Q.1. Apply single linkage clustering and draw dendrogram on the following data. Suppose we have a six objects (with name A, B, C, D, E, F) and each object has two measured features ($x_1$ and $x_2$)

| | $x_1$ | $x_2$ |
|---|---|---|
| A | 1 | 1 |
| B | 1.5 | 1.5 |
| C | 5 | 5 |
| D | 3 | 4 |
| E | 4 | 4 |
| F | 3 | 3.5 |

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 0.71 | 5.66 | 3.61 | 4.24 | 3.20 |
| B | 0.71 | 0 | 4.95 | 2.92 | 3.54 | 2.50 |
| C | 5.66 | 4.95 | 0 | 2.24 | 1.41 | 2.50 |
| D | 3.61 | 2.92 | 2.24 | 0 | 1.0 | 0.50 |
| E | 4.24 | 3.54 | 1.41 | 1.0 | 0 | 1.12 |
| F | 3.20 | 2.50 | 2.50 | 0.50 | 1.12 | 0 |

$d(A,B) = \sqrt{(1.5-1)^2+(1.5-1)^2} = 0.71$

$d(A,C) = \sqrt{(5-1)^2+(5-1)^2} = 5.66$

$d(A,D) = \sqrt{(3-1)^2+(4-1)^2} = 3.60$

$d(A,E) = \sqrt{(4-1)^2+(4-1)^2} = 4.24$

$d(A,F) = \sqrt{(3-1)^2+(3.5-1)^2} = 3.20$

$d(B,C) = \sqrt{(5-1.5)^2+(5-1.5)^2} = 4.95$

$d(B,D) = \sqrt{(3-1.5)^2+(4-1.5)^2} = 2.92$

$d(B,E) = \sqrt{(4-1.5)^2 + (4-1.5)^2} = 3.54$

$d(B,F) = \sqrt{(3-1.5)^2 + (3.5-1.5)^2} = 2.50$

$d(C,D) = \sqrt{(3-5)^2 + (4-5)^2} = 2.24$

$d(C,E) = \sqrt{(4-5)^2 + (4-5)^2} = 1.41$

$d(C,F) = \sqrt{(3-5)^2 + (3.5-5)^2} = 2.50$

$d(D,E) = \sqrt{(4-3)^2 + (4-4)^2} = 1.$

$d(D,F) = \sqrt{(3-3)^2 + (3.5-4)^2} = 0.50$

$d(E,F) = \sqrt{(3-4)^2 + (3.5-4)^2} = 1.12$

$d(B,B) = 0$

$d(A,A) = 0$

$d(C,C) = 0$

$d(D,D) = 0$

$d(E,E) = 0$

$d(F,F) = 0$

min Distance (Single Linkage)

Dist

| | A | B | C | D,F | E |
|---|---|---|---|---|---|
| A | 0 | 0.71 | 5.66 | ? | 4.24 |
| B | 0.71 | 0 | 4.95 | ? | 3.54 |
| C | 5.66 | 4.95 | 0 | ? | 1.41 |
| D,F | ? | ? | ? | 0 | ? |
| E | 4.24 | 3.54 | 1.41 | ? | 0 |

$d(D,f) \to A = \min(d_{DA}, d_{FA}) = \min(3.61, 3.20) = 3.20$
$d(D,f) \to B = \min(d_{DB}, d_{FB}) = \min(2.92, 2.50) = 2.50$
$d(D,f) \to C = \min(d_{DB}, d_{FB}) = \min(2.24, 2.50) = 2.24$
$d_E \to (D,f) = \min(d_{ED}, d_{Ef}) = \min(1, 1.12) = 1$

Updated distance matrix becomes

| Dist | A | B | C | (D,f) | E |
|---|---|---|---|---|---|
| A | 0 | 0.71 | 5.66 | 3.20 | 4.24 |
| B | 0.71 | 0 | 4.95 | 2.50 | 3.54 |
| C | 5.66 | 4.95 | 0 | 2.24 | 1.41 |
| D,f | 3.20 | 2.56 | 2.24 | 0 | 1 |
| E | 4.24 | 3.54 | 1.41 | 1 | 0 |

| Dist | A,B | C | (D,E) | E |
|---|---|---|---|---|
| A,B | 0 | ? | ? | ? |
| C | ? | 0 | 2.24 | 1.41 |
| (D,f) | ? | 2.24 | 0 | 1 |
| E | ? | 1.41 | 1 | 0 |

$d_C \to (A,B) = \min(d_{CA}, d_{CB}) = \min(5.66, 4.95) = 4.95$
$d_{(D,f)} \to (A,B) = \min(d_{DA}, d_{DB}, d_{FA}, d_{FB}) = \min(3.61, 2.92, 3.20, 2.50) = 2.50$
$d_E \to (A,B) = \min(d_{EA}, d_{EB}) = \min(4.24, 3.54) = 3.54$

Updated Matrix

| Dist | A,B | C | (D,F) | E |
|---|---|---|---|---|
| A,B | 0 | 4.95 | 2.50 | 3.54 |
| C | 4.95 | 0 | 2.24 | 1.41 |
| D,F | 2.50 | 2.24 | 0 | 1 |
| E | 3.54 | 1.41 | 1 | 0 |

min Dist

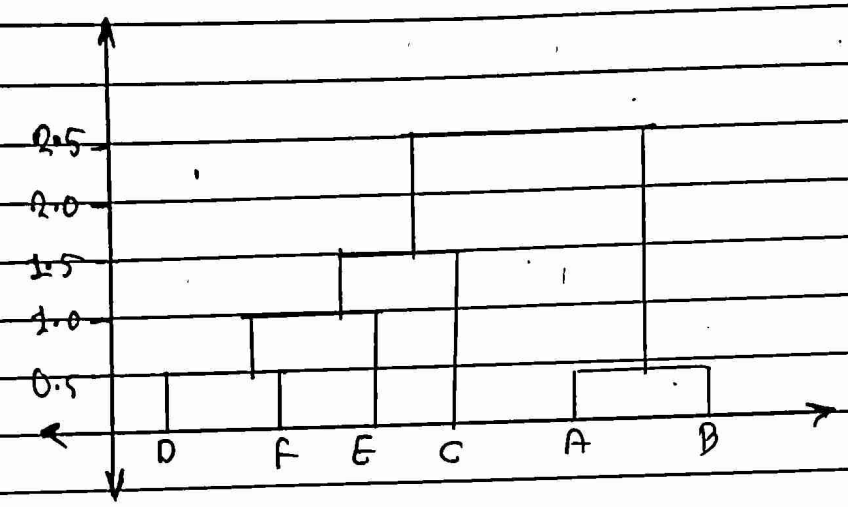
| Dist | (A,B) | C | (D,F),E |
|---|---|---|---|
| (A,B) | 0 | 4.95 | 2.50 |
| C | 4.95 | 0 | 1.41 |
| (D,F),E | 2.50 | 1.41 | 0 |

$d(D,F),E) \to C = \min(d_{DC}, d_{FC}, d_{EC}) = \min(2.24, 2.50, 1.41) = 1.41$

min dist

| Dist | (A,B) | (D,F),E),C |
|---|---|---|
| (A,B) | 0 | 2.50 |
| (D,F),E),C | 2.50 | 0 |

$d((D,F),E),C) \to (A,B) = \min(3.61, 2.92, 3.20, 2.50, 4.24, 3.54, 5.66, 4.95) = 2.50$




Q.2. Apply Single link approach and complete link approach to plot a dendrogram following fig. Contains sample data items indicating the distance between the elements

| Items | E | A | C | B | D |
|---|---|---|---|---|---|
| E | 0 | 1 | 2 | 2 | 3 |
| A | 1 | 0 | 2 | 5 | 3 |
| C | 2 | 2 | 0 | 1 | 6 |
| B | 2 | 5 | 1 | 0 | 3 |
| D | 3 | 3 | 6 | 3 | 0 |

→ Distance matrix :-

| | E | A | C | B | D |
|---|---|---|---|---|---|
| E | 0 | | | | |
| A | 1 | 0 | | | |
| C | 2 | 2 | 0 | | |
| B | 2 | 5 | 1 | 0 | |
| D | 3 | 3 | 6 | 3 | 0 |

$dist((E,A),C) = Min(dist(E,C), dist(A,C))$
$= min(2,2)$
$= 2$

$d((EA),B) = min(d(E,B), dist(A,B))$
$= min(2,5)$
$= 2$

$d((E,A),D) = min(dist(E,D), dist(A,D))$
$= min(3,3)$
$= 3$

| | E,A | C | B | D |
|---|---|---|---|---|
| E,A | 0 | | | |
| C | 2 | 0 | | |
| B | 2 | 1 | 0 | |
| D | 3 | 6 | 3 | 0 |

Step 2: Consider the dist matrix obtain in step 1 ; Since B,C dist is minimum, we combine B & C

$d((B,C),(E,A)) = min(dist(B,E), dist(B,A), dist(CE), dist(CA))$
$= min(2,5,2,2)$
$= 2$

$dist((B,C),D) = min(dist(B,D), dist(C,D))$
$= min(3,6)$
$= 3$

Q.2-

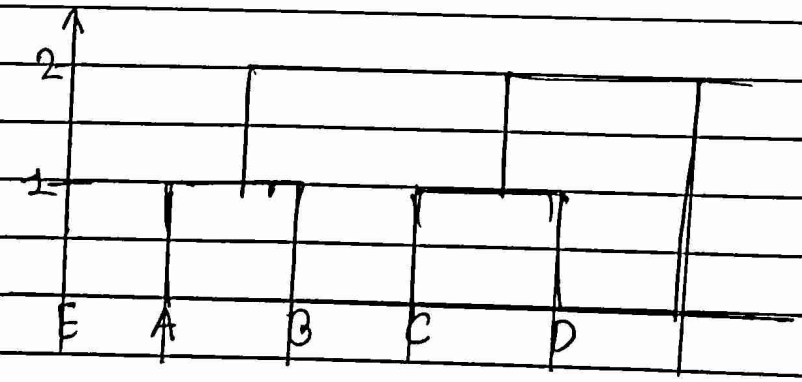
| | E,A | B,C | D |
|---|---|---|---|
| E,A | 0 | | |
| B,C | 2 | 0 | |
| D | 3 | 3 | 0 |

Step 3:- Consider the dist matrix obtained in Step 2. Since (E,A) & (B,C) dist is minimum, we combine them

$$dist((E,A),(B,C)) = min(dist(E,B), dist(E,C), dist(A,B), dist(AC))$$
$$= min(2, 2, 2, 5, 2)$$
$$= 2$$

| | E,A,B,C | D |
|---|---|---|
| EABC | 0 | |
| D | 2 | 0 |

Step 4:- Combine D with (EABC)



Experiment No. Date :

Q.4. Write a Short note on :-

a] Accuracy and Error Measures in Model Evaluation :-

→ Accuracy of classifier M, acc(M) is the percentage of test set tuples that are Correctly classified by Model M

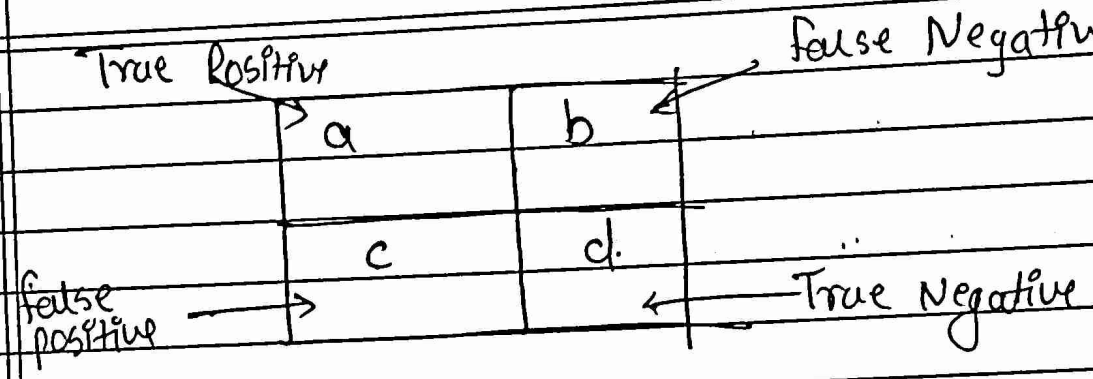
Training Set :-
Training Set is Subset of data used to train/ build the model.

Validation Set :-
It is used for parameter tuning but it can not be the data. Validation data can be training data, or a Subset of training data.

Test Set :-
It is set of instances that have not been used in training process. The model's performance is evaluated on unseen data.

The accuracy metrics are calculated with help of machine learning Confusion Matrix.

$$\text{Accuracy} = \frac{\text{No. of Correct predictions}}{\text{Total of all cases to be predicted}}$$




accuracy, recognition rate $= \frac{TP + TN}{P + N}$

error rate $= \frac{FP + FN}{P + N}$

Specificity $= \frac{TN}{N}$

Precision $= \frac{TP}{TP + FP}$

Experiment No. Date :

b] Cross validation :-

→ A Statistical Method or a resampling procedure used to evaluate the skill of machine learning models on a limited data sample.

Use Cross-validation to detect overfitting i.e failing to generalize a pattern

Steps involved in Cross validation are as follows:-
1. Reserve some portion of sample data-set
2. Using the rest data-set train the model
3. Test the model using the reserve portion of data set

Common Method used for Cross validation :-
1. Validation set approach
2. Leave-p-out Cross validation
3. leave-one out Cross validation
4. k-fold cross validation
5. Stratified k-fold cross validation

Application :-
- It has scope in medical research field
- It can also be used for the meta-analysis as it is already being used by data scientists in field of medical statistics

c] Bootstrap :-

→ - Works well with small data sets.
- bootstrapping is a way to quantify the uncertainty in your model while Cross validation used for model selection and measuring predictive ~~area~~ accuracy.
- Samples the given training tuples uniformly with replacement

i.e each time a tuple is selected, it is equally likely to be selected again and re-added training set

- Several bootstrap methods, and a common one is 632 bootstrap.
- Suppose we are given a data set of d tuples. The data set is sampled d times with replacement resulting in a training set of d samples. The data tuples that did not make it int training set end up forming the test set

Repea

$$acc(M) = \sum_{i=1}^{k} (0.632 \times acc(Mi)_{test\_set} + 0.368 \times acc(Mi)_{train\_set})$$