ML ALGOM

PAGE NO.:
DATE: / /

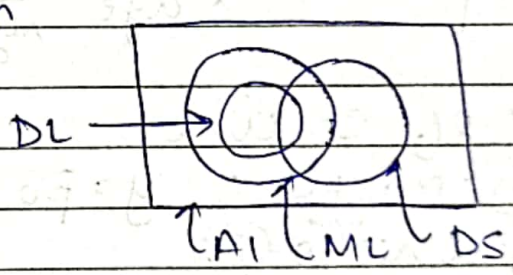
DAY-1 INTRODUCTION TO ML ALGO

1. AI VS ML VS DS VS DL

- AI app^{n} : able to its task without any human intervention
  ex: Netflix → Recommendation system
  Amazon, self driving cars.

- ML : subset ~~tool~~ of ai ; state tool to analyse, visualize, prediction, forcasting.

- DL : Mimic human brain; mimic human brain



Machine learning

- Supervised L.
  - Regression
  - Classification
    - Binary
    - Multi class
- Unsupervised L.
  - Clustering
  - Dimensionality Reduction

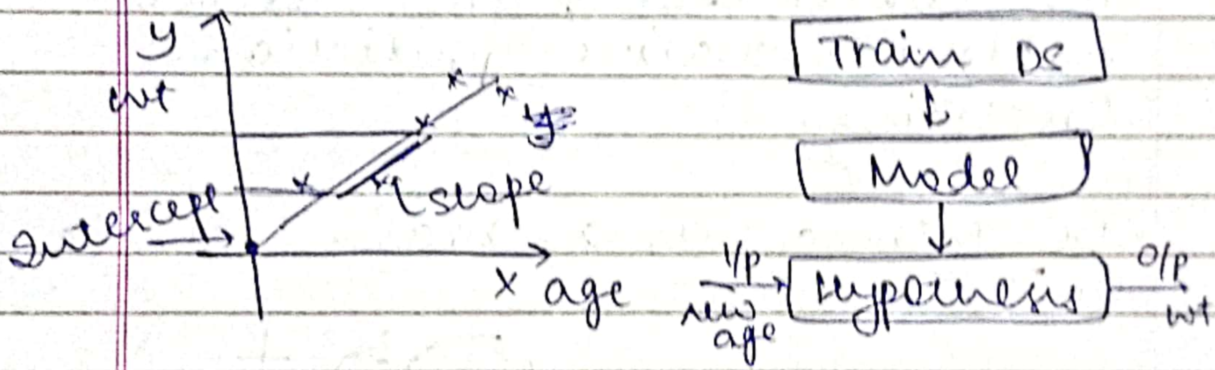
* Supervised Algo^{m}
  linear Regression, Ridge & Lasso, Logistic Regression, Decision Tree, Adaboost, Random Forest, Gradient Boosting,

Module 1M

XGBoost, Naive Bayes, SVM, KNN

*. Unsupervised Algo^{m}
KMeans, DBScan, Heirarchial clustering, K-Nearest Neighbour Cluster, PCA, LDA

1. Linear Regression



Eq^{n} of St. line
$y = mx + c$; $y = \beta_0 + \beta_1 x$;
Hypothesis
$h_\theta(x) = \theta_0 + \theta_1 x$

θ_0 = Intercept
{ when $x = 0$, $\theta = \theta_0$
at what pt. we get y axis
θ_1 = slope or coef.

Best fit line: when all the data pts are at min distance from the prediction line.

Cost f^{n}:
$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$ ⟶ Squared Error Function
↳ m: average
↳ derivation purpose / Only for simpler cal^{n}

PAGE NO.:
DATE: / /

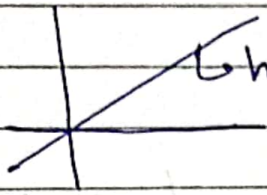
What we need to solve?

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

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

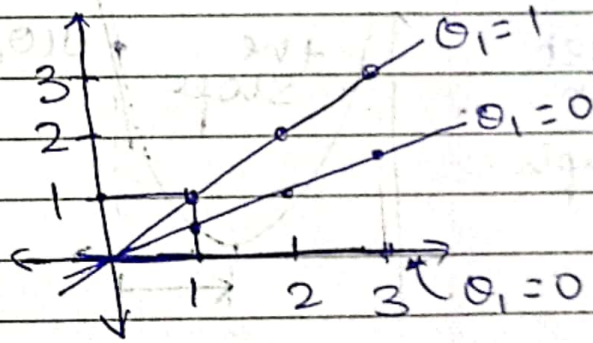
If $\theta_0 = 0$

Intercept is passing through origin.



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

(1,1) } pts
(2,2)
(3,3) }

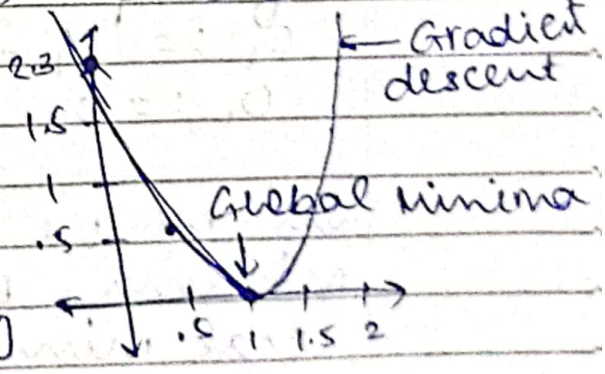


$\theta_1 = 1$

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

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

$J(\theta_1) = 0$



$\theta_1 = .5$

$J(\theta_1) = \frac{1}{2m} [(.5-1)^2 + (1-2)^2 + (1.5-3)^2]$

$= \frac{1}{2 \times 3} [.25 + 1 + 2.25]$

$= .58$

$\theta_1 = 0$

$J(\theta_1) = \frac{1}{2m} [(0-1)^2 + (0-2)^2 + (0-3)^2] = 2.3$

PAGE NO.:
DATE: / /

Convergence Algo^m

Repeat until convergence

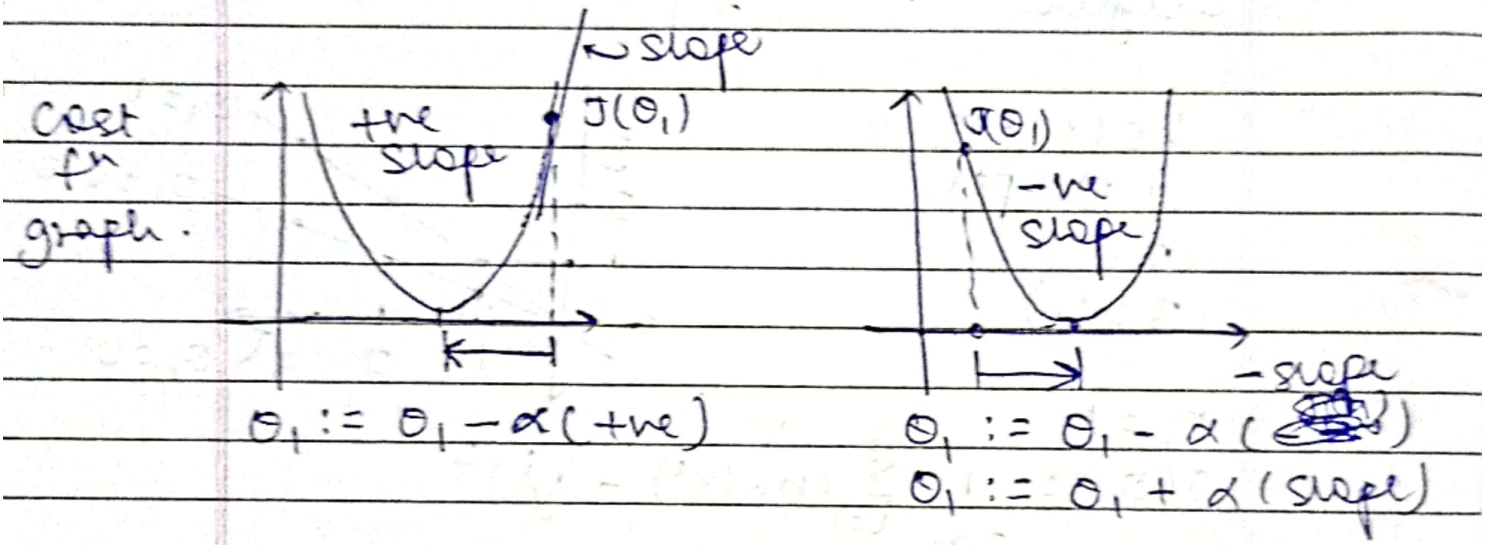
{ ↓ learning Rate

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

↳ derivative of J

↳ slope.

}



$\theta_1 := \theta_1 - \alpha(+ve)$

$\theta_1 := \theta_1 - \alpha(-ve)$

$\theta_1 := \theta_1 + \alpha(slope)$

If slope = 0

$\theta_1 := \theta_1 - \alpha(0)$

$\theta_1 := \theta_1$

In ML we don't have any local minima. But in DL we may have local minima and we are stuck.

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

PAGE NO.:
DATE: / /

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

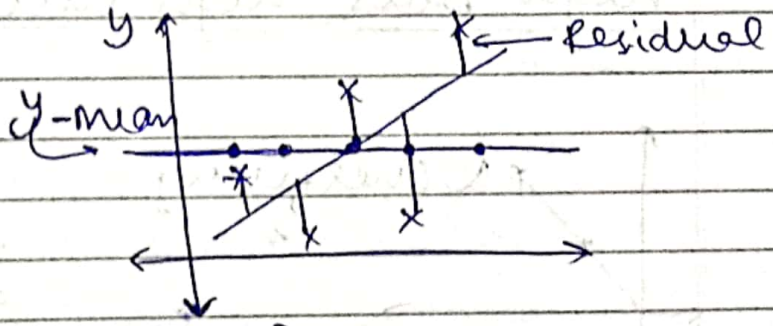
$$j=1 \Rightarrow \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i) \cdot x^i.$$

Performance Matrix

↳ R^2 and Adjusted R^2.

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}} = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

($\hat{y}$ ← $h_\theta(x)$)



↑ small no.

$= 1 - \frac{(low)}{(high)}$ = Big NO = 90%

If we add a feature then our R^2 will inc. But if we add a ~~completely independent~~ feature that is not correlated then also the R^2 inc. This is a problem.

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

P: features or predictors
N: No of data points

PAGE NO.:
DATE: / /

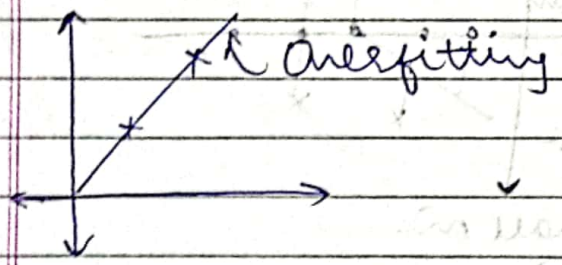
DAY:2 Ridge, Lasso Logistic Regression

Ridge & Lasso Regression

Overfitting: Model performs well with training data but fails to perform well ~~in training~~ with test data.
Low Bias, High variance

~~Model~~

Underfitting: Model accuracy is bad with training data and model accuracy is also bad with testing data.
High bias, high variance



Let, $\lambda = 1$

$h_\theta(x) = \hat{y}$

~~$J(\theta_1) = 0$~~

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

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

In ridge (L2 Regularization)

$= J(\theta_1) + \lambda (\text{slope})^2$

$= 0 + 1(2)^2$

$= 4 \downarrow\downarrow$

$= \underbrace{\frac{\Sigma(\hat{y} - y)^2}{2m}}_{\text{small value}} + \lambda (\text{slope})^2$

$= (\text{small value}) + \lambda (\text{slope})^2$

$= \frac{\text{SV}}{3} (\text{SV}) + 1(1.5)^2 \downarrow\downarrow$

PAGE NO. :
DATE : / /

- Ridge helps in creating generalized model. Helps avoid overfitting.

$\lambda \rightarrow$ hyperparameter
↳ steepness of the slope

LASSO (L1 Regularization)

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

↳ feature selection

$h_\theta(x) = \hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$

$|\text{slope}| = |\theta_0 + \theta_1 + \dots + \theta_n|$

(i) Prevents overfitting
(ii) Feature selection

Ridge Regression (L2 Reg.)

$CF = (h_\theta(x)^i - y^i)^2 + \lambda (\text{slope})^2$

- Prevents overfitting

Lasso Regression (L1 Reg.)

$C.f = (h_\theta(x)^i - y^i)^2 + \lambda |\text{slope}|$

- Prevents overfitting
- Feature selection

#. Assumptions of linear Regression

(i) If our features are in Normal or Gaussian distribution, our model will be trained well.

(ii) Standardization (scaling the data by z-score) (wherever there is gradient descent it's good to do Standardization.

(iii) It works w.r.t linearity. If our data is linearly seperable then it will be able to give good ans.

(iv) Multicolinearity

Logistic Regression (classification)

- Works well with binary classification
- Can also use for multi class.

Ex: Predict P/F based on no. of hrs study.

Q. Why can't we use linear regression for classification?

- Because of an outlier our line is completely shifted.
- Sometimes we are getting value greater than 1 & less than 0 (max & min are 1 & 0)

∴ we need to squash this f^n.

We use sigmoid activation f^n.

Decision Boundary Logistic Regression

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

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

Let $z = \theta_0 + \theta_1 x$

$h_\theta(x) = g(z)$

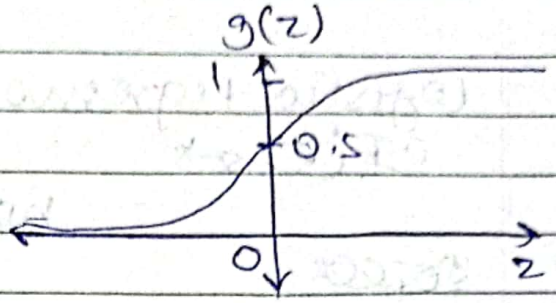
$h_\theta(x) = \frac{1}{1+e^{-z}}$ → Sigmoid or logistic f^n

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

PAGE NO.:
DATE: / /

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

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



- Training set:

$\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$

$y \leftarrow \{0, 1\}$ (2 o/p)

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

change parameter θ_1.

C.F of linear Reg.

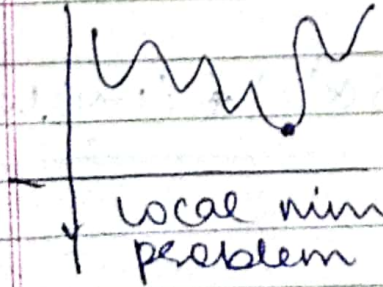
$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$

Logistic Regression

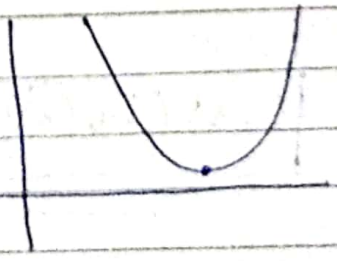
C.F $= \frac{1}{2} (h_\theta(x^i) - y^i)^2$] we cannot use this as C.F.

~~$= \frac{1}{2} (\frac{1}{1+e^{-(\theta_1 x)}} - y^i)^2$~~

non-convex f^n



convex f^n



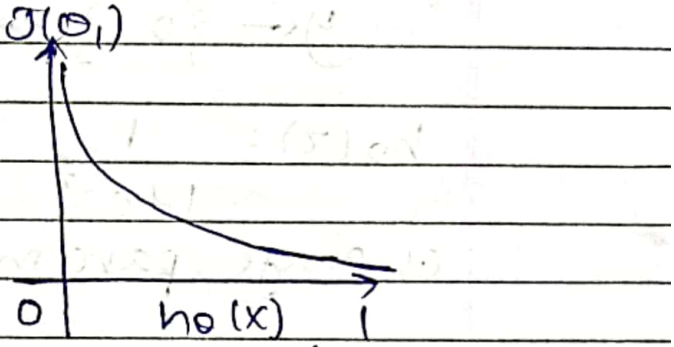
PAGE NO.:
DATE: / /

Logistic Regression CF.

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

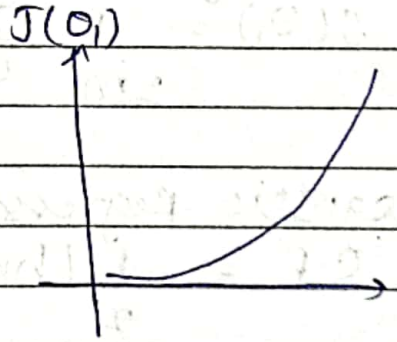
$$J(\theta_1) = \begin{cases} -\log(h_\theta(x)), & y=1 \\ -\log(1-h_\theta(x)), & y=0 \end{cases}$$

↑ Gives global min.



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

(classification)



if $y=0$

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

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

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

↑ Cost fn.

convergence algo^m

repeat until

{

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

}

Performance Matrix {classification problem}

x_1	x_2	Actual y_1	Actual y_2	{Pred}	1	0
-	-	0	1	1	3	2
-	-	1	1	0	1	1
-	-	0	0			
-	-	1	1			
-	-	1	1			
-	-	0	1			
-	-	1	0			

CONFUSION MATRIX

TP	FP
FN	TN

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

= 4/7 = 57%.

① Precision ⟶ TP / TP + FP

② Recall ⟶ TP / TP + FN

③ F-score ⟶ ~~$\frac{(P^2+1)PR}{P^2R+P}$~~

ex: Spam classification: Precision

Person has cancer or not: Recall

Stock market will crash: depends on problem statement.

↳ F-score

PAGE NO.:
DATE: / /

$$F\text{-score} = \frac{(1+\beta^2)\,\text{Precision} \times \text{Recall}}{\beta^2 \times (\text{Precision} + \text{Recall})}$$

F1 score If $\beta = 1$, $\frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$

Harmonic Mean

↳ $\frac{2pr}{p+r}$

F.5 score $\beta = 0.5$ FP ≥ FN → less important

↳ More importance

$$\frac{(1+.5^2)\,p.r}{(.5)^2(p+r)}$$

F2 score $\beta = 2$. FN > FP → less imp.

↳ More imp

PAGE NO.:
DATE: / /

DAY-3 KNN & NAIVE BAYES ALGOM

Naive Bayes Infusion (classification) by Bayes Th.

$$P(B|A) = \frac{P(B) * P(A|B)}{P(A)}$$ → Bayes Theorem

independent I/P ⟵ dependent O/P

X1 X2 X3 X4 .... Xn Y

$$P(Y|X_1, X_2, \dots, X_n) = \frac{P(Y) * P(X_1, X_2, \dots, X_n|Y)}{P(X_1, X_2, \dots, X_n)}$$

$$= \frac{P(Y) * P(X_1|Y) * P(X_2|Y) * \dots * P(X_n|Y)}{P(X_1) * P(X_2) * \dots * P(X_n)}$$

Y ⟷ yes / no

$$P(yes|X_i) = \frac{P(yes) * P(X_i|yes) * \dots * P(X_n|yes)}{P(X_1) * \dots * P(X_n)}$$ # fixed

$$P(no|X_i) = \frac{P(no) * P(X_i|no) * \dots * P(X_n|no)}{P(X_1) * \dots * P(X_n)}$$ # fixed

∴ we ignore denominator

$P \le 0.5 = 0$
$P \ge 0.5 = 1$

$P(yes|X_i) = .13$
$P(no|X_i) = .05$

$$\text{Normalization} = \frac{0.13}{0.13 + 0.05} = 0.72 = 72\%$$ ⟹

$P(yes)$

$P(no) = 1 - 0.72 = 0.28 = 28\%$

PAGE NO.:
DATE: / /

Dataset:

Day	Outlook	Temp	Humidity	Wind	Play(o/p)

#Outlook

P(Sunny | yes)

		Yes	NO	P(Y)	P(N)
Cal. of Outlook	Sunny	2	3	2/9	3/5
	Overcast	4	0	4/9	0/5
	Rain	3	2	3/9	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

TestCase → (Sunny, Hot)

$$P(Yes/Sunny, Hot) = \frac{P(Yes) * P(Sunny/Yes) * P(Hot/Yes)}{P(Sunny) * P(Hot)}$$

$$= \frac{\cancel{9} * \cancel{2} * 2}{\cancel{14} \; 7 \; \cancel{9} \quad 9}$$

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

$$P(NO/Sunny, Hot) = \frac{\cancel{9}\cancel{5} * 3 * 2}{\cancel{14} \; 7 \; \cancel{5} \quad 5} = \frac{3}{35} = 0.085$$

$P(Yes/x_i) = .031 \quad 1 - .73 = 27\%$

$P(No/x_i) = .085 \quad = \frac{.085}{.085+.031} = 73\%$

→ (Sunny, Hot) → NO
↳ pred^.

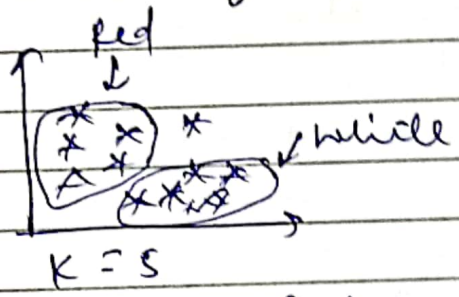
→ (Overcast, Mild)

$P(Y/x_i) = \frac{2}{9} \times \frac{4}{7} \times \frac{9}{14} \times \frac{4}{9} = \frac{8}{63}$

$P(N/x_i) = 0$

KNN Algo^m
↳ classification
↳ Regression

(i) Classification



K = 5

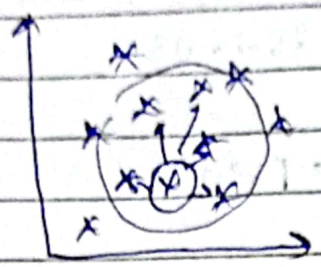
Max no Red = 3
White = 2

we calculate dist using eucledian dist^n & Manhadtan dist^n.

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

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

(ii) Regression



$k = 5$

↳ hyperparameter

Avg of all pts is o/p.

KNN worst cases

(i) outliers

(ii) Imbalanced dataset

PAGE NO :
DATE : / /

DAY-4 DECISION TREE & ENSEMBLE ALGO^m

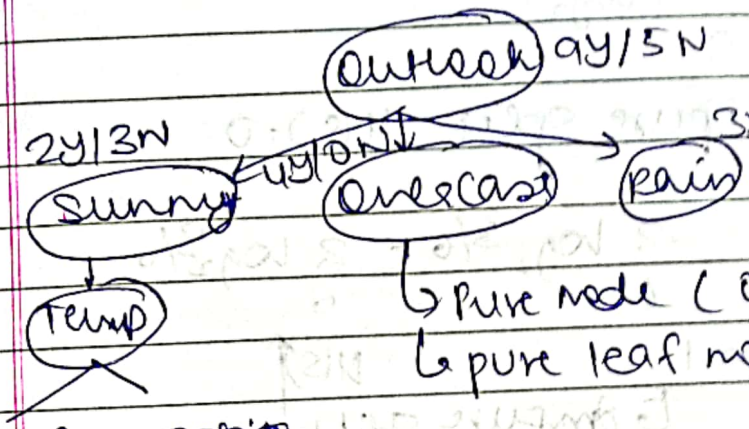
Decision Tree
- Regression
- classification

- we use nested if-else conditions

(i) classification

Dataset.

Day	outlook	Temp	Humidity	Wind	o/p



~~Pure split~~

we will split till we get a leaf node.

① Purity:
- Entropy
- Gini coefficient / Impurity

② How the features are selected?
- Information gain

① Entropy

$H(S) = -P \log_2 P - R \log_2 R$

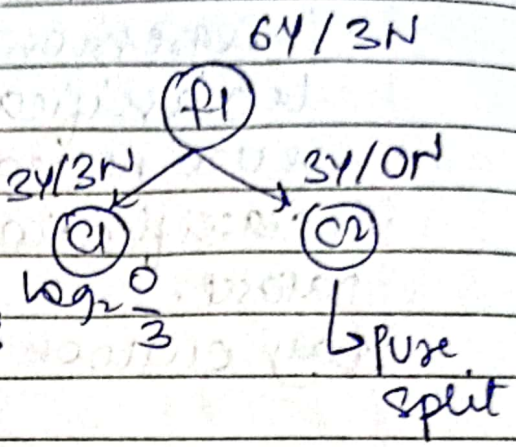
② ~~Gini Impurity~~

PAGE NO.:
DATE: / /

① Entropy

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

$P_+ \rightarrow$ Prob. of Yes.



$H(S)_{C_2} = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$

$= -1\log_2 1$

$H(S) = 0$

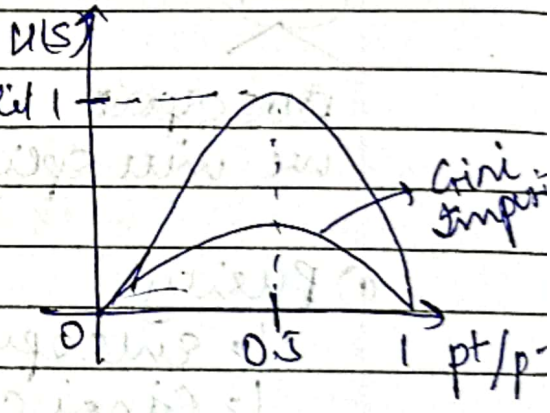
∴ pure split $H(S) = 0$

$H(S)_{C_1} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$

$H(S) = 1$

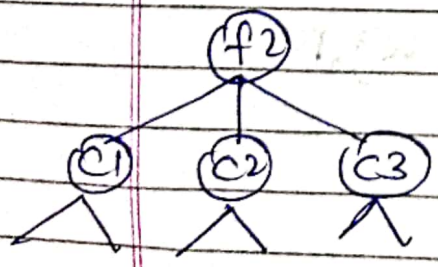
∴ impure split

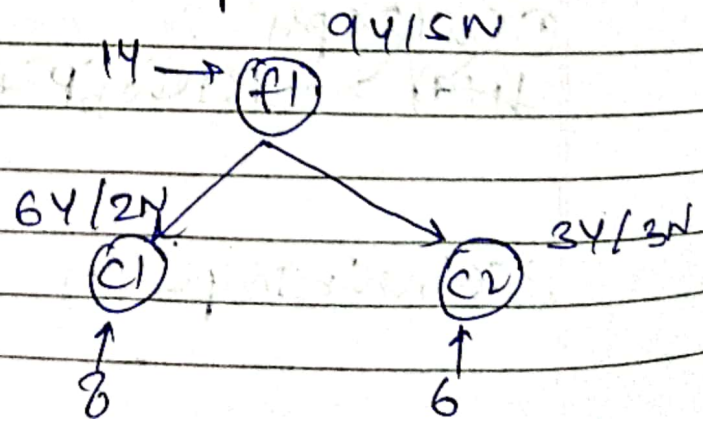
$0 \leq H(S) \leq 1$



Purity test is done with entropy

② which feature to split





PAGE NO.:
DATE: / /

Information Gain

$$Gain(S, f1) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} \cdot H(S_v)$$

$$H(S) = \frac{-9}{14}\log_2 9/14 - \frac{5}{14}\log_2 5/14 \approx 0.94$$

$$H(S_{v_1}) = -6/8 \log_2 6/8 - 2/8 \log_2 2/8 \approx 0.81$$
$$H(S_{v_2}) = -3/6 \log_2 3/6 - 3/6 \log_2 3/6 = 1$$

$$Gain(S, f1) = 0.94 - \left[\frac{8}{14} * 0.81 + \frac{6}{14} * 1\right]$$

$$Gain(S, f1) = 0.049$$
$$Gain(S, f2) = 0.051$$

$G(S, f2) > G(S, f1)$
∴ we will start the feature split with f2.

② Gini Impurity

$n = 2$ o/p ⇒ Yes / No

$$G_I = 1 - \sum_{i=1}^{n} (p)^2$$

$$= 1 - [(p_+)^2 + (p_-)^2]$$
$$= 1 - [(1/2)^2 + (1/2)^2]$$
$$= 0.5$$

2Y/2N

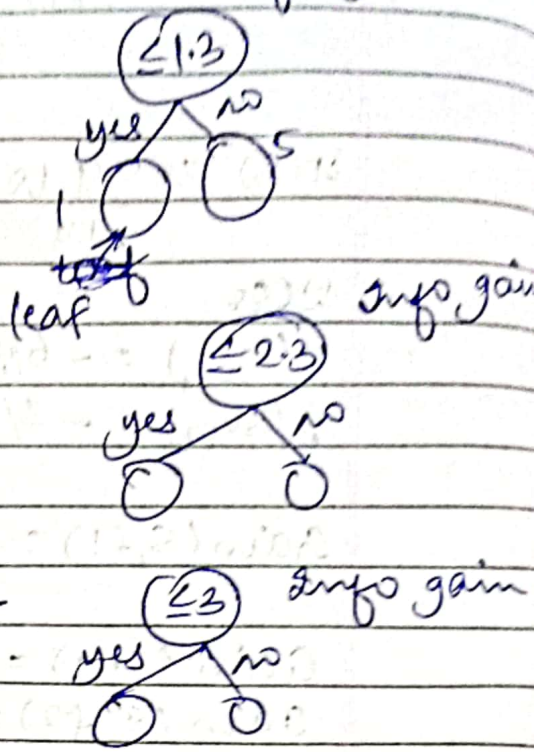
Impure split

$$0 \leq G_I \leq 0.5$$

* Gini Impurity is faster than entropy.
No. of features is greater → gini
less → entropy

If f1 is numerical rather than object.

f1	o/p	f1
2.3		1.3
1.3		2.3
4		3
3		4
5		5

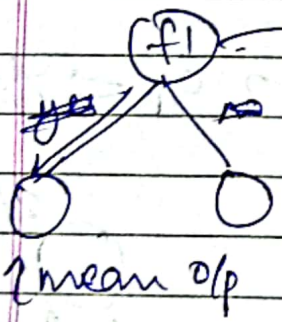


It will cal information gain for all the records and the best will be selected.

Decision Tree Regressor

f1 → continuous

↳ mean MSE or MAE



↳ $\frac{1}{2m}\sum_{i=1}$ $(y_i - y_i)^2$

- In regressor instead of using entropy or gini impurity we use MSE or MAE.

PAGE NO.:
DATE: / /

Hyperparameters

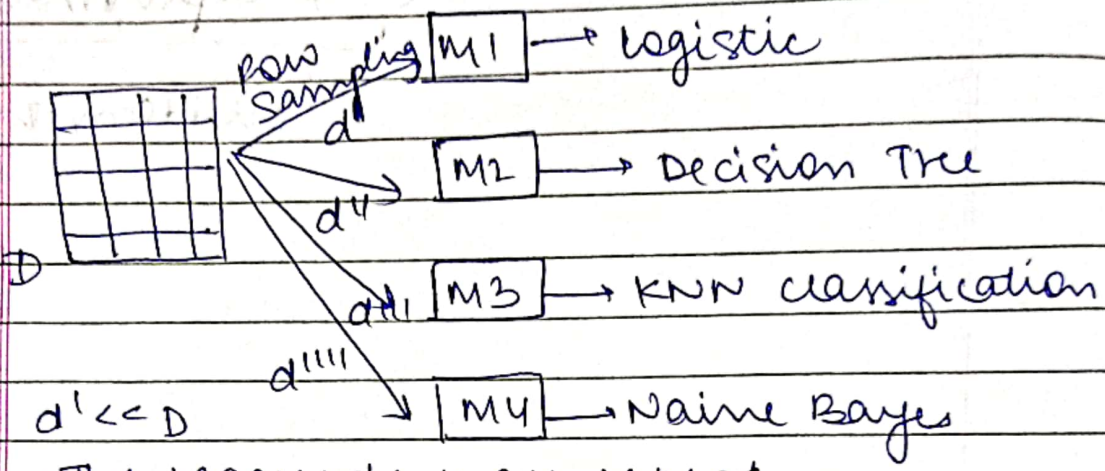
Decision → Overfitting
↳ Post pruning
↳ Pre pruning
(max-depth, max-leaf)
↳ GridSearchCV

PAGE NO.:
DATE: / /

DAY:5 Adaboost, Random Forest, XGBoost

Ensemble Techniques
↳ Multiple algo to solve a problem.

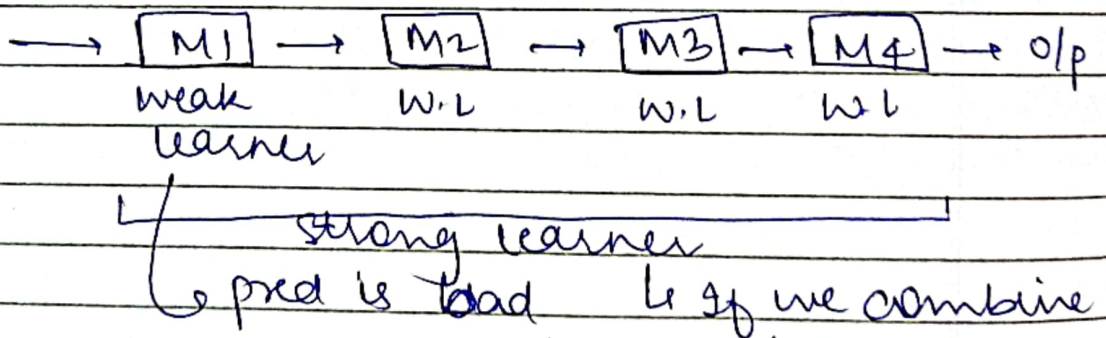
(i) Bagging - classification



- The records may repeat.
- There will be 100 to 200 models.

- Regression
- Average/mean will be taken.

(ii) Boosting



↳ If we combine multiple W.L then we have strong learner.

- It's a sequential set of all the models combined together and these models are weak learner. Then

PAGE NO.:
DATE: / /

when we combine them they gives us strong learners.

BAGGING	BOOSTING
1. Random Forest classifier	1. Adaboost
2. Random Forest Regressor	2. Gradient
	3. xgboost

1. Random Forest classifier

Q. what is the main disadv. of decision tree?

Ans Overfitting

∴ low bias & high variance.

- Normalization is not required in Decision Tree.
- Random forest is not affected by the outliers

Adaboost

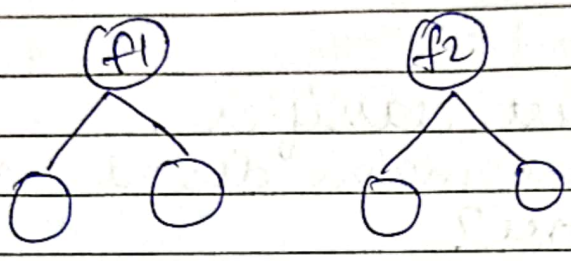
f1	f2	f3	f4	o/p	weight
–	–	–	–	Yes	1/7
–	–	–	–	No	1/7
–	–	–	–	–	1/7
					⋮

$\sum \frac{1}{7}$

overall sum of wt = 1

PAGE NO.:
DATE: / /

- We create decision trees ~~on~~ by selecting the features with the help of information gain or gini impurity.
- In random forest we create the decision trees upto only one levels, known as STUMPS.



{weak learner}

Step I: ~~Total error~~ If only one record was predicted wrong.

∴ Total Error = 1/7.

Step II. Performance of a Stump

$$= \frac{1}{2}\log_e\left(\frac{1 - T.E}{T.E}\right) \rightarrow \text{Total error.}$$

$$= \frac{1}{2}\log_e\left(\frac{1 - 1/7}{1/7}\right)$$

$$\approx 0.895$$

Step III. Update all the weights

performance of Stump ↑

For correct records = wt $* e^{-Ps}$

$$= \frac{1}{7} \times e^{-0.895} \approx 0.05$$

For incorrect records: wt $* e^{Ps}$

$$= \frac{1}{7} \times e^{0.895} \approx 0.349$$

PAGE NO.:
DATE: / /

- The wt of incorrect records are higher than the correct records bcc. if they have lower weights then they will train more by passing them through more weak learners.

Σ New wt $\neq 1$ to 1
but we know the wt is always equal,

∴ we find the normalized wt

divide all the wt. by the sum of new. wt.

New wt : 0.05, 0.05, 0.05, 0.349, 0.05, 0.05, 0.05
normalized wt : $\frac{0.05}{0.649}, \frac{0.05}{0.649}, \frac{0.349}{0.649}, \ldots$
: 0.07, 0.07, 0.07, 0.537, 0.07 ... 0.07

new buckets will be created

[0 - 0.07], [0.07 - 0.14], [0.14 - 0.21],
[0.21 - 0.747], [0.747, - 0.754], ...
↳ wrong bcc

- In classification problem we take the majority ans.
- In regression we take the mean / avg of all the pred / o/p.

PAGE NO.:
DATE: / /

- Black Box Model: Here we are not able to see how all the weights are updated. The maths behind the algom is complex.
  ex: Random Forest, ANN
- White Box Model: The maths is simple, we can visualize how all the weights are updated.
  ex: Linear Regression, Decision Tree

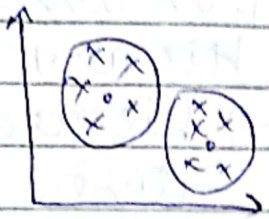
PAGE NO.:
DATE: / /

DAY-6 Kmeans, Hiraerical clustering, Sil hewth score, DBScan cluster.

#. Kmeans clustering

K = centroids

for K=2



Step I: we try with diff. K values

K = 1, 2 . . . n.

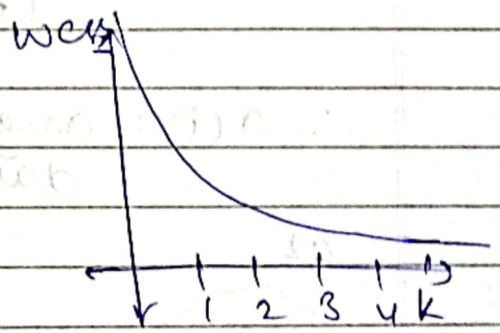
Step II. initialize K no. of centroids

Step III: compute the avg. W update centroide.

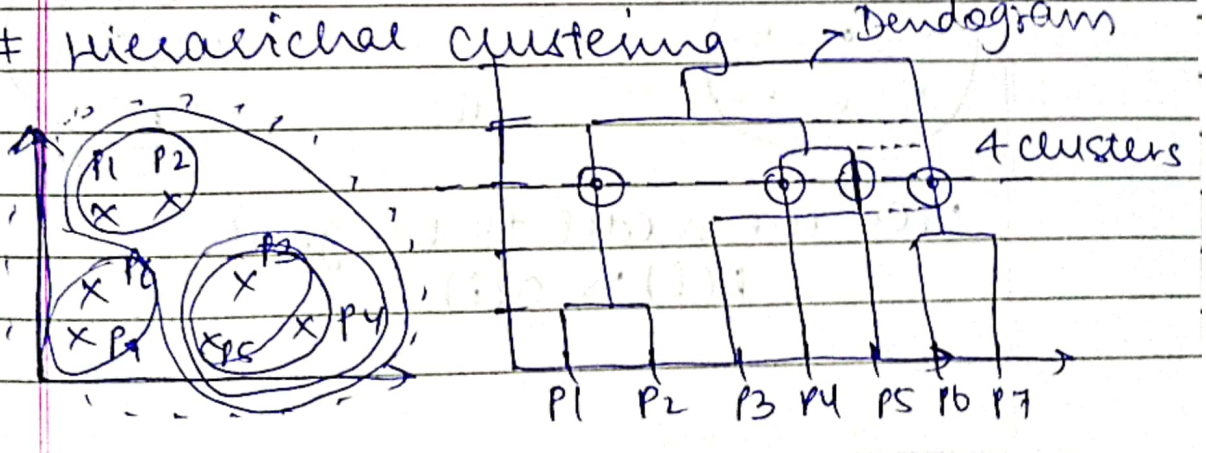
- ELBOW Method (K value)

WCSS - within cluster sum of square

for i = 1 – 10



Hieraerichal clustering



- You need to find the longest vertical line data has no horizontal line passed through it.

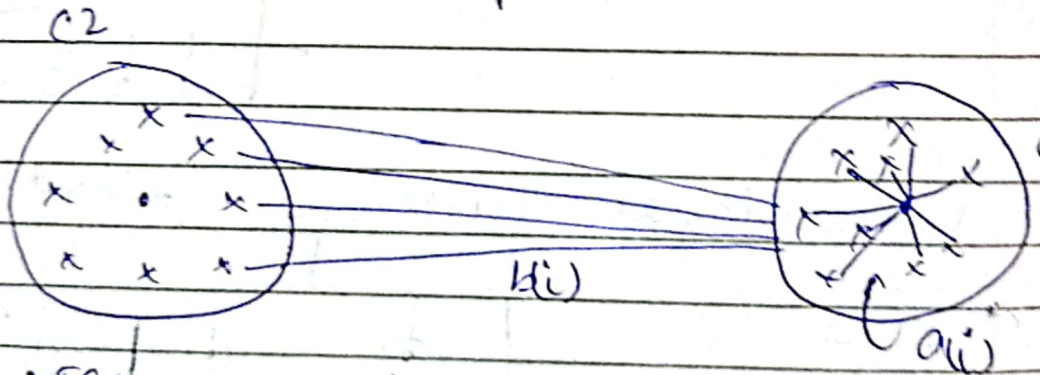
Q. Max time is taken by Kmeans or Hierarchial clustering?
→ Hierarchial clustering will take more time.

- If ds is small → ~~Hcit~~ Hierarchial
- If ds is large → K Means

Validation of Clustering Model
↳ Silhouette ~~SS~~ score

$$a(i) = \underbrace{\frac{1}{|C_i| - 1}}_{\text{↳ average}} \underbrace{\sum_{j, i \neq j} d(i,j)}_{\text{↳ dist bet centroid \& data pts}}$$

∴ a(i) = avg of dist bet centroid & data points.



- For the model to be good
b(i) >> a(i)

$$b(i) = \min \frac{1}{|C_J|} \Sigma d(i,j)$$

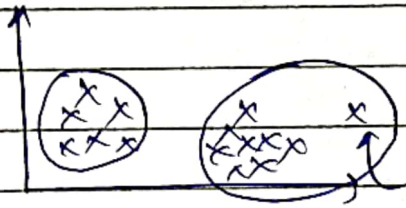
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{if } |C_I| > 1$$

$$-1 \leq s(i) \leq 1$$

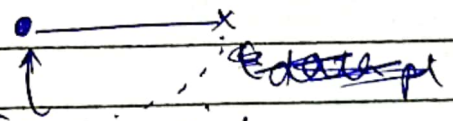
↳ better model

DBScan clustering
(Density Based spatial clustering of Application with Noise)

- we name the outliers seperately and this outliers are known as noise pts.



- epsilon ↳ radius of the circle created by ~~data pt~~ the neighbourhood around a pt x.



- minpts: The density threshold. If a neighbourhood includes at least minpts. it will be considered as a dense region. Alternatively, a point

will be considered as dense if there are atleast the value of minpt in its ε neighbourhood. These dense pts are called core points.

- Border Point: has ε neighbourhood that contains less than minpts but it belongs to the ε neighbourhood of another core pt.

Bias is a phenomenon that skews the results of an algo^m in favor or against an idea (training ds.).

Variance refers to the changes in the model when using different proportion of training or test data.

DAY-7 XGBOOST, SVM, SVR

↳ Extreme Gradient Boosting

XGBOOST classifier:

Salary	Credit	Approval	Residual
<=50	B	0	-.5
<=50	A	1	.5
<=50	A	1	.5
>50	B	0	-.5
>50	A	1	.5
>50K	BN	1	.5

Base Model → pr=0.5

↳ probability of approval

step I: create a binary decision tree using the features.

step II: Cal Similarity wt = $\frac{\Sigma(\text{Residual})^2}{\Sigma(Pr(1-Pr)+\lambda)}$

step III: Calculate information gain

[-.5, .5, .5, -.5, .5, .5, -.5]

Salary ≈ .14

<=50 | >50

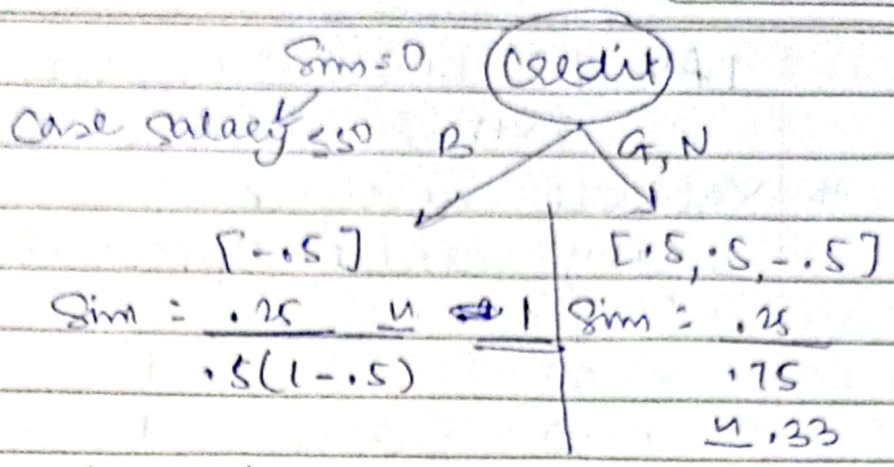
0

$\frac{(-.5+.5+.5-.5)^2}{4(.5)(1-.5)}$ | $\frac{(-.5+.5+.5)^2}{3(.5)(1-.5)}$

≈ .33

Info^{M} gain = 0 + 0.33 - 0.14 = 0.19

PAGE NO.:
DATE: / /



Info^n gain = 1 + 0.33 - 0 = 1.33

* $\sigma[0 + \alpha(Sim)]$
  - α → learning rate
  - σ → sigmoid f^n.

$\sum_{i=0} \sigma_i$

- xgboost → Black box Model

xgboost Regression

Exp^n	Gap	Salary	Rs
2	Y	40K	-11K
2.5	Y	48K	-9
3	N	52K	1
4	N	60K	9
4.5	Y	62K	11

Base Model = avg O/p = 51K

PAGE NO.:
DATE: / /

(-11, -9, 1, 9, 11)

Eqn → Sim ≈ 1/6

≤ 2 | > 2

(-11) | (-9, 1, 9, 11)

Sim = |2|/2 ≈ 60.5 | Sim ≈ 28.8

Sim wt: $\frac{\Sigma(\text{residual})^2}{\text{no. of residual} + \lambda}$

↳ Hyperparameter

Info^n gain = 60.5 + 28.8 − 1/6 ≈ 89.13

Eqn → Sim ≈ 1/6

$\frac{-11-9}{2} = -10$ ≤ 2.5 | > 2.5

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

Sim ≈ 133.33 | Sim ≈ 110.25

Info^n gain = 133.3 + 110.95 − 1/6

$o/p = 5l + \alpha_1(-10) + \alpha_2(DT_2) + \alpha_3(DT_3) + \dots + \alpha_n(DT_n)$

SVM.

Hard/Soft → in case of overlapping marginal plane

↳ Hyperplane

$w^T x + b = -1$

$w^T x_1 + b = 1$

$(-)\ w^T x_2 + b = -1$

$w^T(x_1 - x_2) = 2$

$\frac{|w|}{|w|} \quad \frac{|w|}{|w|}$

↳ $w^T x + b = +1$

↳ $w^T x + b = 0$

{∵ $y = mx + c$}

we have to maximise $2/|w|$.

such that,

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

or

$$y_i * (w^T x_i + b) \geq 1$$

↳ for correct points

$$\max_{(w,b)} \frac{2}{||w||} \Rightarrow \min_{(w,b)} \frac{||w||}{2}$$

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

↳ sum of dist. of the wrong pts.

SVR

↳ similar to SVM with some minor changes in the part

$$C_i \sum_{i=1}^{n} \eta_i$$

S