Performance of Perceptron on Breast Cancer Wisconsin dataset

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

[be innovative but precise with the title]

***Abstract*—In machine learning, the perceptron is an algorithm for supervised learning of binary classifiers (functions that can decide whether an input, represented by a vector of numbers, belongs to some specific class or not).[1] It is a type of linear classifier, i.e. a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector. The algorithm allows for learning, in that it processes elements in the training set one at a time.We study the performance of the perceptron for the classification of the Breast Cancer Wisconsin Dataset. The system developed performs with 88.55% insample accuracy in a two class classification**

1. Introduction

In machine learning the task of classification is to learn a decision boundary from the given data such that the data can be separated into appropriate classes. For example in case the training data belonged to two classes A & B, the idea is to determine an appropriate plane or straight line (in case of 2D data) such that all training samples belonging to class A lie on one side of the plane while the sample points belonging to the other class lie on the other side of the plane. This is done using a learning algorithm like the perceptron learning. The objective of this learning is to ensure that once a decision boundary is learnt, an unknown future 'test' sample point will be correctly classified as belonging to class A/B.

2. Methodology

The preceptron is a basic building block of most neural networks. The Artificial neuron is a threshold logic that accepts multiple inputs, which are weighted and summed. If the sum exveeds a certain threshold, the perceptron fires and an output value is transmitted to the next unit.

The Artificial neuron (Pitts 1943) [**?**] has inputs $x_1, x_2, ...x_N$. The weighted inputs are connected to the output. The inputs are generally real values. The weights $w_1, w_2, ...w_N$ are also real and could be negative. A bias w_0 is introduced as an intercept.

If the weighted input exceeds sum threshold the neuron fires producing ± 1 which can be considered as a class label. The perceptron can model a linear discriminant function which is a straight line in 2D, and a plane in 3D or hyperplane in higher dimensions. The perceptron cannot solve problems like the XOR function. For such functions, multilayer feed forward neural networks modeled on perceptrons can be used.

3. Database - Breast Cancer Wisconsin Dataset(Original)

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used,to distinguish malignant (cancerous) from benign (non-cancerous) samples.This dataset consist of 699 instances and 11 attributes that would help the classiy the data into the two classes.The attributes are described in the fig.1

#	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:	(2 for benign, 4 for malignant)

Figure 1. Breast Cancer Wisconsin Dataset Attributes

4. Algorithm

[Discuss in brief the algorithm the equation, the working and the crux of the algorithm why it is more suited for this task]

The perceptron learning algorithm is as follows:

For a single neuron the output y is

$$y = \sum_{j=1}^{M} w_i x_i + w_0 \quad (1)$$

$$= \mathbf{w^T x} + w_0 \quad (2)$$

$$= \mathbf{w^T x} \quad (3)$$

where w and x are the agumented vectors that include the bias weight and threshold.

During training the weight of the connections are adjusted until the output of the perceptron matched the desired output of the training data. Initially random weights are assigned, then the predicted output $\hat{y_k}$ is computed and the weight w_j is updated.

$$w_j^{(k+1)} = w_j^{(k)} + \eta(y_i - \hat{y}_i^{(k)})x_{ij} \tag{4}$$

where $w_i^{(k)}$ is the weight associated with the i^{th} input at the k^{th} iteration. η is known as the learning rate and w_{ij} is the value of the j_{th} attribute of training parameter $\mathbf{x}_i$.

5. Experiments

The data has two classes Benign and Malignant are taken for training. The perceptron algorithm was run for 100 epochs. The squared error criteria was chosen as the error function. The training was stopped when the error reached 0.001. The system was tested using the test data which we seprated from the dataset using crossvalidation. The performance of the system is shown in Fig.3 with respect to error. The decision surface is seen in Fig.4.

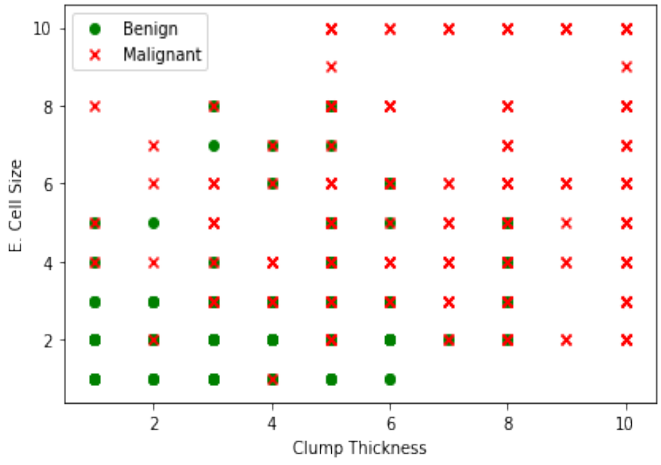

Figure 2.

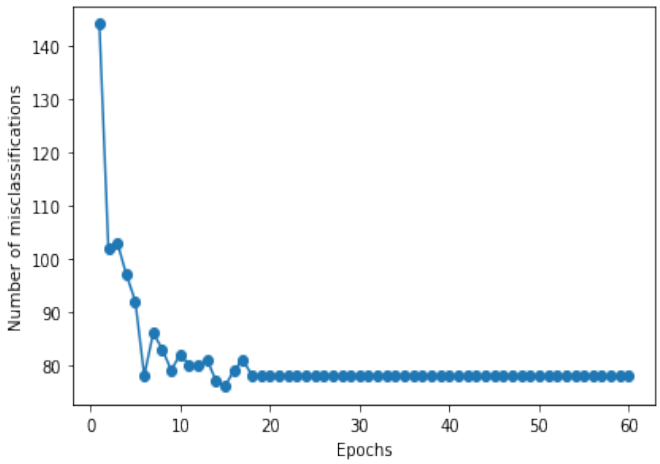

Figure 3. Train error vs Number of Epochs.

After the classification we can see there are many datapoint which belong to wrong category.As we can see in Fig.3 the Perceptron worked fine to classify the dataset by minimizing the mean squared error.So over the iterations we can see that the number of misclassified data decreases from about 150 to 80 and this accounts for aprroximatelty 11% cases

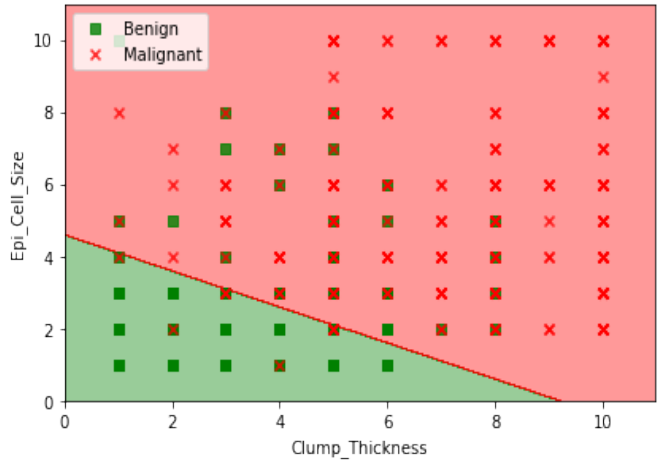

Figure 4. Two class classification performance

6. Conclusion

After the classification we can see there are many datapoint the classification was apt although there were datapoints which belong to wrong category.This indeed suggest that the perceptron model is a simple model which helps to classifify the data into two class Benign and Malignant but it is not the best model to classify the Breast Cancer Dataset. The system developed performs with 88.55% insample accuracy in a two class classification