Breast Cancer Tumor Cells Classification using Logistic Regression

Mansi Wagh
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
mansiwag@buffalo.edu

Abstract

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society. The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modeling. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. The aim of this project is to classify suspected fine needle aspirate (FNA) cells as Benign (class 0) or Malignant (class 1) using logistic regression as the classifier. In machine learning, classification a two class problem. The features used for classification are pre-computed from images of a FNA of a breast mass. The dataset used is this project is Wisconsin Diagnostic Breast Cancer (wdbc.dataset).

1.Introduction

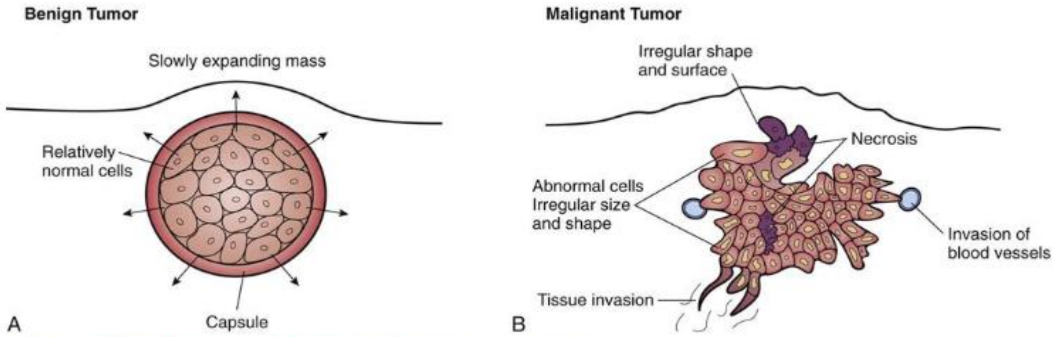
There are multiple causes of breast cancer which involves family history, obesity, hormones, radiation therapy, and even reproductive factors. Every year, one million women are newly diagnosed with breast cancer, according to the report of the world health organization half of them would die, because it's usually late when doctors detect the cancer . Breast cancer is caused by a typo or mutation in a single cell, which can be shut down by the system or causes a reckless cell division. If this condition persist for a few months, masses(tumors) are formed from cells containing wrong instructions.

Malignant tumors expand to the neighboring cells, which can lead to metastasize or reach other parts and damage them, whereas benign masses can't expand to other tissues, the expansion is then only limited to the benign mass. Detection of BC may be hard at the beginning of the disease, due to the absence of symptoms, after some clinical tests, the accurate diagnosis should have the ability to distinguish the benign and malignant tumors.

Machine learning(ML) is a set of tools utilized for the creation and evaluation of algorithms that facilitate prediction, pattern recognition, and classification. ML is based on four steps: Gathering data, selecting the model, training the model, testing the model. The relation between BC and ML is not recent, it had been used for decades to classify tumors and other malignancies, predict sequences of genes responsible of cancer and determine the medical prognostic. The aim of the classification is to put each observation in a category that it belongs to. In this project, we use the Wisconsin breast cancer database to predict the likelihood of cancerous(malignant) or harmless(benign) cells based of the features provided in the data set. The performance of this will be evaluated in terms of accuracy, training process and testing process.

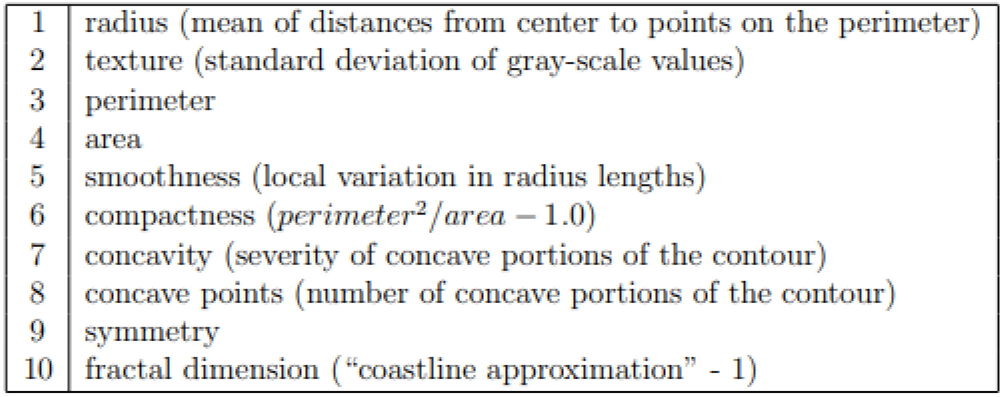
2. Dataset Description:

One application of machine learning in a healthcare context is digital diagnosis. ML can detect patterns of certain diseases within patient electronic healthcare records and inform clinicians of any anomaliesThis project is to illustrate how useful machine learning can be as a medical diagnosis tool, by examining its use in breast cancer detection using a publicly available Breast Cancer Wisconsin (Diagnostic) Data Set. This data set consists of several instances of tumors. Tumors can either be benign (non-cancerous) or malignant (cancerous). Benign tumors grow locally and do not spread. As a result, they are not considered cancerous. However, they can still pose a danger, especially if they press against vital organs like the brain. Malignant tumors, in contrast, have the ability to spread and invade other tissues. This process, known as metastasis, is a key feature of cancer.



Characteristics of benign and malignant tumors.

Wisconsin Diagnostic Breast Cancer (WDBC) dataset will be used for training, validation and testing. The dataset contains 569 instances with 32 attributes (ID, diagnosis (B/M), 30 real-valued input features). Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Computed features describes the following characteristics of the cell nuclei present in the image:



1	radius (mean of distances from center to points on the perimeter)
2	texture (standard deviation of gray-scale values)
3	perimeter
4	area
5	smoothness (local variation in radius lengths)
6	compactness ($perimeter^2/area - 1.0$)
7	concavity (severity of concave portions of the contour)
8	concave points (number of concave portions of the contour)
9	symmetry
10	fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

3. Steps in Dataset Preprocessing:

Step 1: Extract features values and Image Ids from the data:

Process the original CSV data files into a Numpy matrix or Pandas Dataframe.

Step 2: Normalizing the Data Set:

In this project, Data is normalized using MinMaxScaler() function imported from sklearn. MinMaxScaler() function transforms features by scaling each feature to a given range.

Step 3: Splitting the Data Set:

In order to get started modeling, the data set was split into three parts :

i. Train set (80%), for choosing models.
ii. Test set (10%), hold out data on which we will see how well models are able to generalize on unseen data.
iii. Validation set (10%) for validating our data.

Step 4: Train using Logistic Regression:

Gradient Descent has been used for logistic regression to train the model using a group of hyper-parameters.(Epoch and Learning rate)

Step 5: Tune hyper-parameters:

Validate the regression performance of your model on the validation set. Change your hyper-parameters. Try to find what values those hyper-parameters should take so as to give better performance on the validation set.

Step 6: Test your machine learning scheme on the testing set:

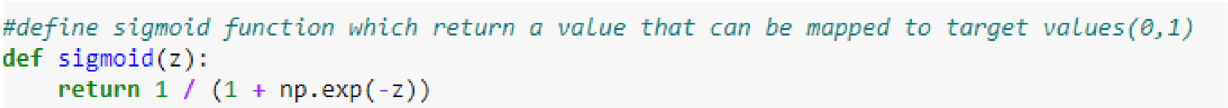
After finishing all the above steps, fix your hyper-parameters and model parameter and test your models performance on the testing set. This shows the ultimate effectiveness of your models generalization power gained by learning.

4. Model Architecture

Logistic Regression forms a model which gives the predicted probability of target variable as a function of input variable X. So, if we consider if X as an input vector then we would want to predict whether particular observation belongs to either class “0” or class “1” (in case of binary classification). We know that linear regression can be expressed by below equation: $y^{\wedge} = w^T .x$

But apparently, this equation does not work well in case of binary classification because we want $y^{\wedge}$ to be the chance that y =1 or 0. If we use linear regression for classification problem our linear regression model will not perform well because of many reasons.

In classification problem where we are predicting the probability of target variable, linear regression will give us values which are above 1 and below 0. Also, the predictions made using linear regression can be highly inaccurate. This is when sigmoid function comes into the picture. The sigmoid function helps in bringing non-linearity in the model. As we are predicting the probability of class 0 and 1 we would not want it to be less than 0 and more than 1. Sigmoid function helps in bounding the probability between 0 and 1. Sigmoid function is defined as follows in our project code:

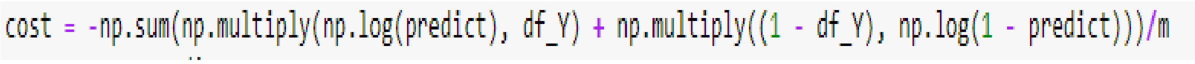


```
#define sigmoid function which return a value that can be mapped to target values(0,1)
def sigmoid(z):
    return 1 / (1 + np.exp(-z))
```

If the value of z is very large positive number then σ (z) = 1/1+0= 1.And if the value of z is very large negative number then σ (z) = 1/1+ (big number) and σ (z) will approximate to 0. So, we can see that the value is bound between 0 and 1.

Now to train our model we need to define the cost function for our logistic regression problem. Before looking at cost function let us see the loss function of logistic regression. The loss function is nothing but the error function which will tell us how good our output variable y^ is when the true label is y. The loss function for logistic regression is given as- **L(y^ ,y) = - (y log y^ + (1 - y) log(1 - y^)).**

Here, we want our loss function to be minimum. The key point here is Loss function is given for single training example,the error for entire training set it is given by Cost function. So, the cost function which is applied to entire training set is given as-



```
cost = -np.sum(np.multiply(np.log(predict), df_Y) + np.multiply((1 - df_Y), np.log(1 - predict)))/m
```

Our aim is to minimize this cost function.

5.Result :

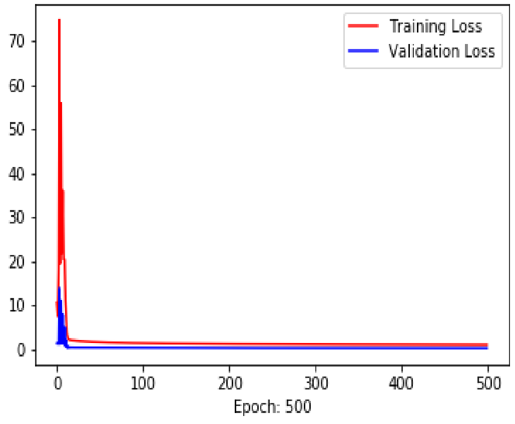
We have to calculate accuracy, precision and recall values to evaluate the efficiency of our model in identifying the tumor as benign or malignant. In this project, accuracy, precision and recall values are calculated by importing accuracy_score(),precision_score() and recall_score() functions from sklearn. Informally, **accuracy** is the fraction of predictions our model got right. Formally, accuracy has the following definition:

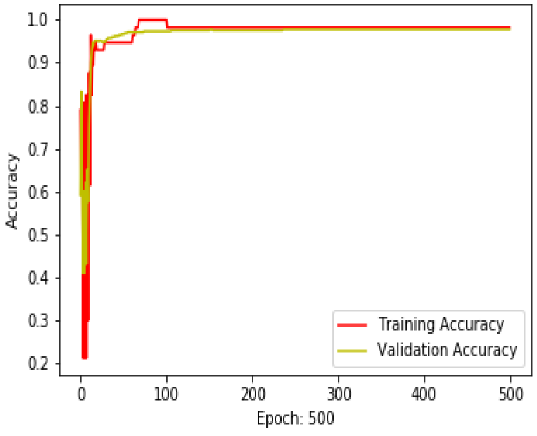
Accuracy=Number of Predictions/Total number of predictions

Precision identifies the proportion of positive identification is actually correct. On the other hand,recall determines what proportion of actual positive outcomes.

This project is expected to plot and demonstrate the learning curves between Accuracy vs Epoch and Loss vs Epoch. Let us take a look at the graphs and try to interpret results

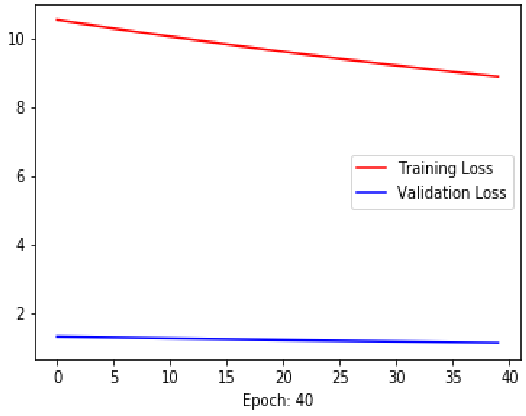
Example 1.

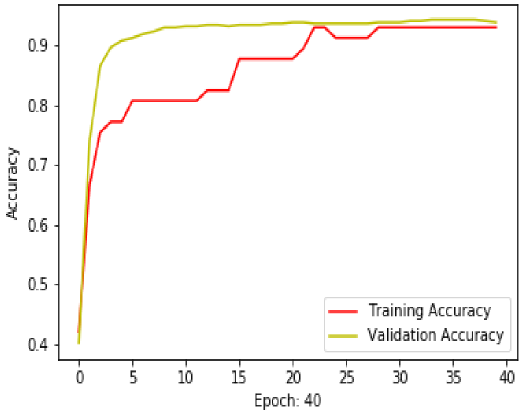




Above plotted graphs have epoch :500 and learning rate:0.7. The first graph indicates that the plots of training and validation loss both decreases to the point of stability and have less gap between them. Thus it is a good fit learning curve. Also, the plots of training and validation accuracy are increasing and gradually become stable.

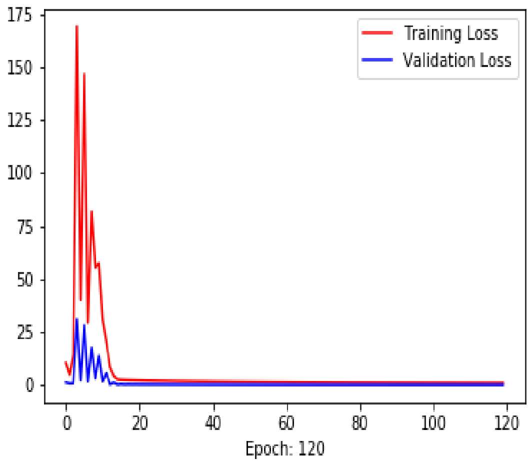
Example 2:

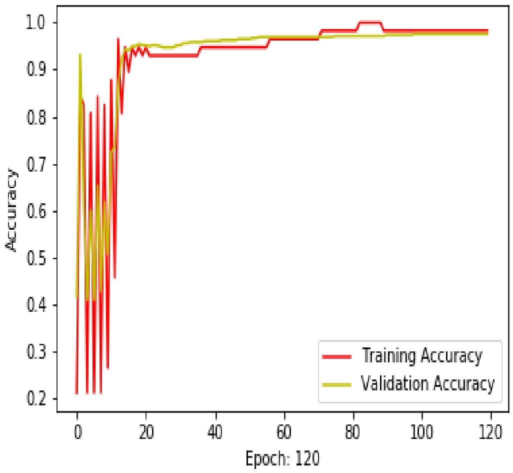




In the first plot , there is large gap between training loss and validation loss. The flat line shows relatively high loss which indicates that the model was unable to learn from training dataset at all. The second plot depicts training and validation accuracies over a period of 40 epochs. The accuracies appear to stabilize after 30 epochs .On comparing it with example 1, we see that the decrease in number of epochs and learning rates impact the evaluation parameters. That implies, decrease in number of epochs and learning rate changes accuracy,precision and recall values of the model significantly.

Example 3:





In this example, the first figure plots a good fitting curve as the gap between training loss and validation loss gets decreased and they appear to converge at some point. The adjacent plot indicates the varying accuracies of the training and validation sets which gradually become stable in the period of 120 epochs.

6.Conclusion:

On the Wisconsin Breast Cancer datasets, logistic regression is used as a classifier to classify the tumor cells as benign or malignant from the breast masses. After splitting the data into training , test and validation sets, logistic regression was used as a classifier that helped distinguish between the cells by comparing each cell against the 30 features given in the dataset.Over a certain period of epoch, the dataset was observed. Values of the losses and accuracies of the training and validation datasets where plotted against the epochs to help predict the class of the cells as benign or malignant. Ultimately the model was evaluated and showed 98.2% accuracy with 93.3% precision for the test dataset.

7.References:

1.https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

2.https://www.nature.com/articles/s41563-019-0360-1

3.https://www.macadamian.com/learn/a-practical-application-of-machine-learning-in-medicine/

4.https://datascienceintuition.wordpress.com/2018/01/16/logistic-regression-as-neural-networks/

5.https://gogul.dev/software/neural-nets-logistic-regression

6.https://developers.google.com/machine-learning/crash-course/classification/accuracy

7.https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

8.https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

9.https://ieeexplore.ieee.org/document/8391453

10.