Pattern Separation Using Logistic Regression

AI2101-Project

April 22, 2022

Team Members

- Amulya Tallamraju - AI20BTECH11003
- Anita Dash - MA20BTECH11001
- Anjali - MA20BTECH11002
- Ruthwika Boyapally - MA20BTECH11004
- Tapishi Kaur - MA20BTECH11017

Introduction

Pattern Recognition and Classification Problem

A Pattern Recognition and Classification Problem is a problem of obtaining a criterion for distinguishing between the elements of two disjoint sets of patterns, that are usually represented by points in a euclidean space.

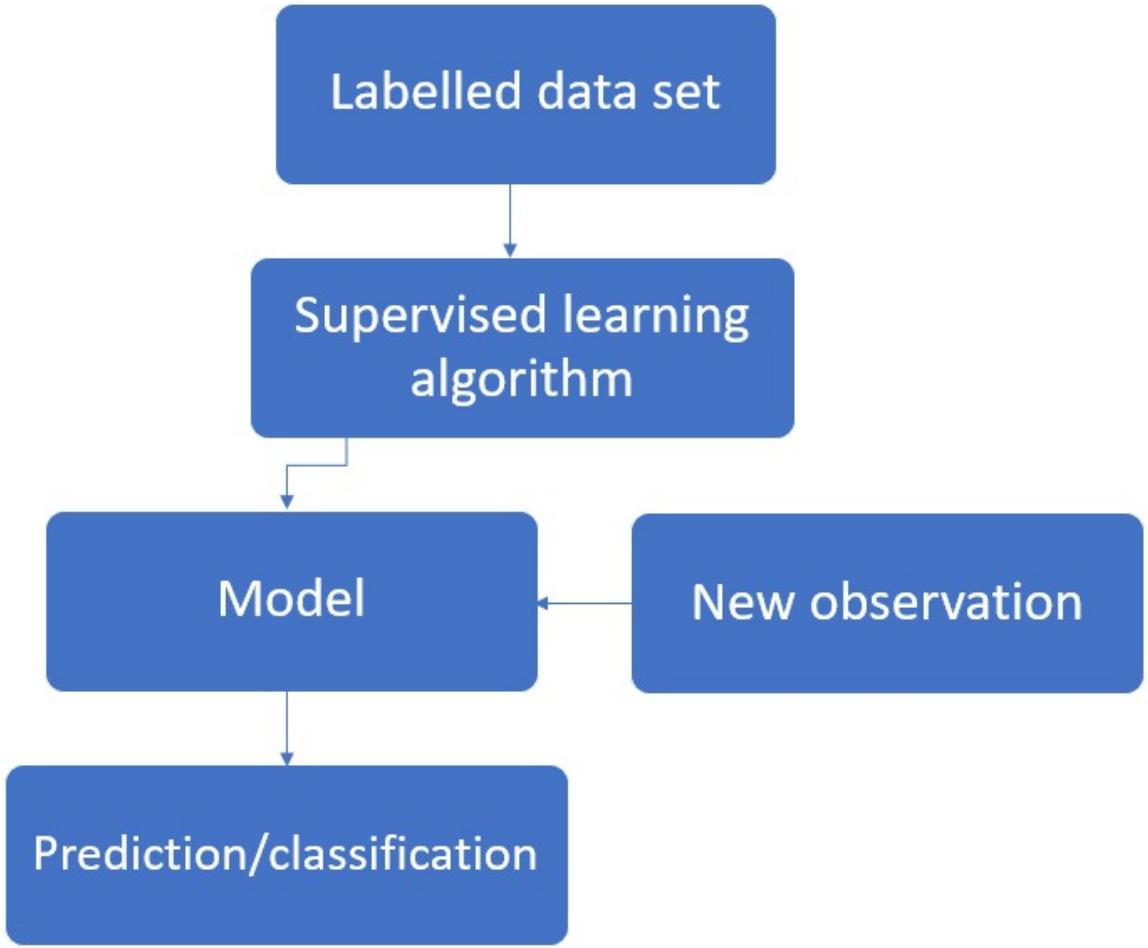


Figure: Classification Flowchart

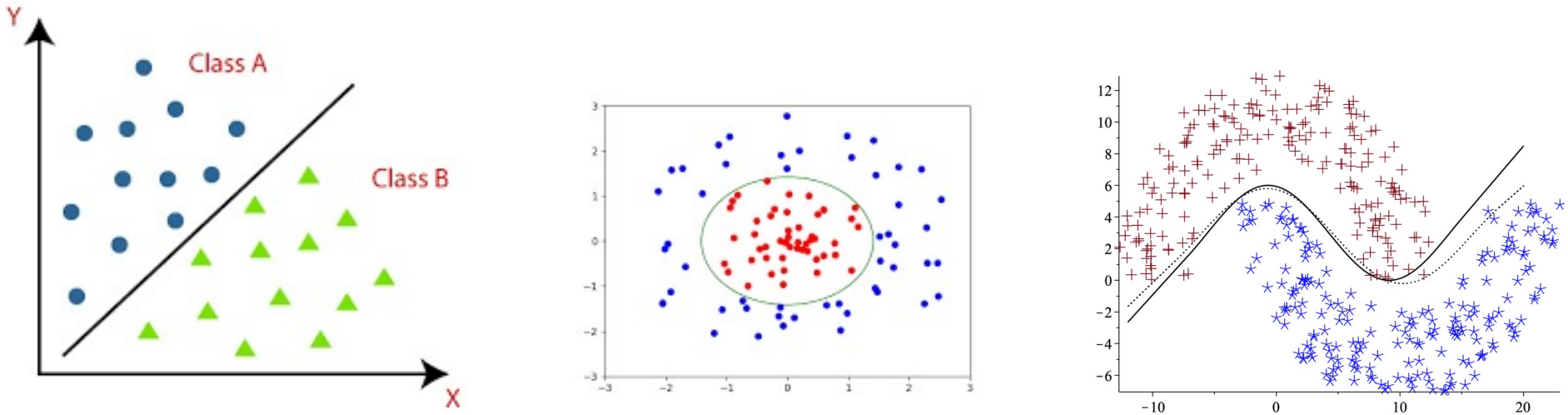


Figure: Linear and Non-Linear classification of a set of data points

Logistic Regression

Definition

Logistic regression is a statistical analysis method used to separate patterns and is based on the idea of maximum likelihood estimation.The classification algorithm Logistic Regression is used to predict the likelihood of a categorical dependent variable.

- Logistic regression is a simple and more efficient method for binary and linear classification problems
- It achieves very good performance with linearly separable classes.

Question

If it is a classification problem, why does the name have regression in it?

Sigmoid Function

A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point and exactly one inflection point.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

- f is a logistic sigmoid function
- f squashes the input between 0 and 1

In order to map predicted values to probabilities, we use the Sigmoid function.

Linear Regression and Logistic Regression

similarities

- Both are supervised machine learning algorithms
- We use linear mathematical equations for both regressions for reasonable predictions.

Linear Regression and Logistic Regression

Linear Regression	Logistic Regression
It is used to handle regression problems	It is used to handle classification problems
Provides continuous output	Provides discrete output
The mean squared error is use to calculate loss function	The maximum likelihood function is used to calculate loss function

Figure: 2

Types of Logistic Regression

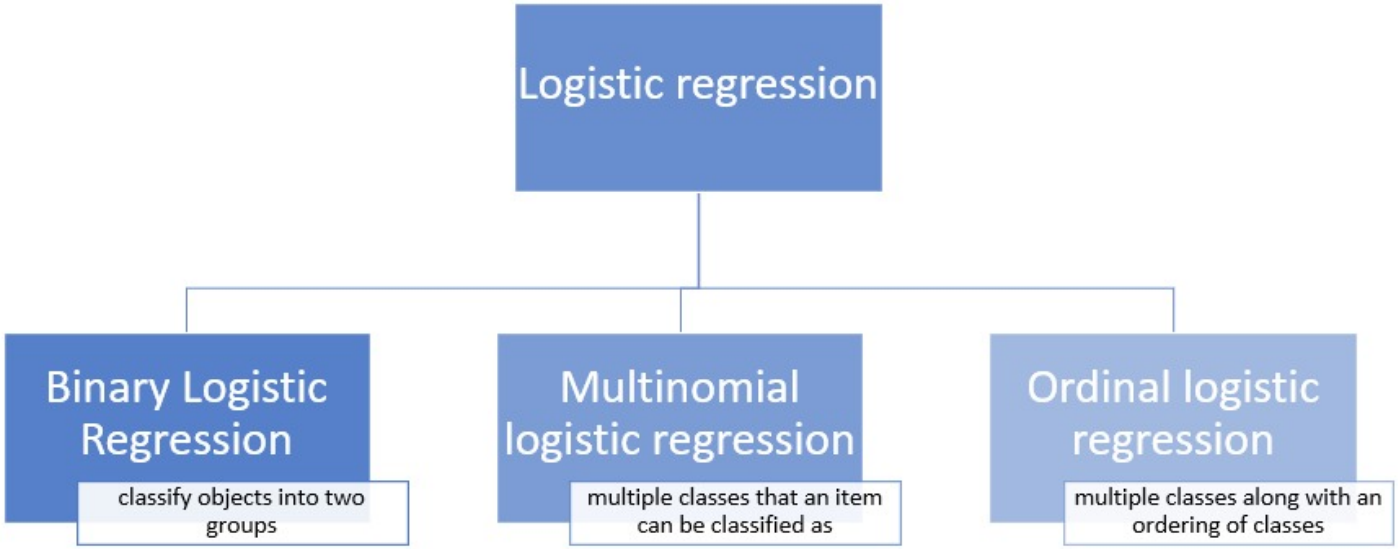


Figure: Since we want to find a hyperplane (n-Dimensional) or line (2-D) that best separates the points in space, we will be going forward with Binary Logistic Regression.

Logistic Regression Problem

We consider the data: x_i, i = 1,...,m. $x_i \in R^n$ are explanatory variables, and their corresponding outcomes y_i, i = 1,...,m. $y_i \in \{0, 1\}$ are their associated Boolean class

Goal

We construct a linear classifier

$$\hat{y} = \mathbf{1}[\beta^T x] \quad \textit{therefore}, \tag{2}$$

$$\hat{y} = \begin{cases} 1 & \text{if } \beta^T x > 0 \\ 0 & \text{if } \textit{otherwise} \end{cases} \tag{3}$$

Goal Contd.

We model the posterior probabilities of the classes given the data linearly, with

$$\log \frac{Pr(Y=1|X=x)}{Pr(Y=0|X=x)} = \beta^T x \tag{4}$$

Therefore, we get:

$$Pr(Y=1|X=x) = \frac{e^{\beta^T x}}{1+e^{\beta^T x}} \quad Pr(Y=0|X=x) = \frac{1}{1+e^{\beta^T x}} \tag{5}$$

We find the maximum likelihood of parameter $\beta \in R^n$, Finding this Maximum Likelihood is sometimes called **Logistic Regression**.

Logistic Regression:Convexity

Proof for Convexity

The log likelihood function $l(\beta)$ is:

$$l(\beta) = \sum_{i=1}^{q} \log(p_i) + \sum_{i=q+1}^{m} \log(1 - p_i) \tag{6}$$

Substituting (5) in (6),

$$l(\beta) = \sum_{i=1}^{q} \beta^T x_i - \sum_{i=1}^{m} \log(1 + e^{\beta^T x_i}) \tag{7}$$

Proof Contd.

- $\sum_{i=1}^{q} \beta^T x_i$ is concave affine
- $log(1 + e^{\beta^T x_i})$ is a concave function (The Hessian of the function is positive semi-definite and the proof for the same has been shown in the report)
- therefore, $(-\sum_{i=1}^{m} \log(1 + e^{\beta^T x_i}))$ is concave

Sum of concave functions is concave therefore, $l(\beta)$ is a concave function

l is concave function of parameter β, the logistic regression problem can be solved as a convex optimization problem.

Geometric Interpretation of Logistic Regression

Let x_i and x_j be two points where $y_i = 1$ and $y_j = -1$. Let the distances be

$$d_i = \frac{W^T x_i + w_0}{\|W\|} \tag{8}$$

$$d_j = \frac{W^T x_j + w_0}{\|W\|} \tag{9}$$

$$d_i > 0, d_j < 0 \tag{10}$$

We can this say that $d_i * y_i > 0$..

$$W, w_0 = argmax(\sum y_i * (W^T * x_i + w_0)) \tag{11}$$

Example data and the need for sigmoid squashing

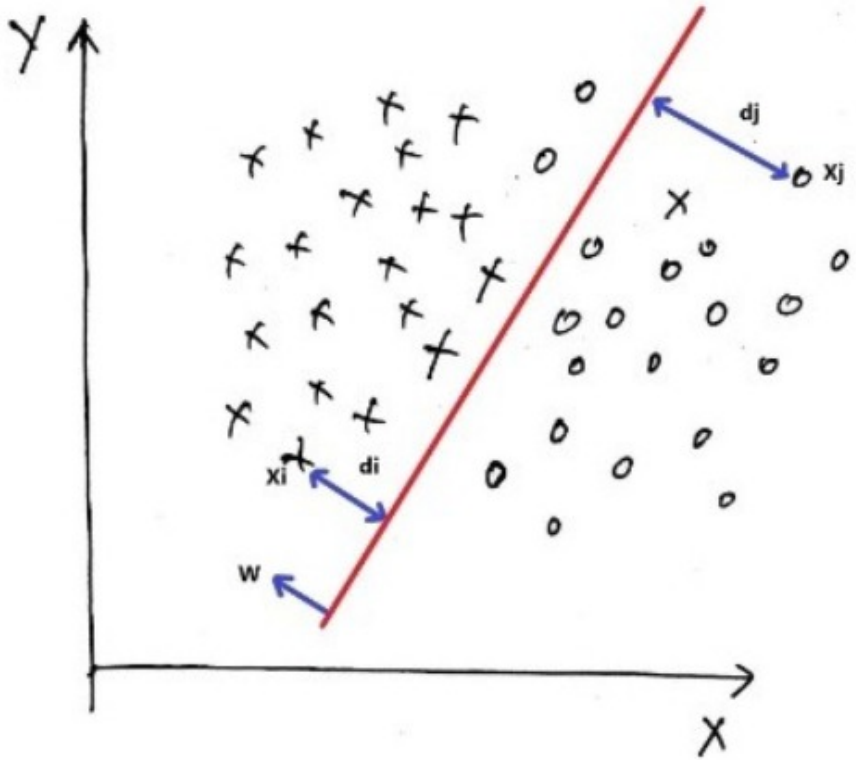


Figure: Xi and Xj are correctly classified points

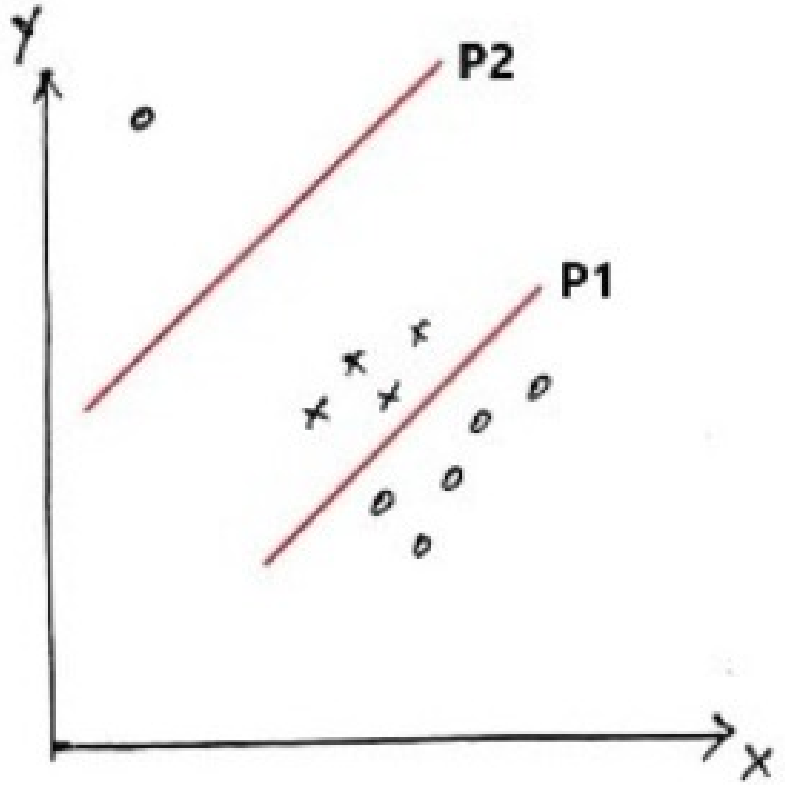


Figure: P1 is a better classifier than P2 but our previous cost function will choose P2

Updated Cost Function

The presence of an outlier or extreme point can affect the plane so we need to find a function that makes the $(y_i * d_i)$ value small if it is too large and if $(y_i * d_i)$ value is small it should remain small. We hence resolve to the Sigmoid function.

Updated Cost Function (cont.)

$$W, w_0 = argmax\left(\sum \frac{1}{1 + \exp\left(-y_i * \left(W^T * x_i + w_0\right)\right)}\right) \tag{12}$$

Since $\log x$ is a monotonically increasing function, maximising $\log x$ is equivalent to maximizing x.Thus,

$$W, w_0 = argmax\left(\sum \log \frac{1}{1 + \exp\left(-y_i * \left(W^T * x_i + w_0\right)\right)}\right) \tag{13}$$

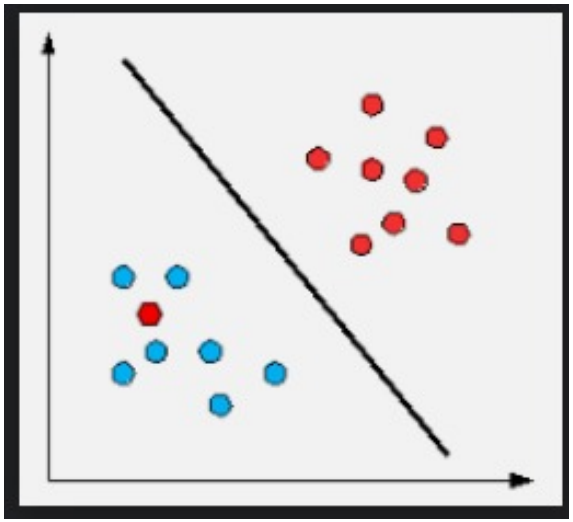
We can re-order the data so for $x_1, ..., x_q$, the outcome is $y = 1$, and for $x_{q+1}, ..., x_m$ the outcome is $y = -1$. The likelihood function then has the same form as the original derived equation.

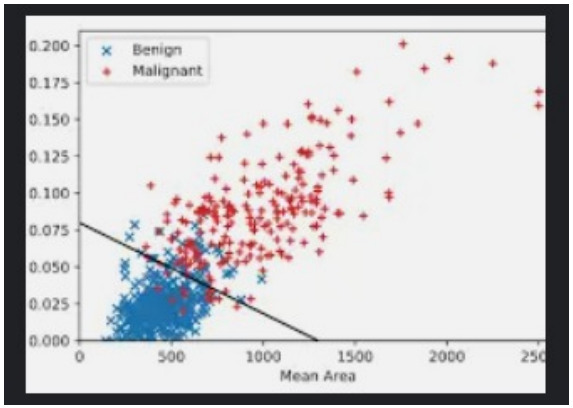
Approximate Linear Discrimination via Logistic Modeling

Linear Discrimination

In linear discrimination, we seek an affine function $f(x) = \beta^T x$ that classifies the points, i.e.,

$$\beta^T x_i > 0, \quad i = 1, .., M \quad \beta^T y_i < 0 \quad i = 1, .., N \tag{14}$$





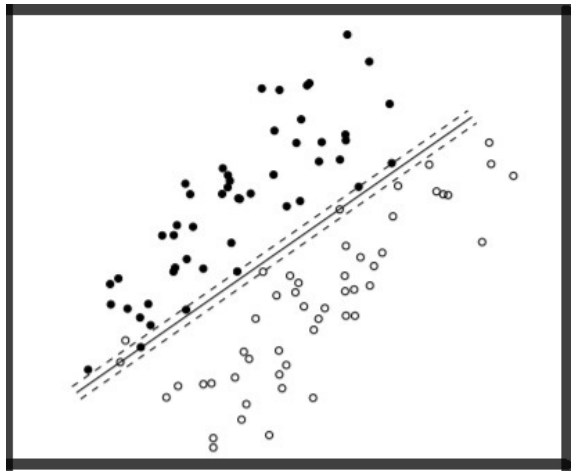


Figure: As we can see from the graphs, sometimes if not most of the time, two sets of points cannot be linearly separated, we seek an affine function that approximately classifies the points, and minimizes the number of points misclassified.

Logistical Modelling

We consider function $L(\beta) = -(l(\beta))$, therefore from (7)

$$L(\beta) = -\sum_{i=1}^{q} \beta^T x_i + \sum_{i=1}^{m} \log(1 + e^{\beta^T x_i}) \tag{15}$$

$$goal = minimize \quad L(\beta) \tag{16}$$

After the maximum likelihood value β has been found by solving the above convex optimization problem, we can form a linear classifier $f(x) = \beta^T x$ for the two sets of points.

The classifier $f(x) = \beta^T x$ has the following properties

- Assuming the data points are generated from a logistic model with parameters β, it has the smallest probability of misclassification, over all linear classifiers.
- The hyperplane $\beta^T x = 0$ corresponds to the points where prob(y = 1) = 1/2, i.e., the two outcomes are equally likely

Regularization

Definition

Regularization is a technique used to prevent overfitting problem. It adds a regularization term to the cost function in order to prevent overfitting of the model.

Definition

L2 regularization adds an L2 penalty equal to the square of the magnitude of coefficients.
By using L2 Regularization eq (15) becomes,

$$L(\beta) = -\sum_{i=1}^{q} \beta^T x_i + \sum_{i=1}^{m} \log(1 + e^{\beta^T x_i}) + \frac{\lambda}{2m} \sum_{i=1}^{m} \beta_i^2 \tag{17}$$

Regularization Parameter

Regularization Parameter

- λ is called the Regularization Parameter
- It controls the trade of between two goals:
  - Fitting the data well.
  - Keeping the parameter β small to avoid overfitting
- we need to be careful while deciding on the value of λ as a large value of λ will lead to the values of β_i shrinking to 0 and taking $\lambda = 0$ will have no regularization effect

Advantages and Disadvantages of Logistic Regression

Advantages

- Easy to implement.
- Efficient training time and doesn't require high computation power.
- Highly scalable and can be easily extended to multinomial regression.
- Very accurate and produces diverse output.

Disadvantages

- Only applicable for linear problems and discrete functions
- Overly sensitive to outliers.
- In case of high dimensional data, overfitting can be a problem
- High data maintenance and requires diverse data sets

Real World Applications

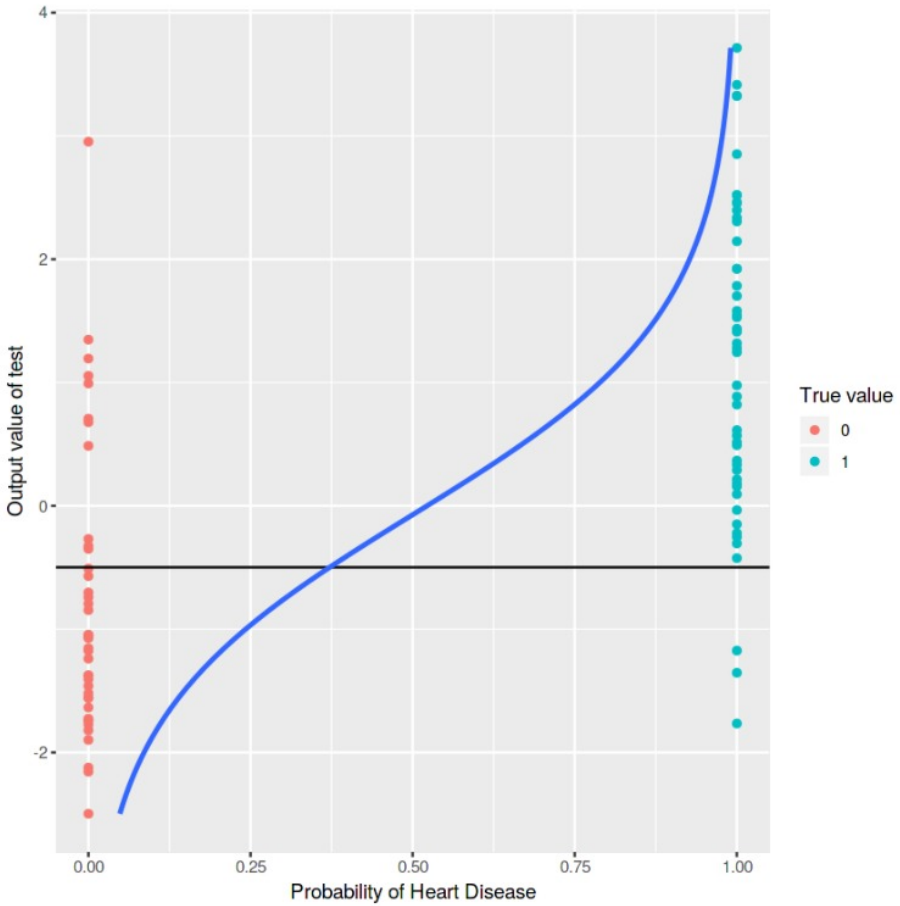


Figure: probability of patients developing a heart disease

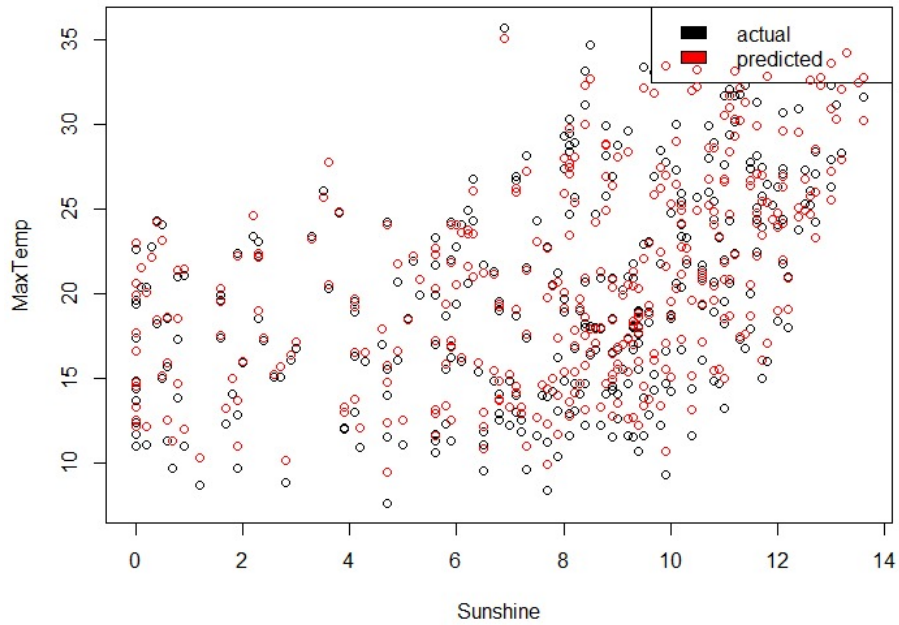


Figure: Weather Prediction

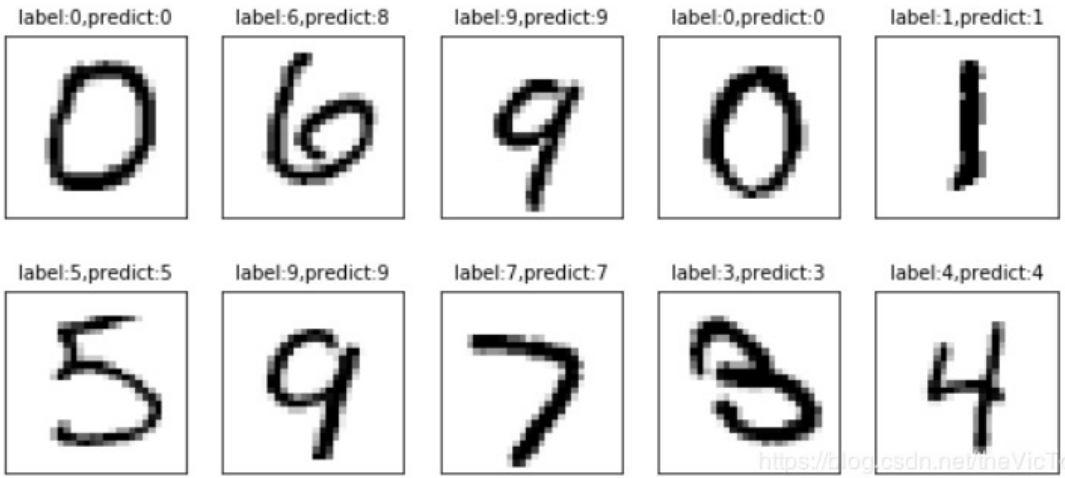


Figure: Handwriting recognition

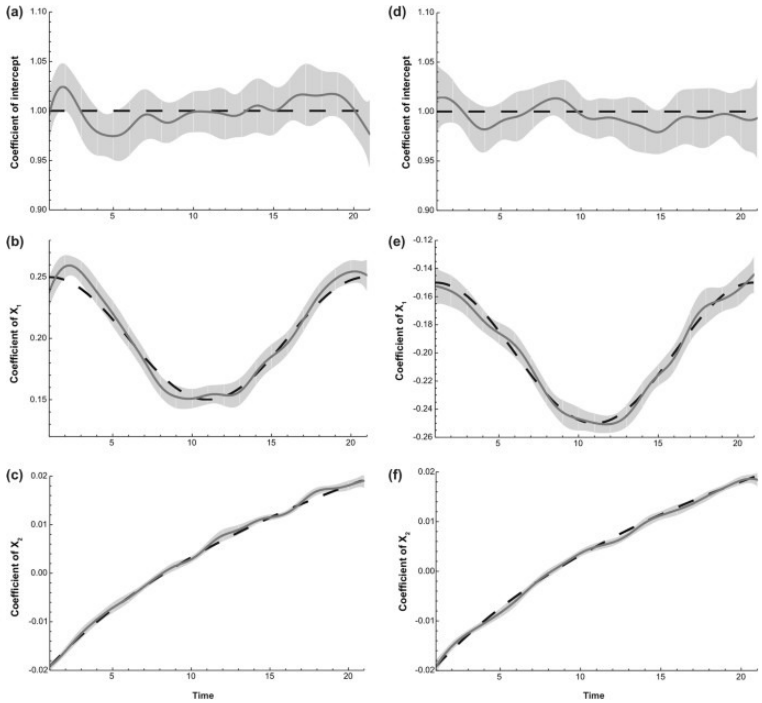


Figure: Geographical Image processing

Logistic Regression using CVXPY

Observations

The jupyter notebooks can be found in the following colab links-

- ▸ Logistic Regression with 1-Dimensional Date
- ▸ Logistic Regression without Regularization
- ▸ Logistic Regression with Regularization
- ▸ LR application for Heart Disease Classification using CVXPY

THANK YOU!