Date :
Page :

Ch⇒ 4

Classification

The linear regression we discussed in last chapter assumes response variable (Y) is quantitative.

But what if that variable is qualitative. Predicting a qualitative response for an observation is reffered to as classification.

An Overview of Classification :

Classification problems occure often, even more than regression problems :

1. Credit Card Fraud Detection : An online banking system uses it to detect frauds.

2. Classifying disease causing cells in a DNA Sequence : On the basis of DNA sequence, a biologist would like to figure out DNA mutations.

3. Classifying an Email is spam or Not : On the basis of text vectory a software developer would like to classify if the mail is spam or not.

Date :
Page :

Given a feature vector X & a qualitative response Y taking values in set C, the task is to build a function C(X) that takes as input the feature vector X and predicts its value for Y; i.e. C(X) ∈ C

Why not Linear Regression:

We have stated that linear regression is not appropriate in the case of qualitative response. Why not?

Let us suppose we can, we'll look at it with an example:

Let us say, we are trying to predict of 3 possible diagnosis:-

$$Y = \begin{cases} 1 : \text{stroke} \\ 2 : \text{drug overdose} \\ 3 : \text{epileptic seizure} \end{cases}$$

Applying LR:

*** This impression of "Y" implies that the difference between a stroke & drug overdose is equal to epileptic seizure & drug overdose.

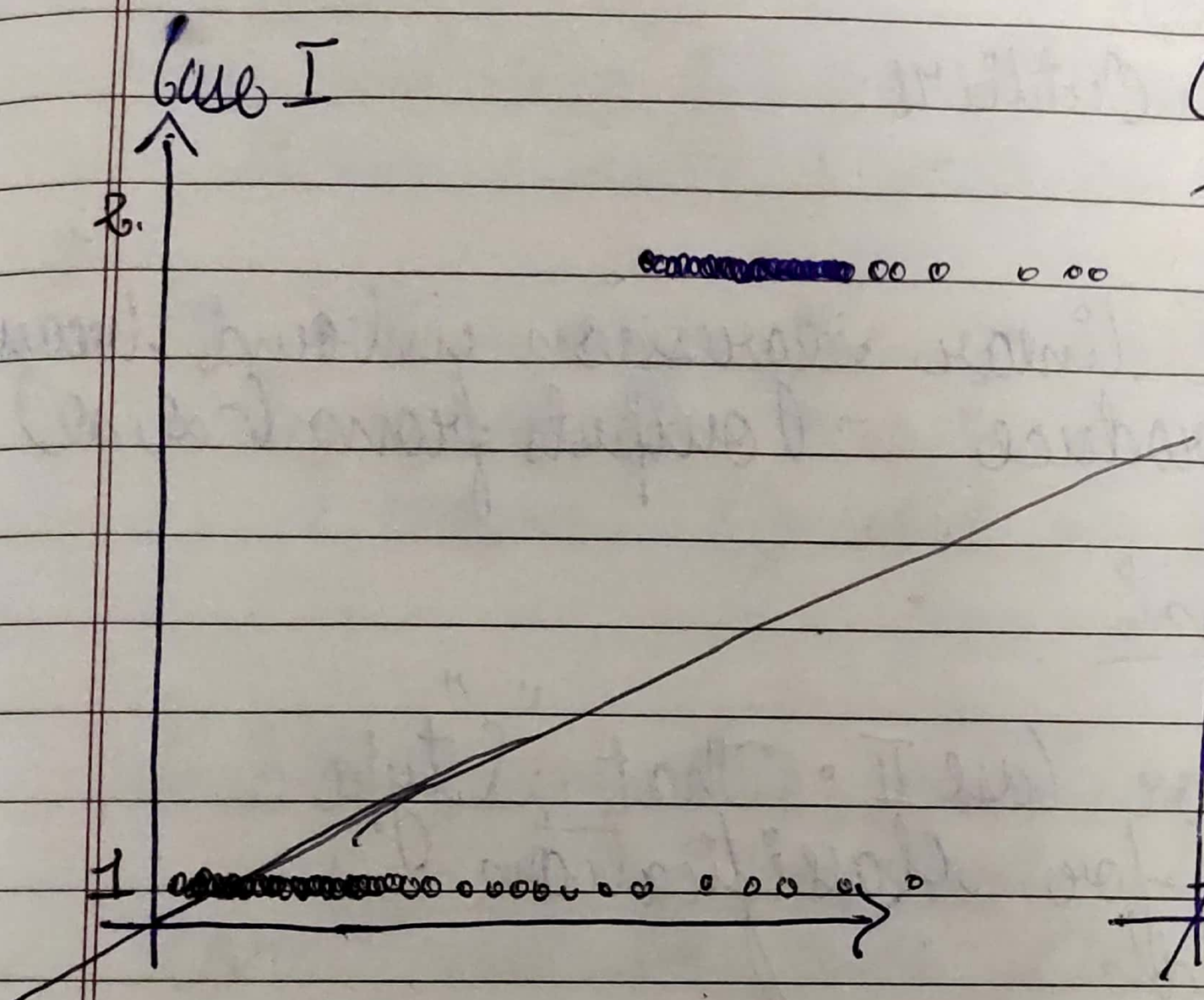
But in real life that's not the case.
For one instance we could encode this as:

$$Y = \begin{cases} 1 : \text{epileptic seizure} \\ 2 : \text{drug overdose} \\ 3 : \text{stroke} \end{cases}$$

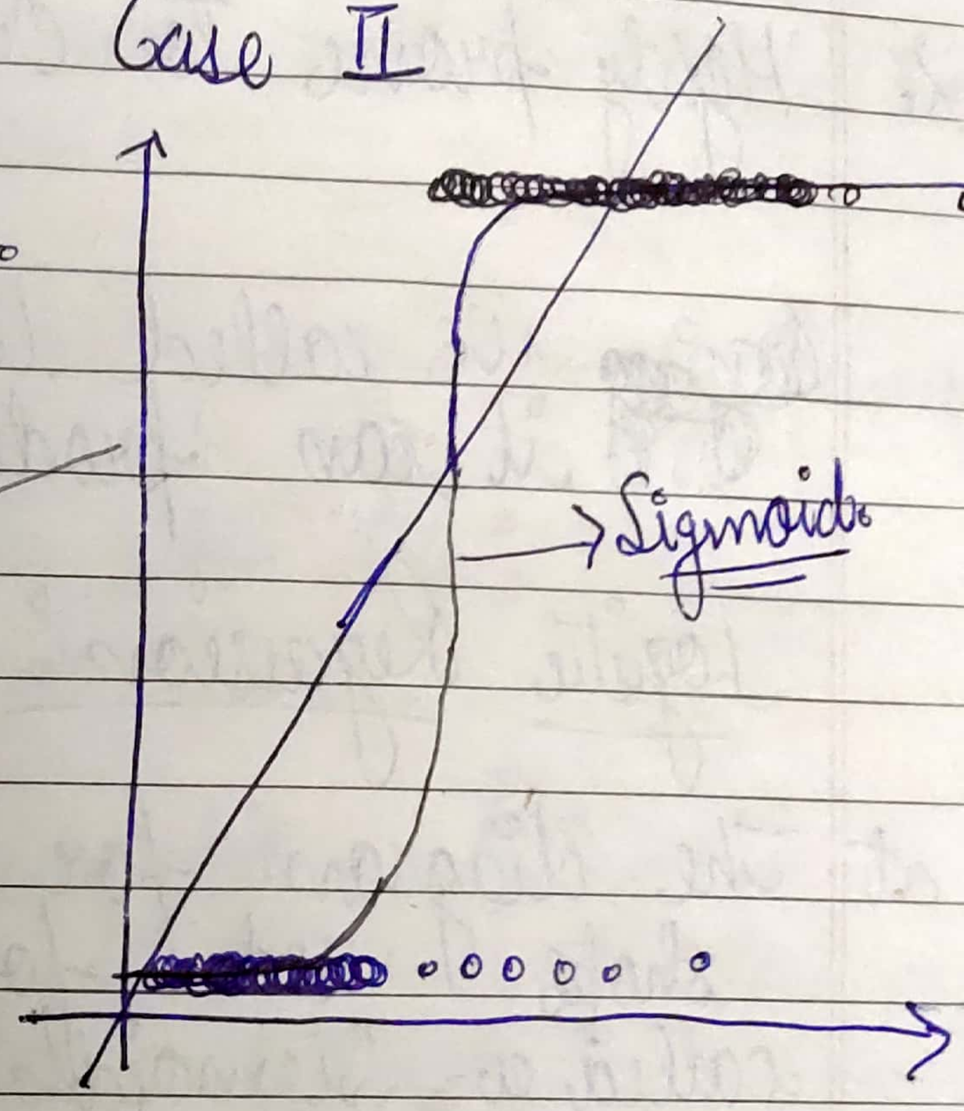
Date :
Page :

Using Linear Regression & Fitting the model & in Case 1&2 would be totally different.

Case I



Case II



- Stroke
- Drug Overdose

The graphs above implies 2 things :-

1. If the encoding / priority changes the graph as well best fit line changes widly. This implies high variation.
2. There is high variance in line (best fit) so the R^2 would be low.

2 major reasons why we can't use linear regression for classification:

1. Linear Regression is Unbound
2. Highly prone to Outliers

We called linear regression unbound because it can produce outputs from $(-\infty, \infty)$

Logistic Regression:

Look at The diagram for Case II. That "S" type shape used for classification is called a Sigmoid.

Let's write $p(x) = P(Y=1|X)$

Logistic Regression uses the form:

$$p(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

where, $e \approx 2.718$

& whatever values β_0, β_1 or X takes, $p(x)$ will have values between 0 & 1.

A piece of rearrangement gives:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X$$

This is called log odds or a logit of $P(x)$

Because of the output of log odds, it is called Regression

where, $\frac{p(t)}{p(1-t)}$ is called odds of t.

As shown in Case II graph, logistic regression ensures that our estimate for $p(x)$ lies between 0 & 1 & fits our data much better than linear Regression.

* It Solves both the problems given by Linear Regression.
where, Sigmoid is given by: $\frac{1}{1+e^{-x}}$

& Because the relationship between x & $p(x)$ is not linear, β_1 does not correspond to one change in $p(x)$ with one unit increase in x as there was the case in linear regression.

For a binary classification:

$p(x) \geq 0.5$; Class = 1
$p(x) < 0.5$; Class = 0

Estimation of Parameters:

The coefficients β_0 & β_1 are unknown and has to be estimated based on the available training data.
We use MAXIMUM LIKLIHOOD ESTIMATION to estimate the parameters of logistic regressio

MAXIMUM LIKLIHOOD ESTIMATION:

MLE is a method of estimating the parameters of a probability distribution by maximizing the liklihood function so that under the assumed model the observed data is most probable, in our case the model is logistic & parameters are β_0 & β_1

TIONAL

HOW DOES A LIKLIHOOD FUNCTION USED/ CALCULATED:

First, of all let's try to understand the difference between liklihood & probability.

probability & frequency/probability gives

Probablity of Y given Parameters is probability at a point "x" where distribution is known.

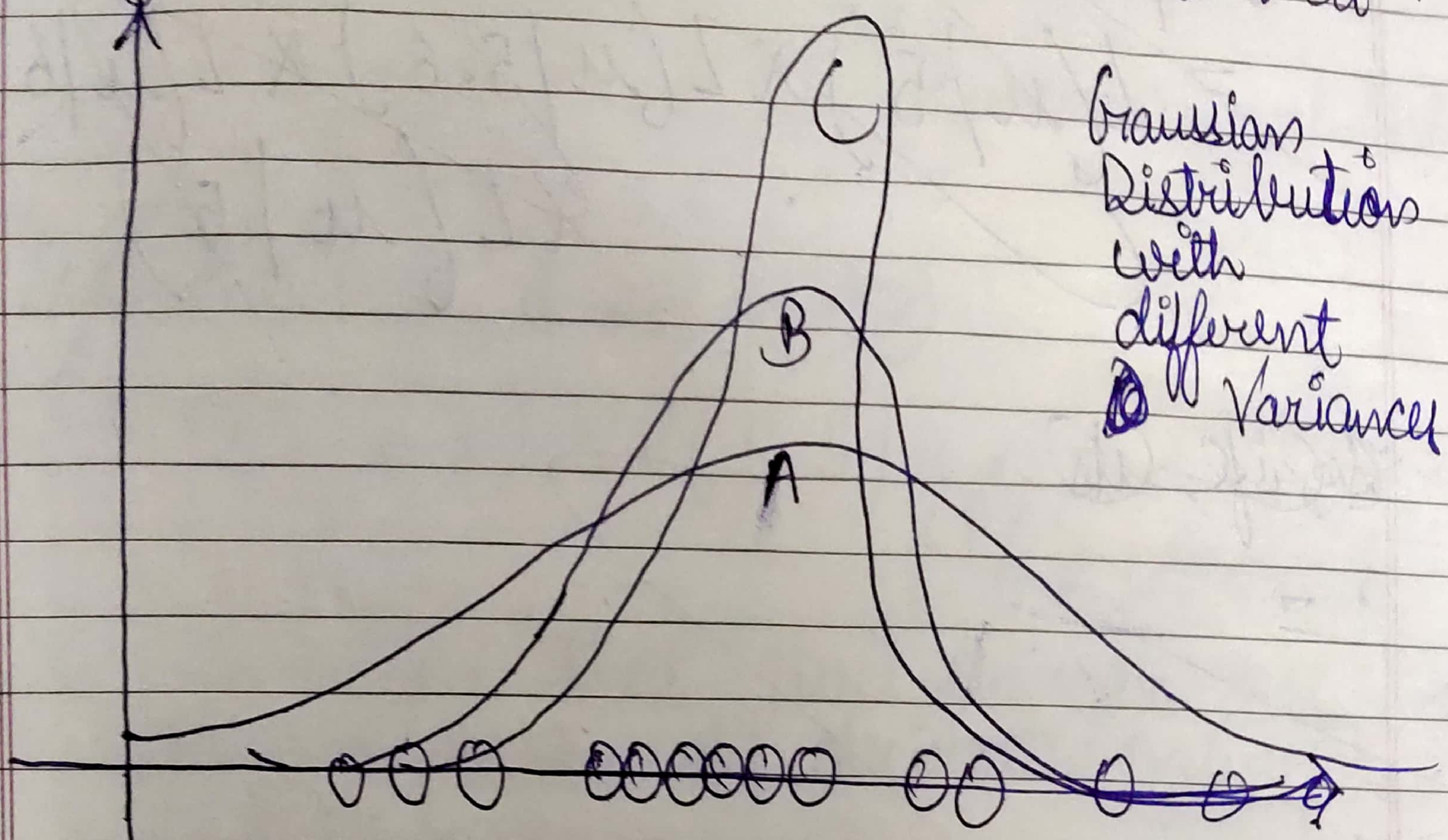
while

Liklihood is probability of data belonging to a certain distributed.

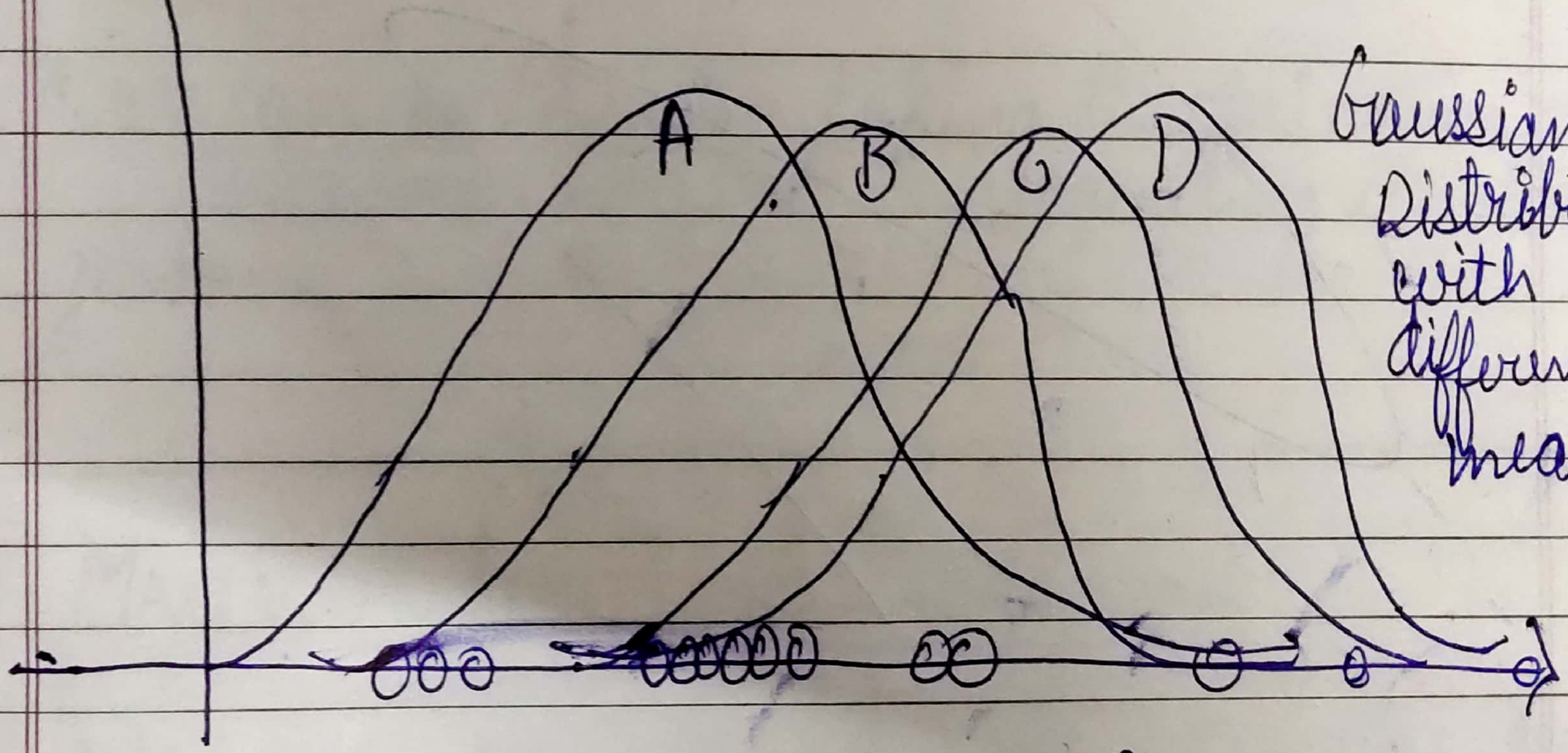
It sounds like a hickey, but you'll understand through an example:

Let us assume we are given a data & we are trying to find a best distributions that fits that specific data.

We decided that we want to fit gaussian Normal distribution to fit it.



Gaussian Distributions with different Variances



Gaussian Distribution with different means.

So, Maximum likelihood estimation lets you get the best fit distribution from infinite distributions with different means & variances.

Lets say 4 data points are given & we have been given a task to decide mean & variance for distribution using MLE.

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

(Gaussian Normal Distribution)

$L(\mu,\sigma \mid X = 5, 5.5, 6, 5)$

$= P(5|\mu,\sigma) \times P(5.5|\mu,\sigma) \times P(6|\mu,\sigma) \times P(5|\mu,\sigma)$

The distribution that gives the max value for Liklihood here will decide the value for μ & σ.

* We can do this by taking the derivative of Liklihood function & equating it to zero.

MAXIMUM LIKLIHOOD ESTIMATION

We use Maximum Liklihood Estimation to estimate the parameters.

$$L(\beta_0, \beta_1) = \prod_{y_i=1} P(x_i) \prod_{y_i=0} (1 - P(x_i))$$

MARINGR PREDICTIONS

Once the coefficients are estimated we put

those estimates in $\hat{p}(x)$

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

if this value is > 0.5 then class-1 is predicted

else class-2 will be predicted

** we can always manipulate this threshold.

MULTI-VARIABLE LOGISTIC REGRESSION:

This changes nothing, the only thing that changes is the number of predictors.

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta} x_3}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x_2 + \hat{\beta} x_3}}$$

MULTI-CLASS LOGISTIC REGRESSION:

$$P(X=k|X) = \frac{e^{\beta_{0k} + \beta_{1k} X_1 + \beta_{2k} X_2 \cdots + \beta_{pk} X_p}}{\sum_{l=1}^{k} e^{\beta_{0l} + \beta_{1l} X_1 + \cdots + \beta_{pl} X_p}}$$

Multiclass Logistic Regression is also called Multinomial Regression

Date :
Page :

✱ Linear Discriminant Analysis:

Here approach is to model the distribution of X in each of the classes seprately & then use Bayes theorem to flip things around and obtain $Pr(Y|X)$.

BAYES THEOREM FOR CLASSIFICATION:

Thomas Bayes was a famous mathematician whose name represents a big subfield of stats & probability.

Here we'll focus on BAYES-THEOREM:

$$Pr(Y=K \mid X=x_0) = \frac{Pr(X=x_0 \mid Y=K) \cdot Pr(Y=K)}{Pr(X=x_0)}$$

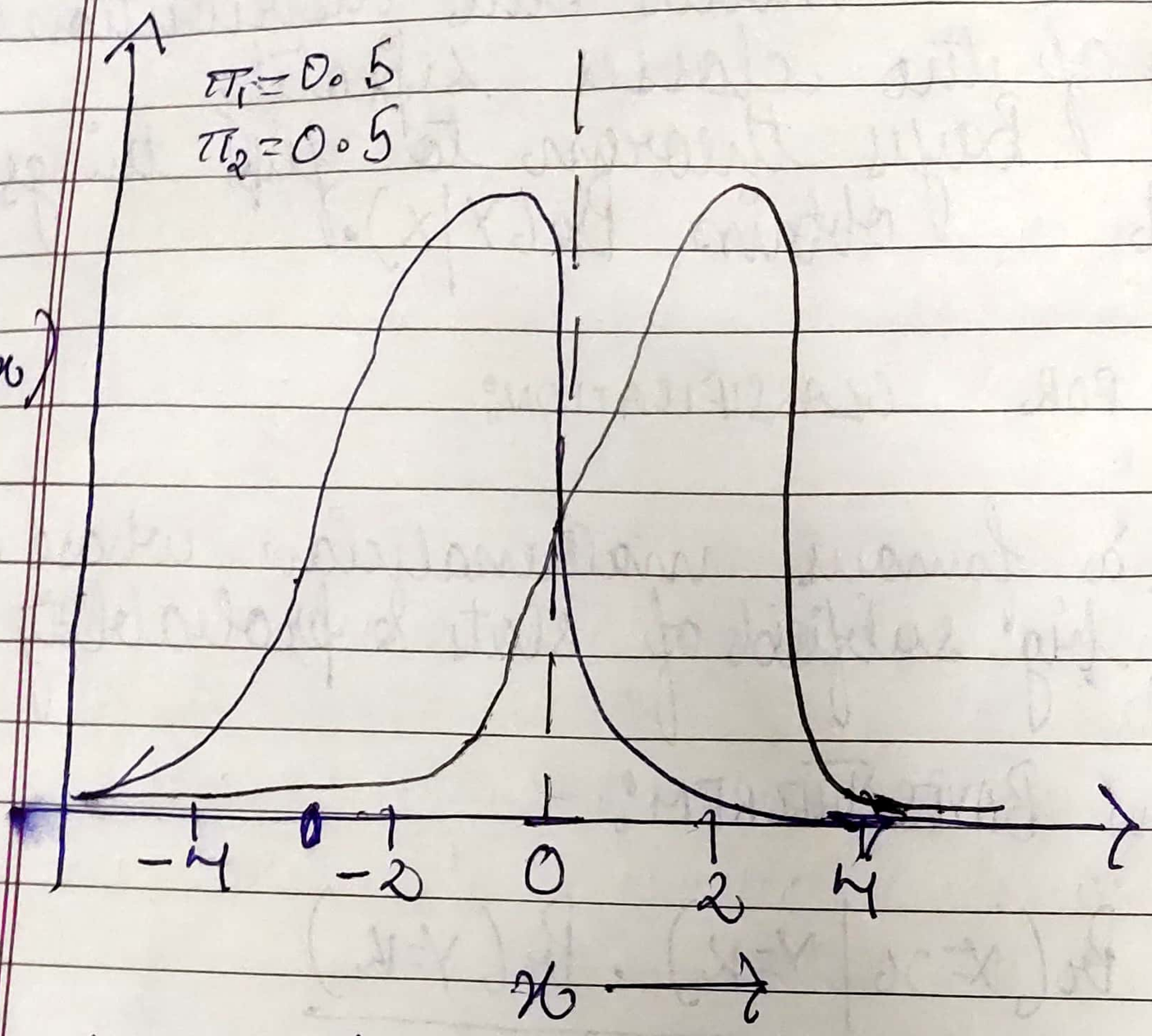
In discriminant Analysis it is written quite differently:

$$Pr(Y=K \mid X=x_0) = \frac{\pi_K f_K(x_0)}{\sum_{l=1}^{K} \pi_l f_l(x_0)}$$ where

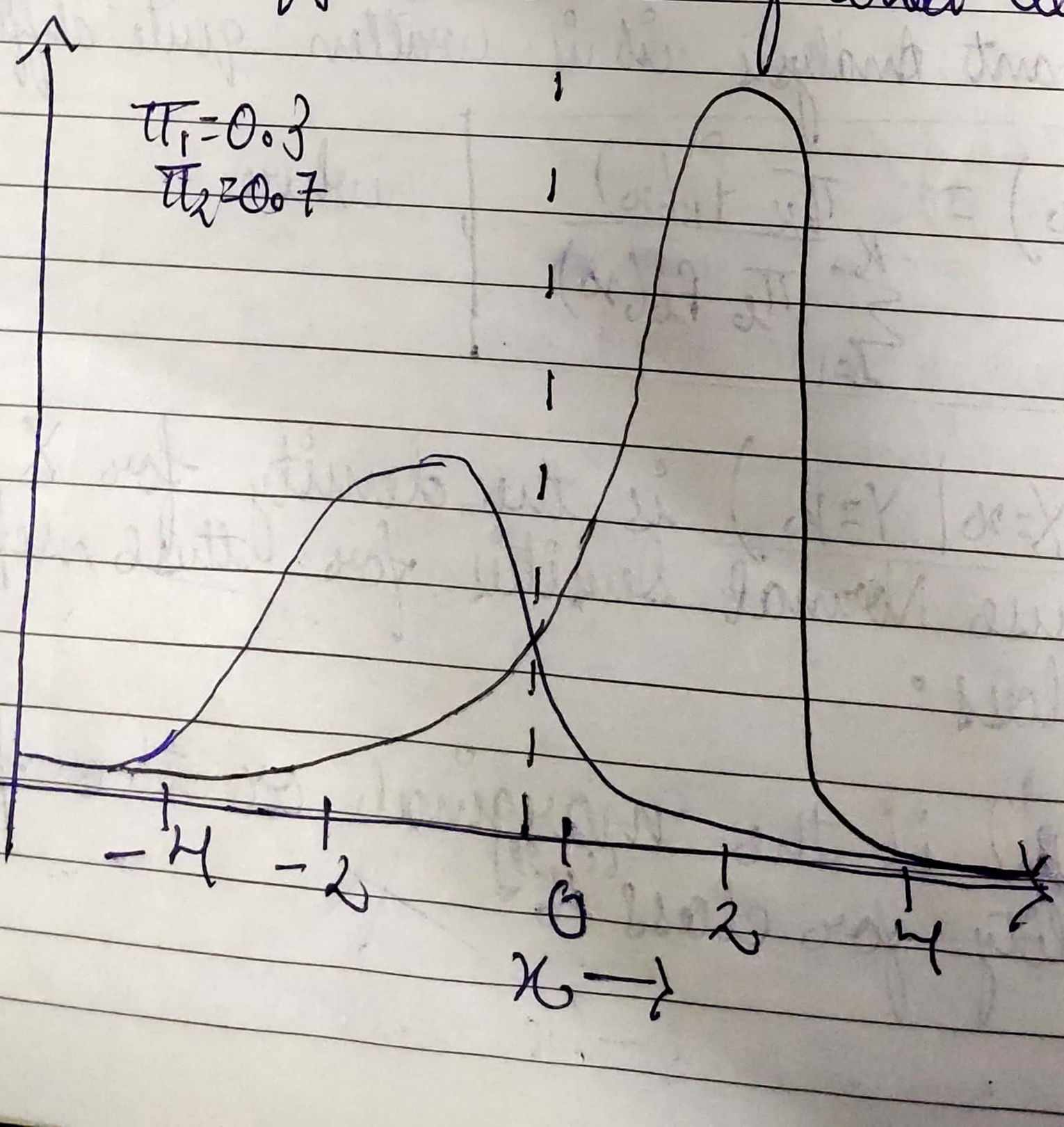
→ $f_K(x) = Pr(X=x_0 \mid Y=K)$ is the density for X in class K. Here, we use Normal Densities for these, seprately in each class.

→ $\pi_K = P(Y=K)$ is the marginal or the prior probability for class "K".

* How do we classify on the basis of density functions



we classify on the bases of what density is best?



When the priors (π_k) were different we had to take them into account.
In the second graph we favor the black class. — the decision boundary was shifted to the left.

Why discriminant Analysis ?

→ When the classes are well separated estimates of Logistic Regression are surprisingly unstable. LDA doesn't suffer from this problem. (Because predictors will overshoot to ∞)

→ If n is small & distribution of predictors are approximately normal, then obviously LDA will perform better.

→ LDA performs better in multiclass - classification, because it also provides low dimensional view of the data.

Linear Discriminant Analysis when $p = 1$.

When there is only one predictor The Gaussian Density has the form:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\,\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$$

LDA takes the assumption that $\sigma_k = \sigma$

Plugging the density functions into the Bayes form gives:

$$P_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

The formula above looks pretty ugly.

What we can do is take the log of probality function & discard all the terms which are independent of μ_k, x.

Why can we discard those terms because we focus on what probality is higher & constant terms doesn't make any change.

This is called a discriminant score and is given by:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

* Note this, the function above is linear in Nature.

We can simplify the application even further when $K=2$ & $\pi_1 = \pi_2 = 0.5$.

$\delta(x)$ will be equal at decision boundary.

$$x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi_1) = x\frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi_2)$$

$$x(\mu_1 - \mu_2) = \frac{\mu_1^2}{2} - \frac{\mu_2^2}{2}$$

$$x = \frac{\mu_1 + \mu_2}{2}$$ } the decision boundary

when $\pi_1 = \pi_2$ & $K=2$ classes.

Estimating the parameters:

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$$

$$\hat{\sigma}_x^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{y_i = k} (x_i - \hat{\mu}_k)^2$$

$$= \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \cdot \hat{\sigma}_k^2$$

where, $\hat{\sigma}_k = \frac{1}{n_{k-1}} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)^2$ is usual formula for the estimated variance in k^{th} class.

Linear Discriminant Analysis when $p > 1$

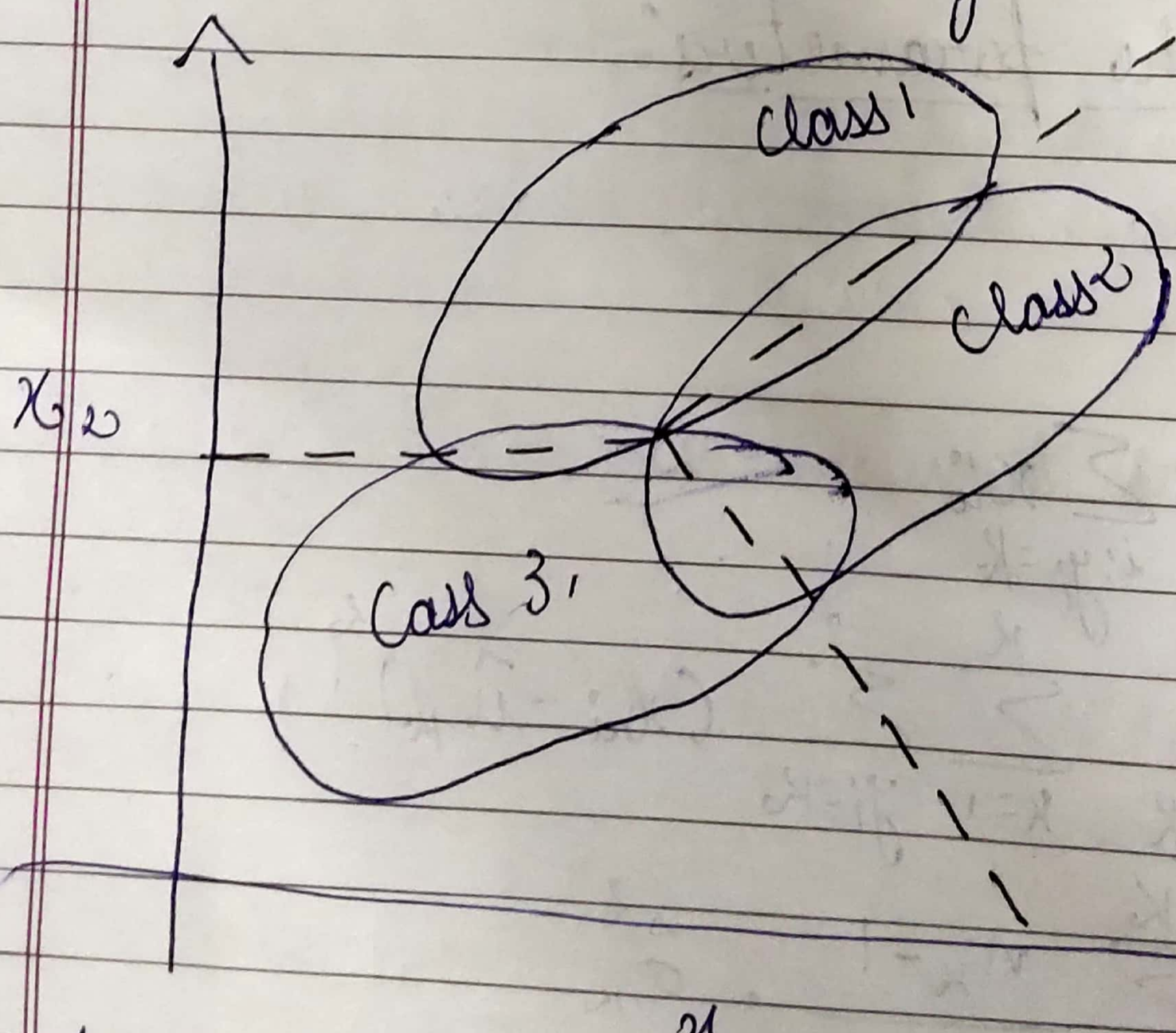
Density: $f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$

The density function changes when we move to multivariate data.

Where, Σ = Covariance Matrix
& p = # predictors.

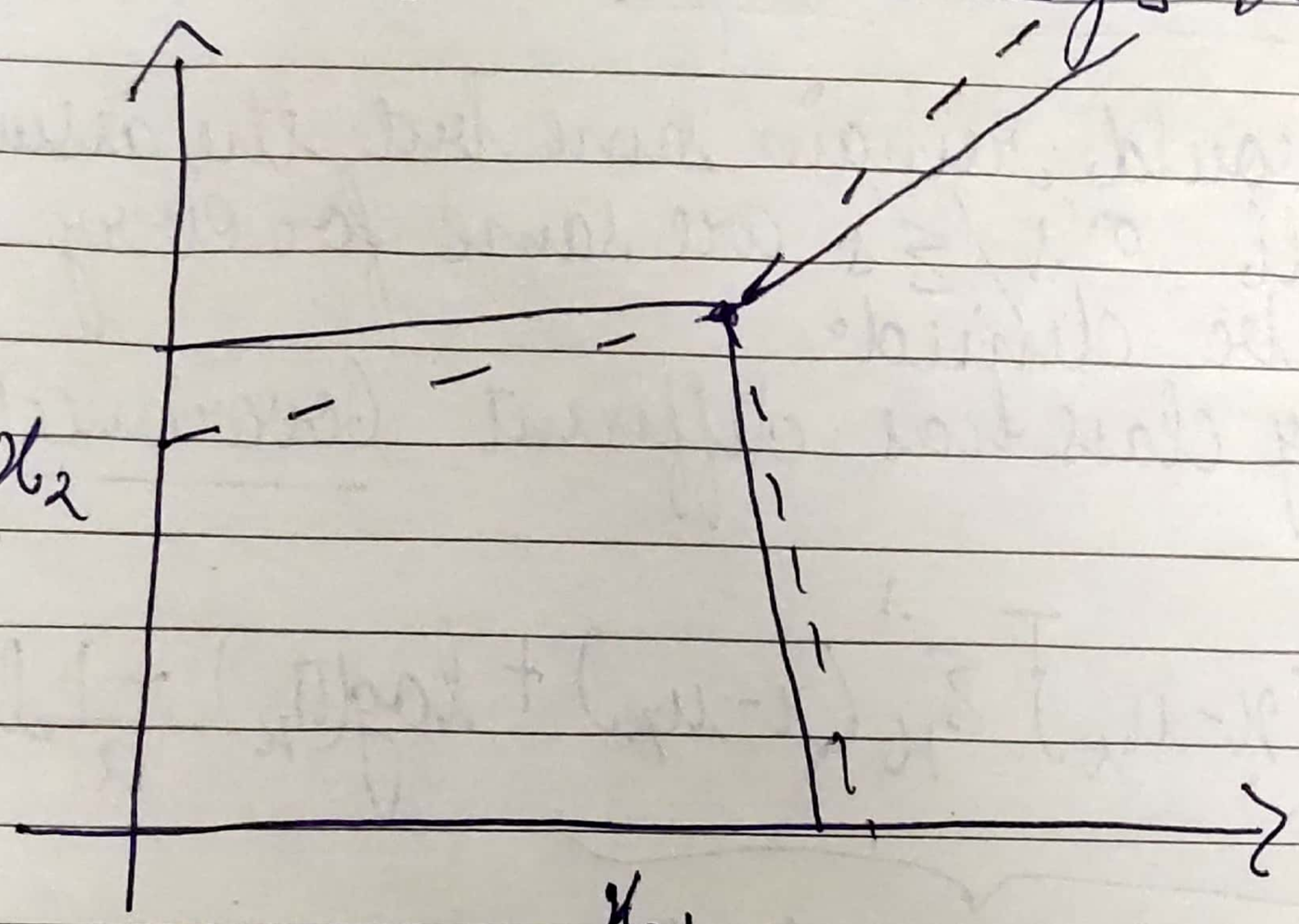
Discriminant function: $x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log$

This is still linear as you can see.



In the plot above we have shown 3 classes as contours of data rather than a scatterplo

The dashed lines are Bayes Decision Boundary.



In the plot above dashed lines are Bayes decision boundary & solid lines are LDA boundries.

From $\delta_k(x)$ to probabilities:

Once we have estimates $\hat{\delta}_k(x)$, we can turn it into estimates of probability.

$$Pr(Y=k \mid X=x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

So, classifying largest $\hat{\delta}_k(x)$ leads to classifying largest $P(Y=k \mid X=x)$

Quadratic Discriminant Analysis:

Everything would remain same but the assumption that all σ's / Σ's are same for every class will be denied.
Now, every class has different Covariance Matrix.

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log(\pi_k) - \frac{1}{2}\log|\Sigma_k|$$

This Quadratic in Nature.
Because, Σ_k is different.

Naive Bayes Classifier

This assumes features are Independent.
This is useful when p is large, here methods like LDA & QDA Breaksdown.

→ Gaussian Naive Bayes assumes Σ_k is diagonal.
Read about it

$$\delta_k(x) \propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right]$$

$$= -\frac{1}{2}\sum_{j=1}^{p}\left[\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log\sigma_{kj}^2\right] + \log\pi_k$$

Date :
Page :

when the data is discrete, $f_{kj}(x_j)$ is replaced by probability mass function (histogram)

Logistic Regression v/s LDA.

For a two class problem, we can show that

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) \text{ is of the form}$$

$$= c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_p x_p$$

So, it has same form as logistic regression.

The difference is how the parameters are estimated:

→ Logistic Regression uses conditional likelihood based on $Pr(Y|X)$ known as discriminative learning.

Vs

→ LDA uses full likelihood based on $P(X,Y)$ known as Generative learning.

Logistic Regression uses only the distributions of Y's
while
LDA uses distributions of both X & Y

Date :
Page :

SUMMARY

→ Logistic Regression works well when $K=2$

→ LDA works well when n is small & classes are well separated. Also when $K>2$

→ Naive Bayes is useful when p is very large.