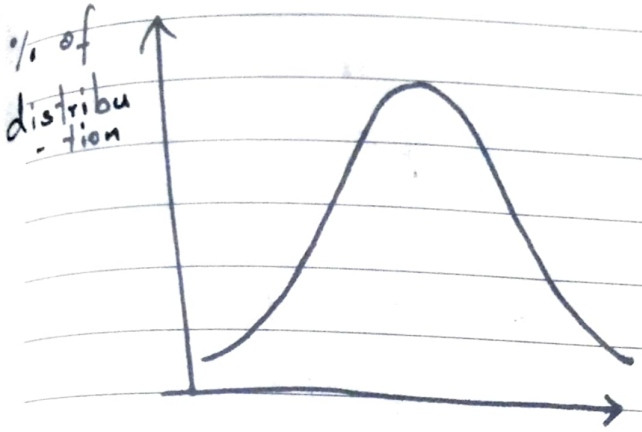
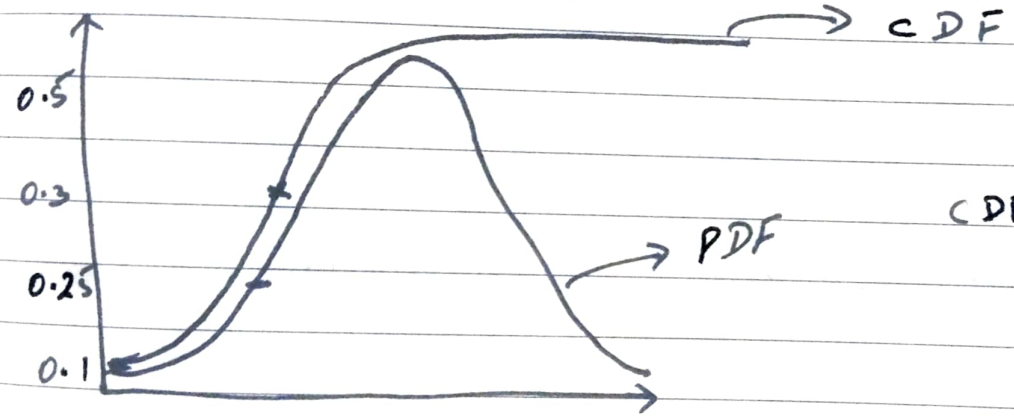
Date

Probability Density function (PDF)



Cumulative Density Function



CDF ⇒ 0.1 + 0.25
= 0.35

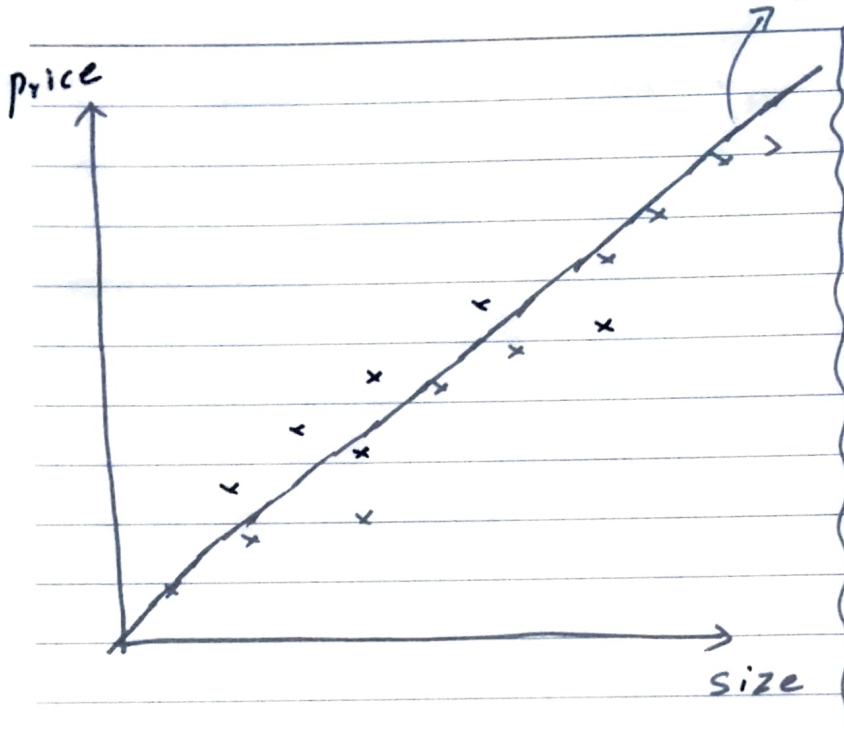
Linear Regression in depth Math

$y = mx + c$ --> Best fit line

m -> slope

c -> intercept

Date:



Cost function

$$= \frac{1}{2m} \sum_{i=1}^{m} (\hat{y} - y)^2$$

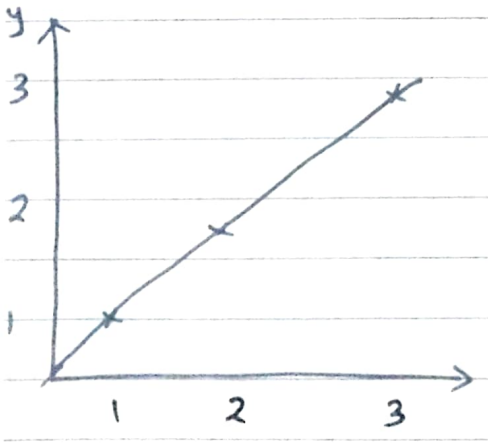
m → No of points

$\hat{y}$ → predicted besfit line

$\hat{y}$ → Actual data point

best fit line is actually a $\hat{y}$

Cost function should be minimum



$\hat{y} = mx + c$

← considering $c = 0$

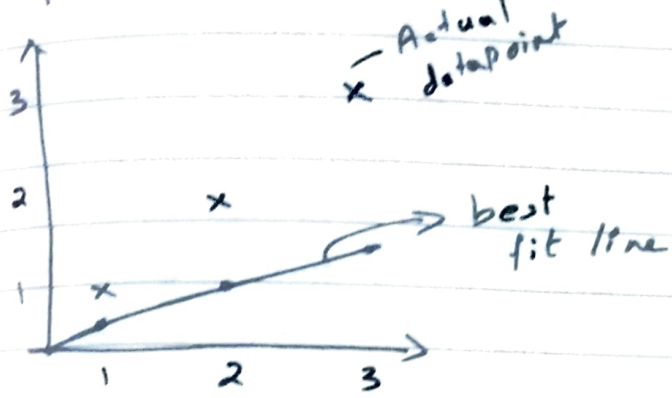
$m = 1$

$$\text{cost function} = \frac{1}{2(3)} \left((1-1)^2 + (2-2)^2 + (3-3)^2 \right)$$

$$= 0$$

Date

if m = 0.5, c = 0



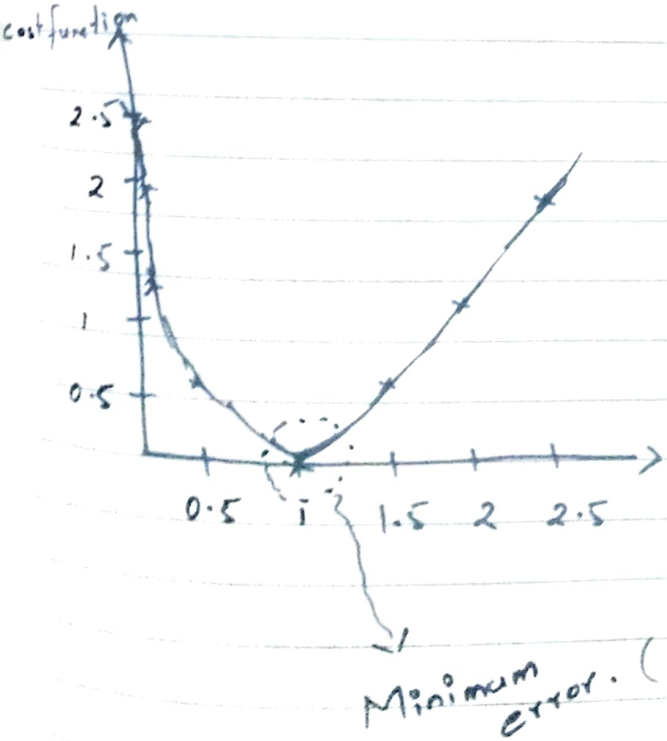
$$\hat{y} = (0.5)(1) + 0$$
$$= 0.5$$

cost function

$$= \frac{1}{2(3)} \times \left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right]$$

$$= 0.58$$

Gradiant Descent



for different slope (m) selection the cost function changes.

← this will help us to selecting (m) value slope value

Date

inorder to move down to reach global minima
we need to write

Convergence theorem

$$m = m - \left(\frac{\partial m}{dm}\right) \times \alpha$$

α → learning rate → learning rate should be small value

$\left(\frac{\partial m}{dm}\right)$ → slope

if the slope is negative

$m = m - (-ve) \times \alpha$

$= m + \alpha$

$=$

greater value → Negative slope → lesser value

Left → greater
Right → lesser

if the slope is positive

$m = m - (+ve) \times \alpha$

$= m - \alpha$

$=$

greater value → positive slope. → lesser vale

Left → lesser
Right → greater.

Date

in each iteration the cost function will reduce and try to attain global minima.

Ridge & Lasso Regression

Ridge if the slope is huge(steep), we try to penalizing with the help of ridge regression with the below formula

$$\text{cost function} + \lambda (\text{slope})^2$$

m^2

[λ is b/w 0 and $+\infty$]

It will help to reduce the steep of the slope.

By this we can reduce overfit condition

Lasso

$$\text{Cost function} + \lambda \, |\text{slope}|$$

$$\downarrow$$

$$\text{Cost function} + \lambda \times |m_1 + m_2 + m_3|$$

Lasso performs well when we have many features.

Date

$$R^2 \text{ score} = 1 - \frac{SS_{res}}{SS_{mean}}$$ (Sum of residual) / (Sum of mean)

$$SS_{res} = \frac{1}{n} \sum_{i=1}^{n} (y - \hat{y})^2$$

$$SS_{mean} = \frac{1}{n} \sum_{i=1}^{n} (y - \bar{y})^2$$

To check Multicollinearity Statsmodel. OLS Summary interpret

1. We need to check $\beta_0, \beta_1, \beta_2, \beta_n$
2. R^2 & Adj R^2 should near 1
3. Std error should be very very less
4. Pvalue should less than 0.05

Date

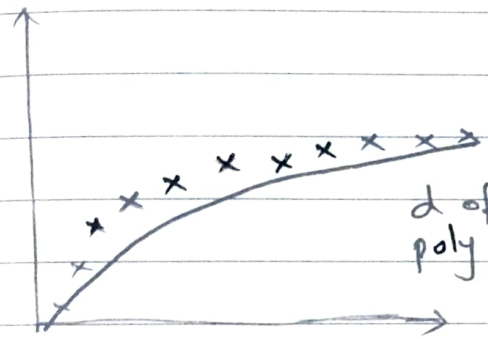
Bias & Variance

In terms of Train data

degree of polynomial = 1

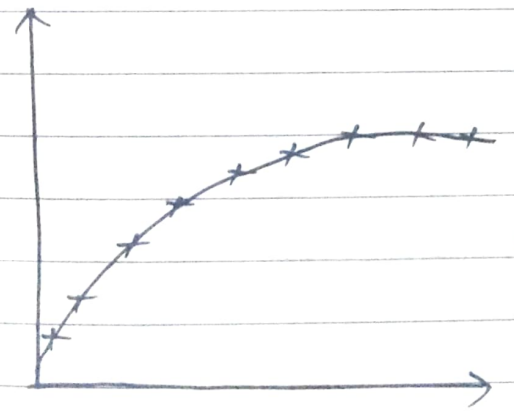
← When the error is high it is underfitted model

← high bias & high variance



d of poly = 2

← low bias & low Variance good model to choose



d of poly = 4

When the error is very very low, it is Overfitted model

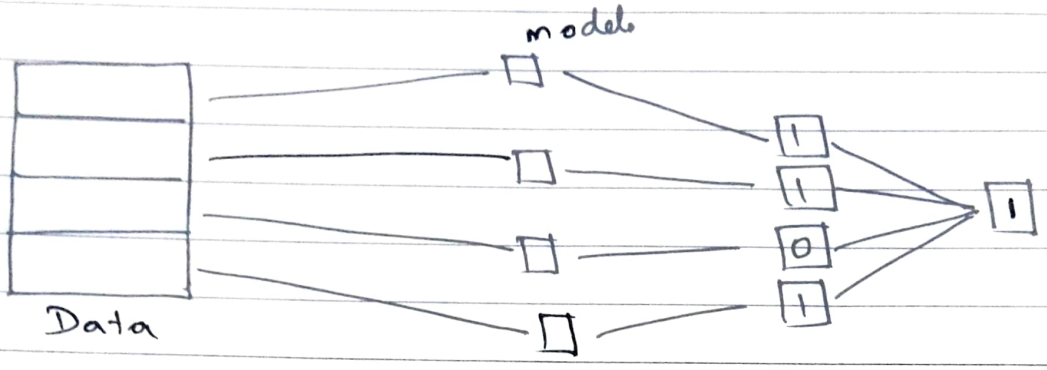
low bias & high variance.

Consider bias → error of train data

Variance → " " test data.

Date

Bootstrap Aggregation



spliting data and passing into various models and getting the highly occured result

eg Random Forest

Decision tree

low Bias, high variance

Random forest

low bias, low variance.

Date

R^2 & Adj R^2

range → (0 to 1)

$$R^2 = 1 - \frac{SS_{res}}{SS_{mean}}$$

(SS_{res} : sum of residual, SS_{mean} : sum of mean)

Sum of residual (or) error

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$\hat{y}$ → Predicted points.

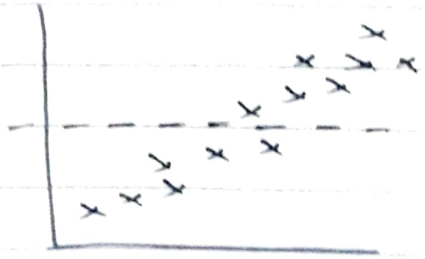
← error sum of this

Sum of mean (or) total average

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$\bar{y}$ → mean

↳



← error sum of this

R^2 can become below 0

When the predicted best fit line is worser than Average line it will become zero.

It means that is the worst model.

Date

Adj R^2

When we are adding more features, the R^2 value will automatically increase, coz it will try to assign some coefficient value, It happens always.

So we need to use Adj R^2

$$\text{Adj } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-P-1}$$

(No of indept fea
P -> No of predictors
N -> No of sample size

It will penelize the attributes that are not correlated with target.

Adj R^2 always less than or equal to R^2

Date

Hypothesis testing

We use this to evaluate 2 mutually excusive statement on Population using sample data.

steps

1. Make initial assumption (H_0)
2. start Collecting data
3. decide to Reject or Not Reject (H_0)

mutually exclusive
2 events can't occur at the same time eg tossing a coin

H_0 → Null hypothesis

	H_0	H_1
Not Reject	Ok (TP)	Type 2 Error (FP)
Reject	Type 1 error (FN)	Ok (TN)

Type I Error
Actually True (Not Reject) by test result false (Reject)

Type II Error
Actually False (Reject) by test result true (Not Reject)

Date

T-Test, chisquare test, ANOVA Test

if we are considering only one categorical column at that time we must use one sample proportion test

eg:-

Gender	AGE	Height
M	45	. .
F	32	. . .
M	17	. . .
M	29	. .
F	19	. .
F	7	. .

if we take this to test

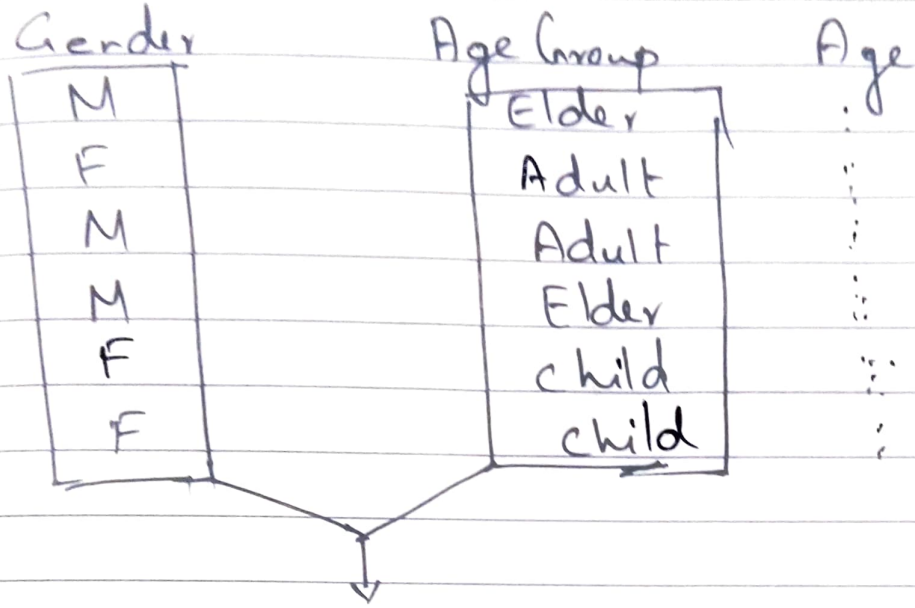
h_0 → No difference in proportion

h_1 → There is a proportion

Pval → ≤ 0.05

if pval less than 0.05 or 5% we will reject h_0 by accept h_1

if we consider 2 categorical features for test then we should use Chi-square test.



h_0 → based on age group No diff in Proportion

h_1 → difference is there

---

for One continuous variable

T-Test

for 2 continuous variable

T-Test / Correlation test

Date

ANOVA Test

One or more Categorical & Numerical features but the Categorical feature have more categories then we should use ANOVA Test.

Metrics in classification problem

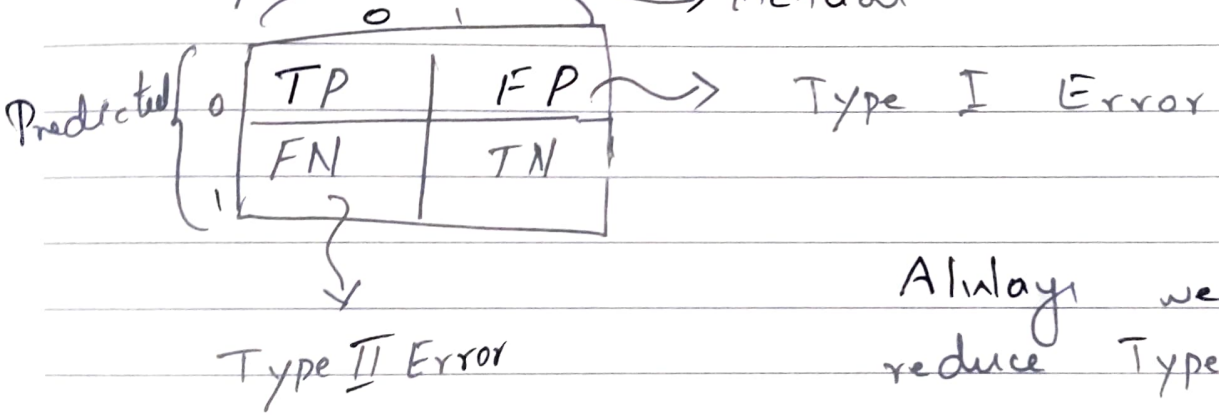
Confusion Matrix
FPR (Type I Error)
FNR (Type II Error)
Recall (TPR, Sensitivity)
Precision (Positive Pred val , Specificity)
Accuracy
F Beta
cohen kappa
ROC AUC
PR Curve

Date

If we have balanced target feature we can consider "Accuracy"

if we don't have balanced target then we can consider Recall, Precision, Fbeta.

Confusion Matrix



Always we should reduce Type I, II Errors.

$$FPR = \frac{FP}{FP+TN}$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

Date

Recall / TPR / Sensitivity	Precision / Specificity / positive Pood value
$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$
Out of all $+^{ve}$ values how many $+^{ve}$ actually predicted correctly	Out of all predicted $+^{ve}$ values, how many are actually positive values.

Whenever FP is much more important go with Precision (eg:- email Spam detection Problem)

Whenever FN is much more important go with Recall (eg:- Cancer detection)

If Both FP & FN is more important then we go with F_β ata score.

Date

$$F_{Beta} = (1+\beta^2)\frac{Precision \times Recall}{\beta^2 + Precision + Recall}$$

$\beta = 1 \rightarrow F_1$ score

$\beta = 0.5 \rightarrow F_{0.5}$ Score

$\beta = 2 \rightarrow F_2$ Score

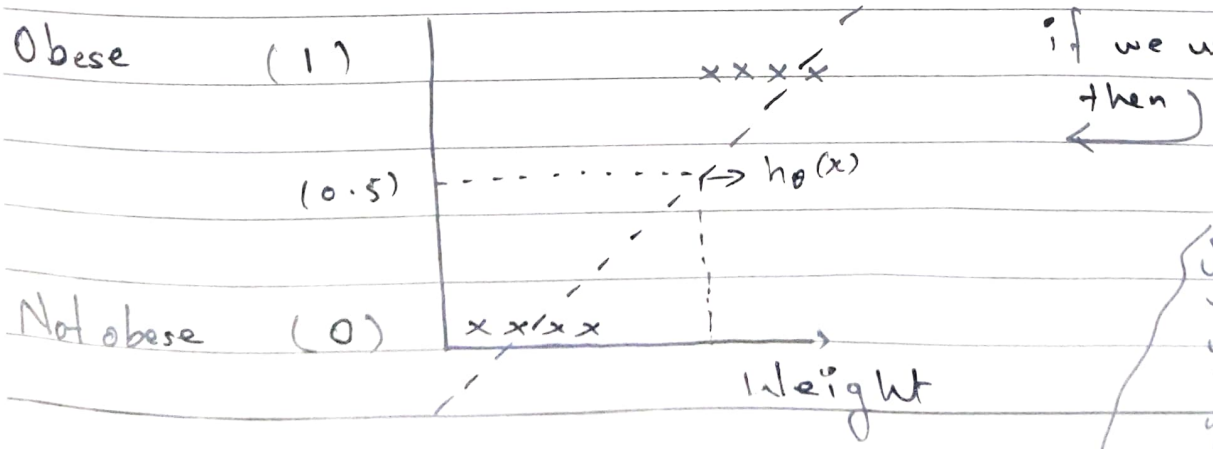
When Both FP & FN are important $\rightarrow F_1$ Score

FP have more impor than FN $\rightarrow F_{0.5}$ Score

FN have more impor than FP $\rightarrow F_0$ $\rightarrow$ greater than 1

Logistic Regresion

* Binary Classification
* Multiclass Classification



if we use LR then

$y = mx + c$

$y = \beta_0 + \beta_1 x$

$y = w^T x + b$

$h_\theta(x) = \theta_0 + \theta_1 x$

Date

Reasons for not using Linear Regression for Classification

Whenever we hv lot of outliers the best fit line completly gets deviated

if we get bestfitline greater than 1 len than 0 wrt datapoint at that time we have no option.

We have to use logistic regression to overcome these problems.

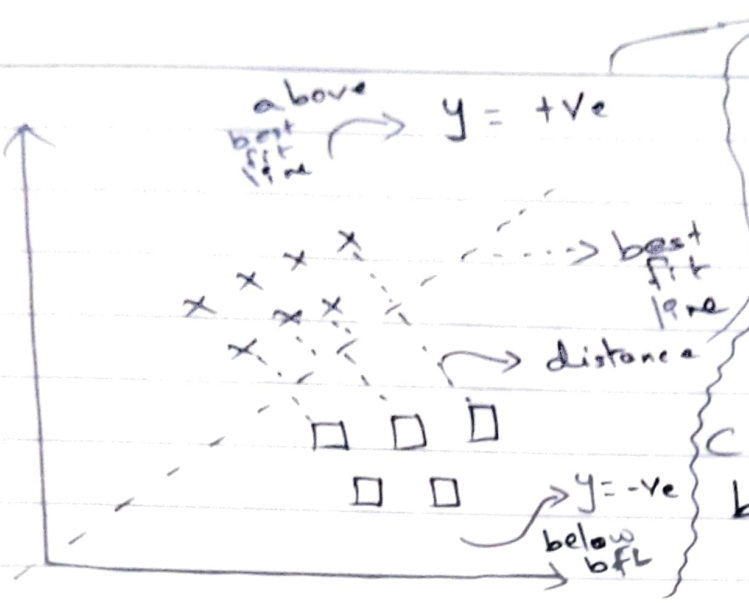
Logistic Regression

Cost function or Optimizer $= \sum_{i=1}^{n} y_i w_i^T x_i$ ⇝ it should be Max

Sigmoid function

$$z = \sum_{i=1}^{n} f(y_i w_i^T x_i)$$

$= \frac{1}{1+e^{-z}}$ → 0 to 1



Multiply $(y_i, w_i^T x_i)$ data

if we have 2 classes we will try to classify using a best fit line

then we will find the distance between the plane (or) best fit line and all the data points

Intercept is 0 in this case bcoz it passes through origin

$y = mx + c$
$y = w^T x + b$ → any
$y = \beta_0 + \beta_1 x$

$w^T x + b$ → Intercept
$w^T x + 0$
↳ it will become distance b/w plane & datapoint

if it is above the plane +ve distance
below " " -ve "

multiply with y_i

$(+ve) + (+ve) = +ve$
$(-ve) + (-ve) = +ve$

$\sum_{i=1}^{n} (y_i, w_i^T x_i)$ ↳ cost function

Date

Cost function is Z

sigmoid $\frac{1}{1+e^{-z}}$ → it will range between 0 & 1

For Multiclass classification

Change the parameter

multiclass : OVR