25-04-2023

Fuzzy Logic :

In real world many problems exist based on fuzzy that means the elements or the attributes which are associated with the problem are not clear. ~~to~~ OR we can say the values are in ambiguous (different person can interpret in different directions).

In this case the elements of the problem are not expressed or represented through classical set.

And also the human thinking & reasoning are changed based upon different situations & the nature will be in fuzzy.

So to express the human thinking & reasoning cannot be possible through classical set theory, which can be expressed through fuzzy set theory.

- Classical set theory allows the membership of the elements in a set in binary (0,1) where as the fuzzy set theory allow the membership value of an element in between 0 to 1. And the membership value of the element is to be assigned through membership function.

- Words like young, tall, good or high are said to be fuzzy & these are known as fuzzy words.

There is no single quantitative value which defines the term young. For some people 25 is young & for others 35 is young.

It has no clear boundary. Age 35 has some possibility of being young & usually depends on context.

→ Fuzzy set theory is an extension of classical set theory where elements have degree of membership

Example: Define a fuzzy set S for young students within a section where the age is from 25 to 28.

Let $S1 = 25$, $S2 = 28$, $S3 = 27.5$, $S_4 = 26$

$S_5 = 25.7$

$S_1 > S_4 > S_3 > S_2$

$S = \{ (S_1, 0.9) (S_2, 0.7) (S_3, 0.75) (S_4, 0.8) \}$

classical set = $\{ S_1, S_2, S_3, S_4 \}$

→ A fuzzy set can be represented using $\tilde{A}$ & the membership value of the elements can be defined through a membership function represented by $\mu_{\tilde{A}}(x)$.

$\tilde{A} = \{ (x_1, 0.9), (x_2, 0.8), (x_3, 0.4) (x_4, 0.1) \}$

$\mu_{\tilde{A}}(x_1) = 0.9$

$\mu_{\tilde{A}}(x_4) = 0.1$

Here 0.1 represents / indicates the belongingness or membership value of the element x_4 in the fuzzy set. A.

→ Fuzzy set theory operations

Given X to be the universe of discourse & $\tilde{A}$ & $\tilde{B}$ to be fuzzy sets with

$\mu_A(x)$ & $\mu_B(x)$. Then the fuzzy set operations.

Union = $\mu_{A\cup B}(x) = \max(\mu_A(x), \mu_B(x))$

Intersection = $\mu_{A\cap B}(x) = \min(\mu_A(x), \mu_B(x))$

Complement = $\mu_A(x) = 1-\mu_A(x)$

Example: $A = \{(x_1, 0.5), (x_2, 0.7), (x_3, 0)\}$
$B = \{(x_1, 0.8), (x_2, 0.2), (x_3, 1)\}$

$A \cup B = \{(x_1, 0.8), (x_2, 0.7), (x_3, 1)\}$
because
$\mu_{A\cup B}(x_1) = \max(\mu_A(x_1), \mu_B(x_1))$
$= \max(0.5, 0.8)$
$= 0.8$

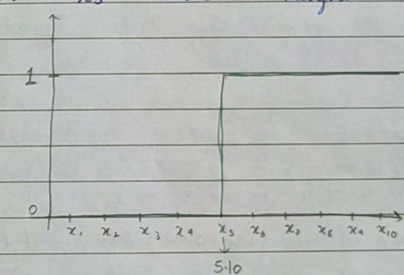
$\mu_{A\cup B}(x_2) = 0.7$ $\quad \mu_{A\cup B}(x_3) = 1$

$A \cap B = \{(x_1, 0.5), (x_2, 0.2), (x_3, 0)\}$
because
$\mu_{A\cap B}(x_1) = \min(\mu_A(x_1), \mu_B(x_1))$
$= \min(0.5, 0.8)$
$= 0.5$

Complement $A^c = \{(x_1, 0.5), (x_2, 0.3), (x_3, 1)\}$
because $\mu_A(x_1) = 1 - \mu_A(x_1)$
$= 1 - 0.5$
$= 0.5$

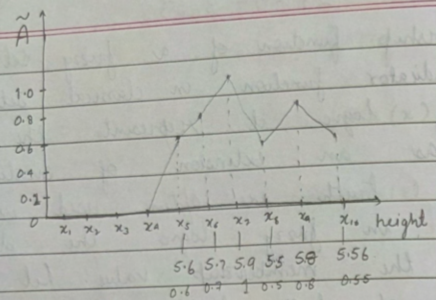
27-04-2023

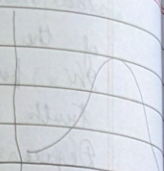
The membership function of a fuzzy set is a generalization of the indicator function in classical set.
In fuzzy logic, it represents the degree of the truth as an extension of evaluation.
Degree of truth are often used with probabilities where as in fuzzy logic the degree of truth represents the membership value. Let us consider, A is a set of tall people. Let
$A = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$
If we consider the tall people having height 5.10 inch. For this case let it be x_5
Then the graph will be
(from x_5 onwards height is greater than x_5)



$A^T = \{x_5, x_6, x_7, x_8, x_9, x_{10}\} \rightarrow$ Set of persons whose height ≥ 5
Here A^T is defined as the set of taller persons having height more than 5.10 inch.

To define a fuzzy set, $\tilde{A}$ which defines the set of taller persons from set A. As height is a fuzzy word. A person is said to be a tallest person depends upon the situation.

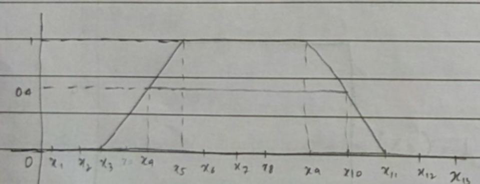




Let us consider:

height: $x_5 = 5.6$, $x_6 = 5.7$, $x_7 = 5.9$, $x_8 = 5.5$, $x_9 = 5.8$
$x_{10} = 5.56$

membership value: $x_5 = 0.6$ $x_6 = 0.7$ $x_7 = 1$ $x_8 = 0.6$ $x_9 = 0.8$
$x_{10} = 0.55$



$\tilde{A} = \{(x_1, 0), (x_2, 0), (x_3, 0), (x_4, 0.4), (x_5, 1)\ (x_6, 1)$
$(x_{7 \ldots 9}, 1), (x_{10}, 0.9), (x_{11}, 0), (x_{12,13}, 0)\}$

11-05-2023

Types of Membership function:

In fuzzy theory, a fuzzy set A of a universe X is defined by function $\mu_A(x)$ called the membership function of A.

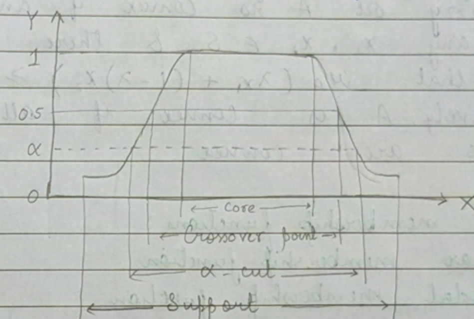
$\mu_A(x) : X \to [0, 1]$ where $\mu_1(x) = 1$. If x is totally in A

$\mu_1(x) = 0$ if x is not in A

1 element with 1 membership value
singleton

$0 < \mu_A(x) < 1$ If x is partly in A.

→ **Basics of fuzzy membership function :-**



1. **Support**: The support of the fuzzy set A is the set of all points $x \in X$ such that $\mu_A(x) > 0$.
Mathematically, we can express
Support $(A) = \{ (x, \mu_A(x)) / \mu_A(x) > 0 \}$

2. **Core**: The core of a fuzzy set A is the set of all elements $x \in A$ such that $\mu_A(x) = 1$
Core $(A) = \{ (x, \mu_A(x))$ where $\mu_A(x) = 1 \}$
$= \{ x \in X \mid \mu_A(x) = 1 \}$

Normality :- A fuzzy set A is normal if its core is non-empty. Core $(A) \neq \phi$.

Singleton :- A fuzzy set A is said to be a singleton fuzzy set if the support of A contains one element or single element x such that $\mu(x) = 1$.

3. α - cut : $\alpha_{cut}(A) = \{ x \in X \mid \mu_A(x) \geq \alpha \}$

Convexity of a fuzzy set:

A fuzzy set A is convex if and only if for any $x_1, x_2 \in S$ & there exist $x = [0,1]$ (constant λ) such that $\mu_A(\lambda x_1 + (1-\lambda) x_2) \geq \min(\mu_A(x_1), \mu_A(x_2))$

Alternatively, A is convex if all its α - cuts are convex.

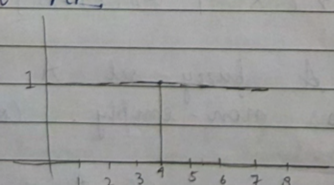
Different membership functions:

1. Triangular membership function
2. Trapezoidal membership function
3. Guassian membership function.
4. Generalized bell MF
5. Sigmoid MF

15-05-2023

Singleton membership function:

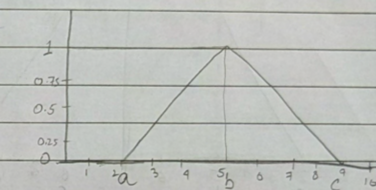
A membership function is said to be singleton MF if the membership value of an element of fuzzy set A is assigned to 1, other elements are assigned to membership value to 0. Then that MF is said to be singleton MF.



$\tilde{A} = \{(0,0)\ (1,0)\ (2,0)\ (3,0)\ (4,1)\ (5,0)\ (6,0)\ (7,0)\ (8,$

2. Triangular MF:

In this MF, three variables are used (a, b & c) are used to define membership function $\mu_A(x)$ of a fuzzy set $\tilde{A}$, where a is the lower bound b is the center, where membership value is 1 & c is the upper bound. (At a & c the membership value is 0).



$\tilde{A} = \{ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \}$

$\tilde{A} = \{ (1,0)\ (2,0)\ (3,0.33)\ (4,0.66)\ (5,1)\ (6,0.75)\ (7,0.5)\ (8,0.25)\ (9,0)\ (10,0) \}$

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & \text{if } x \leq a \\ \dfrac{x-a}{b-a} & \text{if } a \leq x \leq b \\ \dfrac{c-x}{c-b} & \text{if } b \leq x \leq c \\ 0 & \text{if } x \geq c \end{cases}$$

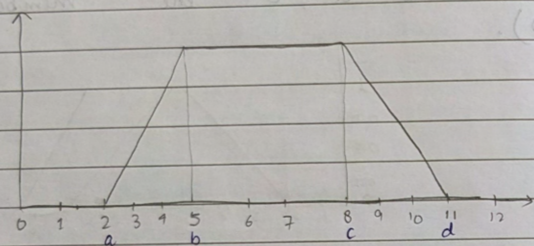
$\mu_{\tilde{A}}(3) = \dfrac{3-2}{5-2} = \dfrac{1}{3} = 0.33$, $\mu_{\tilde{A}}(7) = \dfrac{9-7}{9-5} = \dfrac{2}{4}$

$\mu_{\tilde{A}}(4) = \dfrac{4-2}{5-2} = \dfrac{2}{3} = 0.66$, $\mu_{\tilde{A}}(8) = \dfrac{9-8}{9-5} = \dfrac{1}{4}$

$\mu_{\tilde{A}}(6) = \dfrac{9-6}{9-5} = \dfrac{3}{4} = 0.75$

3) Trapazoidal membership F:

This MF is defined with four variables a, b, c & d where a is the lower bound & d is the upper bound where membership value is 0. Between b & c, the membership value is 1.



$\tilde{A} = \{(1,0)\ (2,0)\ (3,0.33)\ (4,0.67)\ (5,1)\ (6,1)\ (7,1)\ (8,1)\ (9,0.67)\ (10,0.33)\ (11,0)\ (12,0)\}$

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & \text{if} \quad x \le a \\ \frac{x-a}{b-a} & \text{if} \quad a \le x \le b \\ 1 & \text{if} \quad b \le x \le c \\ \frac{d-x}{d-c} & \text{if} \quad c \le x \le d \\ 0 & \text{if} \quad x \ge d \end{cases}$$

$\mu_3(x) = \frac{3-2}{5-2} = \frac{1}{3}$ $\qquad$ $\mu_9(x) = \frac{11-9}{11-8} = \frac{2}{3}$

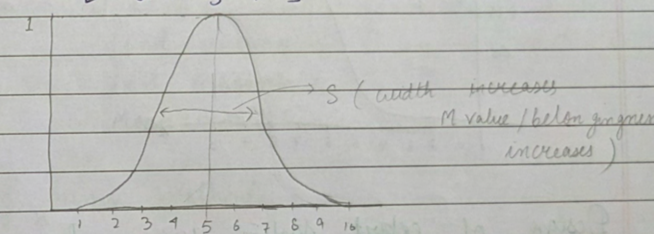
$\mu_4(x) = \frac{4-2}{5-2} = \frac{2}{3}$ $\qquad$ $\mu_{10}(x) = \frac{11-10}{11-8} = \frac{1}{3}$

or $\mu_{\tilde{A}}(x) = \text{Max}\left(\text{Min}\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right)$

4) Guassian MF:

This MF is ~~used as~~ usually represented by using ~~two~~ ~~three~~ three variables c, s, n. where c represents the mean (or center), s represents the std deviation or width & n represents the fuzzification ~~function~~ factor.

$$\mu_{\tilde{A}}(x) = \exp\left[-\frac{1}{2}\left|\frac{x-c}{s}\right|^m\right]$$

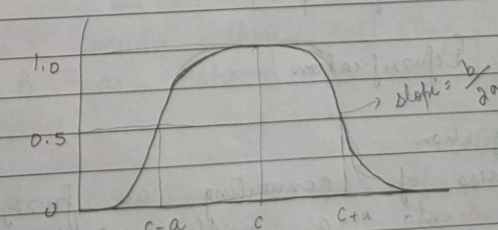


5) Generalized bell membership function.

A Generalized bell MF has three parameters

1. A → ~~A is~~
   a → 'a' is responsible for width
2. b → 'b' is responsible for its slope
3. c → 'c' is " " " center.

$$\text{gbell } f(x; a, b, c) = \frac{1}{1 + \left|\frac{x+c}{b}\right|^{2b}}$$

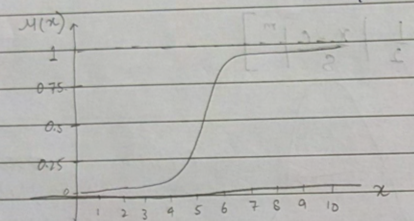


16-05-2023

7. Sigmoid MF:

The Sigmoid MF has two parameters : a & c where a is responsible for its slope & c is the center or Crossover point. The M.F of the Sigmoid is represented as

$$sig\,f\,(x : a, c) = \frac{1}{1 + e^{-a(x-c)}}$$



Design of expert system using fuzzy logic

Mamdani model :

Crisp Input
↓
Fuzzyfication
↓
Fuzzy Input
↓
Rule evaluation
↓
Fuzzy output
↓
Defuzzification

Mamdani

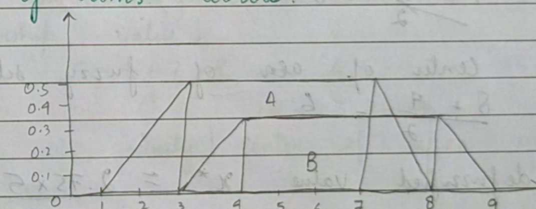
Defuzzification:

A process of converting a fuzzyfied obs output into a single crisp value wrt to a fuzzy set. The defuzzified value is used to represent the action to be taken by the controlling parameter of the process for an expert system.

There are 5 different defuzzification methods

1. Center of sums method (COS)
2. Center of gravity (COG) / Center of area
3. Bisector of area method (BOA)
4. Weighted average method
5. Maxima method :
   - first maxima
   - Last maxima
   - Mean maxima

1. Center of Sums method:



The defuzzified value x^* is defined as

$$x^* = \frac{\sum_{i=1}^{k} A_i \times \bar{x_i}}{\sum_{i=1}^{k} A_i}$$

Let us consider there are two fuzzy sets A & B as shown in the graph.

Here A_i represents the area of i^{th} rules & k is the total no. of rules defined and $\bar{x_i}$ represents the center of area.

The aggregated fuzzy set of the given two fuzzy set A & B can be defined through the area of A & B.

Area of A $= \frac{1}{2} \times 2 \times 0.5 + (7-3) \times 0.5 + \frac{1}{2}(8-7) \times 0.5$

$= 0.5 + 2 + 0.25$

$= 2.75$

Area of B $= \frac{1}{2} \times 1 \times 0.3 \times 2 + (8-4) \times 0.3$

$= 0.3 + 1.2$

$= 1.5$

Now, the center of area of fuzzy set A

$= \frac{7+3}{2} = 5$

the center of area of fuzzy set B

$= \frac{8+4}{2} = 6$

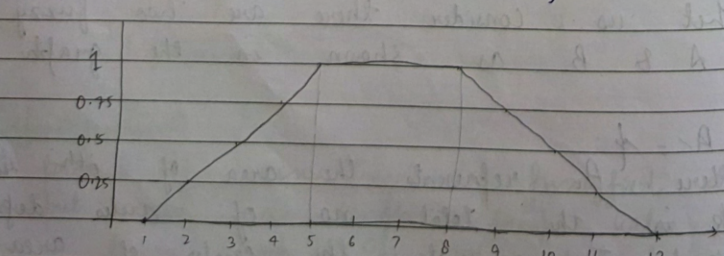
The defuzzified value $x^* = \frac{2.75 \times 5 + 1.5 \times 6}{2.75 + 1.5}$

$= 5.35$

2. Maxima Method:

Let us consider a fuzzy set $\tilde{A} = \{1,2,3,4,5,6 \ldots 10$

$\{(1,0)\ (2,0.25)\ (3,0.5)\ (4,0.75)\ (5,1)\ (6,1)\ (7,1)\ (8,1)\ (9,0.75)\ (10,0.5)\ (11,0.25)\ (12,0)\}$



This maxima method considers values with maximum membership.

1) First Maxima Method:
This method determines the smallest value of the domain with maximum membership value ie 5
Here the ~~first~~ defuzzified value $x^* = 5$

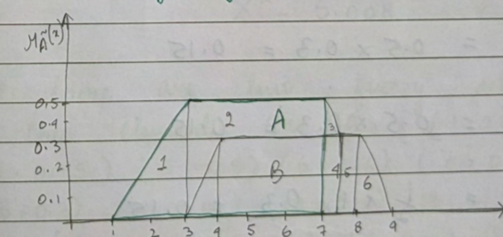
2) Last Maxima Method:
This method determines the highest value from the domain with maximum membership value is 8

3) Mean maxima method:
This method determines the mean of all the elements from the domain with maximum membership value.

ie $\frac{5+6+7+8}{4} = 6.5$

18-05-2023

3. Center of gravity | center of area method:



This method provides a crisp or defuzzified value based on the center of gravity of the fuzzy set.
The total area of membership function distribution is used to represent the combined control action taken by the which is divided into no of subareas. Then the area & the center of

gravity or centriod of ~~are~~ each subarea is calculated.
- The summation of all these subareas are considered as the defuzzified value.

Let us consider two fuzzy set : A & B as shown in the above figure.
There are 6 subareas as shown in the figure based on the fuzzy set elements & the membership value respectively.

Now we have to calculate the area & centroid of area of each subareas

Area of 1 $= \frac{1}{2} \times 2 \times 0.5 = 0.5$

Area of 2 $= 4 \times 0.5 = 2$

Area of 3 $= \frac{1}{2} \times 0.5 \times 0.2 = 0.05$

Area of 4 $= 0.5 \times 0.3 = 0.15$

Area of 5 $= 0.5 \times 0.3 = 0.15$

Area of 6 $= \frac{1}{2} \times 1 \times 0.3 = 0.15$

Now the centriod of the subareas are :

Centriod of 1 $= \frac{3+1+3}{3} = \frac{7}{3} = 2.33$

$\frac{0+0+0.5}{3} = 0.16$

Centroid of 2 : $\left(\frac{3+3+7+7}{4}\right)$ ~~8.8~~ $= \frac{20}{4}$

$= 5.$

Centroid of 3 $= \frac{7+7.5+7}{3}$

$\frac{21.5}{3} = 7.167$

Centroid of 4 $= \frac{7+7.5+7+7.5}{4} = 7.25$

Centroid of 5 $= 7.75$

Centroid of 6 $= \frac{8+8+9}{3} = 8.33$

$$X^* = \frac{\sum_{i=1}^{N} A_i * \bar{x}_i}{\sum_{i=1}^{N} A_i} = \frac{0.5 \times 2.33 + 2 \times 5 + 0.05 \times 7.16 + 0.15 \times 7.25 + 0.15 \times 7.75 + 0.15 \times 8.33}{0.5 + 2 + 0.05 + (0.15 \times 3)}$$

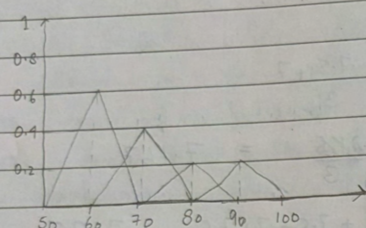
$X^* = 5.008$

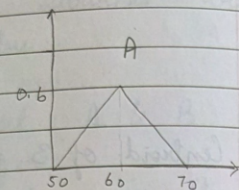
Q Let there are two fuzzy sets A & B. In A the elements are A = {(1, 0.1) (2, 0.3) (3, 0.4) (4, 0.45) (5, 0.5) (6, 0.6) (7, 0.6) (8, 0.6) (9, 0.45) (10, 0.3) (11, 0.04) (12, 0.3) (13, 0)}
B = {(4, 0.4) (5, 0.5) (6, 0.5) (7, 0.5) (8, 0.5) (9, 0.5) (10, 0.39) (11, 0.32) (12, 0.2) (13, 0.15) (14, 0.1) (15, 0)}

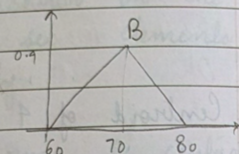
4. Weighted average method
This method is valid for fuzzy set with symmetrical output membership function & this method is efficient as compared to all others method is (Each membership function is weighted by (in terms of computation).

its maximum membership value.

Here $X^* = \frac{\sum \mu(x) \cdot x}{\sum \mu(x)}$



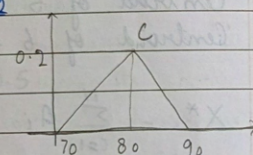


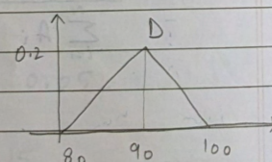


$x^* = \frac{60 \times 0.6 + 70 \times 0.4 + 80 \times 0.2 + 90 \times 0.2}{0.6 + 0.4 + 0.2 + 0.2}$

$= \frac{36 + 28 + 16 + 18}{1.4}$

$= 70$





Example: Let us consider, there are 3 fuzzy sets A, B, C. The elements of A = {(5, 0) (6, 0.25) (7, 0.5) (8, 0.8) (9, 0.5) (10, 0.25)}

Let us consider the problem to analyze the risk factor for completion of a project

- The risk factor of a project can be analyzed based on 2 linguistic variables or parameters like budget & manpower
- The fuzzy variable budget can be adequate, marginal & inadequate. The fuzzy variable manpower can be large & small.

- The above problem can be characterized based on 3 rules.

Rule 1: If X is A3 (project funding is adequate)
OR Y is B1 (project manpower is small)
Then Z is C1 (Then risk is low)

Rule 2: If X is A2 (project funding is marginal)
AND Y is B2 (project manpower is large)
Then Z is C2 (Then risk is normal)

Rule 3: If X is A1 (If project funding is inadequate)
Then Z is C3 (Then risk in high)

- Based on the rules (1, 2 & 3) the fuzzy set A represents the budget & the fuzzy set B represents the manpower.

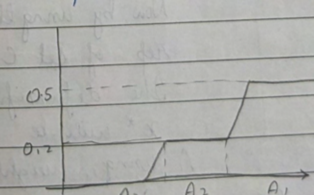
$\tilde{A} = \{A_1, A_2, A_3\}$

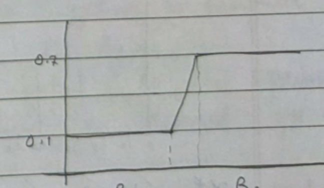
$\tilde{B} = \{B_1, B_2\}$

$\tilde{C} = \{C_1, C_2, C_3\}$

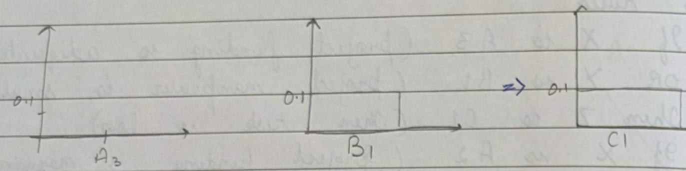
- Now based on the rules defined, the fuzzy value of A_1, A_2, A_3 can be 0.5, 0.2, 0. & for the set B, $B_1 = 0.1$, $B_2 = 0.7$.

Now represent the fuzzy set A & B.



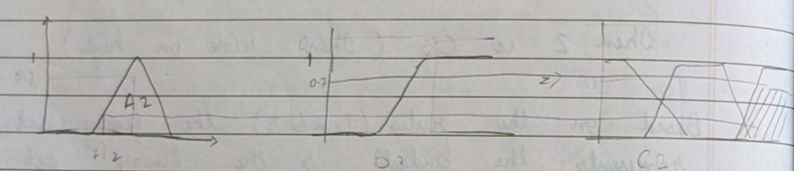


Rule 1: If $x \in A_3$ — 0.1 OR & $y \in B_1$



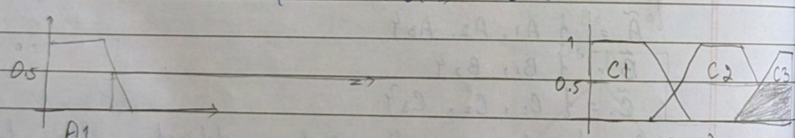
$$Z = Max(\mu_A(x), \mu_B(x)) = 0.1$$

Rule 2: $x \in A_2$ (0.2) & $y \in B_2$ (0.7)



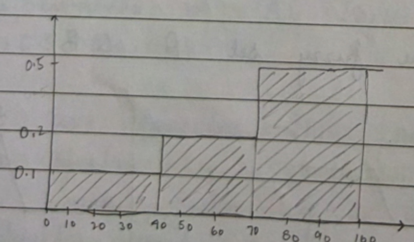
$$Z = \min(\mu_A(x), \mu_B(x)) = 0.2$$

Rule 3:



$$Z = C3\,(0.5)$$

Now,
$\tilde{C} = \{(C_1, 0.1)\ (C_2, 0.2)\ (C_3, 0.5)\}$



Now, by using the graph rep of set C is this. The defuzzified value x^* will be calculated by using weighted avg method

$$\cancel{x^* = (10+20+30+40)0.1 + (50+60+70)0.2 + (80+90+100)0.5}$$

$$x^* = \frac{(10+20+30)0.1 + (40+50+60)0.2 + (70+80+90+100)0.5}{3(0.1) + (0.2)3 + 0.5(\cancel{3}\,4)}$$

$$x^* = 71.03$$

diks

The above method is said to be Mamdani-style rule evaluation or mamdani method

25-05-2023

Machine Learning :-

ML is a subfield of computer science which is generated from the study of pattern recognized & computation learning theory in AI.
Using these algorithms, implicitly program is generated from the data set with the help of iterative learning.

- Learning is any process by which a system improves the performance from the experience.
- ML is the study of algorithms that improve the performance (P) of a task (T) with experience (E) Three tuple (P, T, E)

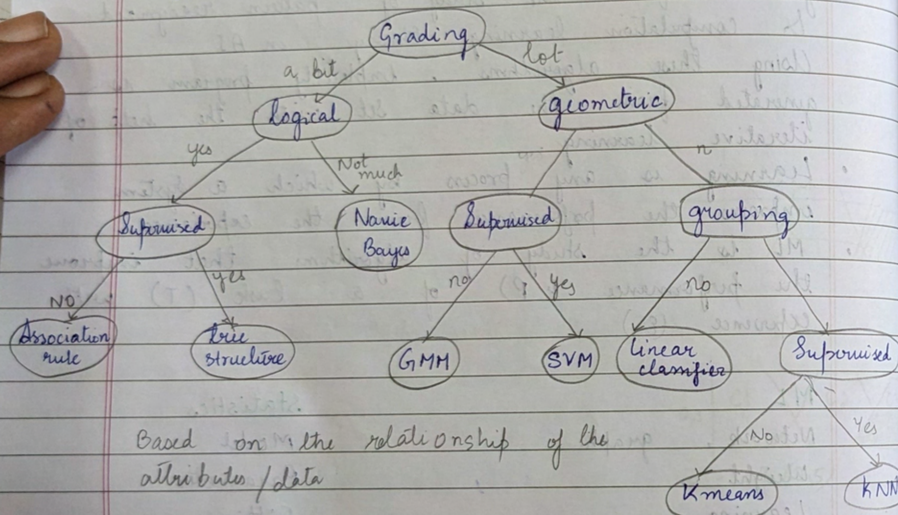
ML	Statistics
Network, graph.	Model
Weight	Parameter
Learning	Fitting.
Supervised learning	Regression / classification
Unsupervised learning	clustering

- Based on the inputs to the system or computer. ML algos are classified into 4 types:

1. Supervised Learning Here the input to the system is training data pnts plus the desired o/p or label.
2. Unsupervised Learning Here the input to the system is only data / trainning data
3. Semi-Supervised Learning Here the input to the system is training with few output / labeled data.

4. Reinforcement :- Here the system will learn from a set of actions or system of actions.

- Based on the nature of the problem, ML algos. can be classified as follows



Based on the relationship of the attributes / data

Unsupervised Machine Learning :-

In case of unsupervised ML the data which are considered as input to the system are not labelled.

This technique is known as clustering.

→ Clustering has been studied extensively in different fields as a result there are different types of clustering techniques :

1. Partitioning method
2. Hierarchical method
3. Density based method
4. Graph based method

29-05-2023

5. Model based clustering

Partitioning method :

K-mean In this method a set of n-distinct object is given within a data set. Based on the number of partition as per the user requirement we have to partition the n no of objects into distinct partitions.

K-means algo is one of the partitioning algo. where K represents the no of partition. (given by the user)

Algorithm :- K-means clustering

Step 1 Randomly chose K-objects from D (dataset) as the initial cluster centroid.

Step 2 For each of the object in D do the following steps :

1. Compute the distance b/w the current objects and K-cluster centroid.
2. Assign the current object to that cluster to which it is closest.

Step 3 Compute the cluster center of each cluster using the mean of all the data points in the cluster. These are the new centroid of the cluster.

Step 4 Repeat step 2 and 3 until a stopping criteria is satisfied.

(Stopping criterias are : 1. no of max iteration reached 2. No change of centroid value of any cluster 3. No movement of object from one cluster to another cluster)

30-05-2023

Example There is a dataset D having 1~~4~~6 different data points with the following coordinates

	A_1	A_2	d_1	d_2	d_3	cluster
1	6.8	12.6	4.03	1.07	5.9	C_2
2	0.8	9.8	3.00	7.4	10.2	C_1
3	1.2	11.6	3.10	6.6	8.5	C_1
4	2.8	9.6	1.044			1
5	3.8	9.9				1
6	4.4	6.5				1
7	4.8	1.1				1
8	6.0	19.9				3
9	6.2	18.5				3
10	7.6	17.4				3
11	7.8	12.2				2
12	6.6	7.7				1
13	8.2	4.5				1
14	8.4	6.9				2
15	9.0	3.4				1
16	9.6	11.1				2

Step 1 Let $k = 3$ that means we have to choose 3 different objects out of the 16 object and these 3 objects are considered as 3 centroid of the three cluster.

Centroid	A_1	A_2
C_1	3.8	9.9
C_2	7.8	12.2
C_3	6.2	18.5

Step 2 In step 2 we have to calculate the distance d_1, d_2, d_3 from an object to C_1, C_2, C_3 respectively. Here the distance metric can be considered here diff. metric can be used to find distance.

hamming, Cosine → for strings

For this case we can use equilidean distance if there are two data points x_1, y_1 & x_2, y_2 then the ~~edi~~ euclidean distance $d = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

$$d_1 = \sqrt{(6.8-3.8)^2+(12.6-9.9)^2} = 4.03$$

$C1 = \{4, 5, 6, 7, 12, 13, 15\}$

$C2 = \{1, 11, 14, 16\}$

$C3 = \{8, 9, 10\}$

There are three objects in cluster 3, four in two & 9 in cluster 1.

Now,

C_3	A_1	A_2
1.	6.0	19.9
2.	6.2	18.5
3.	7.6	17.4
mean :-	6.6	18.6

Similary, in other clusters

New centroid	A_1	A_2
C_1	4.6	7.1
C_2	8.2	10.7
C_3	6.6	18.6

A_1	A_2	d_1	d_2	d_3	cluster
6.8	12.6	5.9	2.36	6	C_2
0.8	9.8	0.	7.54	10.5	
1.2	11.6				
2.8	9.6				
3.8	9.9				

The performance of the algo basically depends upon the nature of the problem or we can say the data types in the data set

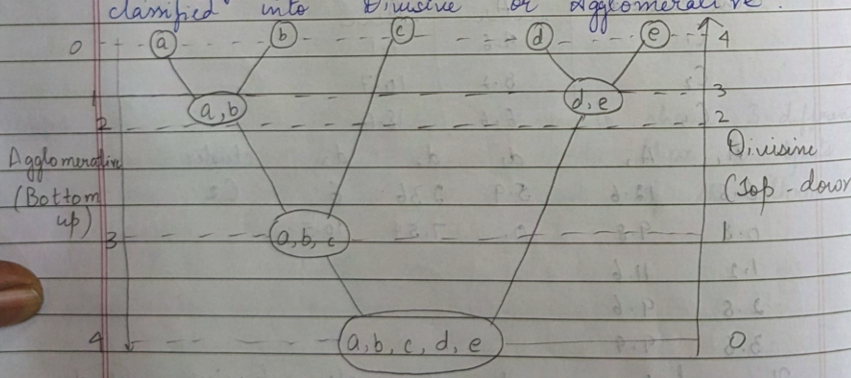
Following are the matrics or the formula used based on the data type on the data set

1. Euclidean distance
2. Cosine distance } → if string len is diff → String datatype
3. Hamming distance } → if length of String is same
4. Manhattan distance
5. Geodesic distance → Image

01-06-2023

Hierarchical clustering Algorithm:

This is an alternative to partitioning algos where the no of partitions are not mentioned prior to the clustering. In hierarchical clustering algo, the hierarchy of the clusters is obtained either by splitting a large cluster into small clusters or by merging small cluster into large clusters. Depending on splitting or merging. It can be classified into Divisive or Agglomerative.



The above representation of the cluster is said to be a Dendagram of datapoints.

Aglometrative hierarchical clustering Algo:

In this technique, the given dataset 'D' are 'n' no of datapoints where each & every datapoint is treated as an cluster. Iteratively we have to generate the clusters based on the minimum/maximum/average distance among all the datapoints.

→ The major steps of this algos are:

S-1 Initialization: Assign each object into a single object cluster. If there are 'n' no of objects such that $x_1, x_2 \dots x_n$, then there are 'n' no of clusters which are represented by using a set $C = \{c_1, c_2, \dots c_n\}$. where the ith object ($x_i$) belongs to cluster c.

S-2 To calculate the inter-cluster distance: In this step we have to calculate the distance b/w each pair of the cluster. if there are 'm' clusters at any instance, then it gives us P no. of distance which is represented as $d_1, d_2, d_3 \dots d_p$ where $P = \frac{m(m-1)}{2}$

S-3: In this step we have to merge the clusters. Here we have to choose the distance with lowest value & merge them into a single cluster. This step reduces the no clusters into one. If more than one pair of clusters have same distance then we have to arbitarily choose any pair of cluster.

S-4: Repeat step 2 & 3 until it reaches our stopping criteria.

S-5: stop.

In this, the major step is merging operation which is based on measuring the distance b/w each pair of cluster. Here $d(x, y)$ represents the distance b/w 2 datapoints x & y. $d^*(C_i, C_j)$ represents the distance b/w 2 clusters C_i & C_j

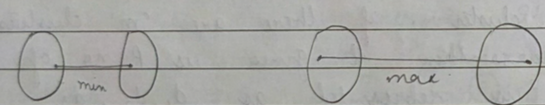
The min distance b/w 2 cluster

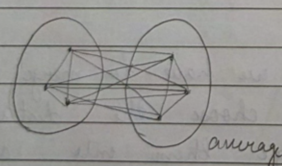
$$d_{min}(C_i, C_j) = \min_{\substack{x \in C_i \\ y \in C_j}} \{d(d\ x, y)\}$$

$$d_{max}(C_i, C_j) = \max_{\substack{x \in C_i \\ y \in C_j}} \{d(x, y)\}$$

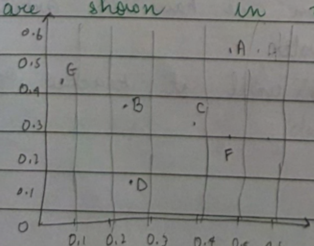
Average distance (CLIQUE)

$$d_{avg}(C_i, C_j) = \frac{1}{|C_i|\,|C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$





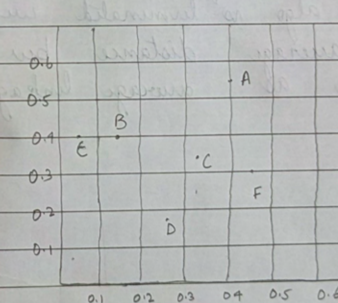
Q Let us consider there is dataset 'D' having 6 no of elements $D = \{A, B, C, D, E, F\}$ & the points are shown in the graph



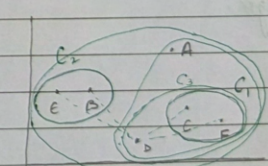
A (0.42, 0.53)
B (0.22, 0.39)
C (0.34, 0.31)
D (0.25, 0.18)
E (0.06, 0.42)
F (0.47, 0.3)

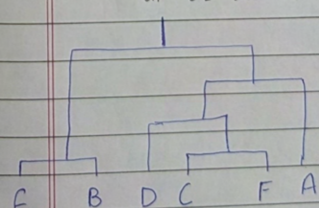
	A	B	C	D	E	F
A	0	0.24	0.22	0.37	0.34	0.23
B	0.24	0				
C	0.22	.15	0			
D	0.37	.20	.15	0		
E	0.34	.14	.28	.29	0	.39
F	0.23	.25	.11	.22	.39	0

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

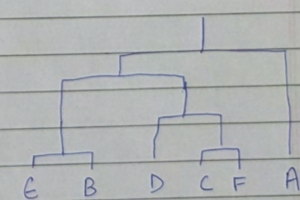


$$d(C_i, C_j) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$





&



→ When this algorithm uses the minimum distance dmin (c_i, c_j) it is called as nearest neighbour clustering algo. Also it is called as MST - Minimum spanning tree clustering algorithm

→ When the algo uses to measure the minimum dist. b/w 2 cluster and the algo is terminated when the distance b/w 2 nearest clusters exceeds the threshold value is called as Single linkage Algo.

→ Whenever if the algo is terminated when the maximum distance b/w 2 nearest clusters exceeds the threshhold value is called the Complete linkage clustering algo

→ Whenever this algo uses average distance b/w two clusters and the algo is terminated with a threshold value (average distance b/w 2 clusters) is known as average linkage algo.