Page No.
Date

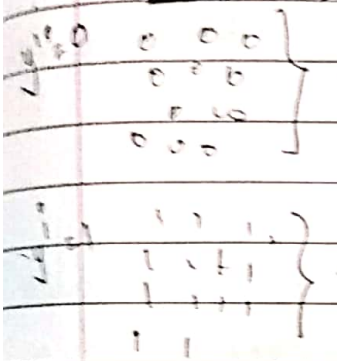
WEEK-7

CLASSIFICATION

⇒ Supervised learning task

$(x_i^{(i)}, y^{(i)})$ $y^{(i)} \Rightarrow$ class label

Goal: Group of i/p sample according to class label



find out the characteristic ~~to from~~ makes those samples from a part. class

use this to classify unseen sample.

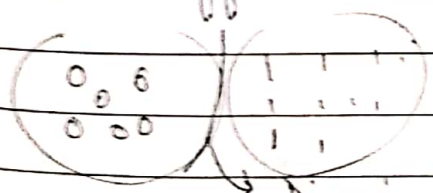
2 Ways of Classification

1) Discriminative

2) Generative

Discriminative

⇒ How different are samples of particular class ~~are diff~~ to other class



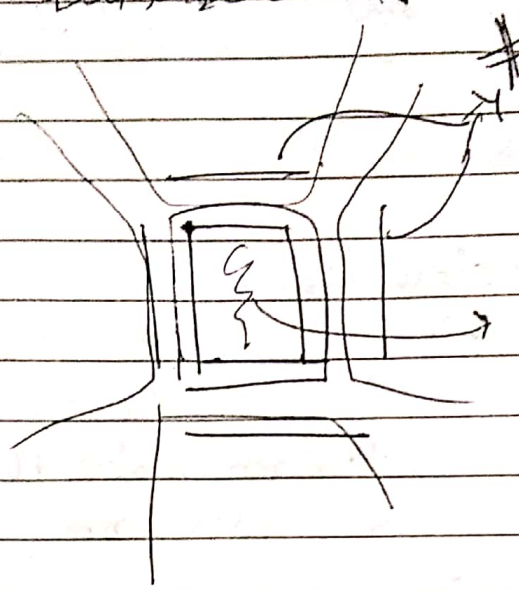
Discrimination of Samples.

Generative

⇒ try to learn Generation Characteristic of Samples.

try to figure of actual Generation process, distribution of Class.

Example



We can see less data is required in case Discriminative

properties of plot is Generative ~~Classifier~~ feat.

By this eg. we can see border pts are main focus to create a differentiated / Discriminative Classifier

Whereas Generative tries to find features of whole Sample space Hence to develop that we require more amount of data.

2) Num classes

Binary V/s Multiclass. (Several binary and true Multiclass)

3) Parametric V/s Non-parametric.

$h_w(x)$

$W = \{w_1, w_2 \dots w_n\}$

parameters.

Eg. Regression, GMM

we estimate model

No - prior model.

we try to learn actual distribution

NN is eg. Non-parametric.

k-means is eg.

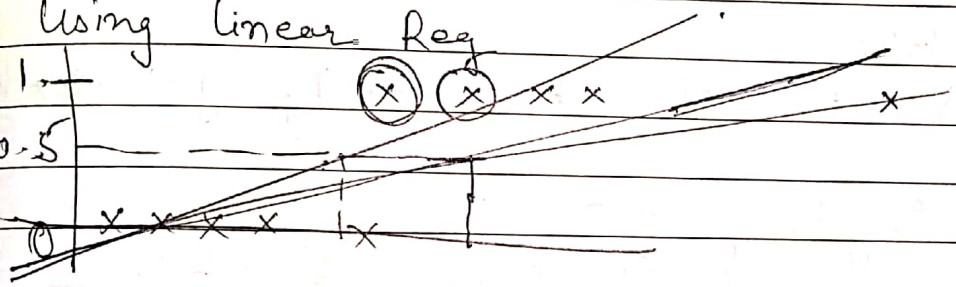
Logistic Regression

a) Starting state for this ~~Regressi~~ classifier was regression.

b) ~~Linea~~ Regression using sigmoidal function.

Binary Classification
Spam / Not Spam
Belign / Malignant
and many more.

Using linear Reg



If we keep threshold as 0.5

$h_\theta(x) \geq 0.5 \Rightarrow 1$

$h_\theta(x) < 0.5 \Rightarrow 0$

In this case this Linear Regression seems to work fine

but if there is one more point.

The circles one will give wrong result

Some data points can be misclassified in Linear regression.

for linear reg

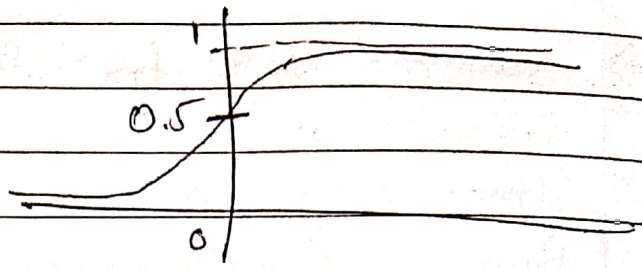
$h_\theta(x) < 0$ & $h_\theta(x) > 1$

Date

Logistic Regression

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

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



Sigmoid function.

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

Interpretation

$h_\theta(x)$ = estimated probability $y=1$ for input x

If $h_\theta(x) = 0.7$

There is 70% chance of y being 1

$P(y=0/x) = 1 - P(y=1/x)$

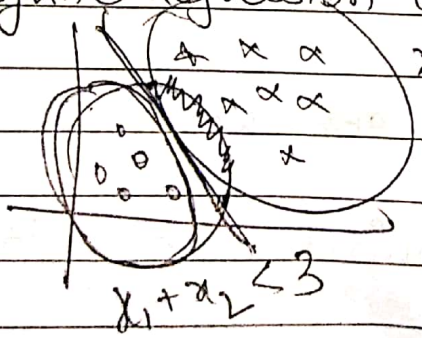
$g(z) \geq 0.5$ when $z \geq 0$

$g(z) < 0.5$ when $z < 0$

$z = \theta^T x$

$\theta^T x \geq 0$

Logistic Regression is Discriminative classifier



$x_1 + x_2 \geq 3$

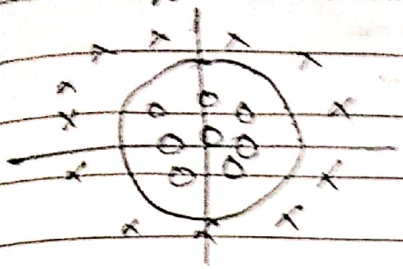
$h_\theta(x) = g(\theta_0 + \theta_1 x + \theta_2 x_2)$

Predict $y=1$

$-3 + x_1 + x_2 \geq 0$

$x_1 + x_2 \geq 3$.

Non-linear decision boundary

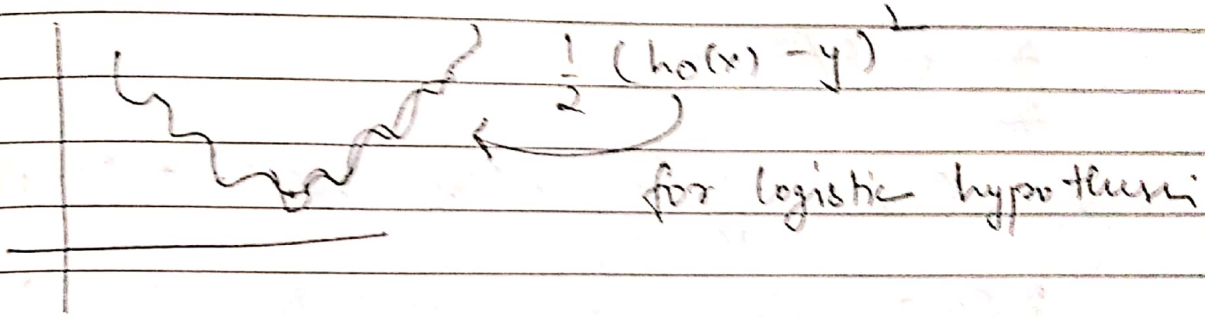


$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1 + \theta_3 x_1^2 + \theta_4 x_2^2)$

Predict $y = 1$ if $-1 + x_1^2 + x_2^2 \geq 0$

$x_1^2 + x_2^2 \geq 1$

How to estimate θ values



MSE may lead to local minimum as nature of hox is not linear but Sigmoidal

New cost function

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

Piece-wise convex nature is exploited

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

when $y = 0$ $\quad$ $y = 1$

$\text{Cost} = -\log(1 - h_\theta(x))$ $\quad$ $\text{Cost} = -\log(h_\theta(x))$

Page No.
Date

Gradient descent

$$\theta_j = \theta_j - \alpha \sum_{1}^{m} (h_\theta(x^i) - y^i)\, x^i_j$$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Derivation

$$\frac{\partial}{\partial w_j} h_w(x^{(i)}) =$$

$$C(w) = \frac{-1}{M} \sum_{i=1}^{M} y^i (\log (h_w(x)) + (1-y) \log (1 - h_w(x^i)$$

$$\frac{\partial C(w)}{\partial w_j} = \frac{-\partial}{\partial w_j} \frac{1}{M} \sum_{i=1}^{M} y^i \log(h_w(x)) + (1-y^{(i)}) \log (1 - h_w(x^i)$$

$$\frac{\partial (C(w)}{\partial w_j^i} = \frac{-1}{M} \sum_{i=1}^{M} y^i \frac{\partial \log (h_w(x^i)))}{\partial w_j^i} + (1 - y^i) \frac{\partial \log (1 - h_w(x^i))}{\partial w_j}$$

$$\frac{\partial (\log (h_w(x^i)))}{\partial w_j^i} = \frac{1}{h_w(x^i)} \times \frac{\partial}{\partial w_j} (h_w(x^i_\theta))$$

$$\frac{\partial (\log (1 - h_w(x^i)))}{\partial w_j} = \frac{1}{1 - h_w(x^i_j)} \times \frac{\partial}{\partial w_j} (-h_w(x^i_\theta))$$

$$h_w(x^i) = \frac{1}{1 + e^{-w^T x}}$$

$$\frac{\partial h_w(x^i)}{\partial w_j^i} = \frac{-1}{(1 + e^{-w^T x^i})^2} (x^i_j)\, e^{-w^T x^i}$$

Page No.
Date

$$\frac{\partial}{\partial w_j} = \sum_{i=1}^{M} y^i \frac{1}{h_w(x^i)} \times \frac{+1 \; e^{-w^T x^i}}{(1+e^{-w^T x^i})^2} + (\quad)(x^i)$$

$$y^i \frac{1}{h_w(x^i)} \times -e^{-w^T x^i} \times (h_w(x^i))^2 \times x_j^i$$

$$+ x_j^i y^i \times \frac{e^{-w^T x^i}}{1+e^{-w^T x^i}}$$

$$(1-y) \frac{1}{1-h_w x} \times \frac{e^{-w^T x^i} \; x_j^i}{(1+e^{-w^T x^i})^2}$$

$$\frac{1}{1-h_w x} = \frac{1}{1-\frac{1}{1+e^{-w^T x_i}}} = \frac{1+e^{-w^T x^i}}{e^{-w^T x_i}}$$

$$(1-y) \times \frac{(1+e^{-w^T x^i})}{e^{-w^T x^i}} \times \frac{e^{-w^T x^i} \; x_j^i}{(1+e^{-w^T x^i})^2}$$

$$- \frac{(1-y) \; x_j^i}{(1+e^{-w^T x^i})}$$

$$x_j^i y^i \times \frac{e^{-w^T x^i}}{1+e^{-w^T x^i}} + x_i y^i - \frac{x_j^i}{1+e^{-w^T x^i}}$$

$$x_j^i y^i \left(\frac{1+e^{-w^T x^i}}{1+e^{-w^T x^i}}\right) - \frac{x_j^i}{1+e^{-w^T x^i}}$$

$$x_j^i \left(y_i - h_w(x_i) \right)$$

$$C(w) = -\frac{1}{M} \sum_{i=1}^{M} (h_w(x^i) - y^i)\, x_j^{(i)}$$

Multiclass Classification

=> Multiclass can be realized as one vs rest. binary classifier

=> Train a logistic Regression classifier $h_w^i(x)$ for each class i to predict the prob $y = i$

=> During Testing, for given x, pick class that maximizes the

$$\max_i h_w^i(x)$$

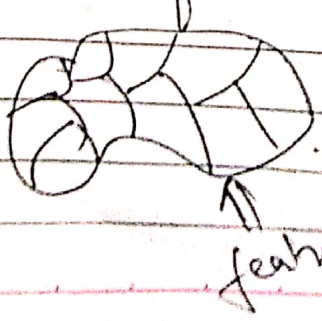
Use of Clustering for Classification.

1st method

In a common feature space clustering is done a label are assigned.

Disadvantage

: There may be overlaps in same feature space soft boundries which may lead to misclassification

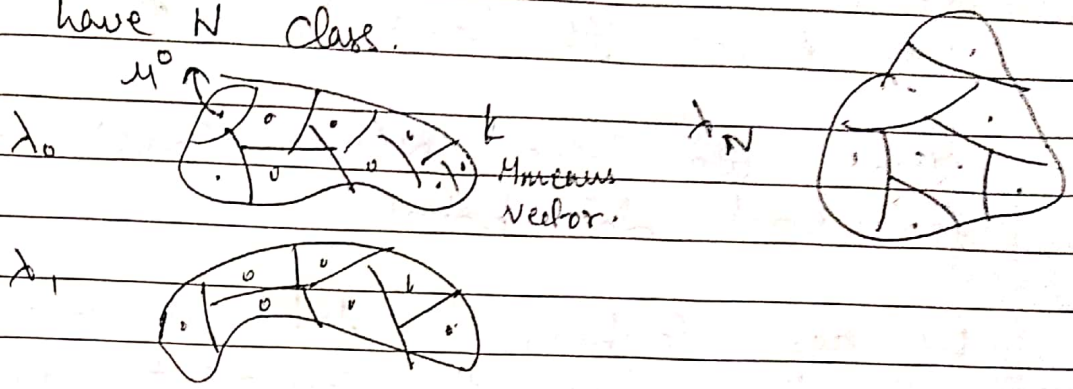


cluster boundry is not proper.

Page No.
Date

Generative Classifier Approach

we have N class.



Comparison

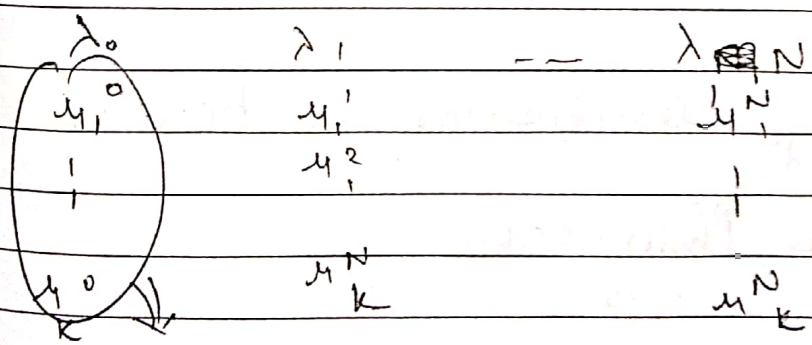
Method we have Separate feature space for each class label

we cluster that particular feature space into k-clusters.

Using this mean vector we try to compare and estimate class of unseen data

Comparison

$x = \{x_1, x_2 \cdots x_N\}$



min of all dis

Page No.
Date

We calculate distance for all class cluster and select the least one.

Then we have N(num class) distance for each data point.

we again compare and to be minimum of all those class, the class that has minimum we classify that point to that class

$$\{ d_0^1 \text{ --- } d_0^N \}$$

$$\min d_0^i$$

i = classes
j = data point.

Why to cluster?

If we don't we have to compare to all data points in that class that is computational hectic.

This way of classification is part of Generative Classification.

No need for other class data to build classifier