POORNIMA FOUNDATION Estd. 2000

POORNIMA FOUNDATION

LECTURE NOTES

Campus: PCE   Course: BTECH in CSE   Class/Section: III Yr. Section- A   Date: 7-03-21

Name of Faculty: Praveen Kumar Yadav   Name of Subject: Machine Learning   Code: 6CS4-02

Date (Prep.): 3-03-21 Date (Del.): 9/4/21 Unit No.: II Lect. No: 19-14

OBJECTIVE: To be written before taking the lecture (Pl. write in bullet points the main topics/concepts etc., which will be taught in this lecture)

K- Means clustering
Hierarchical clustering

IMPORTANT & RELEVANT QUESTIONS:

what is K in K-Means clustering Algom?

FEED BACK QUESTIONS (AFTER 20 MINUTES):

How to determine the best K in K-Means clustering Algon

OUTCOME OF THE DELIVERED LECTURE: To be written after taking the lecture (Pl. write in bullet points about students' feedback on this lecture, level of understanding of this lecture by students etc.)

good

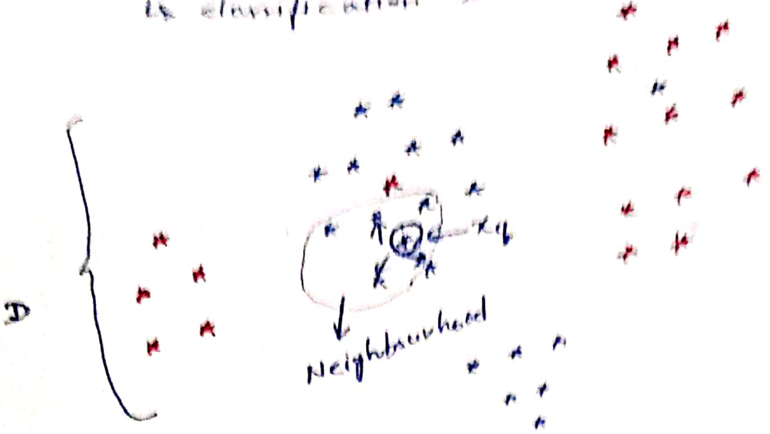
REFERENCES: Text/Ref. Book with Page No. and relevant Internet Websites:

scikit learn with ML.

K-Nearest NN Algorithm :-

i.e. K-Nearest Neighbours Algorithm

i.e. classification use case :-



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

Manhattan distance :-

$*$; +ve data point

$*$; -ve data point

$D = \{x_i\ y_i \mid x_i \in R^2,\ y_i \in \{0, 1\}\}$

y_i: 0 → -ve, 1 → +ve

which are geometrical algos

$D \rightarrow ML \rightarrow$ fun^n

$x_q \rightarrow \boxed{f} \rightarrow$ +ve / -ve

$x_q \rightarrow y_q$

Take points to x_q :

conclude $x_q \rightarrow$ blue (+ve)

given Task $x_q \rightarrow y_q \in \{0, 1\}$

0 → -ve, 1 → +ve

1. Find k - nearest points to x_q in D

let k = 3

(x_1, x_2, x_3) - 3 NN to x_q

$\downarrow \quad \downarrow \quad \downarrow$

$y_1 \quad y_2 \quad y_3$

2. Take their class-labels $\{y_1, y_2, y_3\}$ → Majority vote

$+ \ + \ + \rightarrow y_q = +ve$

$+ \ + \ - \rightarrow y_q = +ve$

k = odd (preferable)

$x_q \rightarrow$ k - NN of x_q

$x_1, x_2, \ldots x_k$

$\downarrow$

$y_1, y_2, \ldots y_k$

$\downarrow$

Majority vote

$\downarrow$

y_q

Simple Idea of KNN.

श्री गणेशाय नमः

Clustering use case -

Unsupervised Learning Algorithm:- using K-means clustering Algorithm:-

Grouping unlabelled items

e.g. Height / weight

Height	weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

K_1 K_2

(185, 72) (170, 56)

ED for - 3rd row

$K_1 \rightarrow \sqrt{(168-185)^2 + (60-72)^2}$

$= 20.80$

$K_2 = \sqrt{(168-170)^2 + (60-56)^2}$

$= 4.48$

Here value 3rd row goes K_2

$K_2 = \{2, 3\}$

New centroid calculation -

$K_2 = \left((170+168)/2 + (56+60)/2 \right)$

$= (169, 58)$

New clusters are -

K_1 K_2

(185, 72) (169, 58)

Euclidean Distance -

$$= \sqrt{(x_0 - x_c)^2 + (y_0 - y_c)^2}$$

observed value. Centroid value

Now Euclidean distance for 4th row. $k_1 = \sqrt{(179-185)^2 + (68-72)^2}$

$= 6.32$

$\rightarrow k_2 = \sqrt{(179-169)^2 + (68-58)^2}$

$= 14.14$

So 4th row goes to cluster k_1 so we again calculate the New centroid for that row.

$k_1 \rightarrow \{1, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

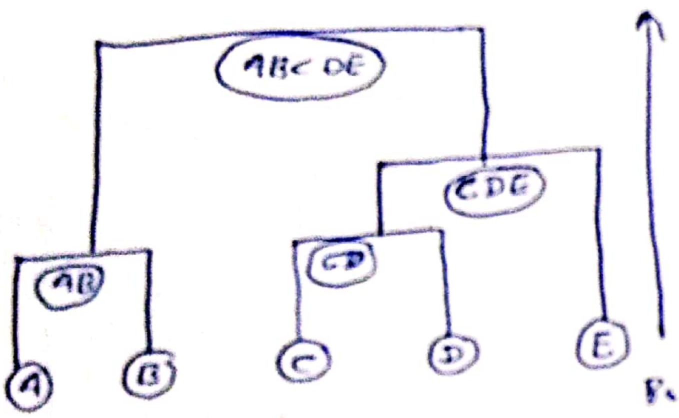
$k_2 \rightarrow \{2, 3\}$

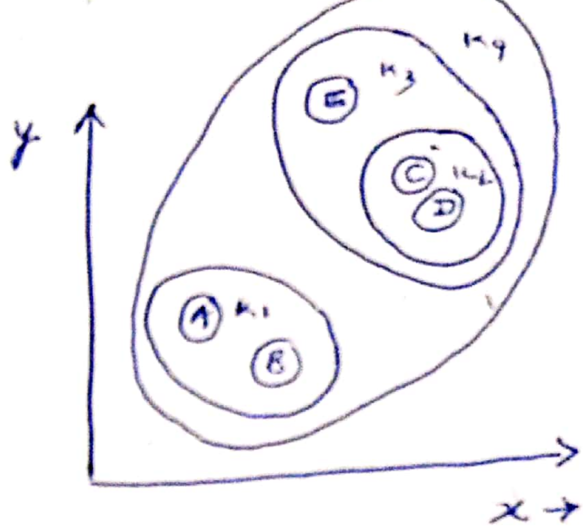
Scanned with CamScanner

Hierarchical clustering:-

Agglomerative clustering -

- starts from Bottom to top.



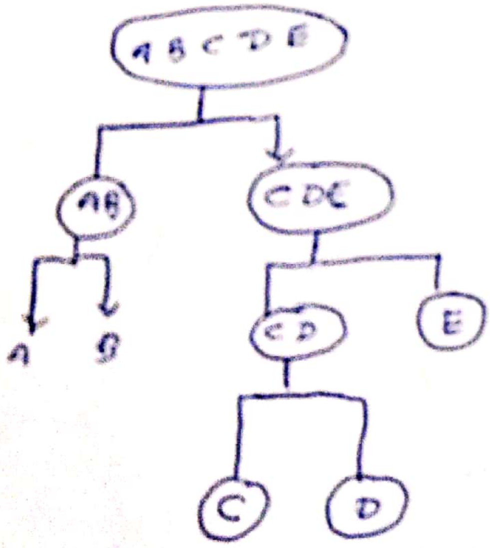


Bottom up representation

(also known as Dendrogram)

Divisive approach -

- Top-to Bottom approach



random Assignment

Agglomerative clustering :-
(with single linkage Technique)

↳ agglomerative clustering starts with each individual data points.

↳ Then we make cluster of these data points until we get single cluster which represent each data point.

step ② P_3 P_5 P_1

Distance MATRIX.

eg -

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	9	0			
P_3	3	7	0		
P_4	6	5	9	0	
P_5	11	10	2	8	0

$P_3 P_5$

⇓

	P_1	P_2	$[P_3\ P_5]$	P_4
P_1	0			
P_2	9	0		
$[P_3\ P_5]$	3	7	0	
P_4	6	5	8	0

$\Rightarrow d(P_1, [P_3\ P_5])$

$\Rightarrow \min(d(P_1, P_3), d(P_1, P_5))$

$\Rightarrow \min(3, 11) \Rightarrow 3$

$\Rightarrow d(P_2, [P_3, P_5])$

$\Rightarrow \min(d(P_2, P_3), d(P_2, P_5))$

$\Rightarrow \min(7, 10) \Rightarrow 7$

$\Rightarrow d(P_4, [P_3, P_5])$

$\Rightarrow \min(d(P_4, P_3), d(P_4, P_5))$

$\Rightarrow \min(9, 8) \Rightarrow 8$

Here we choose minimum as 3 and make a cluster between P_3 P_5 P_1

updated distance matrix are as-

	[P1 P3 P5]	P2	P4
[P1 P3 P5]	0		
P2	7	0	
P4	6	5	0

take min - 5 so make cluster (P2 P4)

$d(P_2, [P_1 P_3 P_5])$

$= \min(d(P_2, P_1), d(P_2, P_3), d(P_2, P_5))$

$= \min(9, 7, 10) \Rightarrow 7$

$d(P_4, [P_1 P_3 P_5])$

$\Rightarrow \min(d(P_4, P_1), d(P_4, P_3), d(P_4, P_5))$

$\Rightarrow \min(6, 9, 8) \Rightarrow 6$

	$[P_1\ P_3\ P_5]$	$[P_2\ P_4]$
$[P_1\ P_3\ P_5]$	0	
$[P_2\ P_4]$		0

$d([P_1\ P_3\ P_5], [P_2\ P_4])$

$= \min(d[P_1\ P_2], d[P_1\ P_4], d[P_3\ P_2], d[P_3\ P_4], d[P_5\ P_2], d[P_5, P_4])$

$\Rightarrow \min(9, 6, 7, 10, 8, 9)$

$\Rightarrow \min 6$

make a single cluster