INDIAN INSTITUE OF TECHNOLOGY(ISM) DHANBAD

UNDER THE GUIDANCE OF

Dr. Haider Banka

Department of Computer Science and Engineering

NAME : BATHINA VISWANATH SRIKANTH

ADMN NO : 15JE001073

SECTION : 1

ROLL NO : 39

(student)

PROJECT : DIABETES ONSET PREDICTION

INTRODUCTION OF DIABETES ONSET PREDICTION :

Diabetes is a chronic disease in which the body's ability to produce or respond to the hormone insulin is impaired, resulting in abnormal metabolism of carbohydrates and elevated levels of glucose in the blood.Given the medical data we can gather about people, we should be able to make better predictions on how likely a person is to suffer the onset of diabetes, and therefore act appropriately to help. We can start analyzing data and experimenting with algorithms that will help us study the onset of diabetes .

ATTRIBUTES :

Pregnancies: Number of times pregnant

Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

BloodPressure: Diastolic blood pressure (mm Hg)

SkinThickness: Triceps skin fold thickness (mm)

Insulin: 2-Hour serum insulin (mu U/ml)

BMI: Body mass index (weight in kg/(height in m)^2)

DiabetesPedigreeFunction: Diabetes pedigree function

Age: Age (years)

ALGORITHM :

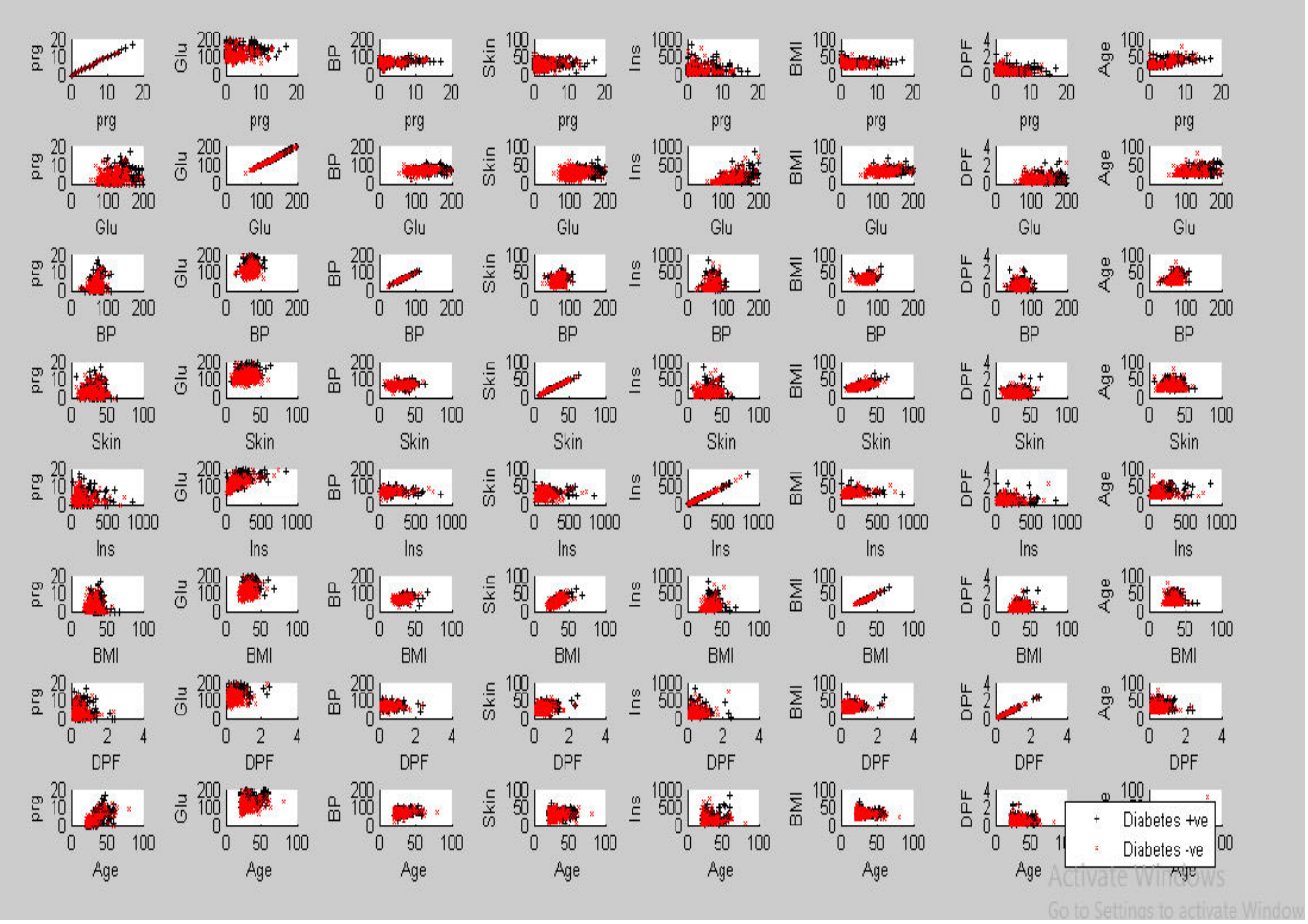
1. Initialize theta to zeros.
2. Get training data from file to matrix X(input) and vector y(output).
3. Plot data w.r.t all parameters.
4. Optimize theta using built-in fminunc function such that cost function is minimized.
5. Find cost and gradient using optimized theta.
6. To check for probability of test data multiply data vector to theta.
7. Find sigmoid of the product which gives probability.
8. Finally use predict function to predict outputs of all data to find accuracy.

METHOD :

1. Initialize theta vector to zero , X matrix to input data and y vector to output data.
2. Plot data with a scatter plot w.r.t all parameters.
3. Check cost and gradient using initialized theta and random theta.
4. Then optimize theta using built in fminunc function such that cost function is minimized.
5. Find cost function and gradient using the optimized theta.
6. Sigmoid function (1/(1+e^(-h))) where h is hypothesis function.
7. Find sigmoid of product of test data and theta to get probability of diabetes.
8. If probability is greater than 0.5 ,predict 1 otherwise predict 0
9. Test the obtained theta using accuracy of obtained theta.

PICTORIAL REPRESENTATION

Scatter plots of the training data w.r.t all parameters.



Here, black + denote that the patient has diabetes.

red x denote that the patient doesn't have diabetes.

The above plot shows that there is no obvious relationship between BP and onset

We can also deduce that larger values of Glucose combined with larger values for age, DPF, mass, insu, skin, BP, and preg tends to show greater likelihood of testing positive for diabetes.

CALCULATIONS :

First we start by taking data from file.

Then X = data(:, 1:8) load first 8 columns (inputs) to X

y = data(:, 9) load last column (outputs) to y

Initialize theta to a vector of 9 elements all initialized to 0.

Add a column of ones to X on left side.

Cost = -(y' * log(h) + (1-y') * log(1-h))/m;

Where h=1/(1+e^(-X*theta)).

We use fminunc function to optimize theta such that cost is optimized.

Then we find cost and gradient with optimized theta.

We take an input vector of 9 elements with 1 as first element and find the probability of onset of diabetes.

If probability is >0.5 person suffers from diabetes.

Else If probability is <0.5 no diabetes.

CODE :

Main function :

```
clear ; close all; clc

data = load('data.txt');

X = data(:, 1:8); y = data(:, 9);

plotData(X, y);

[m, n] = size(X);

X = [ones(m, 1) X];

fprintf('initializing theta to zeros\n');

init_theta = zeros(n + 1, 1);

[cost, grad] = costFunction(init_theta, X, y);

fprintf('Cost at initial theta (zeros): %f\n', cost);

fprintf('Gradient at initial theta (zeros): \n');

fprintf(' %f \n', grad);

test_theta = [-10; 0.1; 0.05;-0.01;0.01;0.001;0.1;1;0.05];

[cost, grad] = costFunction(test_theta, X, y);

fprintf('\nfor a random test theta:');

fprintf('\nCost at test theta: %f\n', cost);

fprintf('Gradient at test theta: \n');
```

```
fprintf(' %f \n', grad);
options = optimset('GradObj', 'on', 'MaxIter', 400);
[theta, cost] = ...
        fminunc(@(t)(costFunction(t, X, y)), init_theta, options);
fprintf('\ntheta after optimization: \n');
fprintf(' %f \n', theta);
fprintf('Cost at optimized theta : %f\n', cost);
fprintf('for a sample test data:\n');
fprintf('Pregnancies: 10\n');
fprintf('Glucose: 150\n');
fprintf('BloodPressure: 50\n');
fprintf('SkinThickness: 25\n');
fprintf('Insulin: 100\n');
fprintf('BMI: 25\n');
fprintf('DiabetesPedigreeFunction: 0.5\n');
fprintf('Age: 25\n\n');
prob = sigmoid([1 10 150 50 25 100 25 0.5 25] * theta);
fprintf('probability of diaabetes for above data ');
fprintf(' %f \n', prob);
if prob>0.5
fprintf('patient suffers from diabetes\n');
else
```

```
fprintf('patient does not have diabetes\n');
end
p = predict(theta, X);
fprintf('Accuracy: %f\n', mean(double(p == y)) * 100);
fprintf('\n');
```

Plot function

```
function plotData(X, y)
   figure;
for i=1 :8
   for j=1:8
      subplot(8,8,8*(i-1)+j);
      hold on;
      pos = find(y==1); neg = find(y == 0);
      plot(X(pos, i), X(pos, j), 'k+','LineWidth', 0.5,    'MarkerSize', 3);
      plot(X(neg, i), X(neg, j), 'rx','LineWidth', 0.5,    'MarkerSize', 3);
      if i==1
            xlabel('prg');
         elseif i==2
            xlabel('Glu');
         elseif i==3
            xlabel('BP');
```

```
        elseif i==4
            xlabel('Skin');
        elseif i==5
            xlabel('Ins');
        elseif i==6
            xlabel('BMI');
        elseif i==7
            xlabel('DPF');
        elseif i==8
            xlabel('Age');
      end
    if j==1
            ylabel('prg');
        elseif j==2
            ylabel('Glu');
        elseif j==3
            ylabel('BP');
        elseif j==4
            ylabel('Skin');
        elseif j==5
            ylabel('Ins');
        elseif j==6
```

```
            ylabel('BMI');
        elseif j==7
            ylabel('DPF');
        elseif j==8
            ylabel('Age');
      end
    hold off;
    end
end
legend('Diabetes +ve', ' Diabetes -ve ')
end
```

Cost function

```
function [J, grad] = costFunction(theta, X, y)
m = length(y);
J = 0;
grad = zeros(size(theta));
h = sigmoid(X* theta);
J= -(y' * log(h) + (1-y') * log(1-h ))/m;
grad = (X')*(h- y);
grad= grad/m;
end
```

Sigmoid function

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

Predict function

```
function p = predict(theta, X)
m = size(X, 1);
p = zeros(m, 1);
o= X*theta;
o=sigmoid(o);
for i=1 : m
    if o(i) >=0.5
        p(i)=1;
    end
end
end
```

OBSEVATIONS:

Based on look up at first 10 columns of data we know that :

The preg and age attributes are integers. The population is generally young, less than 50 years old. Some attributes where a zero value exist seem to be errors in the data (e.g. plas, pres, skin, insu, and mass).This data shall be either removed or handled carefully.

Reviewing scatter plots of all attributes in the dataset shows that:

There is no obvious relationship between age and onset of diabetes. There is no obvious relationship between pedi function and onset of diabetes .This may suggest that diabetes is not hereditary, or that the Diabetes Pedigree Function needs work. Larger values of plas combined with larger values for age, pedi, mass, insu, skin, pres, and preg tends to show greater likelihood of testing positive for diabetes.

CONCLUSION:

In this project , we paid close attention to logistic regression and observed its performance through various metrics. This work gave me a better understanding of machine learning applications in medical diagnosis. This is also focused on data transforms and algorithm analysis. I believe that this is good start for building methods that help predicting the onset of diabetes.

REFERENCE :

1. PIMA Indian heritage UCI Machine Learning Repository .
2. www.machinelearningmastery.com