Date: 15 oct 24

Quiz-4 (set A)

① Shape input is (64, 64, 1)

Layer	Calculation	O/P shape	Total Parameter (weight + bias)
* Conv 2D Filter-64, Kernel size = 3	$W/H = \left\lfloor \frac{W/H - K + 2P}{S} \right\rfloor + 1$ $= \left\lfloor \frac{64 - 3 + 0}{1} \right\rfloor + 1$ $= 62$	62, 62, 64 no of filters	Weight = filter size × no of channels × no of filters. bias = no of filters $\{(m \times n \times d) + 1\} k$ $\{3^2 \times 1 + 1\} 64$ $= 640$
* Max pooling 2D Pool size = 2, stride = 2	$W/H = \left\lfloor \frac{62 - 2 + 0}{2} \right\rfloor + 1$ $= 31$	31, 31, 64	0
* Conv 2D filter = 64, kernels = 3	$W/H = \left\lfloor \frac{31 - 3 + 0}{1} \right\rfloor + 1 = 29$	29, 29, 64	$\{(3 \times 3 \times 64) + 1\} \times 64$ $= 36928$
maxpooling 2D Poolsize = 2 stride = 2	$W/H = \left\lfloor \frac{29 - 2 + 0}{2} \right\rfloor + 1 = 14$	14, 14, 64	0
Flatten	14 × 14 × 64	12544	0
Dense-1 units-128	128 × 1	128	$(12544 \times 128) + 128$ $= 1605760$
Dense-1 units-4	4 × 1	4	$(128 \times 4) + 4$ $= 516$

Q2: Is transfer learning strategy always more cost-effective than naive CNN approach? Does transfer learning always improve the result? Give valid justification of your response.

Ans: Transfer learning is not always more cost-effective or guaranteed to improve results compared to a naive CNN approach. It's effectiveness depends on the task, data & model used.

Cost Effectiveness: Transfer learning can be more cost effective when the task has limited data, as it leverages pre-trained models, reducing the need for extensive training. However for task with abundant, domain-specific data, ~~treaing~~ training a naive CNN from scratch may yeild better performance at a comparable computation cost.

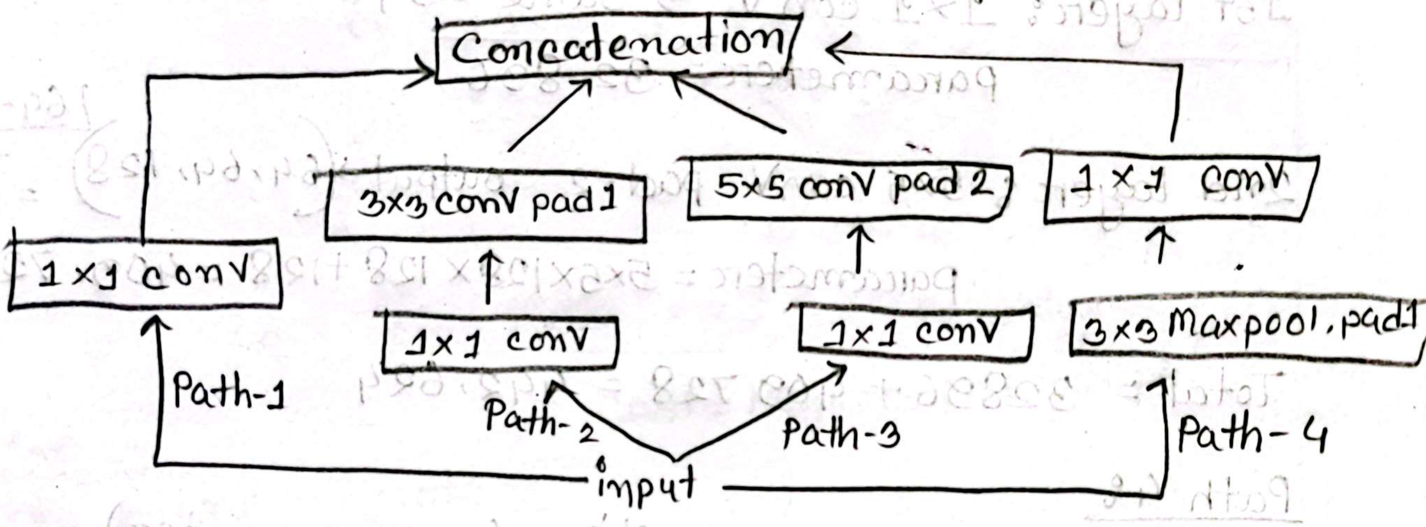
Performance: TL improves result when the pre trained model's knowledge is relevant to the target task, like classifying dog breeds, using a pre-trained model like ResNet (trained on image) can save time and give high accuracy. Other side, For task that are significantly different from the source domain, a naive CNN perform better. Classifying a medical x-rays, a naive CNN might outperform transfer learning from a general model like ResNet, as the pre-trained model may not have learned relevant medical features.

So it is not guaranteed that TL always outperform or more cost effective than a naive CNN.

(setB) → can't find.

set-C

The following inception block consists of 4 parallel paths. Find out the total number of training parameter after all the operations and calculate the dimension of the concatenation layer. The input shape is [64, 64, 256] and all the operations have 128 filters.



Ans:

Path 1: 1×1 conv ; input shape = 64×64×256

output shape = 64×64×128

Parameters = $1 \times 1 \times 256 \times 128 + 128 = 32896$

Path 2:

→ 1x1 conv followed by 3x3 conv pad→1

→ 1st Layer: 1×1 conv

⇒ input = 64x64x256 ; output = 64x64x128

⇒ Parameter = $1 \times 1 \times 256 \times 128 + 128 = 32896$

2nd layer: 3x3 conv pad-1 $\left\lfloor \frac{64-3+2}{1} \right\rfloor + 1 = 64$

→ input = 64x64x128 output → 64,64,128

→ parameter = 3x3x128x128 + 128 = 147,584

Total for Path 2 = 32896 + 147,584 = 180,480

Path 3:

1st layer: 1x1 conV → same as previous.

parameter = 32,896

2nd layer: 5x5 conV pad-2 output → (64, 64, 128) $\left\lfloor \frac{64-5+4}{1} \right\rfloor + 1 = 64$

parameter = 5x5x128x128 + 128 = 409,728

Total = 32896 + 409728 = 442,624

Path 4:

1st layer → 3x3 max pooling (no parameter)

output = 64x64x256

2nd layer → 1x1 conV input = 64x64x256 out = 64x64x128

parameter = 1x1x256x128 + 128 = 32896

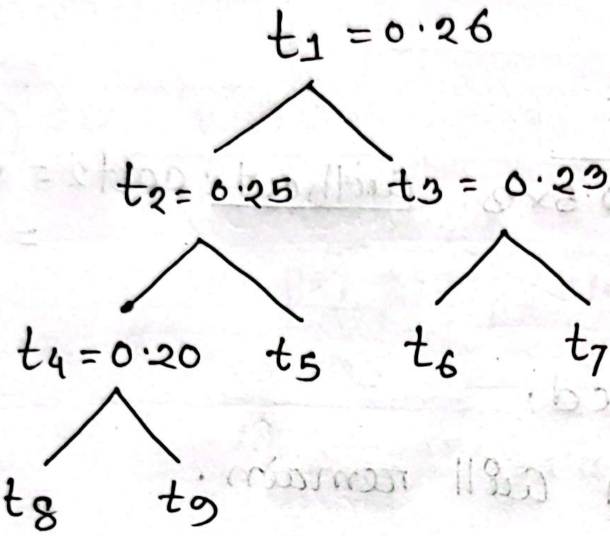
Total Parameters = 32896 + 180480 + 442624 + 32896

= 688,896

Concatenation layer → the output from all paths is concatenated along the channel dimension.

Output shape: 64x64x(128+128+128+128) = 64x64x512.

2. Let's the following tree has a regression value = 0.26 at t_1, 0.25 at t_2, 0.23 at t_3 and 0.20 at t_4. Find out the optimal tree by using cost complexity pruning. $\alpha = 0.5$



Value or

Cost = Regression Error + (α x no of leaf node)

Step 1:

t_4 prune or not:

with t_4 node

$$\text{Cost} = \text{Value} + (\alpha \times \text{no of Leaf node})$$

$$= 0.20 + 0.5 \times 5 \quad (t_8, t_9, t_5, t_6, t_7)$$

$$= 2.7$$

without t_4 node

$\longrightarrow t_4, t_5, t_6, t_7$

$$\text{Cost}_2 = \text{Value} + (\alpha \times \text{no of Leaf node})$$

$$= 0.23 + 0.5 \times 4$$

$$= 2.23$$

$\text{cost} > \text{cost}_2 \rightarrow$ Prune.

t_3 prune or not

with: $cost = 0.23 + 0.5\times 4$
$= 2.23$

without: $cost2 = 0.25 + 0.5\times$
$= 1.75$

t_3 will be pruned

t_2 prune or not:

with: $cost = 0.25 + 0.5\times 3$
$= 1.75$

without: $cost2 = 0.26 + 0.5\times 2$
$= 1.26$

t_2 will be pruned.

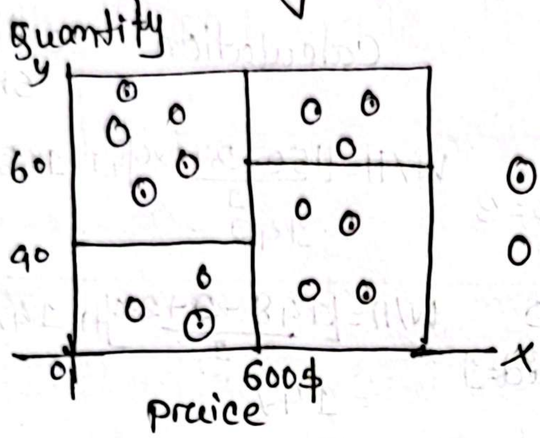
Only the root t_1 will remain.

Set D

1. Shape of input is (150, 150, 3)

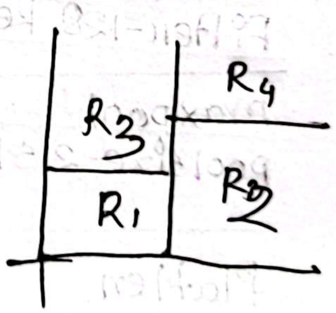
Layers	Calculation	output shape	Parameter $\{(m\times m\times d)+1\}\times k$
ConV 2D filter=16 kernels=3	$W/H=\left\lfloor\frac{150-3+2\times0}{1}\right\rfloor+1 = 148$	148,148,16	$\{(3\times3\times3)+1\}\times16 = 448$
maxpooling 2D pool size=2 strid=1	$W/H=\left\lfloor\frac{148-2+2\times0}{1}\right\rfloor+1 = 147$	147,147,16	0
ConV 2D Filter=64 kenel=3	$W/H=\left\lfloor\frac{147-3\times2\times0}{1}\right\rfloor+1 = 145$	145,145,64	$\{(3\times3\times16)+1\}$ 64 $= 9280$
maxpooling 2D Poolsize=2 Straide=1	$W/H=\left\lfloor\frac{145-2*2\times0}{1}\right\rfloor+1 = 144$	144,144,64	0
Con2D Filter-128 kernel-3	$W/H=\left\lfloor\frac{144-3+0}{1}\right\rfloor+1 = 142$	142,142,128	$\{(3\times3\times64)+1\}\times128 = 73856$
maxpool poolsize-2 strid=1	$\left\lfloor\frac{142-2+0}{1}\right\rfloor+1 = 141$	141,141,128	0
Flatten	141x141x128	2544768	0
Dense 1 unint=512	512x ~~2544768~~ 1=	512	2544768x512+512 =1302921728
Dense 1 unit=1	1x1	1	512x1 +1 =513

2. Explain in which condition the following partitions/ regions are created optimally to formulate a regression tree.



Ans: In the above image, the goal is to explain the optimal conditions for partitioning a data set to create regression tree. Where price and Quantity at x-axis, y axis respectively and there are created 4 region with various threshold values.

The idea is to divide the features such that the sum of squared error (SSE) is minimized in each region.



Now for spliting Region.

$R_1(j,s)$ correspond to the set of data points where feature X_1 (Price) is less than some threshold s ($X_1 < s$) (600$) and $R_2(j,s)$ will be $X_1 > s$

$R_1(j,s) = \{X \mid X_1 < s\}$ $R_2(j,s) = \{X \mid X_1 > s\}$

Now here we seek the value of j and s that minimize the equation →

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \quad \text{——} (1)$$

Again for X_2 (Quantity) feature there are 2 more region for threshold $s = 40$ R_1 & R_3

$$\text{minimize} \sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_3(j,s)} (y_i - \hat{y}_{R_3})^2 \quad \text{——} (ii)$$

$s = 60$

$$\text{minimize} \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 + \sum_{i: x_i \in R_4(j,s)} (y_i - \hat{y}_{R_4})^2 \quad \text{——} (iii)$$

Here $\hat{y}_{R_1}, \hat{y}_{R_2}, \hat{y}_{R_3}, \hat{y}_{R_4}$ are mean predicted values for region, R_1, R_2, R_3, R_4.

Repeat this process, looking for the best predictor and best cutpoint in order to split the data further so as to minimize the RSS within each of the resulting regions.

1. YOLO is the multi-task learning model. How does it perform multi-task learning to detect object. Explain your answer with proper mathmatical equations.

Ans: YOLO is the multi-task learning mode. It is used to detect object.
Suppose an image has several instances. In YOLO the image is divided into several cells. Each cell contains a number of bounding boxes. And the object is detected with bounding box. Here, confidence is $P_r(\text{object}) \times IOU \rightarrow$ (Intersection over Union)

$$IOU = \frac{\text{Area of overlap}}{\text{Area of union}}$$

& The object is chosen by the highest confidence from each cell. Here 3 types of loss is calculated →

Classification loss:

$$\sum_{i=0}^{S^2} \mathbb{1}_i^{obj} \sum_{c \in classes} (P_i(c) - \hat{P}_i(c))^2$$

where, $\mathbb{1}_i^{obj}$ is 1 if an object is in cell, otherwise 0.

$\hat{P}_i(c) \rightarrow$ the probability of having class c in cell i.

Localization Loss: This loss measures the error of the predicted bounding boxes with respect to the expected ones.

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] +$$

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \right]$$

Here $x_i, y_i \rightarrow$ Center coordinates of the bounding boxes.
$\hat{x}_i, \hat{y}_i \rightarrow$ Predicted " " " "

$\lambda_{coord} \rightarrow$ hyperparameter that adjusts the importance of localization error, giving more or less emphasis compared to other components of the total loss, such as confidence loss or classification loss.

$\sqrt{w_i}, \sqrt{h_i} \rightarrow$ Width and height $\rightarrow$ square root is used to reduce the impact of large bounding box.

Confidence Loss: it measures the error when deciding if an object is in the box or not.

$$\sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} (c_i - \hat{c}_i)^2$$

$\hat{c}_i \rightarrow$ Confidence of the box j in cell i

$\mathbb{1}_{ij}^{obj} \rightarrow 1$ if jth box in cell i is responsible for detecting the object.

Since most cells doesnot contain an object so we have to do further computation.

$$\lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj} (c_i - \hat{c}_i)^2$$

Total Loss → সব যোগ

Ques:

(2)

		Actual		
		Se	Ve	Vi
Predicted	Se	16 (C-1)	0 (C-2)	0 (C-3)
	Ve	0 (C-4)	15 (C-5)	2 (C-6)
	Vi	0 (C-7)	3 (C-8)	12 (C-9)

	Actual Y	Actual N
Predict Y	TP	FP
Predict N	FN	TN

or

	Predict Y	Predict N
Actual Y	TP	FN
Actual N	FP	TN

Ab:

	Se	Ve	Vi
TP	$C_1 = 16$	$C_5 = 15$	$C_9 = 12$
FP	$C_2 + C_3$ 0	$C_4 + C_6$ $0+2$ $= 2$	$C_7 + C_8$ $0+3$ $= 3$
FN	$C_4 + C_7$ $0+0$ $= 0$	$C_2 + C_8$ $0+3$ $= 3$	$C_3 + C_6$ $= 0+2$ $= 2$
TN	Se বরাবর row col বাদ দিয়ে, $C_5 + C_6 + C_8 + C_6$ $= 15 + 2 + 3 + 12$ $=$	$C_1 + C_3 + C_7 + C_9$ $= 16 + 0 + 0 + 12$ $=$	$C_1 + C_2 + C_4 + C_5$ $= 16 + 0 + 0 + 15$

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}, \quad F_1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Macro-Averanging (Precision)} = \frac{\text{Precision}_{(se)} + P_{(ve)} + P_{(vi)}}{3}$$

$$\text{" " (Recall)} = \frac{\text{Recall}_{(se)} + R_{(ve)} + R_{(vi)}}{3}$$

$$\text{Micro-Averanging (Precision)}: \frac{TP_{(se)} + TP_{(ve)} + TP_{(vi)}}{(TP_{se} + FP_{se}) + (TP_{ve} + FP_{ve}) + (TP_{vi} + FP_{vi})}$$

$$\text{" " Recall}: \frac{TP_{(se)} + TP_{(ve)} + TP_{(vi)}}{(TP_{se} + FN_{se}) + (TP_{ve} + FN_{ve}) + (TP_{vi} + FN_{vi})}$$

4x4 Confusion Matrix:

	A	B	C	D
TP	C_1	C_6	C_{11}	C_{16}
FP	$C_5+C_9+C_{13}$	$C_2+C_{10}+C_{14}$	$C_3+C_7+C_{15}$	$C_4+C_8+C_{12}$
FN	$C_2+C_3+C_4$	$C_5+C_7+C_8$	$C_9+C_{10}+C_{12}$	$C_{13}+C_{14}+C_{15}$
TN	row, col বাদ দিয়ে সব			

Predict

Actual	A	B	C	D
A	C_1	C_2	C_3	C_4
B	C_5	C_6	C_7	C_8
C	C_9	C_{10}	C_{11}	C_{12}
D	C_{13}	C_{14}	C_{15}	C_{16}

P

A	Y	N
Y	TP	FN
N	FP	TN

Set-F

1. What are the challenges of R-CNN? How those challenges are solved by Faster R-CNN and YOLO models Explain.

→ R-CNN Challenges:

1. Slow → Uses selective search for region proposals, making it too slow.

2. Redundant Computations → Extracts features separately for each region proposal.

3. High storage → Storing features for each region consuming memory.

4. Complex training → Not end to end, requires multiple stages for training.

Faster R-CNN Solution:

1. Region Proposal Network (RPN) → Replaces selective search, speeding up region proposal generation.

2. Shared feature → Uses one CNN for the whole image reduc redundant calculation.

3. End to End Training → Trains region proposals and detection together, simplifying the process.

YOLO Solution:

1. Unified Model: Predicts both bounding boxes & classes in a single forward pass, no region proposals needed.

2. Real Time speed: Extremely fast, suitable for real time application (up to 45fps).

3. Global Context: looks at the entire image at once, improving detection for large objects.

Set G

1. Using numerous Inception blocks caused GoogleNet to encounter difficulties. What was the problem? How was this problem resolved? Give an example using figure.

Ans:

Problems:

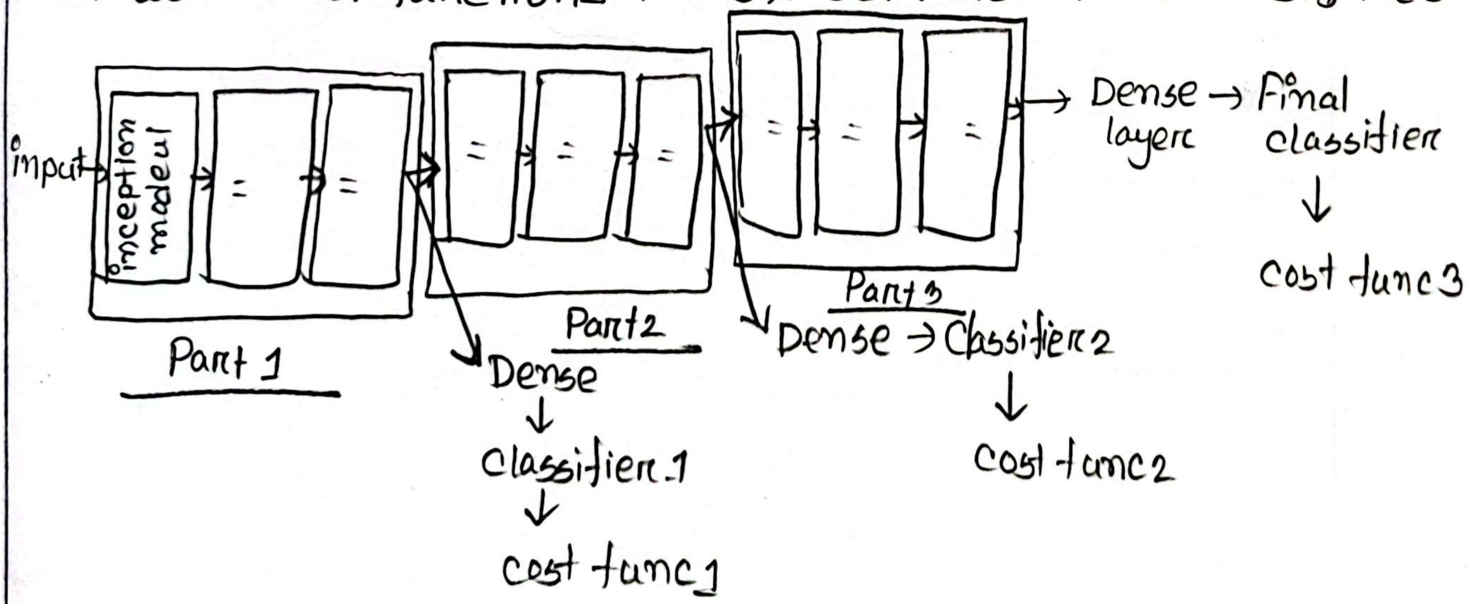
* As depth increased with inception blocks, GoogleNet faced vanishing/exploding gradients.

* Deeper network also tended to overfit or become difficult to train without additional regularization techniques.

Solution:

To tackle this issue, GoogleNet introduced two intermediate loss functions (auxiliary classifiers) at earlier points in the network. These helped propagate stronger gradients to the middle layers. The total loss was calculated as a weighted sum of all losses.

Total loss = Cost function1 + 0.3 x Cost func2 + 0.3 x Cost func3



* The part1 of the network contains three inception moderles, ~~followey~~ followed by Dense layers that connect to classifier 1 cost function 1 is calculated at this point, ensuring that meaningful gradients are backpropagated through the network.

* Similarly Part2 and Part3 have more inception modulos and calculates cost func 2 and 3 where ~~Part2~~ cost func 2 provides further gradient updates for the middle layers and cost fun 3 give the final output.

Thus introducing these auxilary loss functions helped stabilize the training of GoogleNet and improved learning in the middle layeres. addressing vanishing gradient problem effectively.

Quiz-3

Q: The equation below assesses the partition quality in Conceptual-based clustering (COBWEB). Explain the significance of the i, j and k variables as well as the terminology used to formulate the equation.

$$\sum_{k=1}^{n} P(C_k) \sum_{i} \sum_{j} P(A_i = = V_{ij} \mid C_k)^2$$

Ans:

* k refers to the clusters or classes in the dataset. C_k represents the k^{th} cluster.

* P(C_k) (Cluster Probability): represents the prior probability or proportion of data points that belong to cluster C_k. it weighs the contribution of each cluster based on its size or likelihood.

* i (Attributes): represents the index for attributes or feature (height, weight, color) of the objects being clustered.
→ For each cluster, the inner summation over i calculates the contribution of each attribute in determining the cluster structure.

* j (Attribute Value): if attribute A_i is 'color' then j values like 'red', 'blue' or 'green'. The summation ~~of~~ over j captures the contribution of all possible values for a given attribute A_i.

⊛ $P(A_i = V_{ij} | C_k)$ Conditional Probability: represent that attribut A_i takes value V_{ij} given that the object belongs to cluster C_k. It describe how well a specific attribute value fits within the cluster C_k

→ Now the square of C.P is used to emphasize the importance of higher probabilities. This gives more weight to stronger attribute-value association within clusters.

— x —

*** The equation measures partition quality in clustering. It assesses how well the attribute value V_{ij} fit within the clusters by summing the square conditional Probability of attribute values within each cluster, weighted by the cluster probability.

→ The higher this value, the better partitioning as it reflects a stronger association between attribute value and clusers.

2. Investigate the equation and find the use of this equation in the Fuzzy-C mean algorithm for clustering a dataset. Explain the significance of the i, t, j, c & m variables, as well as the terminology used to formulate the equation.

$$A_i^{(t+1)}(x_k) = \left[\sum_{j=1}^{c} \left(\frac{\| x_k - v_i^{(t)} \|^2}{\| x_k - v_j^{(t)} \|^2} \right)^{\frac{1}{m-1}} \right]^{-1}$$

Ans: This equation is from the Fuzzy C-Means (FCM) algorithm, used to calculate the membership value of a data point x_k to a cluster i at iteration $t+1$.

Variable:

→ i: cluster index

→ j: Summation over all clusters. c

→ t: Iteration number

→ c: no of cluster

→ m: Fuzziness parameter controlling the degree of overlap between clusters.

→ x_k: The k-th data point.

→ $v_i^{(t)}$: Center of cluster i at iteration t.

Significance of FCM:

* FCM assigns each data point a membership value in all clusters rather than hard assignment.

* This equation is used to update the membership values iteratively, ensuring that points closer to a cluster center have higher membership values for that cluster, while points farther away have lower membership values.

* repeated until the membership value and cluster center converge, resulting in fuzzy partitions of the dataset.

(A)

(2) Data points $A(2,2)$ $B(3,2)$ $C(1,1)$ $D(3,1)$ $E(1.5, 0.5)$

Forgy's: Step1: Let's A & B two centroid of two cluster

data point	Distance from A $\lvert x_2-x_1\rvert + \lvert y_2-y_1\rvert$	Distance from B $\lvert x_2-x_1\rvert + \lvert y_2-y_1\rvert$	Cluster
A(2,2)	$\lvert 2-2\rvert+\lvert 2-2\rvert = 0$ ✓	$\lvert 3-2\rvert+\lvert 2-2\rvert = 1$	A(2,2)
B(3,2)	$\lvert 3-2\rvert+\lvert 2-2\rvert = 1$	$\lvert 3-3\rvert+\lvert 2-2\rvert = 0$ ✓	B(3,2)
C(1,1)	$\lvert 2-1\rvert+\lvert 2-1\rvert = 2$ ✓	$\lvert 3-1\rvert+\lvert 2-1\rvert = 3$	A(2,2)
D(3,1)	$\lvert 3-2\rvert+\lvert 1-2\rvert = 2$	$\lvert 3-3\rvert+\lvert 1-2\rvert = 1$ ✓	B(3,2)
E(1.5,0.5)	$\lvert 2-1.5\rvert+\lvert 2-0.5\rvert = 2$ ✓	$\lvert 3-1.5\rvert+\lvert 2-0.5\rvert = 3$	A(2,2)

{A, C, E} {B, D}

new center $C_1 = \left(\frac{2+1+1.5}{3}, \frac{2+1+0.5}{3}\right) = (1.5, 1.167)$

$C_2 = \left(\frac{3+3}{2}, \frac{2+1}{2}\right) = (3, 1.5)$

Step2:

	Distance from C_1 (1.5, 1.167)	D from C_2 (3, 1.5)	Cluster
A(2,2)	$\lvert 2-1.5\rvert+\lvert 2-1.167\rvert = 1.33$	$\lvert 2-3\rvert+\lvert 2-1.5\rvert = 1.5$	A(2,2)
B(3,2)	$\lvert 3-1.5\rvert+\lvert 2-1.167\rvert = 2.33$	$\lvert 3-3\rvert+\lvert 2-1.5\rvert = 0.5$	B(3,2)
C(1,1)	$\lvert 1-1.5\rvert+\lvert 1-1.167\rvert = 0.667$	$\lvert 1-3\rvert+\lvert 1-1.5\rvert = 2.5$	A(2,2)
D(3,1)	$\lvert 3-1.5\rvert+\lvert 1-1.167\rvert = 1.667$	$\lvert 3-3\rvert+\lvert 1-1.5\rvert = 0.5$	B(3,2)
E(1.5,0.5)	$\lvert 1.5-1.5\rvert+\lvert 0.5-1.167\rvert = 0.667$	$\lvert 1.5-3\rvert+\lvert 0.5-1.5\rvert = 2.5$	A(2,2)

After step 2 the cluster doesnot change so stop here.

$C_1 = \{A, C, E\}$ $C_2 = \{B, D\}$

Cluster C_1	Cluster C_2
A(2,2)	B(3,2)
C(1,1)	D(3,1)
E(1.5,0.5)	

Step 3: Intra-cluster Diameter / distance C_1

manhattan (as given question)

Distance of $D_{AC} = |2-1| + |2-1| = 1+1 = 2$

$D_{AE} = |2-1.5| + |2-0.5| = 0.5+1.5 = 2$

$D_{CE} = |1-1.5| + |1-0.5| = 0.5+0.5 = 1.$

$max(D_{AC}, D_{AE}, D_{CE}) = 2$

$C_2 \rightarrow D_{BD} = |3-3| + |2-1| = 0+1 = 1.$

$max(D_{BD}) = 1.$

max 2

Step 4: Inter-cluster Distance.

Cluster $C1$: A(2,2) C(1,1) E(1.5,0.5)

$C2$: B(3,2) D(3,1).

$A < \begin{matrix} B \\ D \end{matrix}$ $C < \begin{matrix} B \\ D \end{matrix}$, $E < \begin{matrix} B \\ D \end{matrix}$

$D_{AB} = |2-3| + |2-2| = 1$

$D_{AD} = |2-3| + |2-1| = 2$

$D_{CB} = |1-3| + |1-2| = 3$

$D_{CD} = |1-3| + |1-1| = 2$

$D_{EB} = |1.5-3| + |0.5-2| = 3$

$D_{ED} = |1.5-3| + |0.5-1| = 2$

$\min(D_{AB}, D_{AD}, D_{CB}, D_{CD}, D_{EB}, D_{ED}) = 1$

$$\text{Dunn Index} = \frac{\min_{0 \le i \le n_c,\, 0 \le j < n_c,\, i \ne j} (d(c_i, c_j))}{\max_{0 \le k < n_c} (\text{diam}(c_i))}$$

$$= \frac{1}{2} = 0.5$$

Ans:

For Davies-Bouldin Index

Step 1

$$S_i = \frac{1}{n_i - 1} \sum_{x \in c_i} d(x, v_i)^2 \qquad V_i = 1.5, 1.167 \text{ (mean ..)}$$

For cluster C_1: $S_1 = \frac{1}{3-1}\Big[(|2-1.5| + |2-1.167|)^2 + (|1-1.5| + |1-1.167|)^2 + (|1.5-1.5| + |0.5-1.167|)^2\Big]$

$= \frac{1}{2} \times 2.667$

$= 1.334$

For Cluster 2 $v_j = (3, 1.5)$

$$S_2 = \frac{1}{2-1}\left[(3-3+2-1.5)^2 + (3-3+1-1.5)^2\right]$$

$$= 1 \times 0.5$$

Step2

$d_{ij} = d(v_i, v_j)$ distance between two centroid.

$$d_{ij} = \sqrt{(3-1.5)^2 + (1.167-1.5)^2}$$

$v_i = (1.5, 1.167)$

$v_j = (3, 1.5)$

$$= 1.537$$

Step3

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} = \frac{1.334+0.5}{1.537} = 1.1932$$

Step 4

$$\text{Davies Bouldin Index} = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i$$

For 2 cluster $R_{ij} = R_i$

$$= \frac{1}{2} \times 1.1932$$

$$= 0.5966$$

Ans:

*** If there are 3 cluster, $\frac{2+1+1.5}{3}, \frac{2+1+0.5}{3}$

Suppose, $C_1 = A(2,2)\ C(1,1)\ E(1.5, 0.5)$; $V_1 = (1.5, 1.167)$

$C_2 = B(3,2)\ D(3,1)$; $V_2 = (3, 1.5)$

$C_3 = F(5,3)\ G(6,3)$; $V_3 = (5.5, 3)$

(1)

$$S_1 = \frac{1}{3-1}\left[\ldots\ldots \right] = 1.334 \text{ (same as previous)}$$

$$S_2 = 0.5$$

$$S_3 = \frac{1}{2-1}\left[(5-5.5+3-3)^2 + (6-5.5+3-3)^2 \right]$$

$$= 0.5$$

(2)

$$d_{ij} = d(v_i, v_j)$$

$$d_{12} = \sqrt{(3-1.5)^2 + (1.5-1.167)^2} = 1.537$$

$$d_{13} = \sqrt{(1.5-5.5)^2 + (1.167-3)^2} = 4.399$$

$$d_{23} = \sqrt{(3-5.5)^2 + (1.5-3)^2} = 2.92$$

(3)

For Cluster C_1:

$$R_1 = \max\left(\frac{S_1+S_2}{d_{12}}, \frac{S_1+S_3}{d_{13}}\right) = \max\left(\frac{1.334+0.5}{1.537}, \frac{1.334+0.5}{4.399}\right)$$

$$= \max(1.1932, 0.417)$$

$$= 1.1932$$

For cluster C2:

$$R_2 = \max\left(\frac{s_1+s_2}{d_{12}}, \frac{s_2+s_3}{d_{23}}\right) = \max\left(\frac{1.334+0.5}{1.537}, \frac{0.5+0.5}{2.92}\right)$$

$$= \max(1.1932, 0.34)$$

$$= 1.1932$$

For cluster C3:

$$R_3 = \max\left(\frac{s_1+s_3}{d_{13}}, \frac{s_3+s_2}{d_{23}}\right) = \max(0.417, 0.34)$$

$$= 0.417$$

4 $$DBI = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i$$

$$= \frac{1}{3}(R_1+R_2+R_3)$$

$$= \frac{1}{3}(1.1932+1.1932+0.417)$$

$$= 0.934$$

Ans:

AUST	Ahsanullah University of Science and Technology		
	Department of Computer Science and Engineering		
	SET A, Class Test #3, Fall 2023		
	Course Code: CSE 4261	Course Title: Data Analytics	
	Time: 20 Minutes	Date:01[illegible]/2024	Full Marks: 10
ID: 20200104023		Name: Bidyarthi Paul	

1. Interpret the followingAutocorrelation and Partial Autocorrelation Functions(Figure 1 and Figure 2) and decide the optimum "p" value for an Autoregression (AR) Model for Figure 1 and the optimum "q" value for aMoving Average (MA) Model for Figure 2.

AR

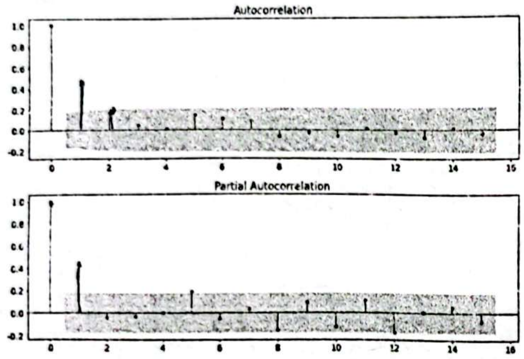


Figure 1: ACF and PACF figures for the AR model

MA

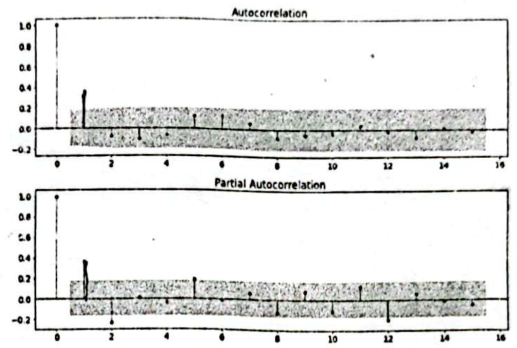
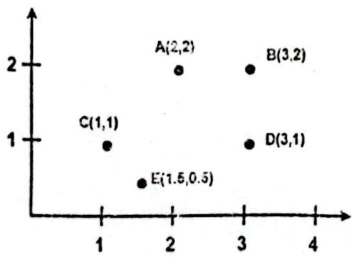


Figure 2: ACF and PACF figures for the MA model

2. Use Forgy's and Kmean's algorithms to create two clusters on the following datasets. The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-D (a, b) = |x2 − x1| + |y2 − y1|. Use the Dunn Index to compare the performance of the two clustering algorithms.



Ans. to the Q.NO.01

From Figure-1;

~~The~~ AR (Autoregression) model uses Partial Autocorrelation (PACF) because AR model predicts the future values from it's past values. And in PACF, The directly or indirect influence of the previous values is taken into consideration for the prediction of values. So, we will use the PACF correlation for the AR model. From the figure we can observe that, only one lag goes beyond the threshold. So, only one ~~Based~~ previous value will have influence on the prediction. AR (P) = AR(1)

CamScanner

From Figure-2;

MA (moving average) uses Autocorrelation because Autocorrelation uses the previous value's error. In ACF, only direct influence is taken into consideration. No indirect values are taken for ACF. So, for the error we need only the ~~pred~~ previous values. So, MA uses ACF.

From the ACF of figure-2; only one lag goes beyond the threshold. $\therefore MA(q) = \cancel{(2)}\,(1)$

1/July/24

Ahsanullah University of Science and Technology		
Department of Computer Science and Engineering		
SET B, Class Test #3, Fall 2023		
Course Code: CSE 4261	Course Title: Data Analytics	
Time: 20 Minutes	Date: 01/6/2024	Full Marks: 10
ID: 20200104038	Name: Jalisha Jashim Era	

1. Interpret the following Autocorrelation and Partial Autocorrelation Functions (Figure 1 and Figure 2) and decide the optimum "p" value for an Autoregression (AR) Model for Figure 1 and the optimum "q" value for a Moving Average (MA) Model for Figure 2.

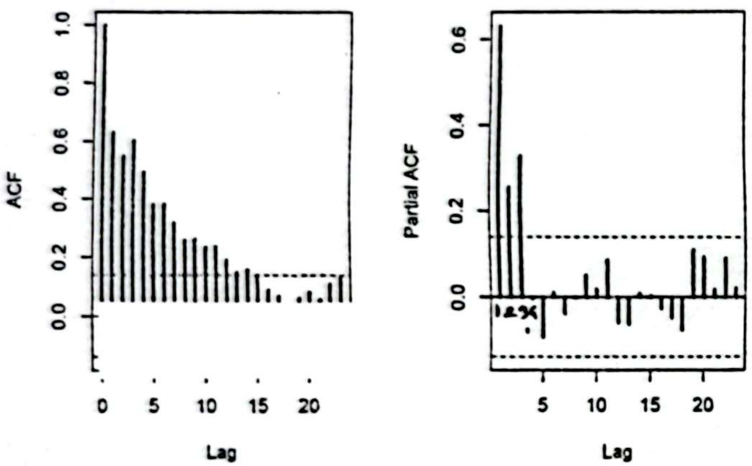


Figure 1: ACF and PACF figures for the AR model

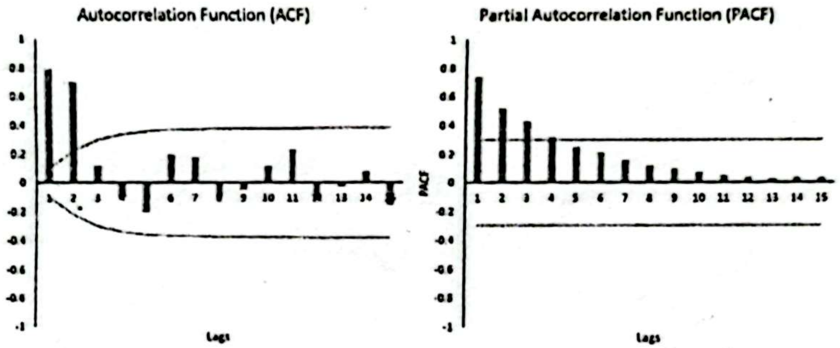


Figure 2: ACF and PACF figures for the MA model

Roll No. .... Course No. .... Course Title ....

Ans: to the Question NO: 1

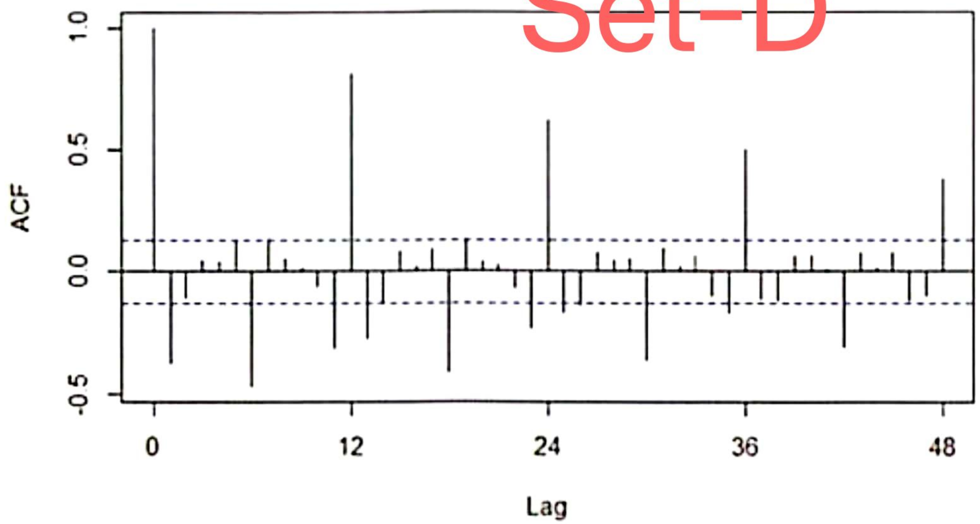
In first Figure-1 ACF vs Lag Here it is noticeable that trend is tailed off gradually with lags while PACF vs lag showing spike along 1, 2, 3 logs and rest of the values in the ~~to~~ confidence level means after lag 3 PACF cuts off rest of them ~~zo~~ close to zero. So Here Autoregression model's order will be 3 AR(3) ~~lo~~ and for PACF perfectly use in the AR (3) p=3 model.

In 2nd figure-2 ACF vs lag ~~6~~ is representing long spike on lags 1, 2 and rest of them are close to zero means ACF cuts off whereas PACF vs Lags showing tailed off gradually so it will not perfectly considarable for moving average model where previous errors are considered for calculation and ACF shows the best ~~p~~ fit for MA (q) ~~o~~ here q = 2 MA(2)

The dashed lines provide upper and lower bounds at a 95% significance level. Any value of the ACF or PACF outside of these bounds indicates that the value is significantly different from zero.

- ACF of the differenced gasoline time series

Set-D

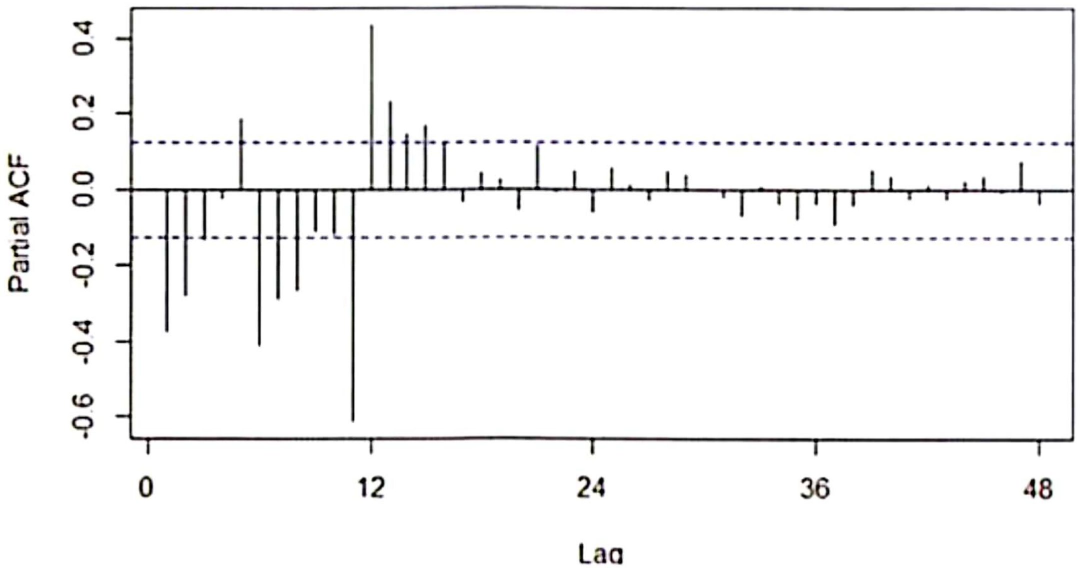


- The figure shows several significant ACF values. The slowly decaying ACF values at lags 12, 24, 36, and 48 are of particular interest.
- The figure indicates a seasonal autoregressive pattern every 12 months

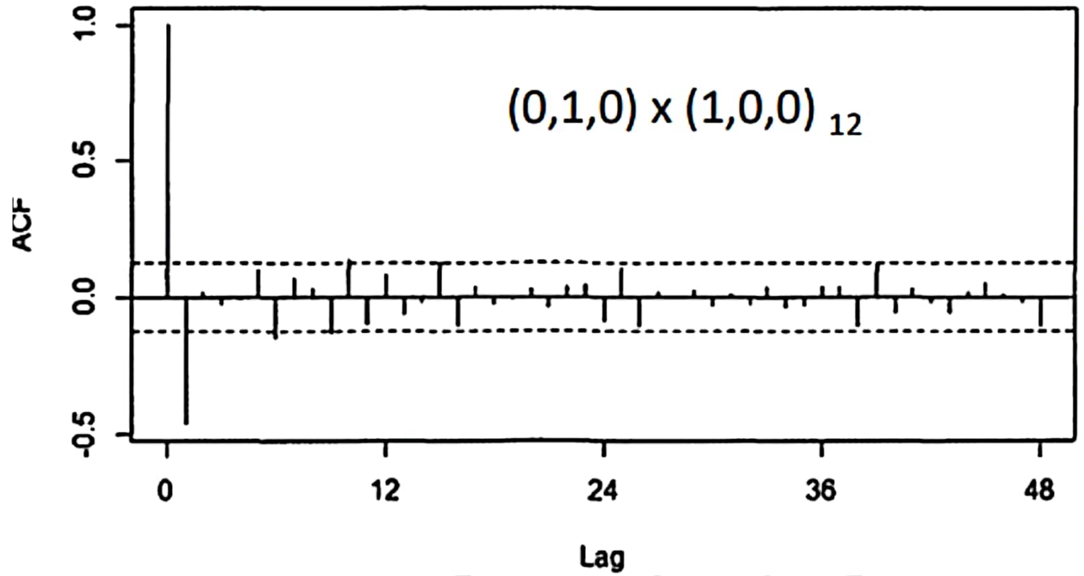
- PACF of the differenced gasoline time series

- Examining the PACF plot in the figure, the PACF value at lag 12 is quite large, but the PACF values are close to zero at lags 24, 36, and 48.

- Thus, a seasonal AR(1) model with period = 12 will be considered.

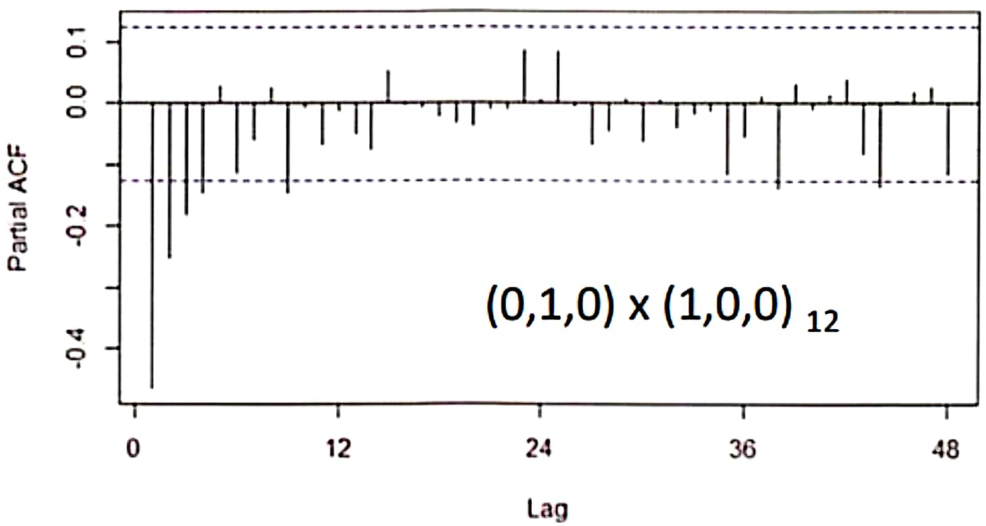


- ACF of residuals from seasonal AR(1) model



The ACF plot of the residuals in Figure indicates that the autoregressive behavior at lags 12, 24, 26, and 48 has been addressed by the seasonal AR(1) term. The only remaining ACF value of any significance occurs at lag 1.

- PACF of residuals from seasonal AR(1) model



The figure shows several significant PACF values at lags 1, 2, 3, and 4. The PACF plot in the figure exhibits a slowly decaying PACF, and the ACF cuts off sharply at lag 1.

MA(1) model should be considered for the nonseasonal portion of the ARMA model on the differenced series. In other words, a (0,1,1) x $(1,0,0)_{12}$ ARIMA model will be fitted to the original gasoline production time series.

Quiz-2

Set A

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})\times(y_i-\bar{y})$	$\hat{y}=A+Bx_i$	$y_i-\hat{y}$	$(y_i-\hat{y})^2$
55	4040	-941	-146918	885481	21584898724	138249838	62693.52	-58653.52	3440235367
77	120120	-919	-30838	844561	950982244	28340122	64757.089	55362.91	3065051937
1212	180180	216	29222	46656	853925284	6311952	171218.497	8961.50	80308534.5
1616	210210	620	59252	384400	3510793504	36736240	209113.13	1096.87	1203122.66
2020	240240	1024	89282	1048576	7971275524	91424768	247007.764	-6767.76	45802628.98
$\bar{x}=$ 996	$\bar{y}=$ 150958			$=3209674$	$=6974376256$	$=301062920$			1326520318

SSX, SST, SSXY, SSE

Now, $B=\frac{SSxy}{SSx}=\frac{301062920}{3209674}=93.799$

$A=\bar{y}-B\bar{x}$

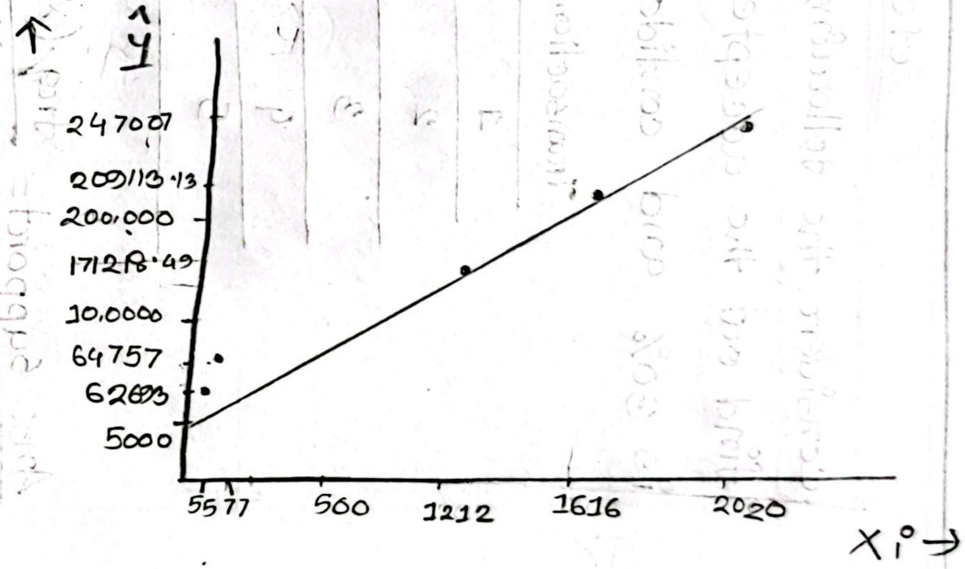
$=150958-(93.799)\times 996$

$=57534.196$

$R^2=1-\frac{SSE}{SST}$

$=1-\frac{1326520318}{6974376256}$

$=0.8098$



Set C

Consider the following dataset with transaction and find out the accepted rules. The minimum support is 30% and confidence is 50%

Transaction	items
1	{A, C, D}
2	{B, C, D}
3	{A, B, C, D}
4	{B, D}
5	{A, B, C, D}

Ans: Support = $\frac{frq(X,Y)}{N}$ N → no of transaction.

Confidence = $\frac{frq(X,Y)}{frq(X)}$

$\geq 30\%$

itemset	frequency	Support(%)
A	3	60%
B	4	80%
C	4	80% ~~100%~~
D	5	100%
{A,B}	2	40%
{A,C}	3	60%
{A,D}	3	60%
{B,C}	3	60%
{C,D}	4	80%
{A,B,C}	2	40%
{A,C,D}	3	60%
{A,B,D}	2	40%
{B,C,D}	3	60%
{A,B,C,D}	2	40%
{B,D}	4	80%

confidence if A is there C will be there $A \to C$

$$\frac{frq(x,y)}{f(x)} = \frac{3}{3} = 100\% \text{ or } \frac{Support(A,C)}{Support(A)} = \frac{60\%}{60\%} = 100\%$$

and greater than 50%

$\{A,B,C\} = \{A,B \to C\}$ or $\{A \to B,C\}$

$$\frac{40\%}{40\%} = 100\% \qquad \frac{40\%}{60\%} = 66.67\%$$

Accepted.

এই math huge বড় হলেও সব combination দেখতে হবে। so normally একটা-২টা দেখালেই হবে ig.

Quiz-1 (Just math)

② A sample of 400 male students is found to have a mean height of 67.47 inches. Can it be reasonably regarded as a sample from a large population with a mean height of 67.39 inches and a standard deviation of 1.30 inches? Test at a 5% level of significance.

z-value

95% →	one tailed :	1.645
	two " :	1.96
99% →	one tailed :	2.33
	two " :	2.58
90% →	one tailed :	1.28
	two tailed :	1.645

Ans:

Given,

no of sample, $n = 400$

mean, $\bar{x} = 67.47$

Standard deviation, $\sigma = 1.30$

level of significance, $\alpha = 5\% = 0.05$

Confidence Interval, CI = 95%

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

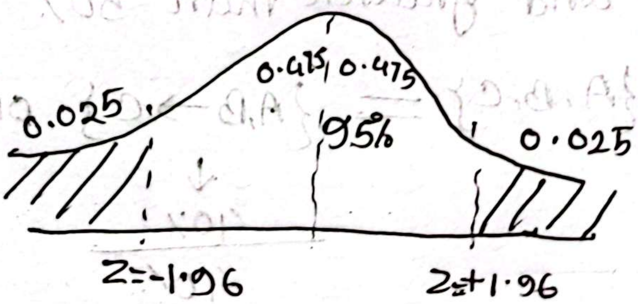
Sample population mean height, $\mu = 67.39$ inch.

$H_0 : \mu = 67.39 \quad H_1 : \mu \neq 67.39$

$$Z_c = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{67.47 - 67.39}{1.30/\sqrt{400}}$$

$= 1.230 < 1.96$ [falls between the interval]



So the null hypothesis is accepted.

<u>Set B</u>

From a random sample of 36 New York civil service personnel, the mean age and the sample standard deviation were found to be 40 years and 4.5 years respectively. Construct a 95% confidence interval for the mean age of civil servants in New York.

Given,

no of sample, $n = 36$; degree of freedom, $DF = n-1 = 35$

mean, $\bar{x} = 40$

S.D from sample, $s = 4.5$ [when, from previous studies σ is unknown]

Confidence Interval, $CI = 95\%$

Level of significance, $\alpha = 5\% = 0.05$

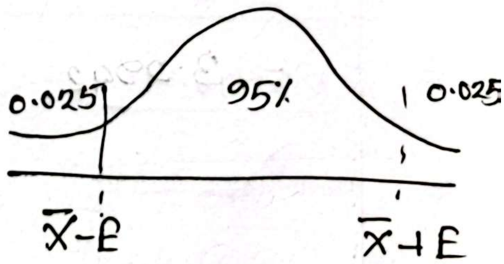
$\alpha/2 = 0.025$

*** For DF 35 and 0.05 (two tail) no t value is present in the table provided in slide.

$E = t \frac{s}{\sqrt{n}}$ just Assume $t = 2.045$

$= 2.045 \frac{4.5}{\sqrt{36}}$

$= 1.533$



$40 - 1.533 < \mu < 40 + 1.533$

$\Rightarrow 38.467 < \mu < 41.533$

Ans

CS CamScanner

Set C

The Foreman of ABC mining Company has estimated the average quantity of iron ore extracted to be 36.8 tons per shift and the sample standard deviation to be 2.8 tons per shift. based upon a random selection of 4 shifts. Construct a 90% percent confidence interval around this estimation

Given.

sample, $n = 4$ $\quad df = 3$

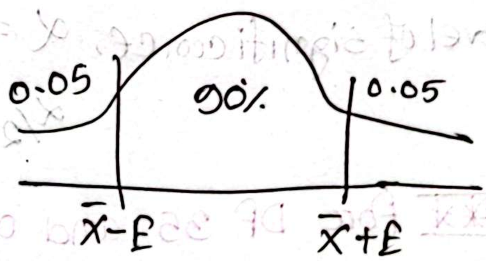
mean, $\bar{x} = 36.8$

SD $\quad s = 2.8$

CI = 90%

$\alpha = 10\% = 0.1 \quad \alpha/2 = 0.05$

$t = 2.353$



$E = t \frac{s}{\sqrt{n}}$

$= \frac{2.353 \times 2.8}{\sqrt{4}}$

$= 3.2942$

$36.8 - 3.2942 < \mu < 36.8 + 3.2942$

$\Rightarrow 33.5058 < \mu < 40.0942$

Set D

The following nine observations were drawn from a normal population: 27 19 20 24 23 29 21 17 27

test the null hypothesis: $H_0: \mu = 26$ against the alternatives hypothesis $H_a: \mu \neq 26$. At what level of significance can H_0 be rejected?

Ans:

$n = 9 < 30$

$\bar{x} = \frac{1}{9}(27+19+20+24+23+29+21+17+27) = 23.$

$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} = 4.09$

$H_0 : \mu = 26$

$H_1 : \mu \neq 26$

$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{23-26}{4.09/\sqrt{9}} = -2.2004.$

For 90% Confidence,

if t value is $t > 1.86$ & $t < -1.89 \rightarrow H_0$ will be rejected.

So $t = -2.2004 < -1.89 \rightarrow$ reject H_0

For 95% Confidence,

if $t > 2.306$ & $t < -2.306 \rightarrow H_0$ will be rejected.

So $t = -2.2004 > -2.306 \rightarrow$ Accept H_0.

For 99%

if $-2.896 < t < 2.896 \rightarrow H_0$ Accepted.

So $t = -2.2004$ is Accepted.

So, H_0 will be rejected only for 90% CI.

Set E

A 10 years old survey of CPAs (Certificate Public Accountants) in the USA found that their average salary was \$60,014. An accounting researcher would like to test whether this average has increased over the years. A sample of 125 CPAs produced a mean salary of \$68,695. Also given that the population standard deviation σ = \$10,530. level of significance = ~~10~~ 1%

Ans: Given,

$n = 125$

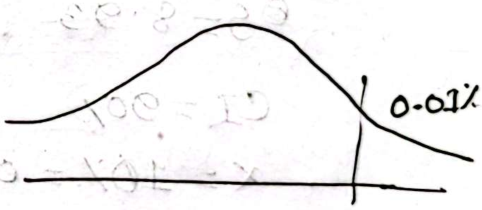
$\mu = 60{,}014$

$\sigma = 10530$

$\bar{X} = 68{,}695$

$H_0: \mu \leq 60014$

$H_1: \mu > 60014$



$\alpha = 1\% = 0.01 \quad \alpha/2 = 0.005$

$z = 2.58$

$$Z_c = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{68{,}695 - 60{,}014}{\frac{10530}{\sqrt{125}}}$$

$$= 9.21 > z$$

Ho null hypothesis is rejected. So The average salary has increased.

Self

A researcher collects 25 examples of products and determines the the percent salt in each. These 25 examples give a sample mean salt content of 40.24 and a sample standard deviation of 8.93. Compute a 90% confidence interval estimate of the true variance of the percentage water for this new process.

Ans:

$n = 25 \quad df = 24$

$\bar{x} = 40.24$

$s = 8.93$

$CI = 90\%$

$\alpha = 10\% = 0.1$

$\alpha/2 = 0.05$

$1 - \alpha/2 = 1 - 0.05$
$= 0.95$

From table (Sir's solve)

$\chi^2_{\alpha/2} = \chi^2_{0.05} = 36.416$

$\chi^2_{1-\alpha/2} = \chi^2_{0.95} = 13.848$

Estimate the interval: Since $n = 25 < 30$ so use,

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$$\Rightarrow \frac{24 \times (8.93)^2}{36.416} < \sigma^2 < \frac{24 \times (8.93)^2}{13.848}$$

$$\Rightarrow 52.56 < \sigma^2 < 138.20$$

A

Fuzzy C-Means Clustering

Step1: Randomly Given Initial membership(fuzzy) values to the Datapoints for each cluster. Sum of the cluster values for each Data points should be 1.

Samples (1,3) (2,5) (4,8) (7,9)

Cluster	{1,3}	{2,5}	{4,8}	{7,9}
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

Step2:

$$V_{ij} = \frac{\sum_{k=1}^{n} [A_i(x_k)]^m x_k}{\sum_{k=1}^{n} [A_i(x_k)]^m}$$

m = fuzziness parameter 2 (always)

$A_i(x_k)$ = membership value of k datapoints.

$$V_{11} = \frac{(0.8^2 \times 1) + (0.7^2 \times 2) + (0.2^2 \times 4) + (0.1^2 \times 7)}{0.8^2 + 0.7^2 + 0.2^2 + 0.1^2} = 1.568$$

$$V_{12} = \frac{(0.8^2 \times 3) + (0.7^2 \times 5) + (0.2^2 \times 8) + (0.1^2 \times 9)}{0.8^2 + 0.7^2 + 0.2^2 + 0.1^2} = 4.051$$

$$V_{21} = \frac{(0.2^2 \times 1) + (0.3^2 \times 2) + (0.8^2 \times 4) + (0.9^2 \times 7)}{0.2^2 + 0.3^2 + 0.8^2 + 0.9^2} = 5.35$$

$$V_{22} = \frac{(0.2^2 \times 3) + (0.3^2 \times 5) + (0.8^2 \times 8) + (0.9^2 \times 9)}{0.2^2 + 0.3^2 + 0.8^2 + 0.9^2} = 8.215$$

Now centroids are $\{1.568, 4.051\}$ $\{5.35, 8.215\}$.

Cluster datapoint	Distance from ① $(1.568, 4.051)$	② $\Rightarrow (5.35, 8.215)$	Cluster
$\{1,3\}$	$D_{11} = \sqrt{(1-1.568)^2+(3-4.051)^2} = 1.2$	$D_{12} = \sqrt{(1-5.35)^2+(3-8.215)^2} = 6.79$	1
$\{2,5\}$	$D_{21} = \sqrt{(2-1.568)^2+(5-4.051)^2} = 1.04$	$D_{22} = \sqrt{(2-5.35)^2+(3-8.215)^2} = 4.54$	1
$\{4,8\}$	$D_{31} = 4.63$	$D_{32} = 1.36$	2
$\{7,9\}$	$D_{41} = 7.34$	$D_{42} = 1.82$	2

Step 3

$$A_i^{(t+1)} x_t = \left[\sum_{j=1}^{c}\left(\frac{\|x_k - v_i^{(t)}\|^2}{\|x_k - v_j^{(t)}\|^2}\right)^{\frac{1}{m-1}}\right]^{-1} \qquad \frac{1}{m-1} = \frac{1}{2-1} = \frac{1}{1} = 1.$$

$$A_{11} = \left[\left(\frac{D_{11}^2}{D_{11}^2}\right)^{\frac{1}{2-1}} + \left(\frac{D_{11}^2}{D_{12}^2}\right)^{\frac{1}{2-1}}\right]^{-1} = 0.97$$

$$A_{12} = \left[\left(D_{12}^2/D_{11}^2\right)^1 + \left(D_{12}^2/D_{12}^2\right)^1\right]^{-1} = 0.03$$

$$A_{21} = \left[\left(D_{21}^2/D_{21}^2\right)^1 + \left(D_{21}^2/D_{22}^2\right)^1\right]^{-1} = 0.95$$

$$A_{22} = \left[\left(D_{22}^2/D_{21}^2\right)^1 + \left(D_{22}^2/D_{22}^2\right)^1\right]^{-1} = 0.05.$$

$$A_{31} = \left[\left(D_{31}^2/D_{31}^2\right)^1 + \left(D_{31}^2/D_{32}^2\right)^1\right]^{-1} = 0.08$$

$$A_{32} = \left[\left(D_{32}^2/D_{31}^2\right)^1 + \left(D_{32}^2/D_{32}^2\right)^1\right]^{-1} = 0.92$$

$$A_{41} = \left[\left(D_{41}^2/D_{42}^2\right)^1 + \left(D_{41}^2/D_{42}^2\right)^1\right]^{-1} = 0.06$$

$$A_{42} = \left[\left(D_{42}^2/D_{41}^2\right)^1 + \left(D_{42}^2/D_{42}^2\right)^1\right]^{-1} = 0.94$$

CamScanner

Regression

Linear regression → relationship between two variables is called a linear regression model.

$y = A + Bx$

- x → independent
- B → slope
- A → y-intercept, Constant
- y → Dependent variable.

$y = A + Bx + e$ (e → error.)

Direct Regression: least Square Error

$$S(A,B) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - A - Bx_i)^2$$

Partial Derivatives with respect to A:

$$\frac{\partial S(A,B)}{\partial A} = -2\sum_{i=1}^{n} (y_i - A - Bx_i)$$

$$\left[\because \frac{\partial S(A,B)}{\partial A} = 0\right]$$

$$\Rightarrow -2\sum_{i=1}^{n} (y_i - A - Bx_i) = 0$$

$$\Rightarrow \bar{y} - A - B\bar{x} = 0$$

$$\Rightarrow \boxed{A = \bar{y} - B\bar{x}}$$

→ Estimation of A.

1 x_i	2 y_i	3 $x_i - \bar{x}$	4 $y_i - \bar{y}$	5 $(x_i - \bar{x})^2$	6 $(x_i - \bar{x})(y_i - \bar{y})$

Again, Partial Derivatives with respect to B.

$$\frac{\partial S(A,B)}{\partial B} = -2\sum_{i=1}^{n}(y_i - A - Bx_i)\,x_i$$

$$\Rightarrow -2\sum_{i=1}^{n}(y_i - A - Bx_i)\,x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} A x_i - \sum_{i=1}^{n} B x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}(\bar{y} - B\bar{x})\,x_i - \sum_{i=1}^{n} B x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y} + \sum_{i=1}^{n} B\bar{x}x_i - \sum_{i=1}^{n} B x_i^2 = 0$$

$$\Rightarrow B\left(\sum_{i=1}^{n}\bar{x}x_i - \sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} x_i\bar{y} - \sum_{i=1}^{n} x_i y_i$$

$$\Rightarrow B = \frac{-\left[\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y}\right]}{-\left[\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n}\bar{x}x_i\right]}$$ [minus common নিয়েছি]

$$= \frac{x_i\left[\sum_{i=1}^{n} y_i - \bar{y}\right]}{x_i\left[\sum_{i=1}^{n} x_i - \bar{x}\right]}$$

$$= \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}$$

So the Ordinary least squares (OLS) estimators of A and B.

$$A = \bar{y} - B\bar{x}$$

$$B = \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Residuals: are the differences between the observed values and the predicted values from a regression model. They provide insight into how well the model captures the data.

Assessing residuals helps determine the goodness of fit of the model.

Total Variability:

Total Sum of Square (SST) = $\sum_{i=1}^{N} (y_i - \bar{y})^2$

$\bar{y}$ → mean of N data observation

y_i ↳ observe value.

Explained Variability:

Regression Sum of Square (SSR) = $\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$

$\hat{y}_i$ ↳ Predicted Value from the model.

UnExplained Variability:

Error Sum of Square (SSE) = $\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

1. Total Variability (SST) = $\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$

2. Partitioning SST = $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$

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

$$\boxed{SST = SSE + SSR}$$

SST → overall measure of variability in the data.

SSE → how much can't explain model (Error)

SSR → how much can explain.

By analyzing these components, we can assess how well the regression model fits the data. A good fit implies that SSR is large (explained variability is high), and SSE is small (unexplained is low). This helps us to understand the effectiveness of the model in capturing the underlying pattern in the data.

$$R^2 = \frac{SSR}{SST}$$

$$= 1 - \frac{SSE}{SST}$$

SSR → close to SST R^2 becomes close to 1 good fit model.

SSE → close to SST R^2 close to 0 not good fit model

Ridge Regression: (L2 regularization): adds a penalty proportional to the square of the coefficients, which helps to shrink the coefficients and mitigate multicollinearity by reducing their variance.

$$\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}\beta_j^2$$

Lasso Regression: (L1 regularization) adds a penalty proportional to the absulate value of the coefficients which not only helps to sherink the coefficients but can also set some of them to zero, effectively performing feature selection.

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}|\beta_j|$$

→ coefficients of the models.

→ regularization parameter that controls the strength of the penalty

~~Regression~~

Regularizations are essential for creating robust regression models, especially when dealing with highly correlated features.