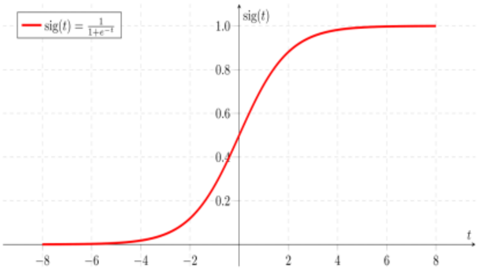


Figure 1: Sigmoid Function

C. *Gradeint Descent*

Gradient Descent is the process by which we update thetas at every iteration we calculate the cost and based on that cost function we update theta if difference between two previous costs is less than 0.001 or some value we select to train than Gradient stops and gives us the predidction features or thetas

Gradient is calculated by these given Eqs

$$\begin{aligned}&Repeat\ \{\\ &\quad \theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\\ &\}\end{aligned}$$

We have to repeat this untill convergance is obtained
I implemented Vactorized form of the above equation which is

$$\theta := \theta - \frac{\alpha}{m} X^T(g(X\theta) - \vec{y})$$

Vactorized form is computationally less expensive because in above snerio we have to implement many many loop but in vactorized form we can get desired result in just one go.

To compute the Cost we have to use this given equation

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1 - y^{(i)})\log(1 - h_\theta(x^{(i)}))]$$

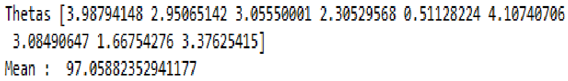
And vatorized form of this above equation is

$$\begin{aligned}&h = g(X\theta)\\ &J(\theta) = \frac{1}{m}\cdot\left(-y^T\log(h) - (1-y)^T\log(1-h)\right)\end{aligned}$$

These Equations were separetly inplemented in functions and than an Evaluate algorithm function do the job

IV. Analysis

After implementing the above equations in python I analyzed the algorithms working as I used K-Fold Data validation scheme I had K results in the end from which best result having best mean is selected over all Mean was 97.43% .one of the resultant learned thetas with the efficiency mean is given below



```
Thetas [3.98794148 2.95065142 3.05550001 2.30529568 0.51128224 4.10740706
 3.08490647 1.66754276 3.37625415]
Mean :  97.05882352941177
```

V. Conclusion

Early stage of cancer can be predicted based on benign and malignant which will be very helpful to doctors. And our model of logistics regression combined with K-Fold validation gave some near to accurate results our accuracy mean was 97%. Our cost also reached near 0.01