Shubham Bhate 8318 TE Comps

DWM Assignment 3

1. Total records = 14

$H(s) = -P(yes)\log_2 P(yes) - P(no)\log_2 P(no)$

$H(s) = 0.94$

① Outlook

$H(\text{sunny}) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0.971$

$H(\text{rain}) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5}) = 0.971$

$H(\text{overcast}) = -\frac{4}{4}\log_2(\frac{4}{4}) = 0$

$IG = 0.94 - (\frac{3}{14} \times 0.971 + \frac{5}{14} \times 0.971) = 0.247$

② Temperature (i) (consider Threshold = 70)

$H(T > 70) = -\frac{4}{9}\log_2(4/9) - 3/9\log_2(3/9) = 0.991$

$H(T \leq 70) = -4/5\log_2(4/5) - 1/5\log_2(1/5) = 0.721$

$IG = 0.94 - [(\frac{9}{14} \times 0.991) + (\frac{5}{14} \times 0.721)] = 0.045$

(ii) Considering Threshold 75

$H(T > 75) = 2/4\log_2(2/4) - 2/4\log_2(2/4) = 1$

$H(T \leq 75) = 7/10\log_2(7/10) - 3/10\log_2(3/10) = 0.72$

$IG = 0.94 - (\frac{10}{14} \times 0.72 + \frac{4}{14} \times 1) = 0.14$

(iii) consider Threshold 80, we get $IG = 0.101$

IG is max for 75, so we consider it.

③ Humidity

(i) Taking 80 as threshold, we get $IG = 0.101$

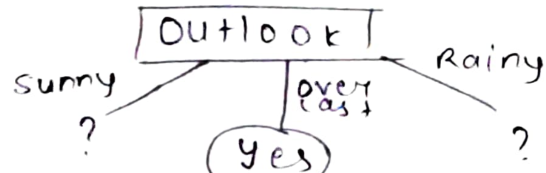
④ Wind

$H(\text{strong}) = -3/6\log_2(3/6) - 3/6\log_2(3/6) = 1$

$H(\text{weak}) = -2/8\log_2(2/8) - 6/8\log_2(6/8) = 0.811$

$IG = 0.94 - (\frac{6}{14} \times 1 + \frac{8}{14} \times 0.811) = 0.049$

Outlook has max IG, so it is the root



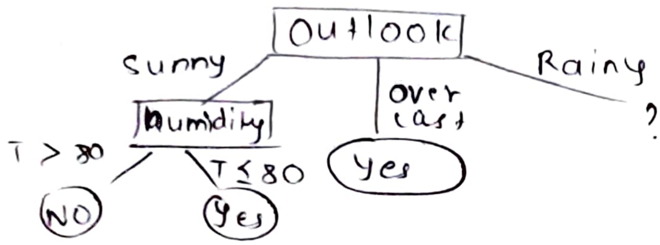
Considering outlook ^sunny calculate IG for all

For humidity,

$H(T \leq 80) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$

$H(T > 80) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) = 0$

$\therefore$ IG = max



For Rain as outlook,

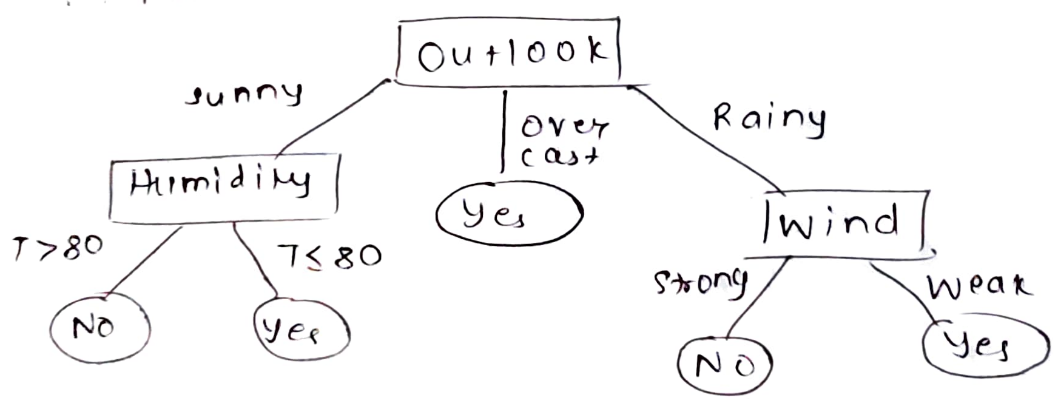
wind

H(weak) = 0

H(strong) = 0

$\therefore$ IG = 0.97 = max

$\therefore$ Final Tree is



2.

Student	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})(x-\bar{x})$
1	95	85	17	8	289	136
2	85	95	7	18	49	126
3	80	70	2	-1	4	-14
4	70	65	-8	-12	64	96
5	60	70	-18	7	324	126
	390	385			730	470

$\bar{x} = 79$, $\bar{y} = 77$, $y = b_0 + b_1 x$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x - \bar{x})^2} = \frac{470}{730} = 0.643$$

$b_0 = \bar{y} - b_1 \bar{x} = 77 - 0.643 \times 79 = 26.84$

$y = 26.84 + 0.643 \times x$

when $x = 80$, $y = 77.98$

3. Initial value of centroids, $x_1 = A_1$, $x_2 = B_2$, $x_3 = C_1$

Iteration 0

using Euclidean Distance

D^0:	A1	A2	A3	B1	B2	B3	C1	C2	Centroid
	0	5	8.48	3.61	7.07	7.21	8.06	2.24	X1
	3.61	4.24	5	0	3.61	4.12	7.21	1.41	X2
	8.06	3.16	7.28	7.21	6.71	5.39	0	7.62	X3

Object clustering

G_0: Group 1 = A1

Grp 2 = A3, B1, B2, B3, C2

Grp 3 = A2, C1

Iteration 1, determine centroid

$X_1 = (2, 10)$

$$X_2 = \frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5} = (6, 6)$$

$$X_3 = \frac{2+1}{2}, \frac{5+2}{2} = (1.5, 3.5)$$

D1:

A1	A2	A3	B1	B2	B3	C1	C2	centroid
0	5	8.44	3.61	7.07	7.21	8.06	2.24	X_1
5.56	4.12	2.83	2.24	1.41	2	6.40	3.61	X_2
6.52	1.58	6.52	5.7	5.70	4.52	1.58	6.04	X_3

G1: Group 1 = A1, C2
Group2 = A3, B1, B2, B3,
Group3 = A2, C1

Iteration 2: centroid

$X_1 = \frac{2+4}{2}, \frac{10+9}{2} = (3, 9.5)$

$X_2 = \frac{8+5+7+6}{4}, \frac{4+8+5+4}{4} = (6.5, 5.25)$

$X_3 = \frac{2+1}{2}, \frac{5+2}{2} = (1.5, 3.5)$

D2:

A1	A2	A3	B1	B2	B3	C1	C2	centroid
1.12	2.35	7.43	2.5	6.02	6.26	7.76	1.12	X_1
6.54	4.51	1.95	3.13	0.56	1.35	6.38	7.68	X_2
6.52	1.58	6.52	5.70	5.70	4.52	1.58	6.04	X_3

G2: Grp 1 = A1, B1, C2
Grp 2 = A3, B2, B3
Grp 3 = A2, C1

Iteration 3: centroid

$X_1 = \frac{2+5+4}{3}, \frac{10+9+8}{3} = (3.67, 9)$

$X_2 = \frac{8+7+6}{3}, \frac{4+5+4}{3} = (7, 4.33)$

$X_3 = \frac{2+1}{2}, \frac{5+2}{2} = (1.5, 3.5)$

D3:

A1	A2	A3	B1	B2	B3	C1	C2	centroid
1.95	4.33	6.61	1.66	5.2	5.52	7.49	0.33	X_1
6.01	5.04	1.05	4.17	0.67	1.05	6.44	5.55	X_2
6.52	1.58	6.52	5.7	5.7	4.52	1.58	6.04	X_3

G3: Grp 1 = A1, B1, C2
Grp2 = A3, B2, B3
Grp 3 = A2, C1

since groups in iteration 2 and 3 are same, we stop
∴ final groups are G3

4. (i) K-Means: In k-means clustering, data objects are classified into attributes or features into k-clusters.
Input : No of clusters (k) and points of problem
shape : spherical
Limitations : cant cluster in geometric shapes. Gives more weight to bigger clusters than smaller ones. Over lapping clusters cant be generated.
Outliers: the cluster centre is pushed towards the outlier, thereby distributing the actual cluster.
(ii) K-mediods : Mediods of values in cluster are used. Attempts to minimize sum of dissimilarities between object labelled to be in a cluster and one of objects designated as the representative of that cluster known as mediod.
Limitation : Does not scale well, or work efficiently for large data sets
Outliers . Partitioning among mediods is more robust than k-~~mediod~~ means in presence of outliers
(iii) CLARA : Extension of K-mediods, can handle large amounts of data .
Limitation : Doesnt guarentee perfect output for localised area .
Uses random samples for neighbors
(iv) OPTICS : Ordering points to identify cluster structure. Does not explicitly segment data into clusters except it produces a visualization of reachability

distance and uses distance to cluster data.
Limitations: No actual clusters, requires more memory compared to other clusters

5. candidate List = { M O N K E Y D A C U I }

C1:

Itemset	Support	Itemset	Support
M ✓	3	D	1
O ✓	3	A	1
N	2	C	2
K ✓	5	U	1
E ✓	4	I	1
Y ✓	3		

As per support = 60% or 3 transactions

L1:

Item set	support
M	3
O	3
K	5
E	4
Y	3

C2:

Itemset	Support	Itemset	Support
M O	1	O E ✓	3
M K ✓	3	O Y	2
M E	2	K E ✓	4
M Y	2	K Y ✓	3
O K ✓	3	E Y	2

L2:

Itemset	support
M K	3
O K	3
O E	3
K E	4
K Y	3

C3:

Itemset	support	Itemset	support
M K O	1	O K E	3
M K O E	1	O K Y	2
M K E	2	O E K Y	2
M K Y	2	K E Y	2

Only {OKE} satisfies support rules

Associations			Confidence	
E O	→	K	3/3	100%
OK	→	E	3/3	100%
KE	→	O	3/4	
O	→	KE	3/3	100%
E	→	OK	3/4	
K	→	E O	3/5	

Final rules are

E, O → K

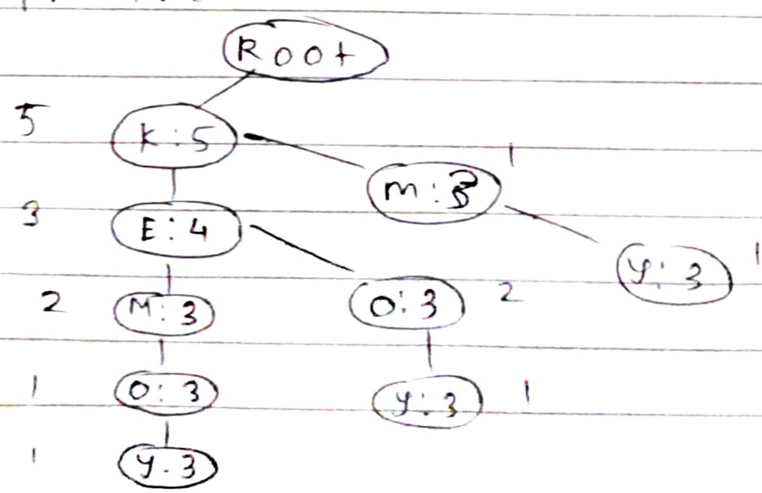
O, K → E

O → K, E

(iii) FP Tree : considering C_1 calculated above, we get in frequency descending order {K E M O Y}

T100	MONKEY	KEMOY
T200	DONKEY	KEOY
T300	MAKE	KEM
T400	MUCKY	KMY
T500	COOKIE	KEO

FP - Tree



(iii) Vertical Data format

A	T_3
C	$T_4\ T_5$
D	T_2
E	$T_1\ T_2\ T_3\ T_5$
I	T_5
K	$T_1,\ T_2\ T_3\ T_4\ T_5$
M	$T_1\ T_3\ T_4$
N	$T_1,\ T_2$
O	$T_1,\ T_2\ T_3$
U	T_4
Y	$T_1,\ T_2\ T_4$

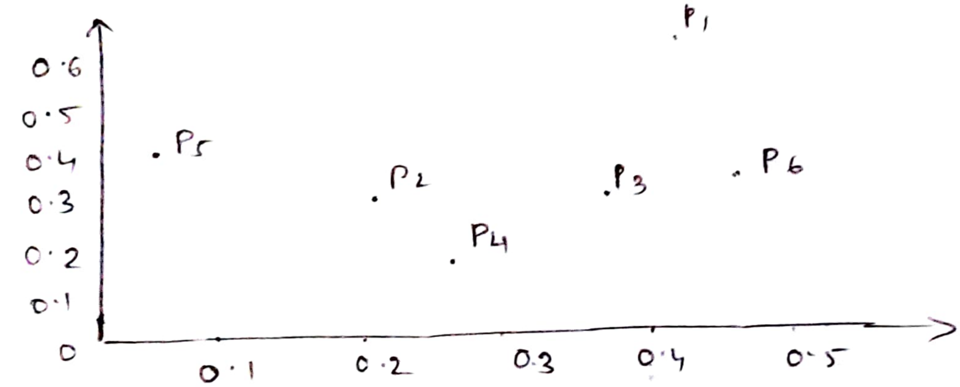
Inter section

AE	T_3
AK	T_3
Am	T_3
CK	$T_4\ T_5$
⋮	
EK	$T_1\ T_2\ T_3\ T_5$
Km	$T_1\ T_3\ T_4$
DK	$T_1\ T_2\ T_5$
KY	$T_1 T_2\ T_4$
EO	$T_1\ T_2\ T_5$

Inter section

EKm	$T_1 T_3$
EOK	$T_1 T_2 T_5$
EKY	$T_1\ T_2$

7. step 1 : plot

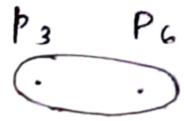


Distance matrix

P1	0					
P2	0.24	0				
P3	0.22	0.15	0			
P4	0.37	0.2	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	$\boxed{0.11}$	0.22	0.39	0
	P1	P2	P3	P4	P5	P6

(P3, P6) has the least distance so make them a cluster

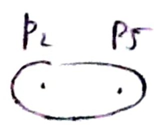
Dendegram



Distance April matrix

P1	0				
P2	0.24	0			
P3 P6	0.22	0.15	0		
P4	0.37	0.28	0.15	0	
P5	0.34	$\boxed{0.14}$	0.28	0.24	0
	P1	P2	P3P6	P4	P5

(P2, P5) is smallest distance so make it a cluster

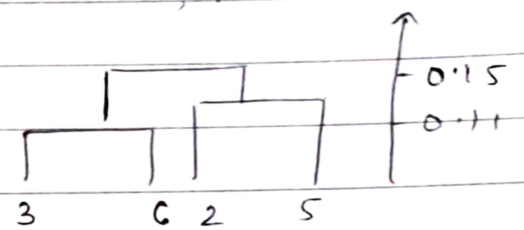


Distance matrix

P_1	0			
$P_2 P_5$	0.24	0		
$P_3 P_6$	0.22	0.15	0	
P_4	0.37	0.20	0.15	0

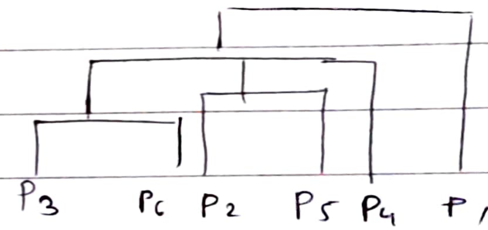
dist 0.15 is min

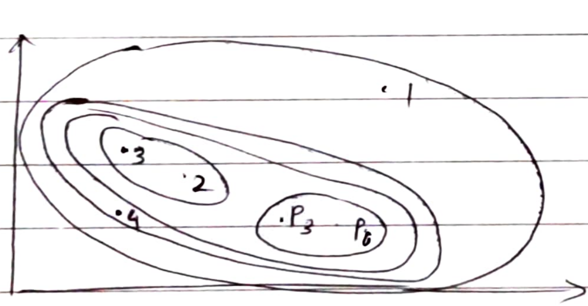
$\therefore P_3 P_6$, $P_2 P_5$ is cluster



Distance matrix

P_1	0		
$P_2 P_5 P_3 P_6$	0.22	0	
P_4	0.37	$\boxed{0.15}$	0
	P_1	P_{--}	P_4





8. We collect a set of attribute value count tables and update counts as each new example streams in. To discover evolution of classification scheme, we can maintain counts for a few classification in parallel. For instance, we can keep one classifier based on history of data, another one based on previous week of data and one based on previous data of data. For, weekly classifiers we can keep count of previous 7 days. At end of each day we discard oldest days count. For daily classifier we maintain separate counts for each hour.