University of Mumbai

Diagnosis of disease using Machine Learning Algorithms.

Submitted at the end of semester VIII in partial fulfillment of requirements

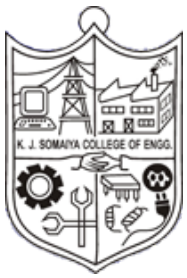
For the degree of

Bachelors in Technology

by

Karan Harjai
Roll No: 1512079
Jay Shah
Roll No: 1512043

Guide- Prof. Sushma Kadge



Department of Electronics Engineering
K. J. Somaiya College of Engineering, Mumbai-77
(Autonomous College Affiliated to University of Mumbai)

Batch 2015 -2019

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Certificate

This is to certify that the dissertation report entitled "**Diagnoses of the disease by Machine Learning Algorithms"** submitted by Jay Shah and Karan Harjai at the end of semester VIII of LY B. Tech is a bona fide record for partial fulfillment of requirements for the degree of Bachelors in Technology in Electronics Engineering of University of Mumbai

_____ _____

Guide Head of the Department

_____

Principal

Date:

Place: Mumbai-77

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Certificate of Approval of Examiners

We certify that this project report entitled "**Diagnoses of the disease by Machine Learning Algorithms"** submitted by Jay Shah and Karan Harjai at the end of semester VIII of LY B. Tech is a bona fide record for partial fulfillment of requirements for the degree of Bachelors in Technology in Electronics Engineering of University of Mumbai

_____ _____

External Examiner Internal Examiner

Date:

Place: Mumbai-77

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

DECLARATION

We declare that this written report submission represents the work done based on our and / or others' ideas with adequately cited and referenced the original source. We also declare that we have adhered to all principles of intellectual property, academic honesty and integrity as we have not misinterpreted or fabricated or falsified any idea/data/fact/source/original work/ matter in my submission.

We understand that any violation of the above will be cause for disciplinary action by the college and may evoke the penal action from the sources which have not been properly cited or from whom proper permission is not sought.

Signature of the Student **Roll No.**	**Signature of the Student** **Roll No.**
Signature of the Student **Roll No.**	**Signature of the Student** **Roll No.**

Date:

Place: Mumbai-77

Acknowledgements:

First of all,we would like to express our deep sense of respect and gratitude towards our advisor and guide Prof. Sushma Kadge, who has been the guiding force behind this work.

We are indebted to her for her constant encouragement for propelling us further in every aspect of our academic life. We consider our good fortune to have got an opportunity to work with such a wonderful person.

We would like to thank all faculty members, staff and HOD of the Department of Electronics Engineering,K.J.Somaiya College of Engineering,Mumbai for their generous help in various ways.

We would also like to thank Prof. Ankita Modi for giving her valuable advices.

We are specially indebted to our parents for their love,sacrifice and support. They are our first teachers after we came to this world and have set great examples for us about how to live,study and work.

We would like to thank all our friends and especially our classmates for all the thoughtful and mind stimulating discussions we had which prompted us to think beyond the obvious.

Abstract

The number and size of medical databases are increasing rapidly but most of these data are not analyzed for finding the valuable and hidden knowledge. Advanced data mining techniques can be used to discover hidden patterns and relationships.

Models developed from these techniques are useful for medical practitioners to make right decisions.

The research studied the application of data mining techniques to develop predictive models for breast cancer recurrence in patients who were followed-up for two years.

The objective of this thesis work is to implement various techniques like logistic regression and Artificial Neural ression, Network (ANN) to develop the predictive models.

The main goal is to compare the performance of some well-known algorithms on our data through sensitivity, specificity, and accuracy.

Key words**:** Artificial neural networks(ANN);
Breast cancer recurrence,
Breast cancer (BC) .

Contents

Nomenclature

Tp True positive

Tn True negative

Fp False positive

Fn False negative

C Category to which documents belong (C_1 or $C_{0)}$

T Task for function

D Decision function

E Experience

P Probability of the program

EP Evaluation parameter combining precision and recall

ID Classification used for a data sample

Aij The element in the ith row and jth column of matrix A.

j=0,1 The feature index number.

Hθ The direction in the step of partial derivative.

vi The element in the ith row of the vector.

MF Multinomial model feature vector for the j^{th} document in testing

Introduction:

There's a science of getting computers to learn without being explicitly programmed.A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T,as measured by P, improves with experience E."
The second major cause of women's death is breast cancer (after lung cancer) 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women.

Information and Communication Technologies (ICT) can play potential roles in cancer care. In fact, Big data has advanced not only the size of data but also creating value from it; Big data, that becomes a synonymous of data mining, business analytics and business intelligence, has made a big change in BI from reporting and decision to prediction results4. Data mining approaches, for instance, applied to medical science topics rise rapidly due to their high performance in predicting outcomes, reducing costs of medicine, promoting patients' health, improving healthcare value and quality and in making real time decision to save people's lives.

You probably use it dozens of times a day without even knowing it.
Each time you do a web search on Google , that works so well because their machine learning software has figured out how to rank what pages.

When Facebook or Apple's photo application recognizes your friends in your pictures,that's also machine learning.Each time you read your email and a spam filter saves you from having to wade through tons of spam, again,that's because your computer has learned to distinguish spam from non-spam email.

So, that's machine learning.

Well what you can do is have the robot watch you demonstrate the task and learn from that.The robot can then watch what objects you pick up and where to put them and try to do the same thing even when you aren't there.

For us, one of the reasons excited about this is the AI,artificial intelligence problem.

Many scientists think the best way to make progress on this is through learning algorithms called neural networks, which mimic how the human brain works,

2. Related work:

Diagnoses is one of the most important and essential task in machine learning and data mining. About a lot of research has been conducted to apply data mining and machine learning on different medical datasets to classify Breast Cancer. Many of them show good classification accuracy.

Vikas Chaurasia and Saurabh Pal compare the performance criterion of supervised learning classifiers; such as Naïve Bayes, RBF neural networks, Decision trees (J48) and simple CART; to find the best classifier in breast cancer datasets.

Djebbari et al,12 consider the effect of ensemble of machine learning techniques to predict the survival time in breast cancer. Their technique shows better accuracy on their breast cancer data set comparing to previous results.

The accuracy of data mining algorithms like logistic regression, Hebbs, Perceptron etc. Pradesh 15. The performance of SMO shows a higher value compared with other classifiers. T.Joachims reaches accuracy of 96.06% with neuro- fuzzy techniques when using Wisconsin Breast Cancer (original) datasets.

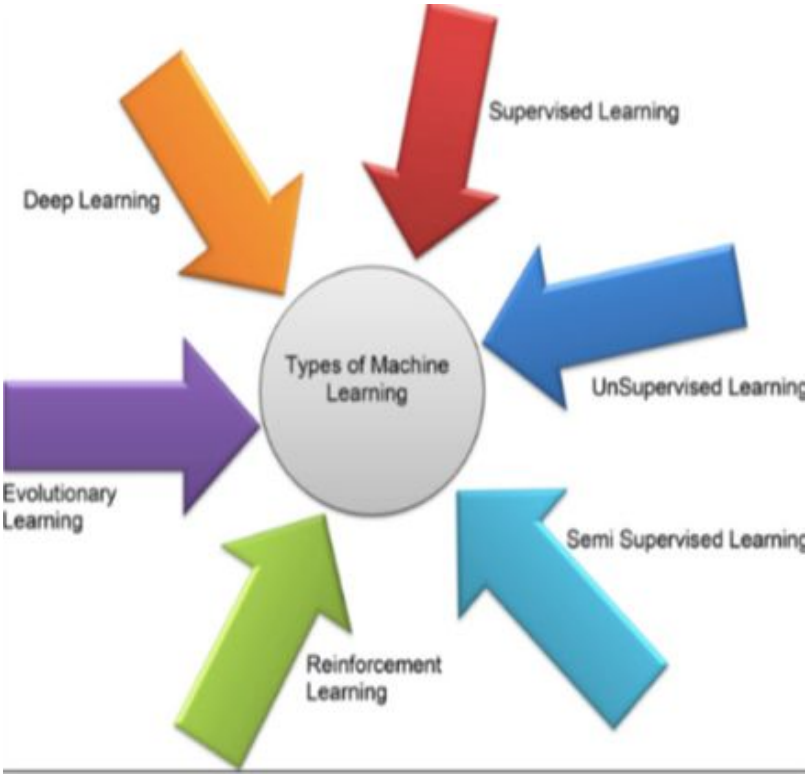
In this study, a hybrid method is proposed to enhance the classification accuracy of Wisconsin Breast Cancer (original) datasets (95.96) with 10 fold cross validation. Liu Ya-Qin's, W.

Cheng, and Z. Lu17 experimented on breast cancer data using C5 algorithm with bagging; by generating additional data for training from the original set using combinations with repetitions to produce multisets of the same size as you're the original data; to predict breast cancer survivability. Delen et al. Lu18 take 202,932 breast cancer patients records , which then pre-classified into two groups of "survived" (93,273) and "not survived" (109,659).

With respect to all related work mentioned above, our work compares the behaviour of data mining algorithms NB, and C4.5 using Breast Cancer datasets in both diagnosis and analysis to make decisions.

The goal is to achieve the best accuracy with the lowest error rate in analysing data.

To do so, we compare efficiency and effectiveness of those approaches in terms of many criteria, including: accuracy, precision, sensitivity and specificity, correctly and incorrectly classified instances and time to build model, among others.



Motivation:

To us it was motivating to learn because finally we could see how all the math we had studied at university is applied in real life, and it's not only interesting, but also very useful.

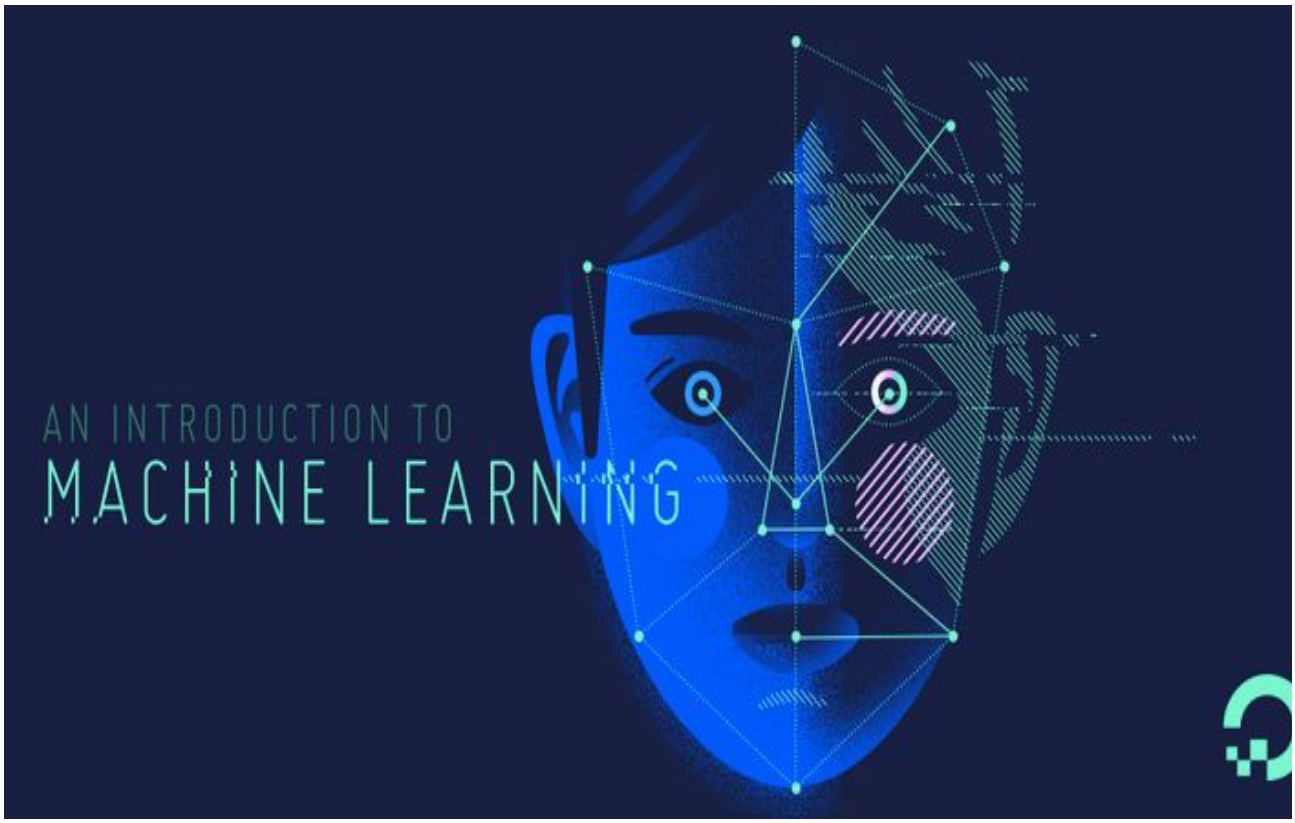
- ➔ Also just the thought that given the data you can extract something useful from it is already very motivating.
- ➔ Machine Learning helps us to understand that instead of writing code, you feed data to the generic algorithm,and it builds logic based on the data given.

Scope of the Project:

Our analysis uses ANN,linear and logistic regression.

The logistic regression classification model predicts breast cancer recurrence with least error rate and highest accuracy(as per literature survey).

The predicted accuracy of the ANN model is the lowest of all. The results are achieved using 10-fold cross-validation for measuring the unbiased prediction accuracy of each model.



Literature Survey:

A literature review showed that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and recurrence using data mining approaches such as ANN.

Delen et al. used artificial neural networks, decision trees and logistic regression to develop prediction models for breast cancer survival by analyzing a large dataset, the SEER cancer incidence database .Lundin et al. used ANN and logistic regression models to predict 5, 10, and 15 -year breast cancer survival. They studied 951 breast cancer patients and used tumor size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis, and age as input variables. Pendharker et al. used several data mining techniques for exploring interesting patterns in breast cancer. In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis, and treatment purposes.

These studies are some examples of researches that apply data mining to medical fields for prediction of diseases.

Corresponding author: Leila Ghasem Ahmad, Department of Management Information Systems, Science and Research Branch, Islamic Azad University of a literature review showed that there have been several studies on the Tehran-Iran, Iran, Survival prediction problem using statistical approaches and artificial.

Received January 28, 2013; Accepted April 18, 2013; Published April 24, 2013 neural networks. However, we could only find a few studies related.

Citation: Ahmad LG, Eshlaghy AT, Pourebrahimi A, Ebrahimi M, Razavi AR to medical diagnosis and recurrence using data mining approaches (2013) Using Three Machine Learning Techniques for Predicting Breast Cancer such as decision trees. Delen et al. used artificial neural networks,

Recurrence. J Health Med Inform 4: 124. doi:10.4172/2157-7420.1000124 decision trees and logistic regression to develop prediction models for determining accuracy and precision of well known algorithms.

J Health Med Inform ISSN: 2157-7420 JHMI, an open access journal.

A computer aided medical diagnosis system generally consists of a knowledge base and a method for solving an intended problem. On the basis of the query posted to the system, it provides assistance to the physicians in diagnosing the patients accurately. The knowledge base of such medical systems relies on the inputs that spring up from the clinical experience of field experts. Knowledge acquisition is the process of transforming human expert knowledge and skills acquired through clinical practice to software. It is quite time consuming and labor intensive task.

Common methods like Case Based Reasoning (CBR) solves the knowledge acquisition problem to some extent because the past records are maintained in a database, including possible remedies, past clinical decisions, preventive measures and expected diagnostic outcome measures.

During Patient diagnosis, the clinical database is matched for analogous past patient‘s record for taking suitable decisions.

Software reliability is defined as the probability that a system will not have a failure over a specified period of time under specific conditions. The knowledge of software reliability is very vital in critical systems because it indicates the design perfection .

In this work, the primary aim is to enhance the software reliability of the computer aided diagnosis systems using machine learning algorithms.A researcher V embandasam yet al. (Vembandasamy.K, 2015) played out a work, to analyze coronary illness. In this the alg Bayes algorithm. In Naïve hypothesis is utilized. Hence effective freedom presumptive collection is gotten from a st driving diabetic research or Tamilnadu. There are more dataset. The device utilized is executed by utilizing 70% exactness offered by Naive Bay The data mining approach applied by the researchers (Chaurasia.V , 2013) to determine information mining device contains an arrangement of ma with the end goal of mining. Bayes, J48 and bagging are informational set is given by that comprises of 76 traits. attributes are used.82.31% pre Bayes. J48 gives 84.35% o exactness is accomplished by l better classification factor.

The researchers Parthib orithm utilized was Naive Bayes algorithm Bayes rth, Naive Bayes has an on. The utilized data ndout amongst the most ganizations in Chennai, han 500 eka and classification is Percentage Split.

In 2006 a researcher M.Peleg,S.tu (M.peleg, 2006) has given a paper named Decision Support ,Knowledge Representation and Management. The clinical decision support is complete program designed to help the health professionals in making clinical decisions. The system has been considered as an active knowledge system. The main objective of the modern clinical system is to assist clinicians at the point of care. The objective of the system is to give the needed information with the health care’s organizational dynamics.

Decision support systems are implemented by standardization in information system infrastructure.

The system give sits support in the complex tasks of differential diagnosis and the therapy planning. The system has to work on the knowledge modeling task in which modelers gives the medical knowledge that enables the system to deliver appropriate decision support system.

Xin Yao et al. 1999 has attempted to implement neural network for breast cancer diagnosi Negative correlation training algorithm was used to decompose a problem automatically ar solve them.

In this article the author has discussed two approaches such as evolutionary approach and ensemble approach, in which evolutionary approach can be used to design compact neura network automatically.

The ensemble approach was aimed to tackle large problems but it was in progress.

Dr.S.Santhosh baboo and S.Sasikala have done a survey on data mining techniques for gene selection classification.

This article dealt with most used data mining techniques for gene selection and cancer classification, particularly they have focused on four main emerging fields.

They are neural network based algorithms, machine learning algorithms, genetic algorithm and cluster based algorithms and they have specified future improvement in this field.

David B.fogel et al. has discussed the evolving neural networks for detecting breast cancer and the related works used for breast cancer diagnosis using back propagation method wit multilayer perceptron.

In contrast to back propagation David B.fogel et al. found that evolution computational method and algorithms were used often, outperform more classic optimization techniques.

Project Design:

Logistic Regression-

Cost Function:

We can measure the accuracy of our hypothesis function by using a cost function.

In linear regression, we have a training set, like the one plotted.

For the parameters theta zero and
theta one so that the straight line we get out of this, corresponds to
a straight line that somehow fits the data well, like maybe that line there.

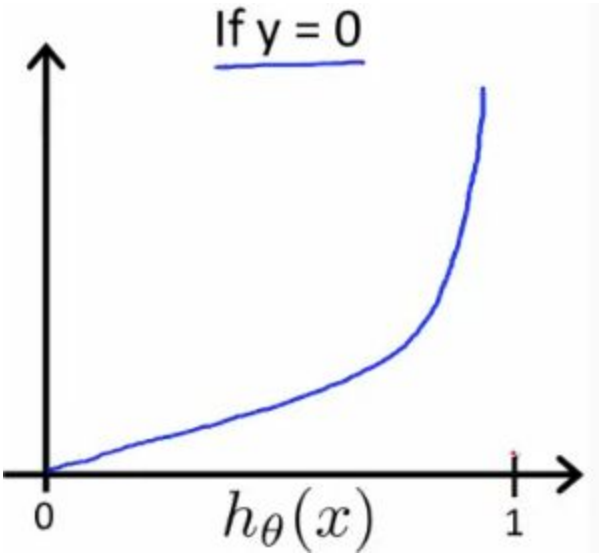
There are other cost functions that will work pretty well.

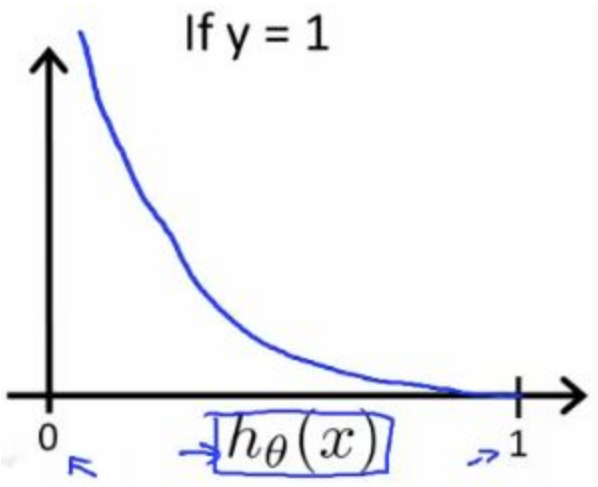
But the square cost function is probably the most commonly used one for regression problems.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if y} = 1$$

$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if y} = 0$$





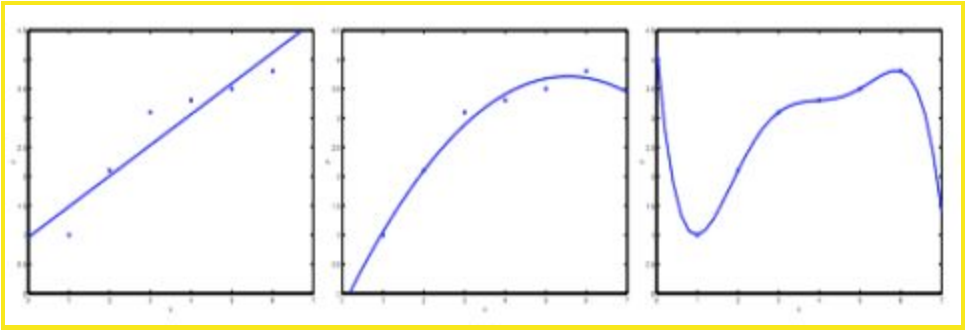
Gradient Descent :

So we have our hypothesis function and we have a way of measuring how well it fits into the data. Now we need to estimate the parameters in the hypothesis function.

That's where gradient descent comes in.

The point of all this is that if we start with a guess for our hypothesis and then repeatedly apply these gradient descent equations, our hypothesis will become more and more accurate.

The Problem of Overfitting:



This terminology is applied to both linear and logistic regression.

There are two main options to address the issue of overfitting:

1) Reduce the number of features:

- Manually select which features to keep.
- Use a model selection algorithm (studied later in the course).

2) Regularization

a) Keep all the features, but reduce the magnitude of parameters.

b) Regularization works well.

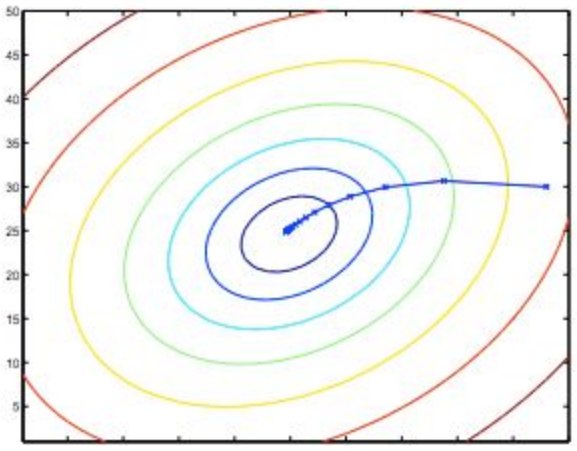
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

repeat until convergence: {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left((h_\theta(x_i) - y_i) x_i \right)$$

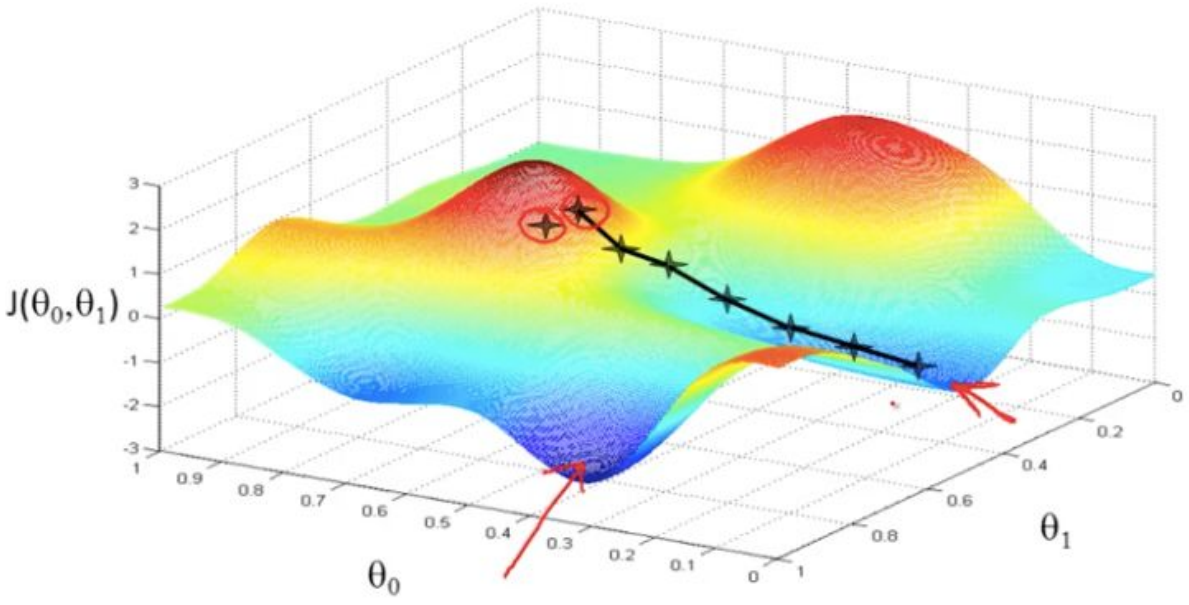
}



The ellipses shown above are the contours of a quadratic function.

Also shown is the trajectory taken by gradient descent, which was initialized at (48,30).

The x's in the figure (joined by straight lines) mark the successive values of θ that gradient descent went through as it converged to its minimum.



The way we do this is by taking the derivative (the tangential line to a function) of our cost function. The slope of the tangent is the derivative at that point and it will give us a direction to move towards. We make steps down the cost function in the direction with the steepest descent. The size of each step is determined by the parameter α, which is called the rate.

Hypothesis Representation:

We could approach the classification problem ignoring the fact that y is discrete-valued, and use our old linear regression algorithm to try to predict y given x.

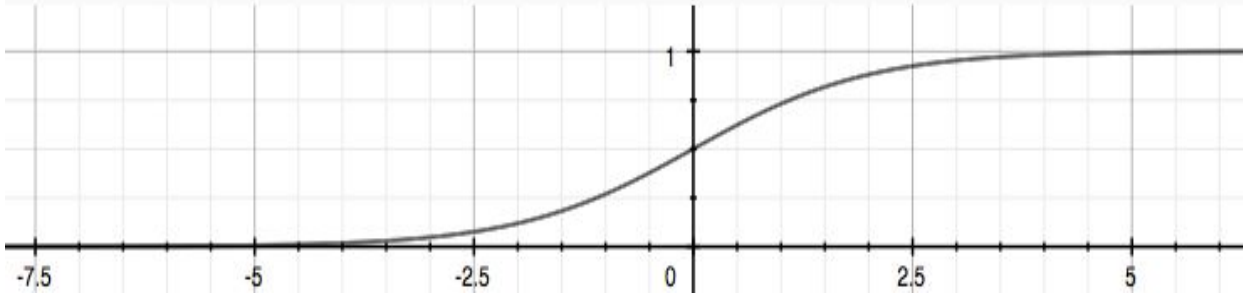
However, it is easy to construct examples where this method performs very poorly. Intuitively, it also doesn't make sense for hθ(x)*hθ*(*x*) to take values larger than 1 or smaller than 0 when we know that y ∈ {0, 1}. To fix this, let's change the form for our hypotheses hθ(x)*hθ*(*x*) to s.

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

$$z = \theta^T x$$

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

The following image shows us what the sigmoid function looks like:



Decision Boundary:

$$h_\theta(x) \geq 0.5 \rightarrow y = 1$$
$$h_\theta(x) < 0.5 \rightarrow y = 0$$

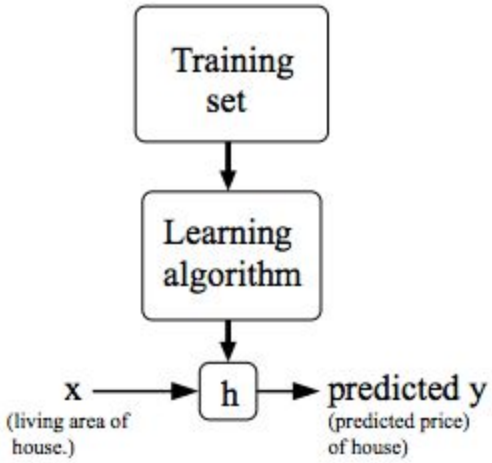
$$g(z) \geq 0.5$$
$$when\ z \geq 0$$

$$z = 0, e^0 = 1 \Rightarrow g(z) = 1/2$$
$$z \rightarrow \infty, e^{-\infty} \rightarrow 0 \Rightarrow g(z) = 1$$
$$z \rightarrow -\infty, e^{\infty} \rightarrow \infty \Rightarrow g(z) = 0$$

Model Representation:

To establish notation for future use, we'll use x(i)$x(i)$ to denote the "input" variables (living area in this example), also called input features, and y(i)$y(i)$ to denote the "output" or target variable that we are trying to predict (price). A pair (x(i),y(i))$(x(i),y(i))$ is called a training example, and the dataset that we'll be using to learn—a list of m training examples (x(i),y(i));i=1,...,m$(x(i),y(i));i=1,...,m$—is called a training set. Note that the superscript "(i)" in the notation is simply an index into the training set, and has nothing to do with exponentiation. We will als use X to denote the space of input values, and Y to denote the space of output values. In this example, X = Y = $\mathbb{R}$.

To describe the supervised learning problem slightly more formally, our goal is, given a training set, to learn a function h : X → Y so that h(x) is a "good" predictor for the corresponding value of y. For historical reasons, this function h is called a hypothesis. See pictorially, the process is therefore like this:



Implementation:

code for gradient descent:

```
function [theta, J_history] = gradientDescent(X, y, theta, alpha, num_iters)
%GRADIENTDESCENT Performs gradient descent to learn theta
%   theta = GRADIENTDESCENT(X, y, theta, alpha, num_iters) updates theta by
%   taking num_iters gradient steps with learning rate alpha

% Initialize some useful values
m = length(y); % number of training examples
J_history = zeros(num_iters, 1);

for iter = 1:num_iters
h = X*theta;
   squ = (h-y);
  theta  = theta - (alpha/m) * (X' * squ);

    J_history(iter) = computeCost(X, y, theta);
end
```

CODE FOR COMPUTING COST FUNCTION:

```
function [J, grad] = costFunctionReg(theta, X, y, lambda)
%COSTFUNCTIONREG Compute cost and gradient for logistic regression with
regularization.
```

```
%   J = COSTFUNCTIONREG(theta, X, y, lambda) computes the cost of using
%   theta as the parameter for regularized logistic regression and the
%   gradient of the cost w.r.t. to the parameters.

% Initialize some useful values
m = length(y); % number of training examples
J = 0;
grad = zeros(size(theta));
h = sigmoid(X*theta);
J = (1/m * (-y'*log(h)-(1-y)'*log(1-h))) + (lambda/(2*m))*sum(theta(2:size(X,2),1).^2) ;
grad(1,1) = 1/m * X(:,1)'*(h-y);
grad(2:size(X,2),1) = 1/m * X(:,2:size(X,2))' *(h-y) + lambda/m * theta(2:size(X,2),1) ;
end
```

UNIVARIATE LINEAR REGRESSION:

Implemented on dataset from a company having knowledge about the population of the city in which it has franchise and whether it have profit or loss in that city.Used to predict whether company will have profit or loss in a new city in which it will be opening its franchise.

OUTPUT :

Running Gradient Descent ...

Theta found by gradient descent:

-3.630291

1.166362

Expected theta values (approx)

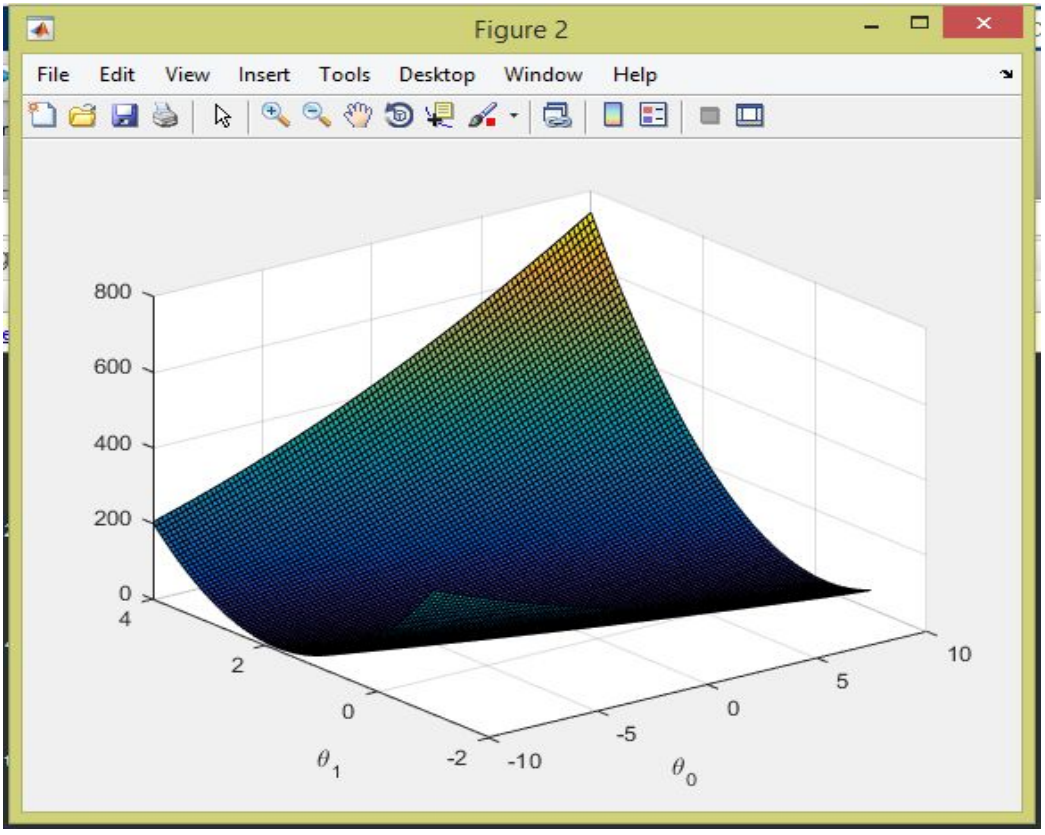
-3.6303

1.1664

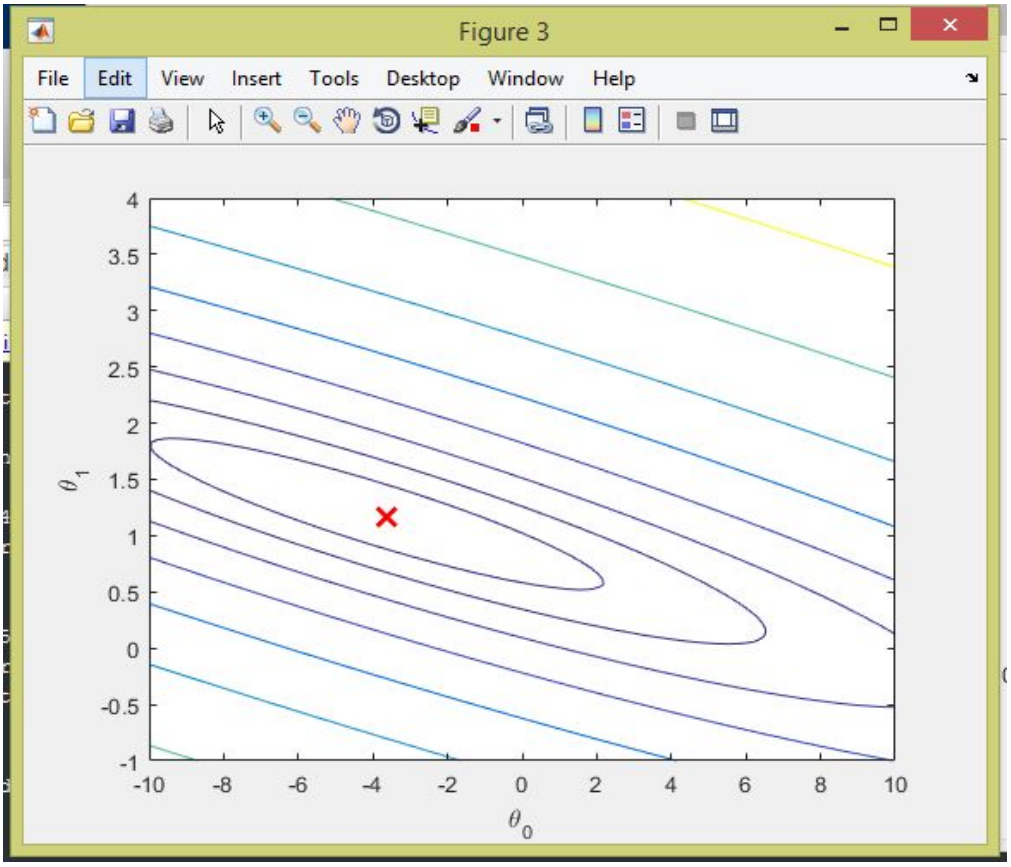
For population = 35,000, we predict a profit of 4519.767868

For population = 70,000, we predict a profit of 45342.450129

Program paused. Press enter to continue.



Visualizing cost function



Contour plot

MULTIVARIATE LINEAR REGRESSION:

Implemented on dataset which contained housing price prediction data.It predicts the price of a house based on the historical data of no of rooms in a house and area of the house.

OUTPUT -

Loading data ...

First 10 examples from the dataset:

x = [2104 3], y = 399900

x = [1600 3], y = 329900

x = [2400 3], y = 369000

x = [1416 2], y = 232000

x = [3000 4], y = 539900

x = [1985 4], y = 299900

x = [1534 3], y = 314900

x = [1427 3], y = 198999

x = [1380 3], y = 212000

x = [1494 3], y = 242500

Program paused. Press enter to continue.

Normalizing Features ...

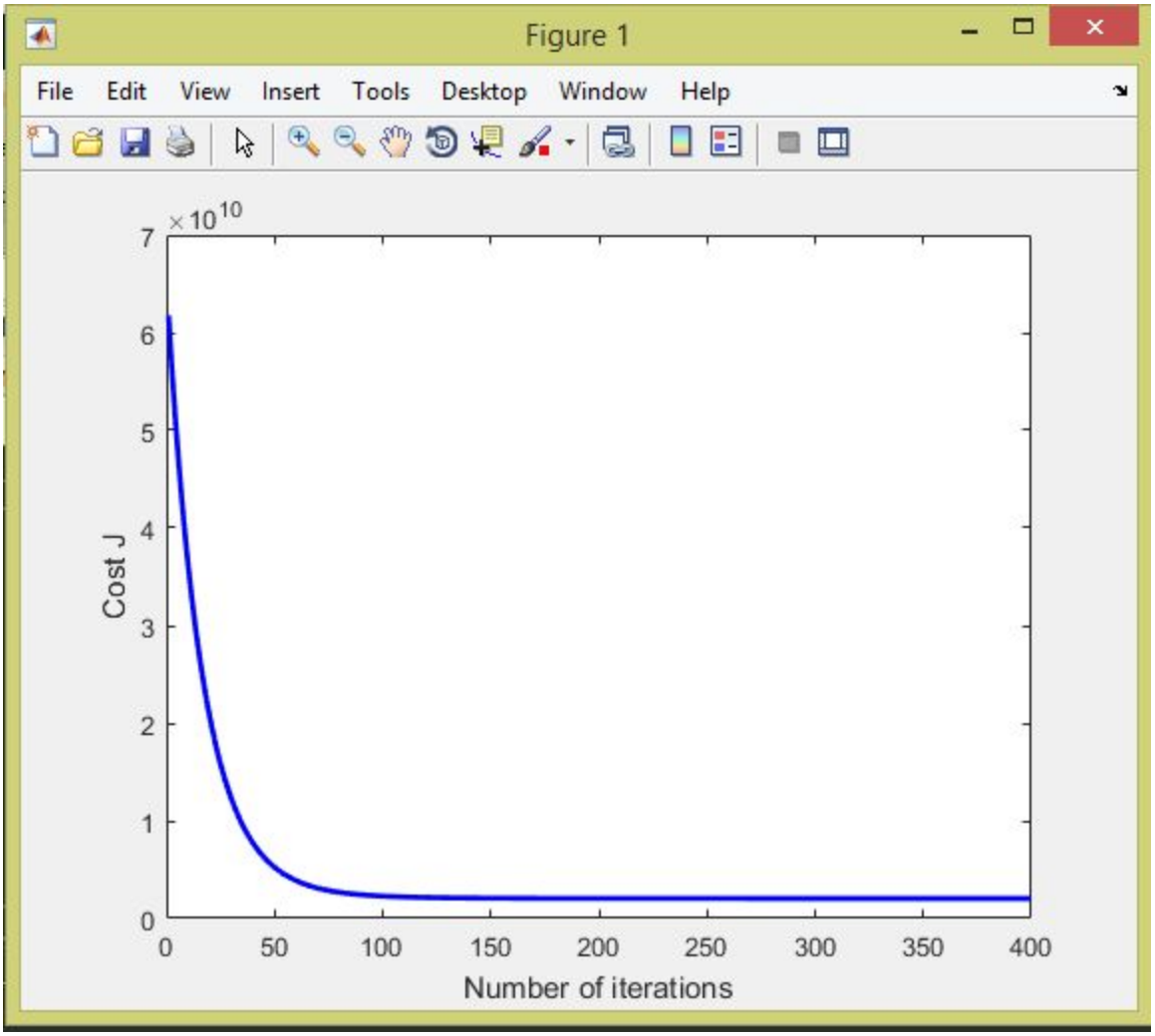
Running gradient descent ...

Theta computed from gradient descent:

340410.918973

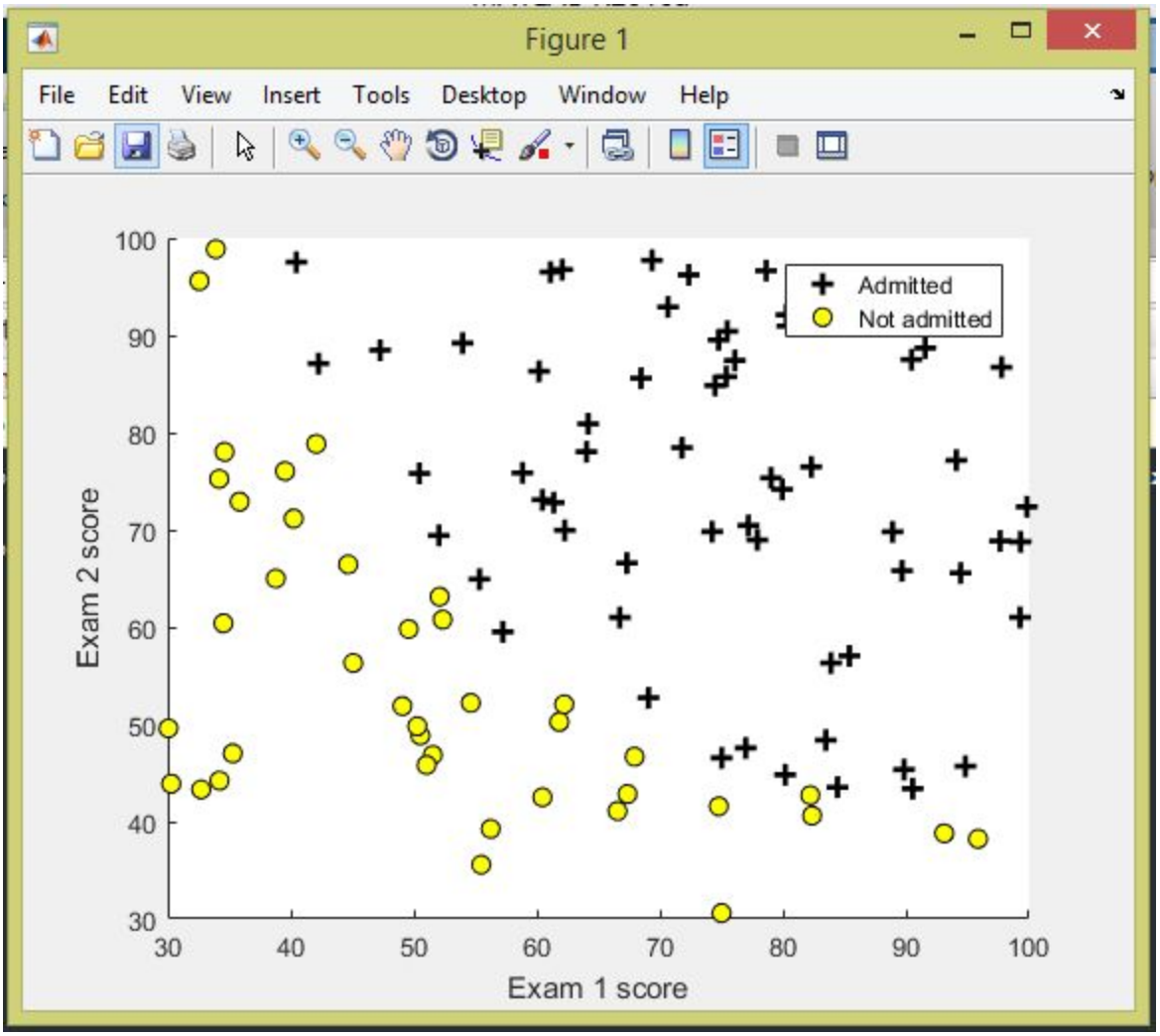
110308.113371

-6326.538108



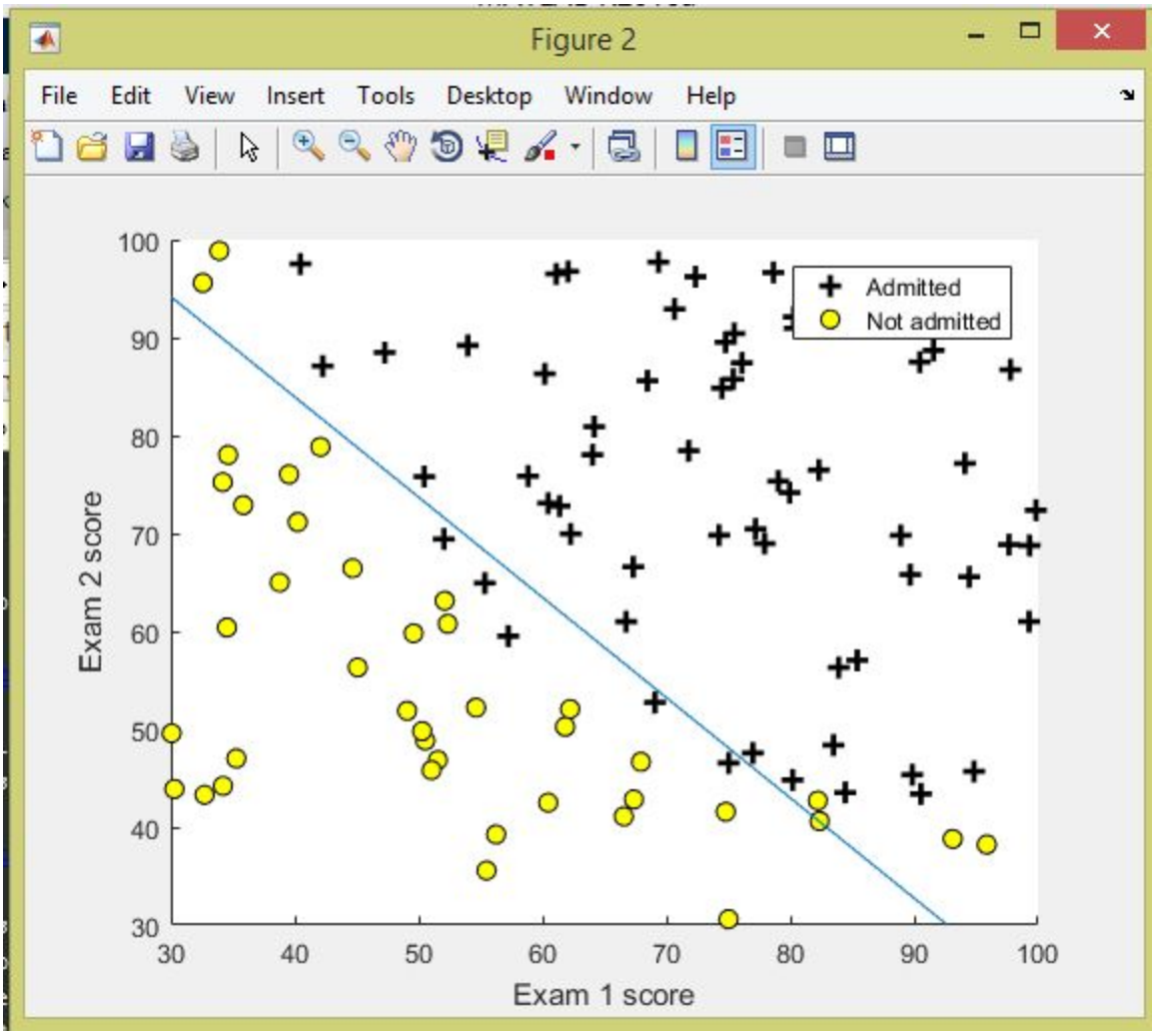
Graph of cost function vs number of iterations.

LOGISTIC REGRESSION: Implemented on data set which contains exam scores of students in two exams and historical data whether a student is admitted to a university or not.



Plotting data with + indicating (y = 1 i.e admitted) examples and o indicating (y = 0 .i.e not admitted) examples.

FINAL PLOT:



Final plot with decision boundary separating the two classes.

LOGISTIC REGRESSION(with batch gradient descent):Implemented on Breast cancer Wisconsin(diagnostic) data set.Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.In total 30 features are present in the data set.

OUTPUT - Elapsed time is 0.405204 seconds.

Train Accuracy: 98.275862

Elapsed time is 0.452496 seconds.

Train Accuracy: 98.275862

Elapsed time is 0.532078 seconds.

Train Accuracy: 98.245614

Elapsed time is 0.472455 seconds.

Train Accuracy: 98.245614

Elapsed time is 0.620096 seconds.

Train Accuracy: 98.245614

Elapsed time is 0.701451 seconds.

Train Accuracy: 98.245614

Elapsed time is 0.698002 seconds.

Train Accuracy: 100.000000

Elapsed time is 0.6818 seconds.

Train Accuracy: 98.214286

Elapsed time is 0.646418 seconds.

Train Accuracy: 94.642857

Elapsed time is 1.3306 seconds.

Train Accuracy: 98.214286

Accuracy: 0.980606

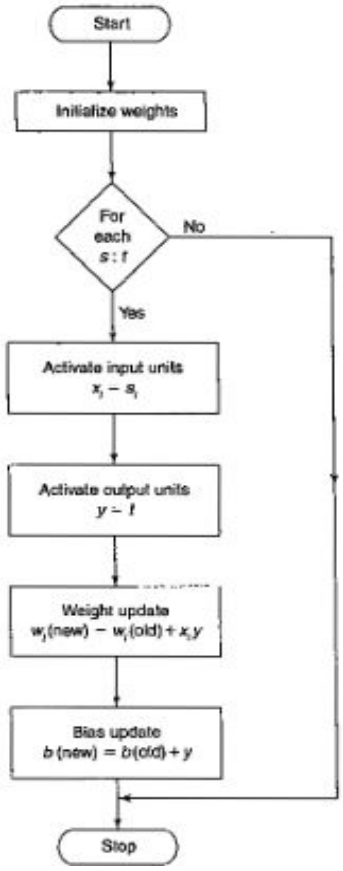
Precision: 0.990909

Sensitivity: 0.957576

Specificity: 0.994365

F1 Score: 0.973249

Hebbs :



Steps:

1) **Start**
2) **Initialize weights**
3) **for each s:t**
4) **Activate input inputs x=s.**
5) **Activate output inputs y=t.**
6) **Weight update w1(new) = w1(old) + X1Y**
7) **Bias update b(new)=b(old)+y**

8) **Stop.**

CODE -

```
m=size(X_train,1);
n=size(X_train,2);
w=zeros(1,n);
b=0;
tic
for i=1:m
   for j=1:n
      w(j)=w(j)+X_train(i,j)*y_train(i,1);
   end
   b=b+y_train(i,1)
end
toc
% Compute accuracy on our training set
R=X_val*w'+b ;
th=mean(R);
RX=R>=th;
%RX= round(sigmoid(R));
train_accuracy = mean(double(RX==y_val) * 100)
```

OUTPUT -

```
Elapsed time is 0.299306 seconds.
train_accuracy =  98.276
```

Elapsed time is 0.295229 seconds.

train_accuracy = 94.828

Elapsed time is 0.314041 seconds.

train_accuracy = 98.246

Elapsed time is 0.333557 seconds.

train_accuracy = 98.246

Elapsed time is 0.334028 seconds.

train_accuracy = 94.737

Elapsed time is 0.310373 seconds.

train_accuracy = 92.982

Elapsed time is 0.292098 seconds.

train_accuracy = 98.246

Elapsed time is 0.342839 seconds.

train_accuracy = 94.643

Elapsed time is 0.363937 seconds.

train_accuracy = 96.429

Elapsed time is 0.299228 seconds.

train_accuracy = 92.857

Accuracy: 0.959488

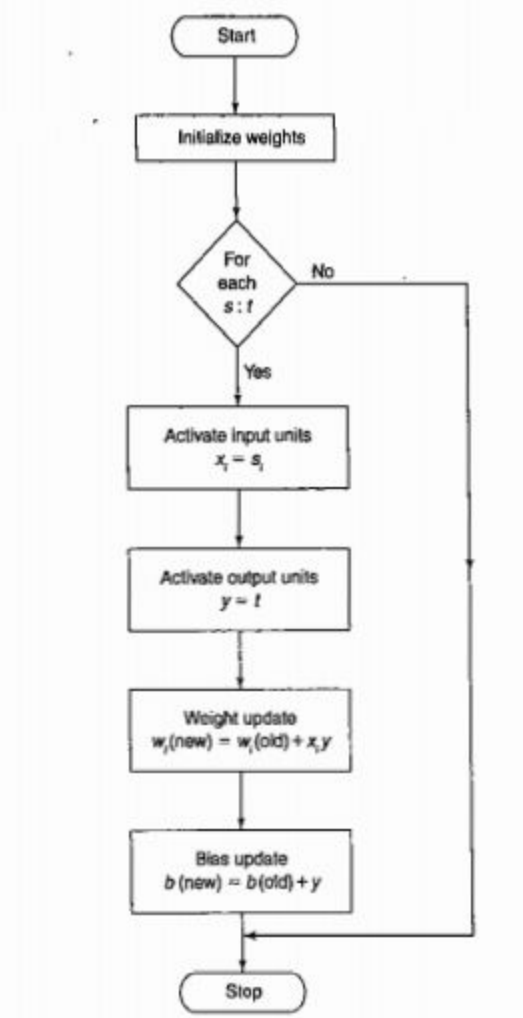
Precision: 0.920880

Sensitivity: 0.976407

Specificity: 0.949524

F1 Score: 0.947547

Perceptron:



Steps:

1) **Start**
2) **Initialize weight and bias.**
3) **Set alpha 0 to 1.**
4) **for each s:t**
5) **If yes,Activate input inputs x=s.**
6) **If no,then stop.**

7) **Calculate net input Yn.**

8) **Apply activation, obtain Y= f(y,)**

9) **Condition if y!=t**

10) **If yes, w1(new) = W1{old)+ atx1 b{new) = b(old) +at**

11) **If weight changes .**

12) **If no, W1(new) = w1{old) b(new) = b(old) .**

13) **Stop.**

CODE -

```
m=size( X_train,1);

n=size( X_train,2);

w=zeros(1,n);

b=0;

%flag=0;

epoch=0;

tic

while  epoch~=1

   %flag=0;

   for i=1:m

      y= X_train(i,:)*w' + b ;
```

```
        %activation function
        yin=y>=7.1162e+08;
        %end of activation function
        %yin=activation func(y);
        %flag=flag+1;
        if yin~=yo(i,1)
            for j=1:n
                w(j)=w(j)+ X_train(i,j)*y_train(i,1);
            end
            b=b+y_train(i,1);
            %flag=0;
        end
        %if i==m
        %end
    end
     epoch=epoch+1;
end
toc
% Compute accuracy on our training set
```

```
R=X_val*w'+b ;

th=mean(R);

RX=R>=th;

%RX= round(sigmoid(R));

train_accuracy = mean(double(RX==y_val) * 100)
```

OUTPUT -

```
Elapsed time is 0.123517 seconds.

train_accuracy =  94.828

Elapsed time is 0.117902 seconds.

train_accuracy =  93.103

Elapsed time is 0.118098 seconds.

train_accuracy =  92.982

Elapsed time is 0.122908 seconds.

train_accuracy =  94.737

Elapsed time is 0.126663 seconds.

train_accuracy =  96.491

Elapsed time is 0.122786 seconds.

train_accuracy =  94.737

Elapsed time is 0.118272 seconds.
```

train_accuracy = 94.737

Elapsed time is 0.119896 seconds.

train_accuracy = 98.214

Elapsed time is 0.12118 seconds.

train_accuracy = 98.214

Elapsed time is 0.130673 seconds.

train_accuracy = 94.643

Accuracy: 0.952687

Precision: 0.917363

Sensitivity: 0.962338

Specificity: 0.947063

F1 Score: 0.938393

Breast Cancer Tumor Diagnosis - Neural Networks,regression:

The dataset we used is the 'Breast Cancer Wisconsin (Diagnostic) Data Set'.

The dataset contains 569 patients with breast tumors. Several measurements and a diagnosis of malignant and benign are available for each patient.

The measurements are used as features to make our predictions on the diagnosis.In total there are 30 features.

Attribute Information:

1) ID number 2) Diagnosis (M = malignant, B = benign) 3-32)

Ten real-valued features are computed for each cell nucleus:

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 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. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recorded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

Results:

Classification	Maximum	Minimum
(logistic regression with batch gradient descent algorithm)	Accuracy: 0.982391 Precision: 0.990909 Sensitivity: 0.962121 Specificity: 0.994365 F1 Score: 0.975934	Accuracy: 0.977128 Precision: 0.986147 Sensitivity: 0.952814 Specificity: 0.991667 F1 Score: 0.968497
(logistic regression with large-scale optimization algorithm)	Accuracy: 0.980730 Precision: 0.990693 Sensitivity: 0.957576 Specificity: 0.994444 F1 Score: 0.973236	Accuracy: 0.977067 Precision: 0.985931 Sensitivity: 0.952597 Specificity: 0.991508 F1 Score: 0.968255
Hebb's rule	Accuracy: 0.966632 Precision: 0.938036 Sensitivity: 0.976623 Specificity: 0.960794 F1 Score: 0.956220	Accuracy: 0.952503 Precision: 0.920273 Sensitivity: 0.961905 Specificity: 0.946905 F1 Score: 0.938534
Perceptron rule	Accuracy: 0.970202 Precision: 0.941629 Sensitivity: 0.981169 Specificity: 0.963651 F1 Score: 0.960758	Accuracy: 0.940250 Precision: 0.914055 Sensitivity: 0.929004 Specificity: 0.946905 F1 Score: 0.920108

Classification:

Hebbs	Perceptron	Logistic Regression
Provides an **algorithm** to update weight of neuronal connection within neural network.	Does not try to optimize the separation "distance".As long as it finds a hyperplane that separates the two sets.	LR is probabilistic method, which produces an inferential and highly interpretable statistical model.
The **Hebbian Rule** works well as long as all the input patterns are orthogonal or uncorrelated.	Perceptrons can be trained online (i.e. their weights can be updated as new examples arrive one at a time)	Logistic regression is used in prediction under certain conditions.
Low as compared to others.	Moderate accuracy	Moderate accuracy
Low as compared to LR.	Moderate Precision	Moderate Precision

Future scope:

1) To visit hospitals like TATA Hospital and get more real world data and testing this algorithms on that data .

2) To create a website and putting algorithms in backend so that people can enter data and get to know that do they have chances of recurrence of the cancer disease.

Bibliography :

- Andrew Ng https://www.coursera.org/learn/machine-learning/lecture/6Nj1q/multiple-features.
  - Principles of Soft Computing by S.N.Sivanandam
  - S.N.Deepa
    - https://www.coursera.org/learn/machine-learning/lecture/2DKxQ/normal-equation.
  - https://www.coursera.org/lecture/assembling-genomes/2011-european-e-coli-outbreak-VgLTB.
  - https://www.omicsonline.org/using-three-machine-learning-techniques-for-predicting-breast-cancer-2157-7420.1000124.php?aid=13087
  - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

- Mehwar Fatima, M. P. (2Ol7). Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, l-l6.

- U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

- 2. Siegel RL, Miller KD, Jemal A. Cancer Statistics , 2016. 2016;00(00):1-24. doi:10.3322/caac.21332. 3. “Globocan 2012 - Home.” [Online]. Available: http://globocan.iarc.fr/Default.aspx. [Accessed: 28-Dec-2015]. 4. Asri H, Mousannif H, El Moatassim H, Noel T. Big data in healthcare: Challenges and opportunities. 2015 Int Conf Cloud