Prediction of Malignant and Benign Tumor using Machine Learning

Ashish Shah
Department of Computer
Science and Engineering
Manipal Institute of Technology,
Manipal University,
Manipal, Karnataka, India

ABSTRACT

Machine Learning is a branch of Computer Science that is concerned with designing systems that can learn from the provided input. Supervised Machine Learning is where the system needs to be first trained using already classified training data as opposed to an unsupervised system where no such training is required. Supervised learning comprises of 2 training techniques. Linear Regression predicts a continuous valued output. Logistic Regression, more commonly known as Classification predicts a discrete valued output. It is the algorithm for identifying to which of a set of categories a new observation belongs. In this paper we aim to assess whether a lump in a breast could be malignant (cancerous) or benign (non-cancerous) by Classification. The 2 features under consideration are Clump Thickness and Marginal Adhesion. Clump Thickness helps us detect cancerous cells as they are often grouped in multilayers whereas benign cells tend to be grouped in monolayers. Normal cells tend to stick together but Cancerous cells tend to lose this ability. So loss of Marginal Adhesion is a sign of malignancy. With the help of the sigmoid function, we find the Cost function of our data and minimize the sum of the squared errors over the training set. Using Gradient Descent we find the global minimum of our Cost function and then calculate the parameters that fit our data. Finally we estimate the probability of the patient’s tumor being malignant or benign based on the values of these 2 features and the parameters.

General Terms

Supervised Machine Learning, Logistic Regression, Gradient Descent, Cost Function

Keywords

Keywords are your own designated keywords which can be used for easy location of the manuscript using any search engines.

1. INTRODUCTION

Machine learning is a field of computer science and statistics that deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions. It is employed in a range of computing tasks where designing and programming explicitly or following rule- based algorithms is infeasible. It is generally divided into 2 categories, Supervised Machine Learning and Unsupervised Machine Learning.

In supervised learning, the computer is presented with example inputs and their desired outputs. The goal is to learn a general rule that maps inputs to outputs. Spam filtering is an example of supervised learning, in particular classification, where the learning algorithm is presented with email messages labeled beforehand as "spam" or "not spam". The objective here is to produce a computer program that labels unseen messages as either spam or not.

In unsupervised learning, no labels are given to the learning algorithm, leaving it on its own to group similar inputs through clustering or density estimates of high-dimensional data that can be visualized effectively. Unsupervised learning can be a goal in itself discovering hidden patterns in data.

Supervised Machine Learning comprises of 2 training techniques, Linear Regression and Logistic Regression. Linear Regression predicts a continuous valued output whereas Logistic Regression, more commonly known as Classification predicts a discrete valued output. Although many factors affect the malignancy of a tumor, the two features generally considered are Clump Thickness and Marginal Adhesion. These 2 features help us in determining the malignancy of a tumor and eventually help us in predicting the probability of the patient’s tumor being malignant or benign. Other features include cell size, cell shape and texture of the chromatin.[1]

Analyzing the dependent variable Y and the independent variables X, the goal is to find the value of ϴ, also known asthe parameter coefficient so that we can fit the model on our data, find the decision boundary and the malignancy of the tumor.

The Cost function is used as a measurement parameter of logistic regression model. We have to keep changing the value of ϴ to minimize the cost. Gradient descent is used to minimize the cost. Every time the value of ϴ is updated until cost is minimized and we arrive at the global minimum. After finding the parameter ϴ, based on the decision boundary we estimate the probability of a patient’s tumor being malignant or benign.[2]

2. RELATED WORK

There are several applications for Machine Learning, the most significant of which is Data Mining. People are often prone to making mistakes during analysis or when trying to establish relationships between multiple features.

There have been many attempts by researchers to build an accurate model for predicting the malignancy of a tumor. One such model was built by Roberto Lopez.[1] His learning algorithm was trained by a dataset from the University of Wisconsin-Madison.[2] His algorithm was trained by approximately 600 patients covering over 10 features with a training accuracy of 71.

Thus the main problem with such training models is the low accuracy achieved due to anomalies in the dataset as well as the gradient descent ending up at the local minima instead of the global minima.

As we increase the number of features to be considered, the complexity of our model increases which may become more difficult to visualize.

In this paper, two features have been considered, namely Clump Thickness and Marginal Adhesion. The data for this problem has been taken from the UCI Machine Learning Repository. Data of over 100 patients was recorded and analyzed by our learning algorithm. We try to estimate the probability and predict whether the patient's tumor is malignant or benign.

3. MACHINE LEARNING ALGORITHMS

Regression is widely used for prediction. Focus of regression is on the relationship between dependent variable and independent variables. Regression analysis helps us understand how the value of the dependent variable can affect the value of another variable. When one of the independent variables is varied, while the other independent variables are held fixed, the value of the dependent variable changes. In regression, the dependent variable is estimated as function of independent variables which is called regression function (Y is function of (X, Ө)). In the regression model, following parameters are used.

1. Independent variables X
2. Dependent variable Y
3. Unknown parameter Ө.

3.1 Linear Regression

Linear regression predicts a continuous valued output. The dependent variable Y is linear combination of the independent variables.[3] The regression model is the straight line so we can predict the value of dependent variable from independent variables.

3.2 Logistic Regression

Logistic regression is a type of statistical classification model which is used to predict a binary or a multiclass response.

It measures the relationship between categorical dependent variable and one or more predictor variables. Here categorical variable might be binomial or multinomial. In case of binomial categorical variable, we have only two categories ('yes' and 'no', 'good" and "bad").

Whereas, in case of the multinomial categorical variable, we have more than two categories ("average", "good" and "best").

We can represent it in following mathematical notation. y = {0, 1}

Let us consider the case of the Spam detector which is a classification problem. Here Detector system will identify whether a given mail is spam or not spam. So our dependent variable will contains only two values "Yes" or "No". In other words, it will be represented in form of positive class (spam) and negative class (not spam). [4]

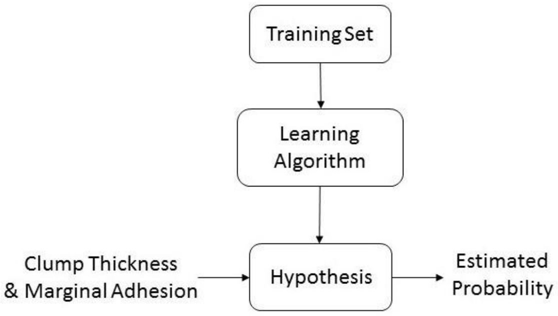

Fig 1: Hypothesis Model

4. HYPOTHESIS REPRESENTATION

We will be building a logistic regression model to predict whether a patient's tumor is malignant or benign based on the values of the following 2 features.

A. Clump Thickness
It helps us detect cancerous cells as they are often grouped in multilayers whereas benign cells tend to be grouped in monolayers.

B. Marginal Adhesion
Normal cells tend to stick together but Cancerous cells tend to lose this ability. So loss of Marginal Adhesion is a sign of malignancy.

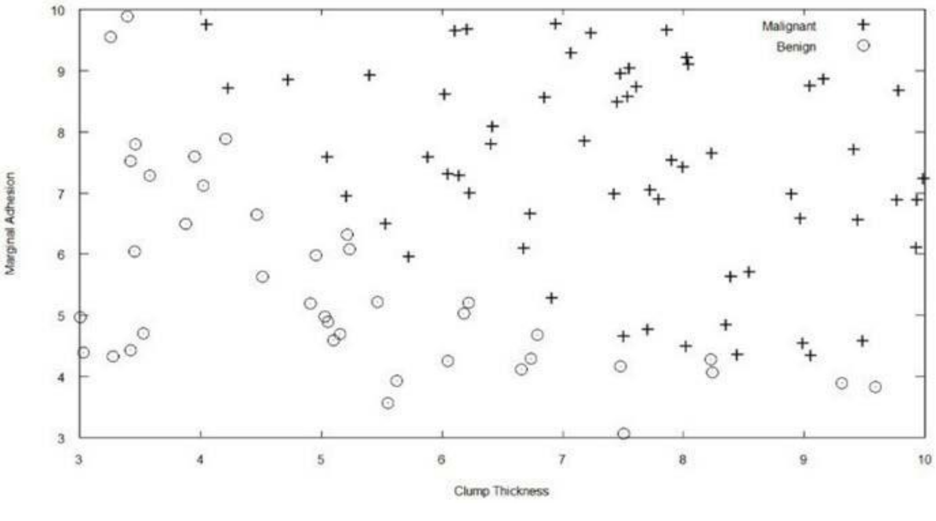

Fig 2 : Plotting the dataset

4.1 Visualizing the Data

Let us understand the data. In the dataset, we have records of previous patients' two test results for Clump Thickness and Marginal Adhesion respectively and a label which indicates whether the patient's tumor is malignant or benign. (0 or 1). A binary classification model has 2 classes.

0: Negative Class (Benign) 1: Positive Class (Malignant)

Before starting to implement any learning algorithm, let us visualize the data. So, let us plot the dataset and try to understand the test results.

4.2 Building the Hypothesis Classifier

In a binary classification model since y can take only 2 values, namely 0 and 1, this indicates that our hypothesis classifier will be in the range 0 to 1. [5]

$$0 \leq h_\theta(x) \leq 1$$

We want prediction in the range 0 to 1. So let us try to interpret the result of hΘ(x). For example, if the output result for our hypothesis of tumor detection equals 0.7, then it represents 70% probability of being malignant. Finally, we want to set some threshold for deciding upon whether given tumor is malignant or benign. Generally, if the probability is greater than 0.5 then it should be classified as malignant otherwise it is classified as benign.

We can say that total probability of tumor being malignant or benign is equal to 1. We can write this in following form.

P(Y=0) + P(Y=1) = 1 So, P(Y=0) = 1 – P(Y=1)

The mathematical definition is denoted as below,

hΘ(x) = P(y=1 | x ; θ)

This is the estimated probability that y=1 in an input given

that x is parameterized by θ.

4.3 Sigmoid Function

Let us discuss on the sigmoid function which is the central part of the logistic regression model.

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

And using this we define our new hypothesis as below.

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

For large positive values of x, the sigmoid should be close to 1, while for large negative values, the sigmoid should be close to 0. Evaluating sigmoid for 0 should give you exactly 0.5.

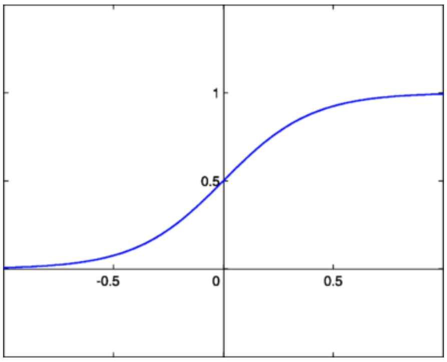

Fig 3: Sigmoid Function

4.4 Decision Boundary

Analyzing the sigmoid function, we arrive at two conditions that help us in predicting the classifier. After estimating the parameters we get a decision boundary which approximately separates the data into 2 classes, namely the positive class and the negative class. [6]

First, if y=1 this means hΘ(x)>0.5. According to the sigmoid

function this indicates g(z)>0.5 when z>0.

Hence, we arrive at the conclusion that when y=1 then ΘTX>0. Likewise, if y=0 this means hΘ(x)<0.5. According to the sigmoid function this indicates g(z)<0.5 when z<0.

Hence, we arrive at the conclusion that when y=0 then

ΘTX<0.

4.5 Cost Function

The goal is to find the value Θ known as coefficient parameter so that we can fit the model on our data. So the cost function helps us find the right Θ in the best possible time so that our decision boundary fits our case. We can predict the value of dependent variable from independent variables.[7] Starting with Θ's value as zero, we find that the difference between actual and predicted value is huge. So the Cost function is used as a measurement parameter of our logistic regression model. Cost function is defined as below.

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

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

Let us now analyze the intuition behind the above stated cost function. If we predict the tumor to be malignant (hΘ(x)=1) and it is indeed malignant (y=1) then the cost will be 0. But if we predict the tumor to be benign (hΘ(x)=0) but it is actually malignant (y=1) then cost will tend towards infinity and we will end up penalizing the learning algorithm by a very large cost.

Now let us see what happens when y=0. If we predict the tumor to be benign (hΘ(x)=0) and it is indeed malignant (y=0) then the cost will be 0. But if we predict the tumor to be malignant (hΘ(x)=1) but it is actually benign (y=0) then cost will tend towards infinity and we will end up penalizing the learning algorithm by a very large cost.

4.6 Gradient Descent

For every value of Θ we get a different value of the cost function (J(Θ)).[8] The optimization objective for our algorithm is to choose a value of Θ which minimizes J(Θ).

We start with some value of Θ. And we keep changing Θ to reduce J(Θ) until we hopefully end up at a minimum. We have to change the values of Θs to minimize cost. Gradient descent is used to minimize the cost.

Where,
$H_\theta(x)=g(\theta^T x)$

And,

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

5. IMPLEMENTING THE MODEL

Let the training set be of the form (x1,y1), (x2,y2)…(xn,yn) with n features and m training examples, X comprises of the values of the 2 features under consideration. Y comprises of the classifier output (0 or 1).

5.1 Initializing the data

```
data = load('data.txt'); X = data(:, [1, 2]);

[m, n] = size(X);

y = data(:, 3);
```

5.2 Implementing the Sigmoid Function

```
g=1./(1+exp(-z));
```

5.3 Implementing the Cost Function

```
m = length(y); J = 0; grad = zeros(size(theta));

J = 1./m * ( -y' * log( sigmoid(X * theta) ) - ( 1 - y' ) * log ( 1 - sigmoid( X * theta)) )

grad = 1./m * X' * (sigmoid(X * theta) - y)
```

5.4 Cost Function & Gradient Descent

```
X = [ones(m, 1) X]; initial_theta = zeros(n + 1, 1); [cost, grad] = costFunction(initial_theta, X, y);
```

5.5 Optimizing using Fminunc

Octave's fminunc is an optimization solver that finds the minimum of an unconstrained function. For logistic regression, you want to optimize the cost function J(θ) with parameters θ.

Concretely, you are going to use fminunc to find the best parameters θ for the logistic regression cost function, given a fixed dataset (of X and y values).[9]

```
options = optimset('GradObj', 'on', 'MaxIter', 400);

[theta, cost] = fminunc(@(t)(costFunction(t, X, y)), initial_theta, options);

plotDecisionBoundary(theta, X, y);
```

6. EVALUATION & PREDICTION

1. Test Case 1

Let us consider an example of a patient whose clump thickness is 4.5 and marginal adhesion is 8.5.

We need to predict the probability of the patient's tumor being malignant or benign based on the values of these 2 features and the parameters.

```
prob = sigmoid([1 4.5 8.5] * theta);
```

We estimate the probability to be 0.776. Thus the patient has a 77.6% chance of having a malignant tumor.

2. Test Case 2

Let us consider an example of a patient whose clump thickness is 5.5 and marginal adhesion is 7.9.

We need to predict the probability of the patient's tumor being malignant or benign based on the values of these 2 features and the parameters.

```
prob = sigmoid([1 5.5 7.9] * theta);
```

We estimate the probability to be 0.896. Thus the patient hasa 89.6% chance of having a malignant tumor.

3. Test Case 3

Let us consider an example of a patient whose clump thickness is 3.0 and marginal adhesion is 4.2.

We need to predict the probability of the patient's tumor being malignant or benign based on the values of these 2 features and the parameters.

```
prob = sigmoid([1 3.0 4.2] * theta);
```

We estimate the probability to be 0.311. Thus the patient hasa 31.1% chance of having a malignant tumor.

Next we find the Train Accuracy of our Logistic Regression model. Train Accuracy is the extent with which we are able to correctly predict whether a tumor is malignant or benign.[10]

```
p = predict(theta, X);

Train Accuracy: mean(double(p == y)) * 100;
```

We were able to achieve a train accuracy of 89 for our implemented Classification model.

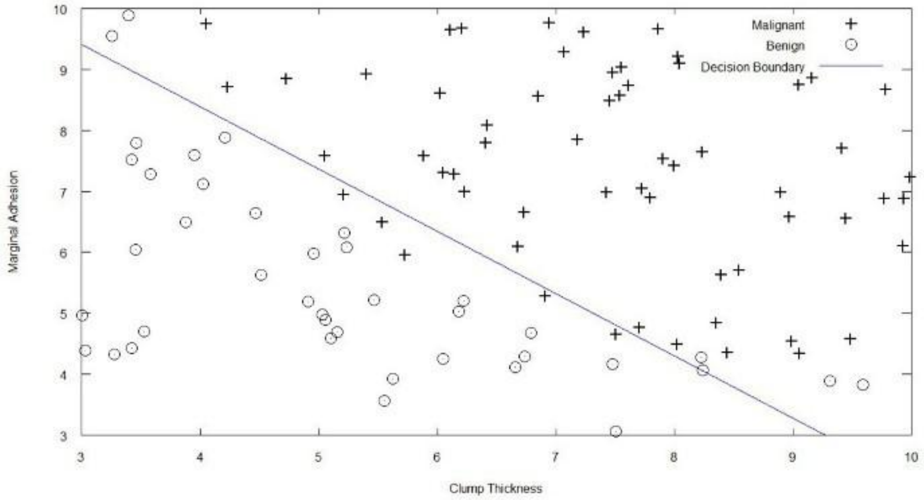

Fig 4 : Division of dataset using gradient descent

7. RESULTS

Table 1 Probability Prediction Table

ID	Clump Thickness	Marginal Adhesion	Probability	Result
1	4.5	8.5	77.6	Malignant
2	5.5	7.9	89.6	Malignant
3	3.0	4.2	31.1	Benign

Our calculation of the probability whether a tumor is malignant or benign is based on 2 features, Clump Thickness and Marginal Adhesion. With appropriate values substituted in the formula, we achieve the probability if a tumor is malignant or benign. If the probability is more than 50%, the tumor is malignant(cancerous), and if the probability is less than 50%, it is benign(non-cancerous). Hence we can see that the first 2 cases have predicted that the tumor is malignant while the last test case has predicted that the tumor is benign. Using the cost function and minimizing the sum of the squared errors over the training set, the probability is achieved. Using Gradient Descent, the global minimum of our Cost function is found and then the parameters that fit our data are calculated. Finally the estimation of the probability of the patient's tumor being malignant or benign is done based on the values of these 2 features and the parameters.

With an accuracy of 89%, the model can successfully predict whether the tumor is malignant or benign.

8. CONCLUSION

We get a new value of cost function for every value of ϴ. Our calculation of the probability whether a tumor is malignant or benign is based on 2 features, Clump Thickness and Marginal Adhesion. With appropriate values substituted in the formula, we achieve the probability if a tumor is malignant or benign. If the probability is more than 50%, the tumor is malignant(cancerous), and if the probability is less than 50%, it is benign(non-cancerous). Using Gradient Descent we try to minimize this value and choose the appropriate parameters which fit out data. With the help of the decision boundary we are able to separate the binary classification into two classes, malignant and benign. This helped us in estimating the probability and predicting whether the tumor is malignant or benign based on the features considered. Also, we were able to achieve a train accuracy of 89 for our implemented model.

The future scope of this model is to incorporate more features into the model which will take into account other parameters and features as well. With more features, the model will be able to predict more accurately whether a tumor is malignant or benign. Integrating other features apart from Marginal Adhesion and Clump Thickness will help doctors and analysts use this model more extensively in their work. With the incorporation of other features which affect the malignancy of a tumor, w can plot a more detailed graph involving the gender and age of the patient as well.

9. REFERENCES

[1] Text Categorization Through Probabilistic Learning:Applications to Recommender Systems, Paul N. Bennett, Department of Computer Sciences, University of Texas at Austin, May 1998.

[2] Reviews of Machine Learning by Ryszard S. Michalski, Jaime Carbonell and Tom Mitchell, Tiago Publishing Company

[3] Dr. BD Prasad, PE Krishna Prasad and Y Sagar, "A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis" – CCSIT 2011, CCIS 131

[4] Anderson, JR & Matessa, M (1992) Explorations of an incremental Bayesian algorithm forcategorization,Machine Learning

[5] Amar Gondaliya, Logistic Regression with R: Step Implmentation 2013.

[6] Pattern Analysis and Machine Intelligence, Stephen Della Pietra, Vincent Della Pietra and JohnLafferty

[7] A Theory of Learning Classification Rules,Dissertation, Dept of Computer Science, University ofTechnology, Sydney.

[8] Duda Ro & Hard PE, Pattern Classification and Scene Analysis, New York

[9] Proc. Of Innovative Application of Machine Learning, Ellen Spertus (1997)

[10] Implementation Of Clustering Through Machine Learning Tool, Sree Ram Nimmagadda, Phaneendra Kanakamedala And Vijay Bashkarreddy Yaramala