Date ___/___/___

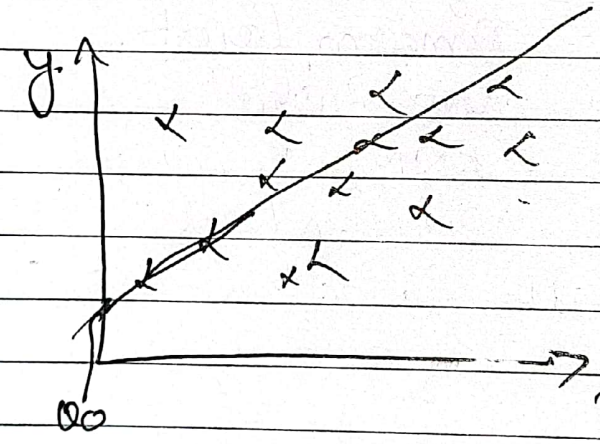
Page _____

Simple Linear Regression

x → independent
y → dependent

Find the best fit line such a way that the difference between the errors and after doing submission should be minimum

whichever is the best fit line and minimal distance is there we can say that model is trained well.



Eq^n of straight line
$y = mx + c$
$y = \beta_0 + \beta_1 x$

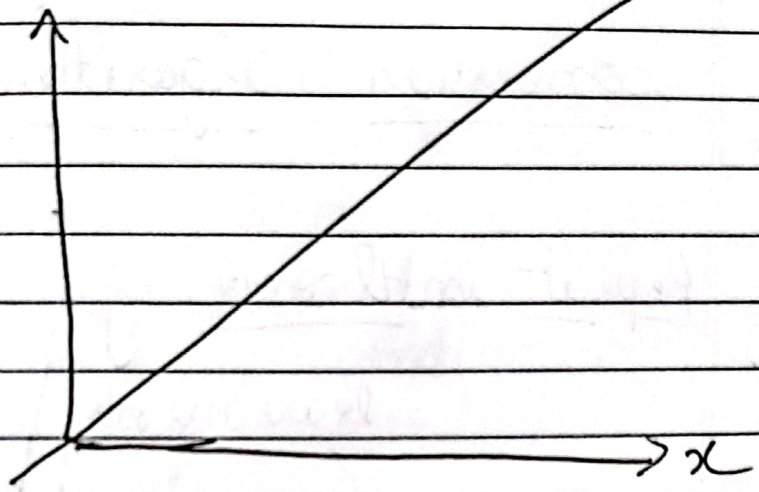
$$h_\theta(x) = \theta_0 + \theta_1 x$$

$\theta_0 \rightarrow$ intercept
$\theta_1 \rightarrow$ slope

Intercept = 0

Page

$\theta_0 = 0$



Cost function

predict Actual

Imp $J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

mean squared error

$m \Rightarrow$ all the datapoints

final aim

minimize $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

$\frac{k}{3}$

Orchies

Date...../...../.....

Page.....

IMP

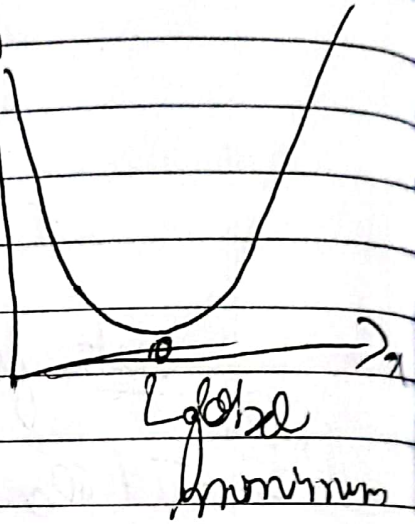
convession algorithm. (optimize the changes of θ_1 value)

Repeat until conv. {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_j)$$

learning rate

}



$$\theta_j := \theta_j - \alpha \text{ (-ve)}$$

$$= \theta_j + \alpha$$

learning rate decides the speed of convergence

Date...../...../.....

Page.....

Q1 What is linear regression?

Ans It is a statistical method that is used for predictive analysis. Linear reg alg shows a linear relationship between a dependent (y) and one or more dependent (y) variables hence called as linear regression.

Q2 how we can calculate error in linear regression?

Ans mean squared error
error b/w predicted value and real error

Q3 diff b/w loss and cost function?

Ans loss is for one

Date...../...../.....

Page.....

Linear Reg Alg.

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

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_n x_n$

Imp **convergence algorithm**

$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - (y^{(i)})^2$$

mean squared error → cost fmctn

loss fmctn Vs cost fmctn

In cost fmction we have to check every observation.

loss fmctn $= (h_\theta(x^{(i)}) - y^{(i)})^2$

$= (\hat{y}^{(i)} - y^{(i)})^2$

predicted value, actual value

Date...../...../.....

Page.....

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_0} \left[\frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x)^{(i)} - y^{(i)})^2 \right]$$

$$= \frac{\partial}{\partial \theta_0} \left[\frac{1}{2m} \sum_{i=1}^{m} \left[(\theta_0 + \theta_1 x)^{(i)} - y^{(i)} \right]^2 \right]$$

$$\Rightarrow \frac{\partial}{\partial \theta_0} \frac{1}{2m} \left[(1 + x)^{(i)} - y^{(i)} \right]^2$$

$=$

$$j=1. \quad = \frac{\partial}{\partial \theta_1} \left[\frac{1}{m} \sum_{i=1}^{m} (\theta_0 + \theta_1)(x)^{(i)} - y^{(i)})^2 \right]$$

$$= \frac{2}{m} \sum_{i=1}^{m} \left[(\theta_0 + \theta_1 x)^{(i)} - y^{(i)} \right] [x].$$

cost functions

1. MSE = mean sq. error
2. RMSE = root " "
3. MAE = mean ~~of~~ absolute error

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

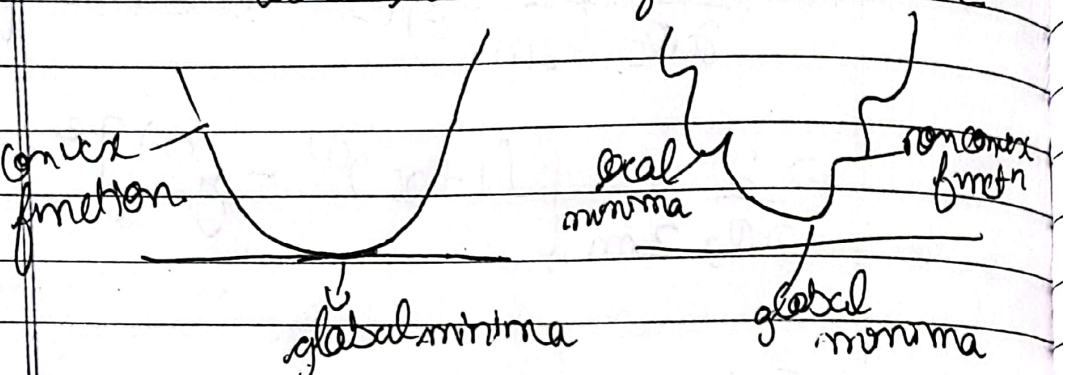
Date...../...../.....

Page.....

① Advantages

(i) This eqn is differentiable

(ii) " " also has one global minima



Disadvantage

(i) This is not robust to outliers

(ii) Penalizing the errors

② MAE [mean absolute error]

$$= \frac{1}{n} \sum_{i=1}^{n} |y - \hat{y}|$$

Date...../...../.....

Page.....

Advantage

(i) Robust to outlier

(ii) It will also be in the same unit.

Disadvantage

1 convergence takes more time
2 optimization is a complex task.

H.W	Huber loss RMSE

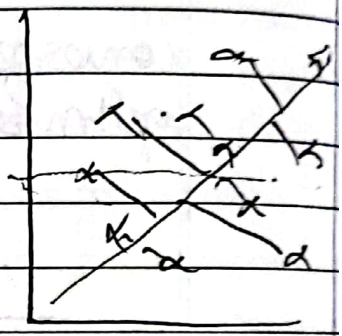
Date...../...../.....

Page.....

Performance Matrix

1. R Squared
2. Adjusted R Squared

Ans R Squared.



$$R\ sq = 1 - \frac{SSRes}{SS\ Total}$$

SSRes - Sum of square residual

SSTot - Sum of square average.

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

$\bar{y}$ → avg of y

R squared = $1 - \{\frac{\text{small numbes}}{\text{Bigger numbes.}}\}$ → small na

Date...../...../.....

Page.....

R squared is used to measure the performance of the model that you have actually created.

IMP ADJUSTED R SQUARED.

(overfitting & underfitting)

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

N = No of data points

P = No of independent features

Date...../...../.....

Page.....

most Important

Overfitting and Underfitting

(Bias & Variance)

Dataset → 1000 datapoints

→ Training (700)

→ Testing (300)

↓

new dataset

Exp | Yes

— | —

< | —

< | .

— | .

Training

500 → Train

200 → Validation

↓ Training the model

↓ Hyperparameter tuning the model

MODEL

TRAIN DATA	very good Accuracy (90%) [Bias]
TEST DATA	" " " (85%) [Var.]

↳ generalised model

Orchies

Date...../...../.....

Page.....

Train — Very good accuracy [90%] (low Bias)

Test — Bad Accuracy [50%] [High var.]

↓

overfitting

↓

hyperparameter tuning.

model low. (high Bias)

accuracy low / high (low or high Bias)

↓

underfitting.

Date...../...../.....

Page.....

Imp

RIDGE REGRESSION
LASSO "
ELASTIC NET. "

① Ridge regression (L2 regularization)

↓ aim

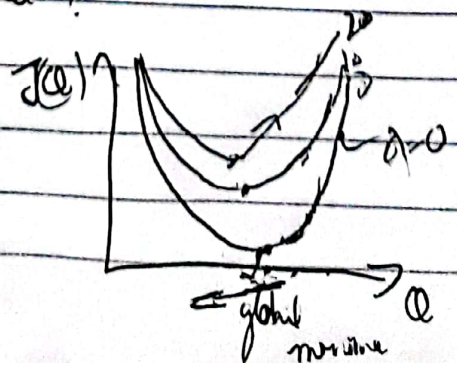
reduce overfitting

we will use same cost fnct^{n}

$$= \frac{1}{n}\sum_{i=1}^{m} [h_\theta(x^{(i)}) - y^{(i)}]^2 + \lambda |\text{slope}|^2$$

λ = hyperparameter.

→ Imp what is the relationship b/w slope & lambda ?



global minima will change

Archies

Date...../...../.....

Page.....

λ is inversely proportional to slope

when using ridge regression it will never overfit any line.

In ridge regression θ will never become zero. because when θ_0 is 0 you will see one feature gets deleted and this will effect our model.

LASSO REGRESSION {L1 Norm {L1 regularization}}

(1) aim

to reduce the feature

(2) feature selection

$$\text{Cost fnctn} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta x^{(i)} - y^{(i)})^2 + \lambda \sum_{i=1}^{m} |\theta_j|$$

Date...../...../.....

combination of ridge & lasso.

ELASTIC NET [L1 & L2 norm]

$$\text{cost fn} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta x^{(i)} - y^{(i)}) + \lambda_1 \sum^{m}(\text{slope}^2) + \lambda_2 \sum_{i=1}^{m} |\text{slope}|$$

Orchies

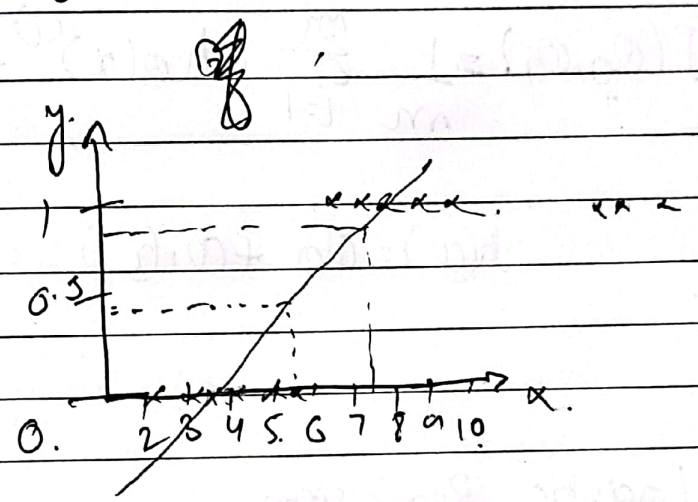
Date...../...../.....

Page.....

LOGISTIC REGRESSION

⇒ It is used to solve classification problem.

with the help of regression we try to make best fit line



If $y \leq 0.5 = 0$

$y \geq 0.5 = 1$

It means 0.5 is my threshold

best fit line $h_\theta(x) = \theta_0 + \theta_1 x$

↓

sigmoid activation

↓

output = 0 to 1

If there is a outlier then the best fit line gets changed and we do not want that. So we use sigmoid activation to avoid this.

Date...../...../.....

Page.....

② Sigmoid fn = $\frac{1}{1+e^{-x}}$ ⇒ 0 to 1

Linear regression cost function.

$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x)^{(i)} - y^{(i)})^2$.

→ mse

((gradient descent ((convergence

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

Logistic Regression

① create a best fit line

② Squashing: → Sigmoid fnctn.

$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x) - y^{(i)})^2$.

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

↓

sigmoid activation.

let $z = \theta_0 + \theta_1$

Date...../...../.....

Page.....

$= \sigma(z)$

$= \frac{1}{1+e^{-z}}$.

$h_\theta(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x)}}$

✱ log loss cost function.

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

y = truth value

$$\text{cost}(h_\theta(x)^{(i)}, y^{(i)}) = -y \log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$$

⇓

This will create a convex function

⇓

never local minima

Archies

Date...../...../.....

Page.....

minimize cost fun^n $J(\theta_0, \theta_1)$ by changing θ_0, θ_1 convergence alg.

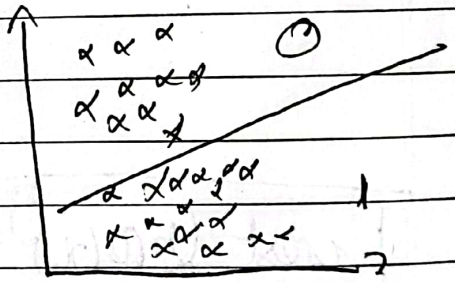
Repeat convergence

{

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad j=0 \text{ and } 1$$

}

Performance metrics.



1. confusion matrix
2. Accuracy
3. precision
4. Recall
5. f-Beta score

Actual	predicted	
f_1	f_2	o/p
-	-	0
-	-	1
-	-	0
-	-	1
-	-	1
-	-	0
-	-	1

Date...../...../.....

Page.....

Confusion matrix

predicted \ Actual value	1	0
1	3	2
0	1	1

	1	0
1	TP	FP
0	FN	TN

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

$$= \frac{3+1}{3+2+1+1} = \frac{4}{7}$$

$$\simeq 57\%$$

* DATASET → BINARY CLASSIFICATION
  - 1000 datapoints.
    - → 900 → 1
    - → 100 → 0
    - } Imbalanced dataset

* Precision : $\frac{TP}{TP+FP}$

out of all the actual values how many are correctly predicted

	Act. 1	Act. 0
pred 1	TP	FP
pred 0	FN	TN

* RECALL $\frac{TP}{TP+FN}$ ⇒ out of all the predicted values how many are correctly predicted

Date...../...../.....

Page.....16 oct.....

* **F-Beta Score :** $\frac{(1+\beta)^2 \cdot \text{Precision} \times \text{Recall}}{(\beta^2 \text{ precision} + \text{recall})}$

① If FP and FN are both important.

$\beta = 1$

$$\text{F1 score} = 2 \frac{P \times R}{P + R}$$

② If FP is more important than FN

$\beta = 0.5$

$$\text{F.5 score} = \frac{1 + 0.25}{(0.25} \frac{P \times R}{P + R)}$$

③ If FN >> FP

$$\text{F2 score} = \frac{(1+4)(P \times R)}{4(P+R)}$$

Date...../...../.....

Page.....

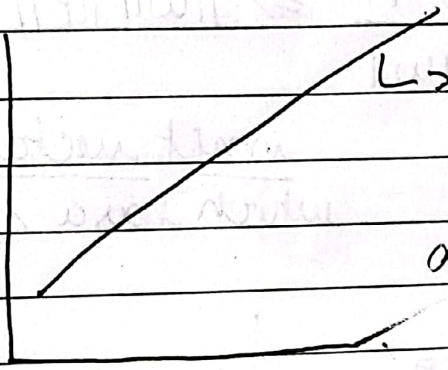
SUPPORT VECTOR MACHINE (SVM)

(support vector classifier)

SVC ① solve both classification and

SVR ② regression problem

(support vector regr)



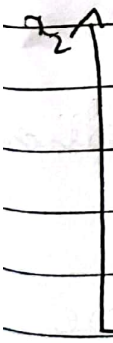
↳ $y = mx + c$

$y = \beta_0 + \beta_1 x$

$ax + by + c = 0$

$y = -\frac{a}{b}x - \frac{c}{b}$

coef intercept



$a x_1 + b x_2 + c = 0$

$w_1 x_1 + w_2 x_2 + b = 0$

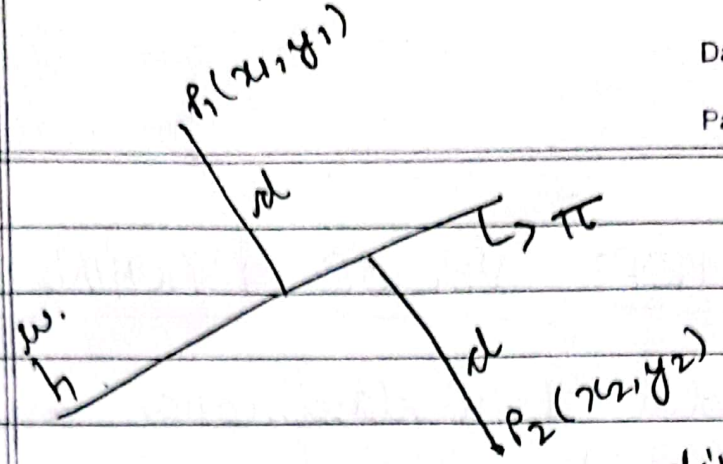
$w^T x + b = 0$

$w^T x = 0$

equation of line passing through origin

Date...../...../.....

Page.....



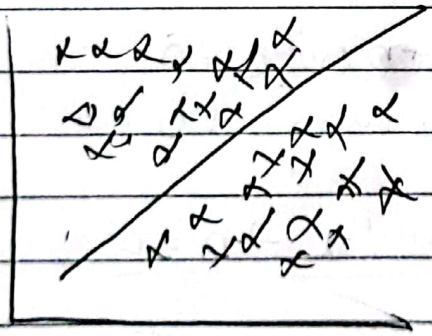
distance of a point to the plane

$$d = \frac{w^T P_1}{\|w\|} \Rightarrow \|w\| \|P\| \cos\alpha$$

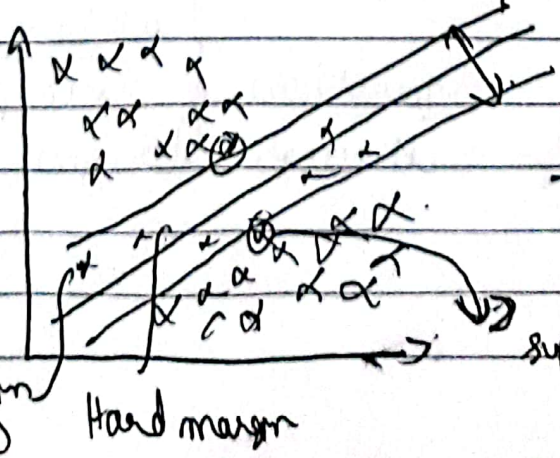
unit vectors => vectors which has a magnitude of 1

~~$\frac{w^T P}{\|w\|} \Rightarrow$ distance~~

Geometric intuition behind SVM



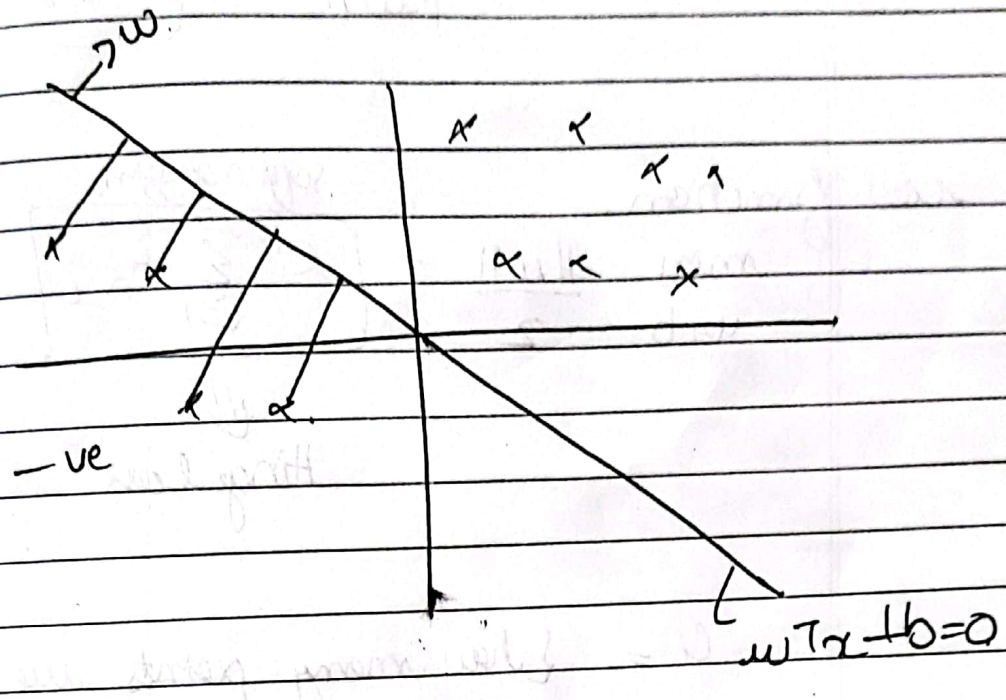
logistic regression



Date...../...../.....

Page.....

② SVM Mathematical Intuition



Cost function.

maximise $\frac{2}{\|w\|}$ $\Rightarrow$ distance b/w marginal planes
w, b

constraints $y_i \begin{cases} 1 & w^T x + b \geq 1 \\ -1 & w^T x + b \leq -1 \end{cases}$

for all correct point

constraints $\rightarrow$ $y_i * (w^T x + b) \geq 1$

Date...../...../.....

Page.....

maximize $\frac{2}{\|w\|}$ $\Rightarrow$ minimize $\frac{\|w\|}{2}$ w, b

cost function.

soft margin.

$\min_{w,b} \frac{\|w\|}{2} + \left[C \sum_{i=1}^{m} \xi_i \right]$ ↗.

Hinge loss

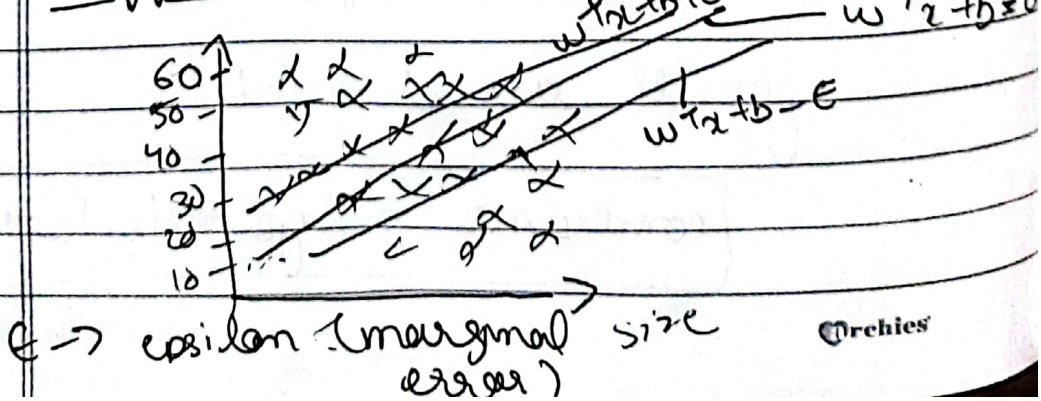
C_i = {how many points we can ignore for misclassification?

data

→ submission of the distance of the incorrect datapoints from marginal points

h

Support vector Regression



Archies

Date...../...../.....

Page.....

constraint

$$|y_i - w_i x_i| \leq \epsilon + \xi_i$$

→ Will SVM impacted by outlies.?
Yes

→ SVM kernel