M.L

Linear Regression

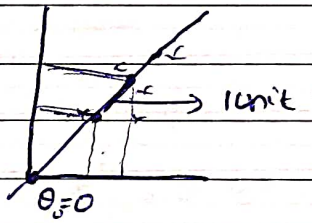
Equation of straight line

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

θ_1 = slope or coefficient

θ_0 = intercept

→ when θ_0 (intercept) is 0 that means the line is intersecting the y-axis at 0.



→ θ_1 = slope, i.e one unit moment towards the x-axis what is the unit movement in y-axis

$$\text{Cost\_function} = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

→ Square Error function

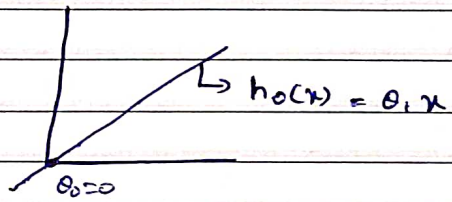
→ $\frac{1}{m}$ gives average of all the values.

What we need to solve

minimize cost function by adjusting θ_0, θ_1

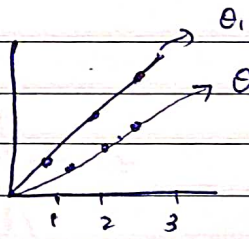
(*) $h_\theta(x) = \theta_0 + \theta_1 x$ if $\boxed{\theta_0 = 0}$

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



Ex:- our data points are

(1) Series$_1$ = (1,1) (2,2) (3,3) $\boxed{\theta_0 = 0}$ $\boxed{\theta_1 = 1}$



(1) $J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{3} (h_\theta(x) - y)^2$

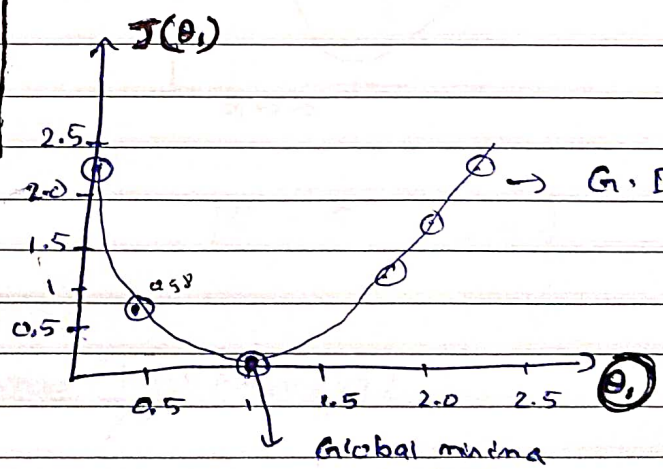
$= \frac{1}{2m} [(1-1)^2 + (2-2)^2 + (3-3)^2]$

$= 0$

(2) $\theta_0 = 0$, $\theta_1 = 0.5$

$J(\theta_1) = \frac{1}{2m} [(0.5-1)^2 + (1-2)^2 + (1.5-3)^2]$

$= 0.58$



Convergence

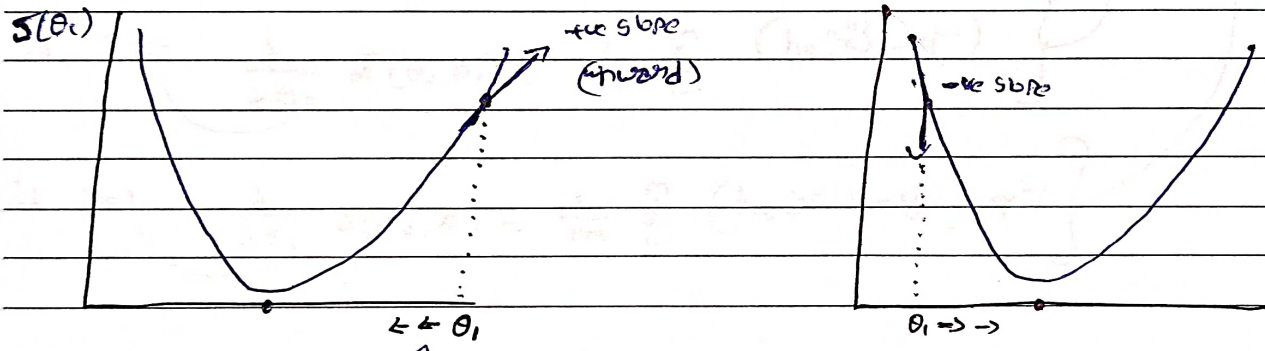
Repeat until convergence

{

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$ → derivative of slope

}

Explanation for derivative of slope.



$\theta_i = \theta_i - \alpha(+ve)$

$\theta_i = \theta_i - \alpha(-ve)$

α → (speed) model must converge (hyper parameter)

Trees are the earth's endless effort to speak to the listening heaven. - Rabindranath Tagore

Notes: | Appointment: | Phones:

Gradient descent Algorithm

Repeat until convergence

{

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

}

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

if $J=0$ $\quad \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)$

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

Performance metrics

R^2 and Adjusted R^2

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

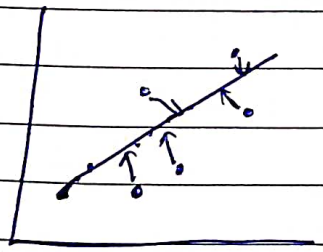
SS_{res} = Sum of residuals

SS_{tot} = Sum of total.

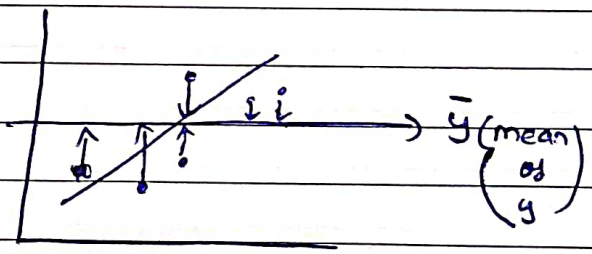
$$SS_{res} = \sum (y_i - \hat{y}_i)^2$$

$$SS_{tot} = \sum (y_i - \bar{y})^2$$

Sum of residuals



Sum of total



$\therefore$ $1 - \frac{\text{low}}{\text{high}}$ → small num b.n → → high

→ Drawback with R^2 is

if we keep on adding new features the R^2 value will keep on increasing irrespective of correlation b/w that new independent & dependent feature.

→ To solve that **Adjusted R^2**

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

P = no. of features
N = no. of data points

Ridge & Lasso regressions

Overfitting

→ Low Bias
→ High Variance

Ridge → L2 regularization.

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

$$(\hat{y}_i - y^{(i)}) + \lambda (\text{slope})^2 \rightarrow \text{Penality term}$$

→ Reduce overfitting.

λ → how fast we want to change the steep

→ making Low Bias & Low variance

I'm torn between the desire to create and the desire to destroy.- Charles M. Schulz

Notes: Appointment: Phones:

Lasso Regression

$$(\hat{y} - y)^2 + \lambda |\text{slope}|$$

$$h_\theta(x) = \hat{y} = \theta_0 + \theta_1 x + \theta_2 x_2 + \theta_3 x_3 + \dots \theta_n x_n$$

$$\text{Slope} = |\theta_0 + \theta_1 + \theta_2 + \dots \theta_n| \rightarrow \text{feature selection}$$

Assumptions of linear regression.

⑦ Error terms are normally distributed

① Normal / Gaussian Distribution → model will get trained well

② Standardization {scaling data} → Z-Score $\mu = 0, \sigma = 1$

③ Model based on Linearity

④ Multicolinearity, Variation Inflation factor can solve multicolinearity.

⑤ X residual must be uncorrelated → Exogenity

⑥ Error term must show case constant → Homoscedasity
Variance

I'm told I'm very charming when people do what I want.- Steven Brust,

Notes : Appointment : Phones:

Logistic Regression (Classification)

→ Why not use linear regression ?

① Linear regression line is can be affected by outliers. This leads to wrong Predictions.

② Sometimes our Prediction can get greater than 1 and less than 0 if we use linear regression.

→ For 2nd reason we have to make a limit to the line without crossing 0.

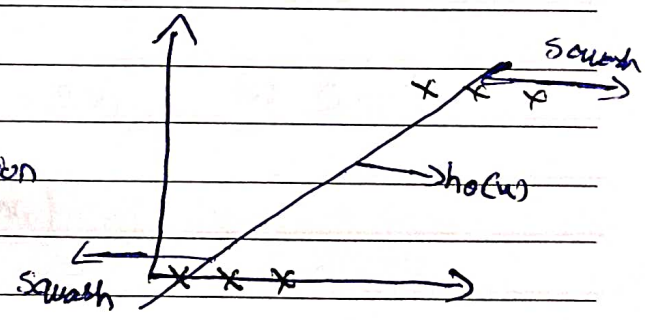
For that we are using Sigmoid

Decision Boundary Logistic Regression.

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

$$\boxed{h_\theta(x) = \theta^T x}$$

→ Apply some function on linear regression to squash this line.



Notes :

Appointment :

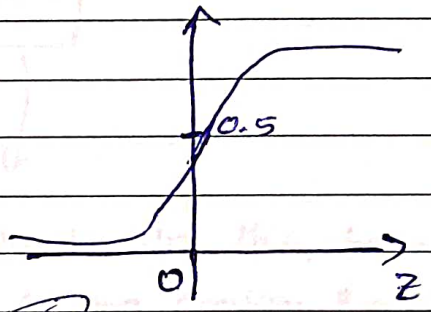
Phones :

$h_\theta(x) = \theta_0 + \theta_1 x_1$

$h_\theta(x) = g(\theta_0 + \theta_1 x_1)$ let $\boxed{z = \theta_0 + \theta_1 x_1}$

→ $\therefore h_\theta(x) = g(z)$ g = Sigmoid or logistic

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



→ Based on the graph

$\left\{ g(z) \geq 0.5 \text{ when } z \geq 0 \right\}$ $\left\{ g(z) \leq 0.5 \text{ when } z < 0 \right\}$

Training Set

$\{(x^1, y^1), (x^2, y^2), (x^3, y^3), \ldots (x^n, y^n)\}$

$y \in \{0, 1\}$ → 2 o/p

$z = \theta_0 + \theta_1 x_1$ if $\theta_0 = 0$

$\boxed{z = \theta_1 x_1}$

$$\boxed{h_\theta(z) = \frac{1}{1+e^{-z}}}$$

MARCH

APRIL

Let your life lightly dance on the edges of time like dew on the tip of a leaf. - Rabindranath Tagore

Notes: Appointment: Phones:

Change Parameter θ_1 to get best fit curve.

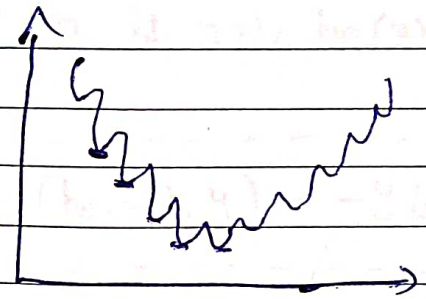
Cost function $J(\theta_1) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}_{\text{Loss}}$

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

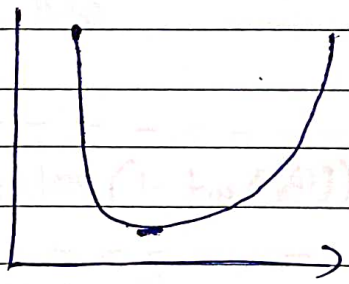
$= \frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

We cannot use this, because it is a non convex function.

Non Convex function

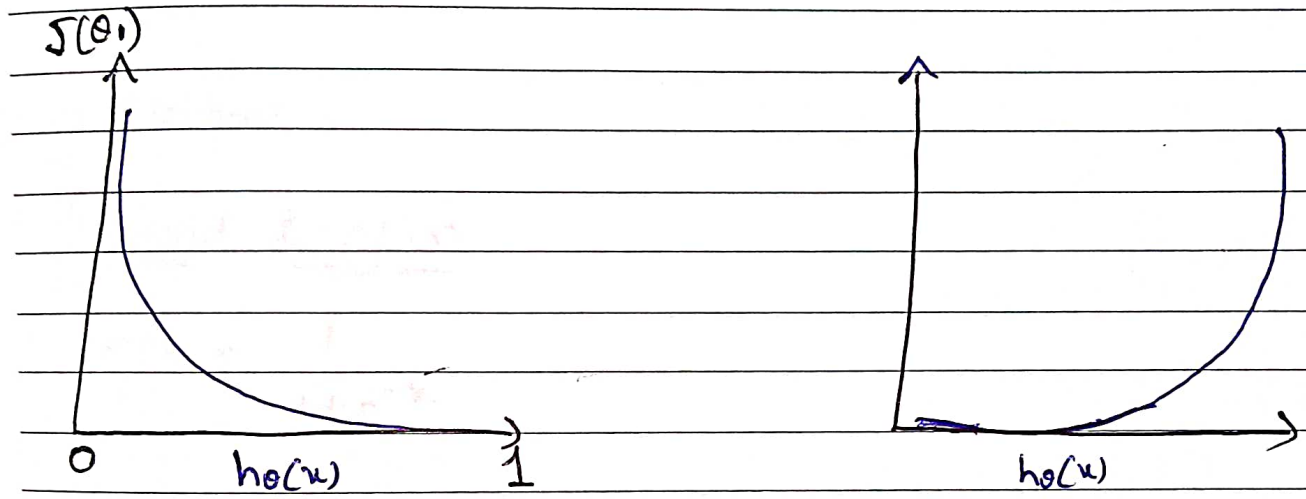


Convex function



Logistic regression cost function.

$$J(\theta_1) = \begin{cases} -\log(h_\theta(x^{(i)})) & y=1 \\ -\log(1-h_\theta(x^{(i)})) & y=0 \end{cases}$$



Cost = 0 if $y=1$, $h_\theta(x)=1$ if $y=0$

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

if $y=0$, ↑ 0 if $y=1$ ↑ 0

$$\frac{1}{2m}\sum_{i=1}^{m}$$

Just like LR

Repeat until convergence

{

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} (J(\theta_1))$$

}

→ log likelihood.

⇒ Sigmoid function

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

~~$g_2 \theta_1 x_1 + \theta_2 x_2 * \theta$~~

$$P(x) = \frac{1}{1+e^{-(\theta_1 x + \theta_2 x_2 + \theta_0)}}$$

or simply

$$P(x) = \frac{1}{1+e^{-(\beta_1 x + \beta_0)}} = \frac{e^{(\beta_1 x + \beta_0)}}{1+e^{\beta_1 x + \beta_0}} \quad - (1)$$

Notes: Appointment: Phones:

Sigmoid = Probability

logit = +ve or -ve.

$$1 - P(x) = 1 - \frac{e^{(\beta_1 x + \beta_0)}}{1 + e^{\beta_1 x + \beta_0}}$$

$$1 - P(x) = \frac{1}{1 + e^{(\beta_1 x + \beta_0)}} \quad \text{--- (2)}$$

$$\beta_1 x + \beta_0 = z$$

$$(1)/(2) = \frac{P(x)}{1 - P(x)} = \frac{\frac{e^z}{1 + e^z}}{\frac{1}{1 + e^z}} = e^z$$

$$\frac{P(x)}{1 - P(x)} = e^z$$

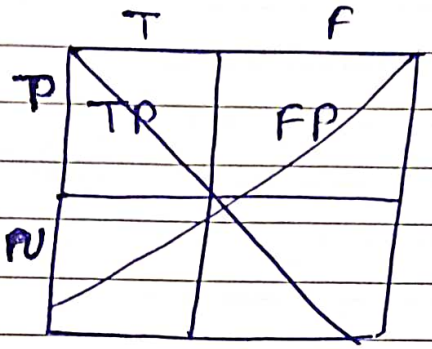
$$\log\left(\frac{P(x)}{1 - P(x)}\right) = z \quad \therefore \quad \boxed{\log\left(\frac{P(x)}{1 - P(x)}\right) = \beta_1 x + \beta_0}$$

$z = -ve$ if $P(x) < 0.5$

$z = +ve$ if $P(x) > 0.5$

Performance metrics {classification}

Predicted \ Actual		
	TP	FP
	FN	TN



TP = Positives values Predicted +ve (correct)

TN = -ve values Predicted -ve (correct)

FP = -ve Predicted as +ve

FN = +ve Predicted as -ve

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

0 → 900 } Imbalanced dataset
1 → 100 }

0 → 600 } Balanced dataset
1 → 400 }

→ So consider Imbalanced dataset 0 : 900
1 : 100
model gives → 0
& that gives 90% accuracy.

→ So we use Precision, Recall, F-score

$$\text{Precision} = \frac{TP}{TP + FP}$$ ex:- Spam classification

$$\text{Recall} = \frac{TP}{TP + FN}$$ ex:- Cancer or not

$$F\text{-}\beta = \frac{(1+\beta^2)(\text{Precision} \times \text{Recall})}{\beta^2 \times \text{Precision} + \text{Recall}}$$ ex:- Stock market crash or not (Lumpsum, People)

I'm not stubborn. My way is just better.- Maya Banks.

Notes : Appointment : Phones :

MARCH
APRIL

Naive Baye's Intution

- Baye's Theorm

x x x
o o

$P(x) = \frac{3}{5}$

$P(o) = \frac{2}{4} = \frac{1}{2}$

x = Red

o = Green

$P(R \text{ and } G) = P(R) \times P(G|R)$ ⟶ Conditional Probability

$P(G \text{ and } R) = P(G) \times P(R|G)$

⇒ $P(A \text{ and } B) = P(B \text{ and } A)$

$P(A) \times P(B|A) = P(B) \times P(A|B)$

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)}$$ ⟶ Baye's theorm

A brother may not be a friend, but a friend will always be a brother.- Benjamin Franklin

Notes: | Appointment: | Phones:

Input

$[x_1 \; x_2 \; \dots \; x_n]$

output

y

$$P(y|x_1, x_2, \dots x_n) = \frac{P(y) \times P(x_1|y) \times P(x_2|y), P(x_3|y) \dots P(x_n|y)}{P(x_1) \times P(x_2) \times P(x_3) \dots P(x_n)}$$

Ex. x_1 x_2 x_3 x_4 y

[ ] Yes

[ ] NO

$$P(y=yes|x_i) = \frac{P(yes) \times P(x_1|yes) \times P(x_2|yes) \times P(x_3|yes) P(x_4|yes)}{P(x_1) \times P(x_2) \times P(x_3) \times P(x_4)} \rightarrow \text{Consted}$$

$$P(y=no|x_i) = \frac{P(no) \times P(x_1|no) \times P(x_2|no) \dots P(x_4|no)}{P(x_1) \times P(x_4)} \rightarrow \text{Consted}$$

is $P(yes|x_i) = 0.13$ $P(no|x_i) = 0.05$

Normalizing

$$P(yes|x_i) = \frac{0.13}{0.13+0.05} = 0.72 \qquad P(no|x_i) = \frac{0.05}{0.05+0.13} = 0.28$$

24 February 2022
Thursday

FEBRUARY'22

SU	MO	TU	WE	TH	FR
		1	2	3	4
6	7	8	9	10	11
13	14	15	16	17	18
20	21	22	23	24	25
27	28				

Week 09 • 055-310

Day	outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	NO
D2	Sunny	Hot	High	Strong	NO
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	yes
D6	Rain	Cool	Normal	Strong	NO
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	NO
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	mild	Normal	Weak	Yes
D11	Sunny	mild	Normal	Strong	Yes
D12	Overcast	mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	mild	High	Strong	NO

There is only one happiness in this life, to love and be loved. - George Sand

Phones:

Appointment:

Notes:

2022 February 25
Friday

056-309 • Week 09

Consider only outlook & temperature.

Outlook (column)

	Yes	No	P(y)	P(N)
Sunny	2	3	P(y|s) = 2/9	P(N|s) = 3/5
overcast	4	0	P(y|o) = 4/9	P(N|o) = 0/5
Rain	3	2	P(y|R) = 3/9	P(N|R) = 2/5
total =	9	5		

Temperature

	Yes	No	P(y)	P(N)
Hot	2	2	2/9	2/5
mild	4	2	4/9	2/5
Cold	3	1	3/9	1/5
total	9	5		

Play

Yes	9	P(yes) = 9/14
No	5	P(no) = 5/14
	14	

The last capitalist we hang shall be the one who sold us the rope. - Karl Marx

Notes: Appointment: Phones:

MARCH
APRIL

Test data (Sunny, Hot) → O/P ??

$$P(\text{yes} \mid \text{Sunny, Hot}) = \frac{P(\text{yes}) \times P(\text{Sunny} \mid \text{yes}) \times P(\text{Hot} \mid \text{yes})}{P(\text{Sunny})\ P(\text{Hot})}$$

$$= \frac{9}{14_7} \times \frac{2}{9} \times \frac{2}{9}$$

$$= \frac{2}{63} = 0.031$$

$$P(\text{No} \mid \text{Sunny, Hot}) = \frac{P(\text{No}) \times P(\text{Sunny} \mid \text{No}) \times P(\text{Hot} \mid \text{No})}{P(\text{Sunny}) \times P(\text{Hot})}$$

$$= \frac{5}{14} \times \frac{3}{5} \times \frac{2}{5}$$

$$= \frac{3}{35} = 0.085$$

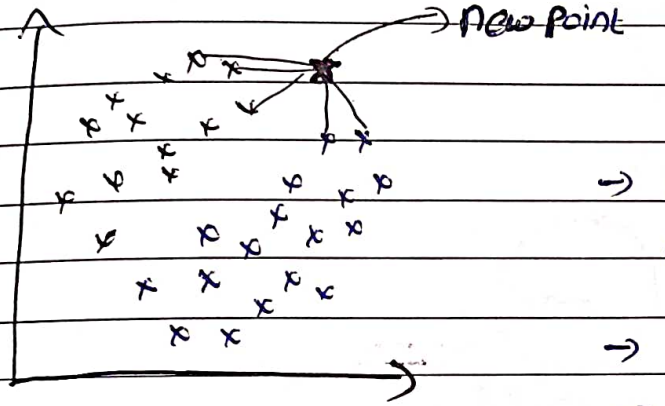
Prediction ↑

$$P(\text{No} \mid \text{Sunny, Hot}) = 0.085 = \frac{0.085}{0.031 + 0.085} = 0.73 = 73\%\ (\text{No})$$

$$P(\text{yes} \mid \text{Sunny, Hot}) = 1 - 0.73 = 0.27 = 27\%\ (\text{yes})$$

K-Nearest Neighbour

- Outliers
- Imbalanced



if $K = 5$

→ Calculate distance btw new Point and nearest Points. (Euclidian distance)

→ Vote

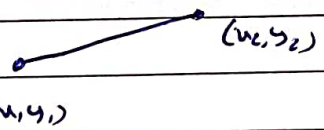
→ 3 Black
2 Blue

→ So new Point goes into Black.

→ In case regression

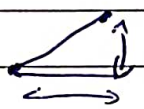
→ Get the average of nearest 5 Points.

Euclidian distance -



$$\sqrt{(x_1 - x_2)^2 + (y_2 - y_1)^2}$$

Manhaton distance



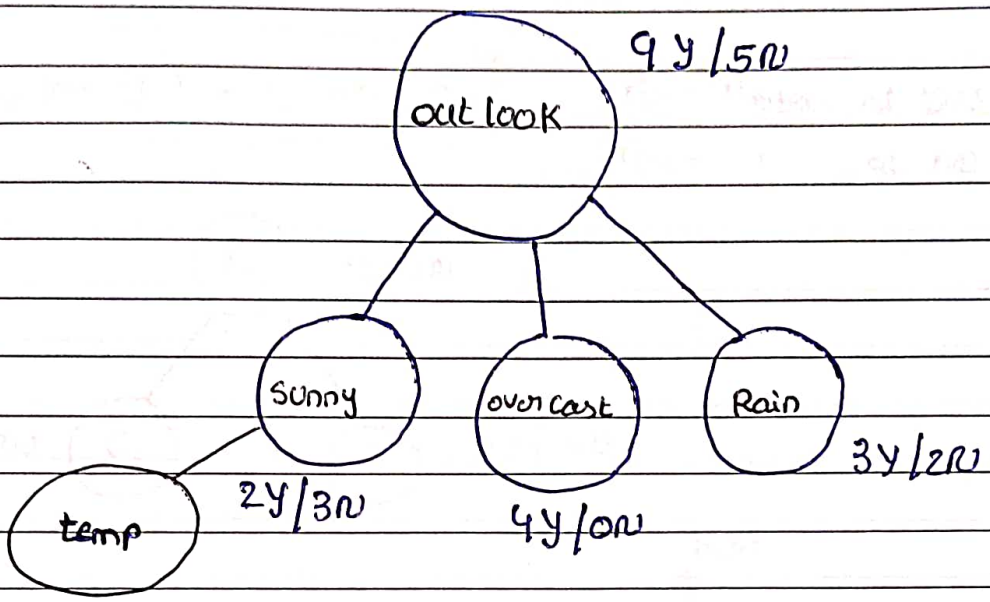
$$|(x_2 - x_1) + y_2 - y_1|$$

Show me a family of readers, and I will show you the people who move the world.- Napoléon Bonaparte

Notes: | Appointment: | Phones:

MARCH
APRIL

Decision Tree



Pure split - Overcast - 4Y/0N → All ways yes (Pure) - No split.

→ How can we say pure split?

→ Entropy

→ Gini Coefficient (or) Gini Impurity

→ How the features are selected

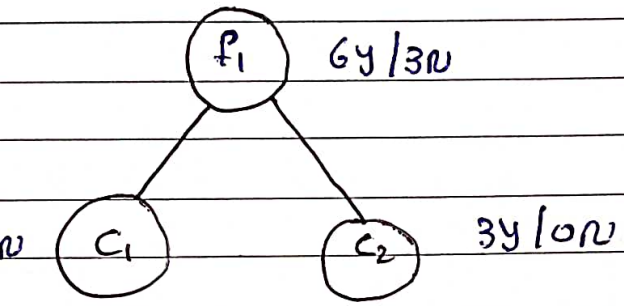
→ Information Gain

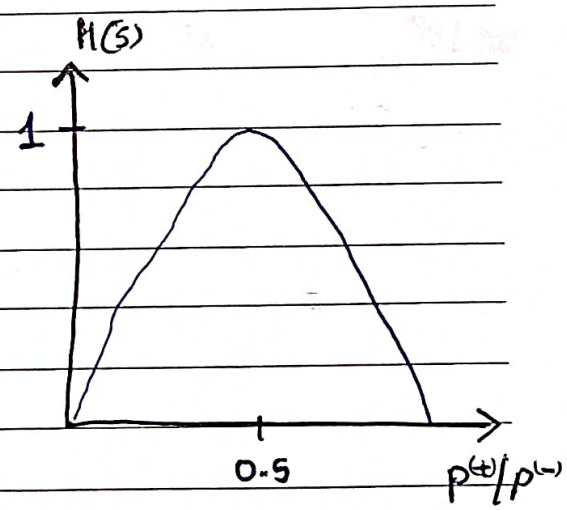
① Entropy:

$$H(S) = - P_{(+)} \log_2(P_{(+)}) - P_{(-)} \log_2(P_{(-)})$$

$P_{(+)}$ = Proba. of yes

$P_{(-)}$ = " of no





for C_2

$$H(S) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$$

$$= -1\log_2(1)$$

$$= \boxed{0} \rightarrow \text{Completely Pure Split}$$

for C_1

$$H(S) = \frac{-3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= \boxed{1} \rightarrow \text{Completely Impure Split}$$

② Which feature to take to split?

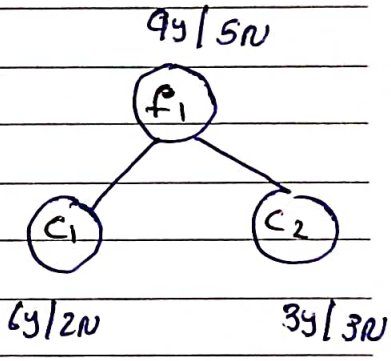
→ Which feature should I take and make the split.

Information Gain

Root node

$$Gain(S, f_i) = H(S) - \sum_{Veval} \frac{|S_v|}{|S|} H(S_v)$$

Sub nodes



$$H(S) = -P_{(+)} \log_2(P_{(+)}) - P_{(-)} \log_2(P_{(-)})$$

$$= -\frac{9}{14} \log_2(9/14) - \frac{5}{14} \log_2(5/14)$$

$$H(S) = 0.94$$

$$H(C_1) = -\frac{6}{8} \log_2\left(\frac{6}{8}\right) - \frac{2}{8} \log_2\left(\frac{2}{8}\right)$$

$H(C_1) = 0.81$ Similarly $H(C_2) = 1$

$$Gain(S, f_1) = 0.94 - \left[\frac{8}{14}(0.81) + \frac{6}{14}(1)\right]$$

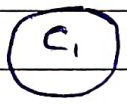
$Gain(S, f_1) = 0.049$ f_1 : Feature 1

Gini Impurity

$$G.I = 1 - \sum_{i=1}^{n} (P)^2$$

$$G.I = 1 - [(P_{(+)})^2 + (P_{-})^2]$$

lets take a node

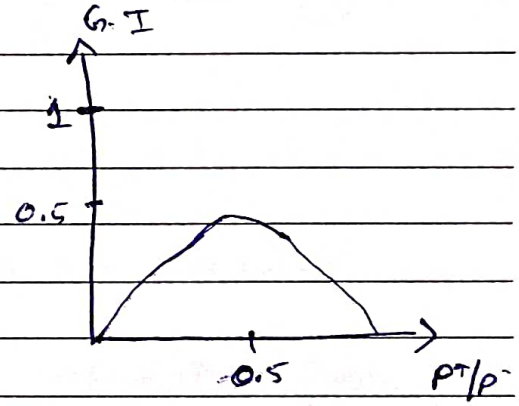


2y/2N

$$G.I = 1 - [(\tfrac{1}{2})^2 + (\tfrac{1}{2})^2]$$

$$= 1 - \frac{1}{2}$$

$$= 0.5$$

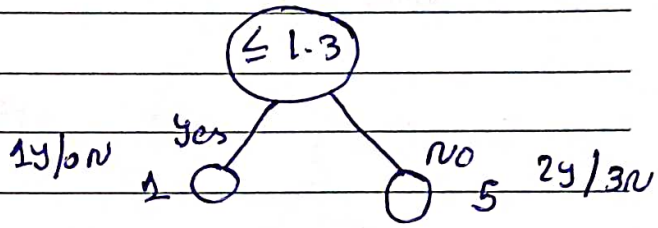


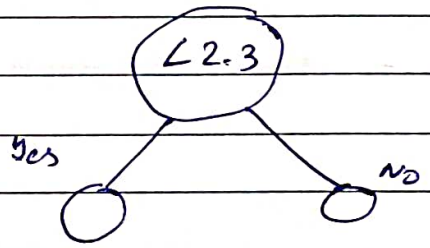
→ Entropy takes more time

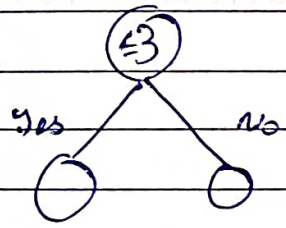
→ Gini impurity have less time

What if Continous variables

f_1	O/P	$\Rightarrow$	Sort f_1
2.3	y		1.3
1.3	y		2.3
4	N		3
5	y		4
7	N		5
3	N		7

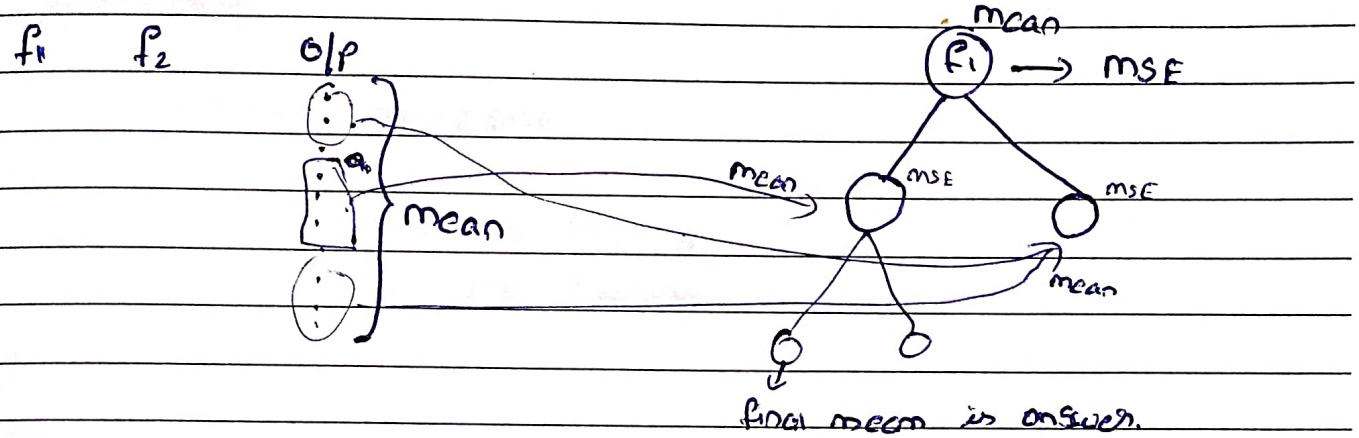






→ like that for every row it will create and calculate Information Gain. Then it will take best information Gain.

Decision tree regressor



→ Check Decision tree by normalized nerd Youtube channel

One must wait until the evening to see how splendid the day has been. - Sophocles

Notes:

Appointment:

Phones:

2022 March 15
Tuesday

⇒ What if there are very large no. of nodes?

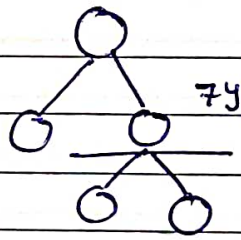
→ Overfitting

Prevent overfitting

→ Post Pruning
→ Pre Pruning

⇒ Post Pruning

So in that node we can say that ^70% accuracy (Yes) So, no need of spliting.



→ This dividing is called Post Pruning.

⇒ Pre Pruning

→ hyperparameters will decide, like 'max_depth', 'max_leaf'

Happiness is not something ready made. It comes from your own actions. - Dalai Lama XIV

Notes: Appointment: Phones:

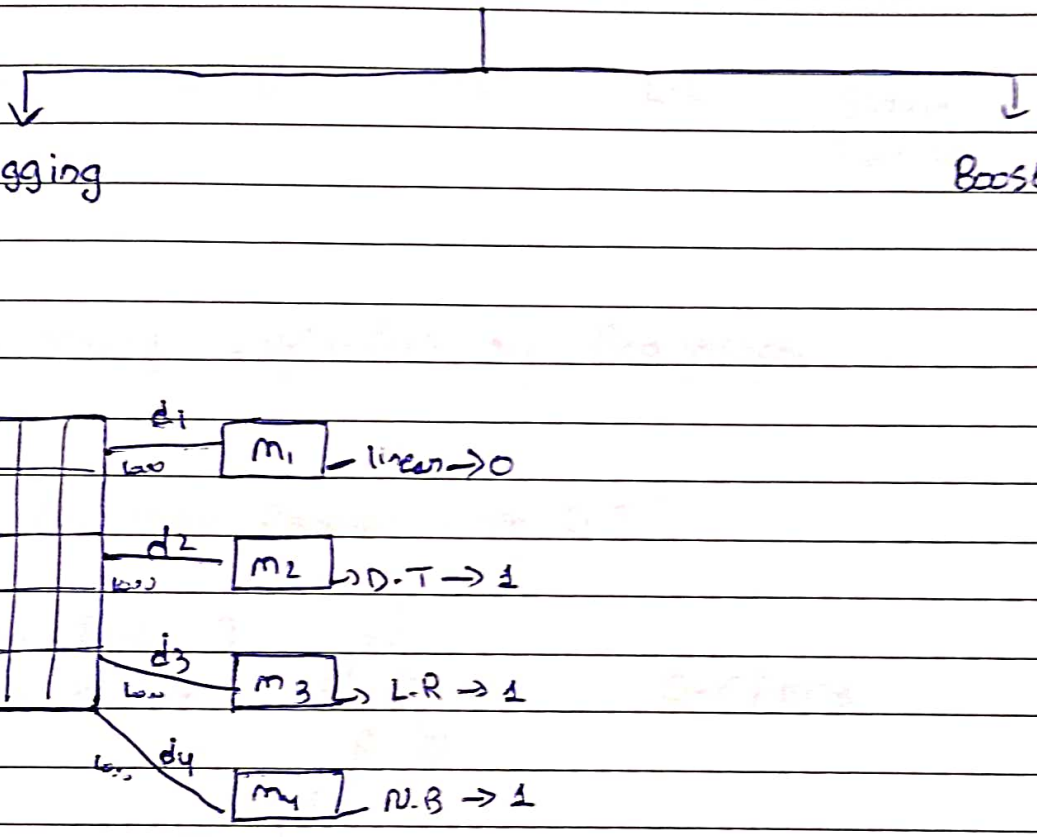
APRIL

2022 March 17
Thursday

Ensemble Techniques

Ensemble Techniques

→



Sample data = d $\boxed{d < D}$

→ Some data can repeat, d_1 in d_2, ind...

→ Majority voting so 1 } Bootstrap Aggregation

→ If regression Average

I am not a product of my circumstances. I am a product of my decisions. Stephen Covey

Notes: Appointment: Phone:

APRIL

2022 March 19
Saturday

Boosting → Sequential set of all models combined together these models are weak learners after that they become strong learner

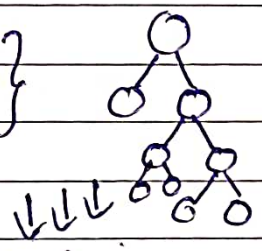
→ [m_1] → [m_2] → [m_3] → [m_4] → o/p

weak learner — W.L — W.L — W.L — Strong learning

Random forest classifier and Regressor

What is the main problem with D.T?

Random forest
- { Low Bias, High Variance } → Overfitting
- { low Bias, low variance }



→ Like Bagging but all are decision trees.

→ Row Sampling, Feature Sampling *** → Selecting some features & distributing

2022 March 21
Monday

080-285 • Week 13

Q) Is normalization required in Random forest.

A) No, because what ever the value is, the decision will split the same.

Q) Is standardization or normalization required?

A) Yes, because is the values are high distance evaluation become complex, if standardized then easy to find distance.

Q) Is random forest impacted by outliers?

A) No, outlier impact is very less on random forest.

because random forest handles outliers by essentially binning them.

Q) Why Navie Bayes is called as 'navie'?

→ Consider apple → red, round, 4 diameter

these all features are related and dependent on existence of other feature.

But Navie Bias consider every feature as independent.

→ It is navie because it make assumptions may or may not correct.

Notes: Appointment: Phones:

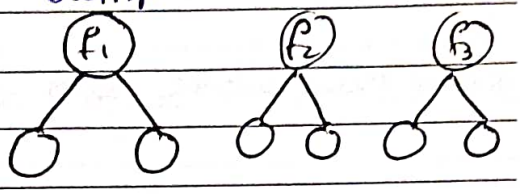
2022 March 23
Wednesday

082-283 • Week 13

Adaboost:

f_1	f_2	f_3	f_4	o/p	Weight	
-	-	-	-	Yes	1/7	
-	-	-	-	No	1/7	
-	-	-	-	Yes	1/7	
-	-	-	-	No	1/7	wrong
-	-	-	-	Yes	1/7	
-	-	-	-	Yes	1/7	
-	-	-	-	No	1/7	

Stump



① Total error = 1/7

② Performance of stump = $\frac{1}{2}\log_e\left(\frac{1-TE}{TE}\right) = \frac{1}{2}\log_e\left(\frac{1-1/7}{1/7}\right) = 0.895$

③ update all these weights

New sample weight = weight * e^{PS} = $\frac{1}{7} \times e^{-0.895} = 0.05$

Incorrect record = weight * e^{+PS} = $1/7 \times e^{0.895} = 0.349$

Weight	Normalized Wgt	Buckets
0.05	0.07	[0 – 0.07]
0.05	0.07	[0.07 – 0.14]
0.05	0.07	[0.14 – 0.21]
0.349	0.537	[0.21 – 0.747] → Big Bucket size.
0.05	0.07	[0.747 – 0.751]
0.05	0.07	
0.05	0.07	
0.649		

We generate fears while we sit. We overcome them by action. – Dr. Henry Link

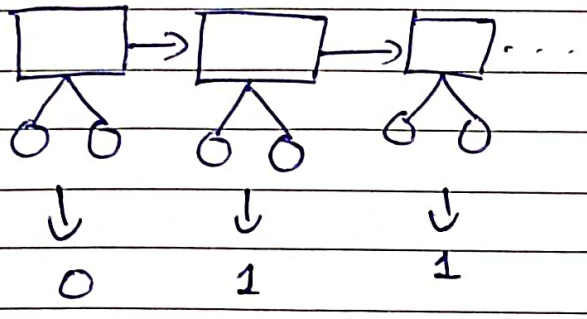
Notes: Appointment: Phones:

APRIL

2022 March 25
Friday

→ So, now random numbers btw 0 to 1 are generated then according to the random numbers the records are selected based on bucket

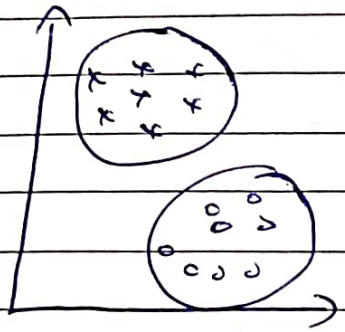
→ The wrong are got more bucket size, so that record will select multiple times.



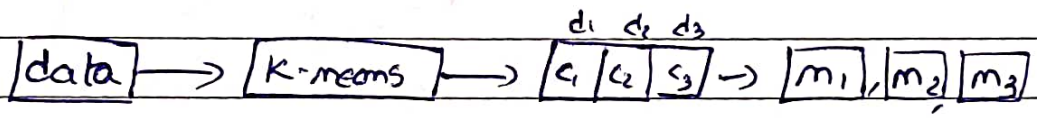
→ Voting - classification

→ Mean - regression

K-means clustering (unsupervised)



Custom ensemble technique → used in industry



→ K-means

K= Centroids

① We try different K-values → K= 1, 2, 3, 4, ... (WCSS)

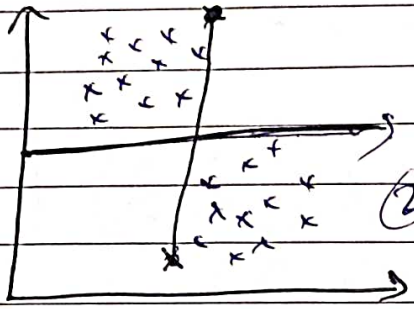
② Initialize K no. of centroids

③ Compute average to update centroid

2022 March 29
Tuesday

088-277 • Week 14

(1)

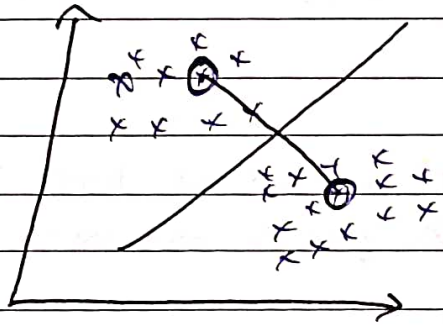


(2)

* = centroid

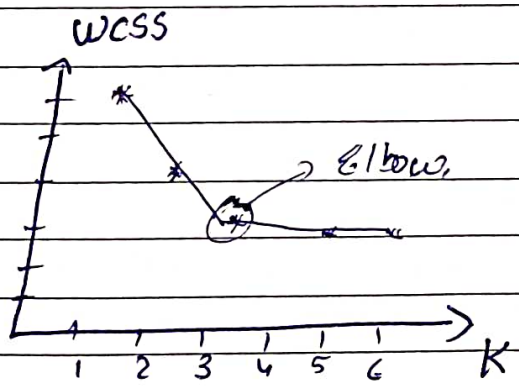
* Kmeans ++ make sure that the centroids initialized at 1st are far away from data.

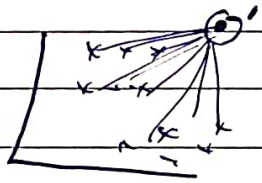
(1)

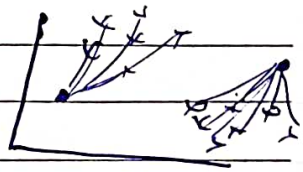


Elbow method

WCSS -> within cluster sum of square = [Centroid to every point / no. of points / Distance b/w]







To believe in something, and not to live it, is dishonest. - Mahatma Gandhi

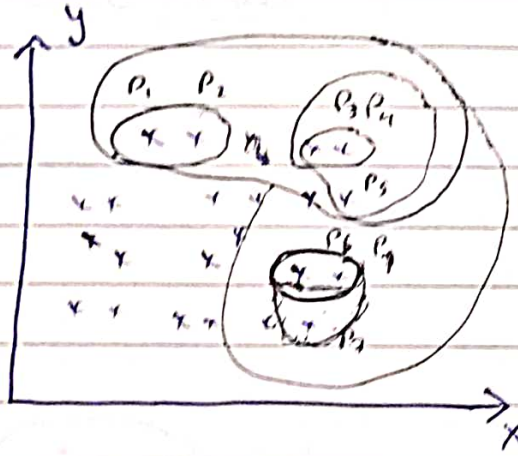
Notes: | Appointment: | Phones:

APRIL

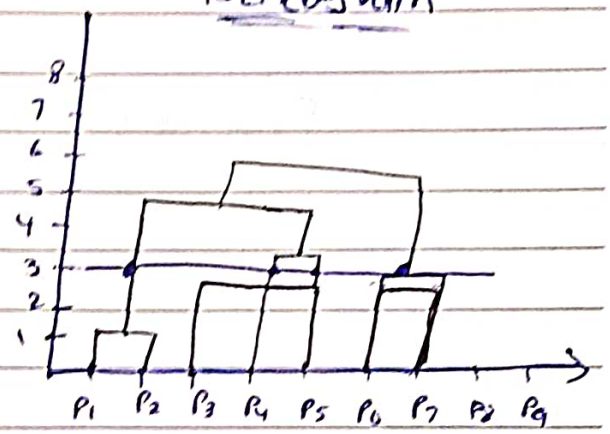
2022 March 31
Thursday

(2) Hierarchical clustering

(1)



Dendogram

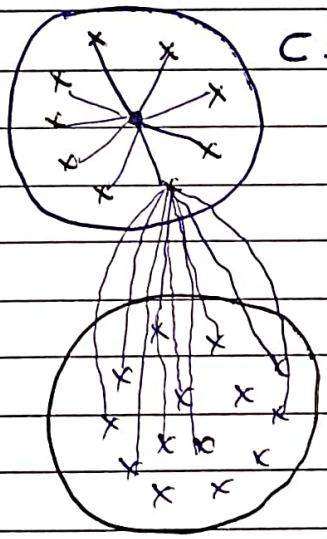


→ We need to find the longest vertical line that has no horizontal line passing through it.

→ Hierarchi: Clustering takes more time than K-means

Validate clustering

→ Silhouette → [-1 to +1]
- → K-means
- → Hierarchical



C1 o = centroid

a(i) = distance blw centroid & each & every point.

b(i) = (distance blw a single point in C_1 to all points in C_2) & their mean.

→ Good model b(i) >> a(i)

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i} d(i,j)$$

$$b(i) = \min_{C_J} \frac{1}{} \sum_{j \in C_J} d(i,j)$$

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad \text{is } |C_I| > 1$$

$S(i) = 0$ is $|C_I| = 1$

DB Scan

→ Density Based Spatial Clustering of Applications with Noise

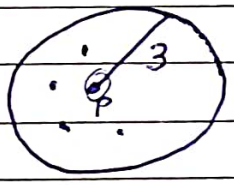
1. Epsilon
2. min Points
3. Core Points
4. Border Points
5. Noise Point

if

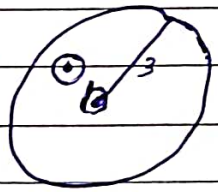
min Points > 4

Epsilon = radius = 3

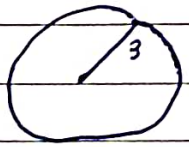
⊙ = Core Point



=> Within this cluster radius if there are minimum points mentioned or more than it is considered as Core Point



→ If the Core Point condition didn't satisfied when drawn another circle then if is another Core Point is presented then it is considered as border Point (b)



→ If Above 2 conditions are satisfied then it is Noise.

Advantages of DBScan

→ Great at seperating clusters of high density Vs clusters of low density within a given dataset.

→ Is great with handling outliers within the dataset.

Disadvantages of DBScan

→ Does not work well when dealing with clusters of varying densities. It seperates low vs high densities but finds difficult to cluster with similar density.

→ Struggles with high dimensional data.

XG Boost

Extreme Gradient Boosting

Salary	Credit	Approval	Residual
<=50	B	0	-0.5
<=50	G	1	0.5
<=50	G	1	0.5
>50	B	0	-0.5
>50	G	1	0.5
>50	N	1	-0.5

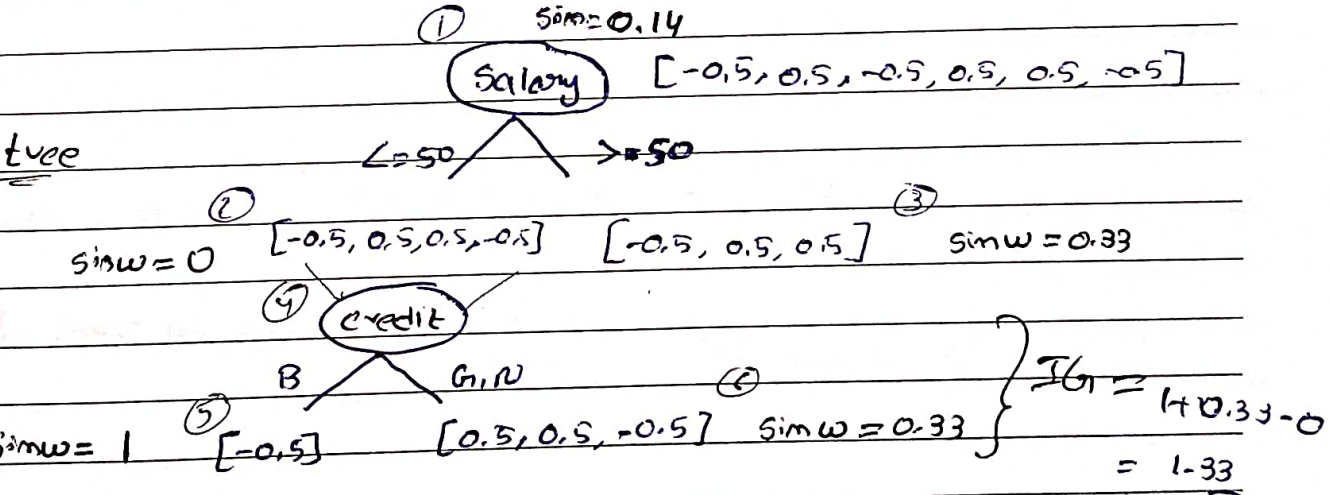
Base model → Pv = 0.5

① Create a Binary decision tree using the feature

② Calculate Similarity weight = $\frac{\Sigma (\text{Residual})^2}{\Sigma (P_v(1-P_v)) + \lambda}$

③ Information Gain

Binary tree



Never bend your head. Hold it high. Look the world straight in the eye.- Helen Keller

Notes:

Appointment:

Phones:

Sum of residual square for ② (Similarity weight)

$$= \frac{[-0.5+0.5+0.5-0.5]^2}{0.5(1-0.5)+0.5(1-0.5)+0.5(1-0.5)+0.5(1-0.5)}$$

$$= 0$$

Sum of Similarity weight for ③

$$= \frac{(-0.5+0.5+0.5)^2}{0.5(1-0.5)+0.5(1-0.5)+0.5(1-0.5)}$$

$$= \frac{0.25}{0.75} = \frac{1}{3} = 0.33$$

Similarity weight for ①

$$= \frac{0.25}{1.75} = \frac{1}{7} = 0.14$$

Information gain = $0 + 0.33 - 0.14 = 0.19$

Only for Base model. $\log\left(\frac{p}{1-p}\right) = \log\left(\frac{0.5}{0.5}\right) = 0$

One can never consent to creep when one feels an impulse to soar.- Helen Keller.

Notes: | Appointment: | Phones:

Inferencing

$[O + \alpha(1)]$ (4)

$= \sigma[O + \alpha(1)]$

$= \sigma[O + \alpha_1(DT_1) + \alpha_2(DT_2) + \alpha_3(DT_3) + \ldots \alpha_n(DT_n)]$

(2) XG Boost Regressor → 1

Exp	Gap	Salary	R_1	Errors
2	Yes	40K	-11K	
2.5	Yes	41K	-9K	
3	No	52K	1K	
4	No	60K	9K	
4.5	Yes	62K	11K	
	Avg =	51K		

Base model = 51K

[-11, -9, 1, 9, 11] 1/6

Exp

≤2 / >2

[-11] [-9, 1, 9, 11]

$\frac{121}{1+1} = 60.5$

$\frac{(-9+1+9+11)^2}{4+1} = \frac{144}{5} = 28.5$

It is not that I'm so smart. But I stay with the questions much longer.- Albert Einstein

Notes: Appointment: Phones:

Similarity weight = $\frac{\sum (\text{residuals})^2}{\text{No. of Res} + \lambda}$

Information Gain = $60.5 + 28.8 - \frac{1}{6} = 89.13$

→ Just like that create for each and every row then using Intuition gain see best split & use it.

→ Inferencing

$o/p = 51 + \alpha_1 (-10) + \alpha_2 (DT_2) + \alpha_3 (DT_3) + \dots \alpha_n (DT_n)$

Don't gain the world and lose your soul, wisdom is better than silver or gold. -Bob Marley

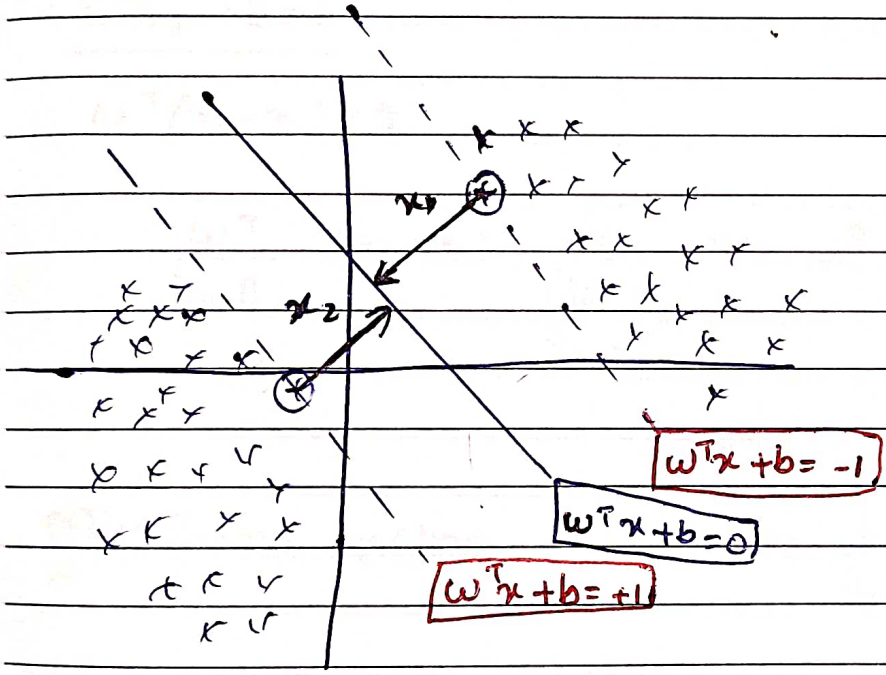
Notes:

Appointment:

Phones:

Support Vector machine

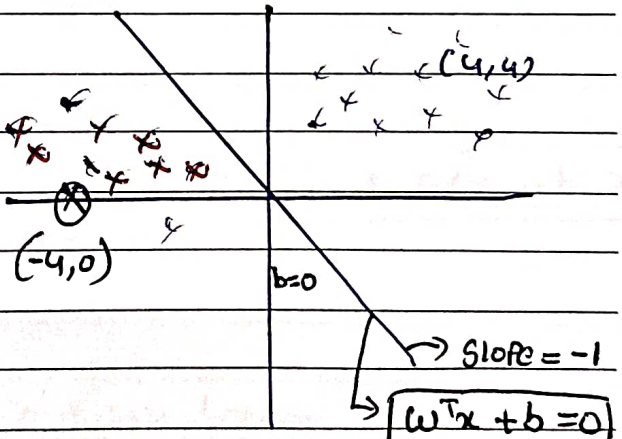
1. Support vectors
2. Hyperplanes
3. Marginal Distance
4. Linear seperable
5. Non - linear seperable.



$y = w^T x + b$ → eq of a line.

$m = -1$, $b = 0$

$y = w^T x + 0$



$\begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} -4 & 0 \end{bmatrix} = 4 \Rightarrow$ +ve value ⇒ going to be always +ve.

If at first you don't succeed, try, try again. Then quit. No use being a damn fool about it. -W.C. Fields

Notes: | Appointment: | Phones:

$y = w^T x$

$= \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 4 & 4 \end{bmatrix} = -4 \Rightarrow -ve$ ⇒ Those side Points are going to be always -ve.

$\Rightarrow w^T x_1 + b = -1$

$\Rightarrow w^T x_2 + b = 1$

$w^T (x_2 - x_1) = 2$ $\quad \|w\| = norm\ of\ w$

$\frac{w^T}{\|w\|} (x_2 - x_1) = \frac{2}{\|w\|}$ maximize?

by dividing w^T by $\|w\|$ whole magnitude will go off. by direction don't so off.

Optimization function (w, b) $\max \frac{2}{\|w\|}$

Such that

$$y_i \begin{cases} 1 & w^T x + b \geq 1 \\ -1 & w^T x + b \leq -1 \end{cases} \Rightarrow \boxed{y_i * w^T x_i + b_i \geq 1} \quad (1)$$

→ If (1) is not ≥ 1 then it is misclassification.

Do what you can with all you have, wherever you are. – Theodore Roosevelt

Notes: Appointment: Phones:

$$(w^*, b^*) = \min \frac{\|w\|}{2} + C_i \sum_{i=1}^{n} \xi_i$$

C_i → How many errors?

ξ → Value of error

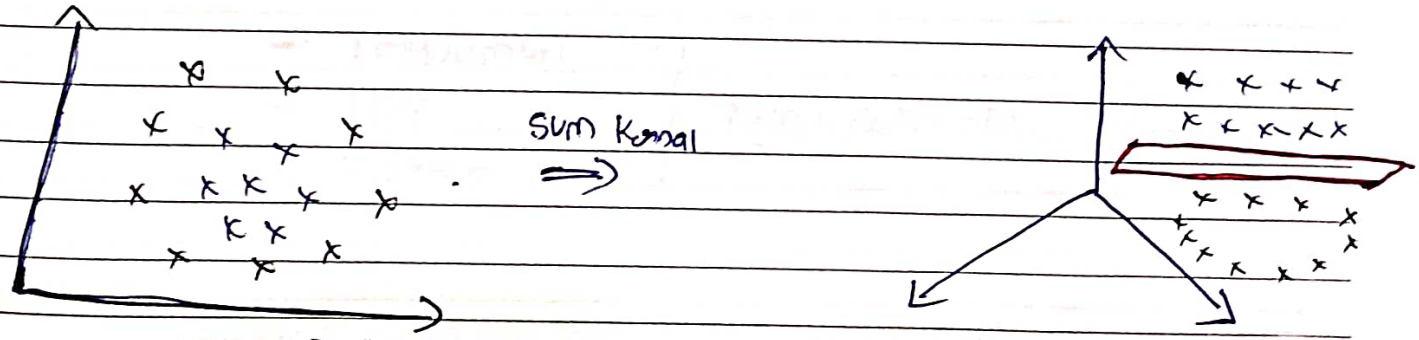
→ If we some points are present in b/w margin & line or hyperplane so those are errors.

→ C_i → How many errors we can have so that the margin doesn't change (Regularization) ↑ C_i

SVM KERNALS

① Soft margin → contains errors

② Hard margin → contains no points in b/w.



Reading is to the mind, as exercise is to the body. – Brian Tracy

Notes: Appointment: Phones:

Polynomial Kernal

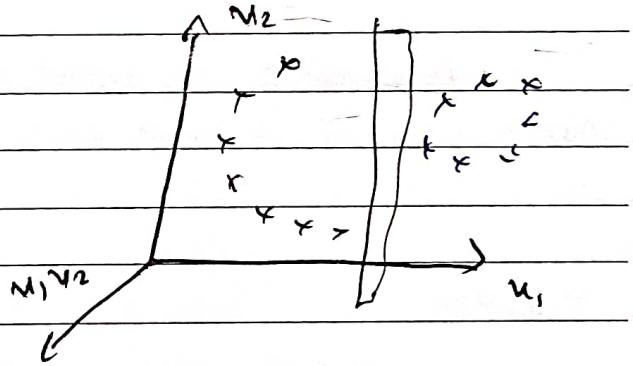
$$\delta(x_1, x_2) = (x_1^T \cdot x_2 + 1)^d$$

d: dimension

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cdot \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1 x_2 \\ x_1 x_2 & x_2^2 \end{bmatrix}$$ Original

So,

x_1	x_2	y_0	x_1^2	x_2^2	$x_1 x_2$



→ 3 types of Kernels

- → Polynomial
- → RBF
- → Sigmoid

} hyperparameter tuning

Feature Engineering

2022 April 27 Wednesday

→ Different types of encoding.

Nominal Ordinal

Nominal → One hot
→ One hot (KDD orange)
→ Mean encoding

Ordinal → Label Encoding
→ Target Guided Ordinal Encoding

Target Guided ordinal → ① mean based on categories.
⇒ ② give rank based on highest mean.

Mean encoding → ① Same as 'target guided' but instead of rank we replace mean values.

→ Why feature scaling?

Features (Male)
→ Magnitude
→ Units

Height Weight BMI

→ low computation
→ Faster convergence

→ Handling missing values in [categorical] variable

① Delete Rows
② Replace with most frequent values
③ Apply classifier algorithm and Predict
④ Apply unsupervised ML
⑤ Mean/median

} only for Categorical

new ⑥ Regression Imputation, interpolation, extrapolation
⑦ Arbitary imputation
⑧ Frequent category imputation
⑨ End of distribution imputation
⑩ Capturing NAN values with new features

}

2022 May 01
Sunday

-> How to handle categorical features many categories.

-> Count / Frequency encoding.

-> The Problem is also called as High Cardinality.

=> => Count / Frequency encoding

-> Replace the categorical variables by its Count.

Advantages

① Very simple to implement
② Does not increase the feature dimensional space

Disadvantages :-

① if some of the labels have same count
② Adds some arbitary numbers, and ∴ weights to be different labels, that may not be related to their Predictive power.

=> consider example of Zomato dataset.

Notes: | Appointment: | Phones:

→ Life cycle of a data science projects.

① Data collection:-

- From company side
- 3rd party Api's
- Surveys

② Feature Engineering ... M

- Handling the missing values.

Different types of missing Data?

→ Missing Completely at Random.

→ If the Probability of being missing is same for all observations.

→ when data is MCAR, there is no relationship between the data missing and any other values observed or missing.

→ Missing Data not at random. - Systematic missing values.

-> When the missing value have no relation with any other column but have relation with the column in which the value is missing.

Example men - hide their salary
women - hide their age.