Q1

(i) AB, AC, AD, BC, BD, ~~CD~~, AE
1 2 3 4 5 6 7

3 item set

1 (AB, AC) → ABC ✓
(AB, AD) → ABD ✓
(AB, BC) → ~~ABC~~
(AB, BD) → ABD ✓
(AB, CD) → α
(AB, AE) → ABE ✓

2 AC, AD → ACD ✓
AC, BC → ~~ABC~~
AC, BD → α
AC, CD → ~~ACD~~
AC, AE → ACE ✓

3. AD, BC → α
AD, BD → ~~ABD~~
AD, CD → ~~ACD~~
AD, AE → ADE ✓

4 BC, BD → BCD ✓
BC, CD → ~~BCD~~
BC, AE → α

5. BD, CD → ~~BCD~~
BD, AE → α

6. CD, AE → α

∴ Candidates after the scan are
ABC, ABD, ABE, ACD, ACE, ADE, BCD
(B) is ~~correct~~ closest option

January						2019
S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

(ii) ABC, ABD, ACD, BCD, BCE, CDE
1 2 3 4 5 6

1. ABC, ABD → ABCD ✓
ABC, ACD → ~~ABCD~~
ABC, BCD → ~~ABCD~~
ABC, BCE → ABCE ✓
ABC, CDE → ABCDE ✗

2. ABD, ACD → ~~ABCD~~
ABD, BCD → ~~ABCD~~
ABD, BCE → ABCDE ✗
ABD, CDE → ABCDE ✗

3. ACD, BCD → ~~ABCD~~
ACD, BCE → ABCDE ✗

ACD, CDE → ACDE ✓

4. BCD, BCE → BCDE ✓
BCD, CDE → ~~BCDE~~

5. BCE, CDE → ~~BCDE~~

∴ ABCD, ABCE, ACDE, BCDE are candidates.

Q2

Naive Bayes

Let $C_i = \{YES, NO\}$ $i \in 1,2$
are two classes for stolen & not stolen

We know $\boxed{P(A|B) = \frac{P(B/A)\,P(A)}{P(B)}}$ BAYES RULE is conditional probability ——— ①

Let feature values $\{Red, Domestic, SUV\} = X$

To classify X, we adopt eqn ①.

Posterior prob $P(C_i|X) = \frac{\overbrace{P(X/C_i)}^{\text{likelihood}}\,P(C_i) \text{ — Prior}}{P(X) \text{ — Evidence}}$

$\propto P(X/C_i)P(C_i)$ considering $P(X)$ as scaling factor and is same for all classes

$\propto P(X_1/C_i)\,P(X_2/C_i) - P(X_n/C_i) * P(C_i)$

This is naive bayes assumption, where each X_i is treated as independent.

$$P(C_i/X) \simeq \prod_{j=1}^{p} P(X_j/C_i)$$

We select where distribution is maximize $\arg\max \left(\prod_{j=1}^{p} P(X_j/C_i)\right)$ ——— ②

January						2019
S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

- Given the data

Stolen? ⓐ

YES 5 NO 5

$P(Y) = \frac{5}{10} = \frac{1}{2}$ $P(N) = \frac{5}{10} = \frac{1}{2}$

Colour = Red

Red ⓑ

Y 3 N 2

$P(R|Y) = \frac{3}{5}$ $P(R|N) = \frac{2}{5}$

Type = SUV ⓒ

Y 1 N 3

$P(SUV|Y) = \frac{1}{5}$ $P(SUV|N) = \frac{3}{5}$

Origin = Domestic ⓓ

Y 2 N 3

$P(Dom|Y) = \frac{2}{5}$ $P(Dom|N) = \frac{3}{5}$

(i) $P(X|Y) = ?$

$\Rightarrow P(\{Red, Dom, SUV\} / Stolen = Y)$

$= P(Red|Y) \cdot P(Dom|Y) \cdot P(SUV|Y) \cdot P(Y)$

$= \left(\frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5}\right)\frac{1}{2} = \frac{6}{125} \cdot \frac{1}{2}$ from ⓐ,ⓑ,ⓒ,ⓓ

$= \underline{0.024}$

(ii) $P(X|N) = ?$

$\Rightarrow P(\{Red, Dom, SUV\} / Not\ stolen)$

$= P(Red|N) \cdot P(Dom|N) \cdot P(SUV|N) \cdot P(N)$

$= \left(\frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5}\right) \cdot \frac{1}{2} = 0.036$ from ⓐ,ⓑ,ⓒ,ⓓ

$\therefore P(X/N) > P(X/Y) \Rightarrow$ Classify it as Not stolen from eq. ②

Q3

Continuing with info in Q2

Colour = Yellow (e)

Y: 2, N: 3

$P(\text{Yellow}|Y) = \frac{2}{5}$ $P(\text{Yellow}/N) = \frac{3}{5}$

Type = sports (f)

Y: 4, N: 2

$P(\text{sports}/Y) = \frac{4}{5}$ $P(\text{sports}/N) = \frac{2}{5}$

Orrgin = Imported (g)

Y: 3, N: 2

$P(\text{Import}/Y) = \frac{3}{5}$ $P(\text{Import}/N) = \frac{2}{5}$

$$\boxed{\text{Entropy} = -\sum_{i=1}^{N} p_i \log(p_i)}$$

Information Gain = Entropy of Attribute − Weighted average of entropy of each child set

$$G(T) = E(T) - E(T, X)$$

January						2019
S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Total Entropy E(T)

$$= -\left[\frac{1}{2}\log\left(\frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{1}{2}\right)\right]$$

$$= -\left[\log(1/2)\right]$$

$$= -\log(2^{-1}) = \log_2 2$$

$$= 1$$

Entropy of Colour E(T, colour)

= E(T, Red) E(T, Yellow)

$$= -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} \qquad -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$$

$$= 0.442 + 0.528 \qquad 0.528 + 0.442$$

$$= 0.97 \qquad 0.97$$

Weighted entropy sum

$$W_1 \times E(T, Red) + W_2 \times E(T, Yellow)$$

$$= \frac{5}{10}(0.97) + \frac{5}{10}(0.97)$$

$$= 0.97$$

$$W_1 = \frac{5 \text{ Reds}}{\text{Records}} = \frac{5}{10}$$

$$W_2 = \frac{5 \text{ Yellow}}{\text{Records}} = \frac{5}{10}$$

Info gain colour $= 1 - 0.97$

$= 0.03$

Entropy of Type E(T, Type)

$= E(T, Sports)$

$= -\frac{4}{6}\log(\frac{4}{6}) - \frac{2}{6}\log(\frac{2}{6})$

$= 0.918$

$E(T, SUV)$

$-\frac{1}{4}\log(\frac{1}{5}) - \frac{3}{4}\log(3/6)$

$= \cancel{0.906}$

$= 0.81127$

Weighted entropy sum

$= W_1 * E(T, Sports) + W_2 * E(T, SUV)$

$= \frac{6}{10}(0.918) + \frac{4}{10}(0.81127)$

$= 0.8753$

Info Gain Type $= 1 - 0.853$

$= 0.124$

Entropy of Origin E(T, Origin)

E(T, Domestic

$-\frac{2}{5}\log(\frac{2}{5}) - \frac{3}{5}\log(\frac{3}{5})$

$= 0.97$

E(T, Imported)

$-\frac{3}{5}\log(3/5) - \frac{2}{5}\log(\frac{2}{5})$

$= 0.97$

Weighted entropy sum

$= W_1 * E(T, Domestic) + W_2 * E(T, Imported)$

$= \frac{5}{10}(0.97) + \frac{5}{10}(0.97$

$= 0.97$

January						2019
S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Info Gain = 0.03

Example Info gain

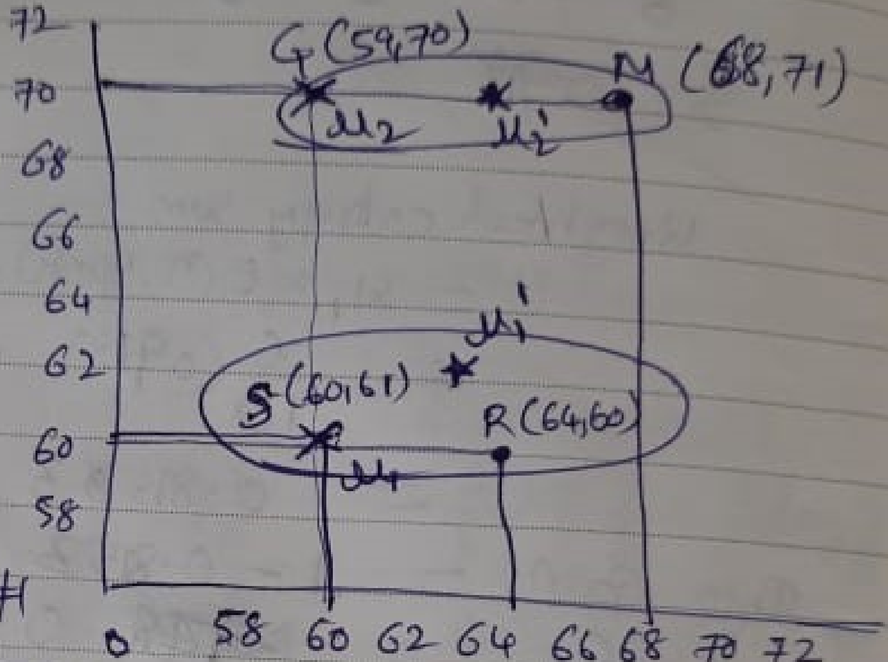
$= \{10 * \Sigma -1 \log 1 + 0 \log 0\}$

$= 0$

∴ Info gain

- Level 1 entropy 1
- Example $1-0=$
- Colour 0.03
- Type ~~0.12~~ 0.14
- Origin 0.03

Q4 :-



K = 1, iteration #1

$(R,S) = \sqrt{(64-60)^2 + (60-61)^2}$

$= \sqrt{4^2 + 1} \approx 4$ approx

$(G,M) = \sqrt{(68-59)^2 + (71-70)^2}$

$= \sqrt{9^2 + 1^2} \approx 9$ approx

$(S,M) = \sqrt{(68-60)^2 + (71-61)^2}$

$= \sqrt{8^2 + 10^2}$

$(S,M) > (G,M)$

$(S,M) > (R,S)$

∴ No of clusters = 2

Cluster 1 = (R,S) & Cluster 2 = (G,M)

(ii) Clustering quality:- $E = \sum_{i=1}^{N} \sum_{x \in C_i} d(x, m_i)^2$

$\underline{k=2}$ Update centroids be μ_1', μ_2'

$\mu_1' \leftarrow \frac{60+64}{2}, \frac{61+60}{2}$

$\mu_1' \leftarrow (62, 60.5)$

$\mu_2^2 \leftarrow \frac{59+68}{2}, \frac{70+71}{2}$

$\mu_2^2 \leftarrow (63.5, 70.5)$

Iteration #2

$(S, \mu_1') = \sqrt{(62-60)^2 + (60.5-61)^2}$

$= \sqrt{2^2 + (\frac{1}{2})^2} = \sqrt{4+0.25} = \approx 2$

$(S, \mu_2') = \sqrt{(63.5-60)^2 + (70.5-61)^2}$

$= \sqrt{(3.5)^2 + (9.5)^2}$

$(S, \mu_2') >> (S, \mu_1')$

$\Rightarrow S \in \mu_1'$

Ily $(S, R) \in$ cluster 1

&

$(G, M) \in$ cluster 2

January						2019
S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Clustering quality can be measured by

MSE – Mean Squared error

$$E = \sum_{i=1}^{N} \sum_{x \in C_i} d(x, m_i)^2$$

where
x – sample
m_i – mean of cluster i
C_i – cluster i

$(S, R) \in$ Cluster #1 with mean μ_1' (62, 60.5)

$(G, M) \in$ Cluster #2 with mean μ_2' (63.5, 70.5)

$MSE = (S, \mu_1') + (R, \mu_1') + (G, \mu_2') + (M, \mu_2')$

$$MSE(E) = (62-60)^2 + (60.5-61)^2 + (62-64)^2 + (60.5-60)^2 + (63.5-59)^2 + (70.5-70)^2 + (63.5-68)^2 + (70.5-70)^2$$

$$= 2^2 + (0.5)^2 + 2^2 + (0.5)^2 + (4.5)^2 + (0.5)^2 + (4.5)^2 + (0.5)^2$$

$$= 8 + 4(0.5)^2 + 2(4.5)^2$$

$$= 8 + 4(0.25) + 2(20.25)$$

$$\therefore MSE = 8 + 1 + 40.5 = 49.5$$

We can also use mean absolute error.

$$MAE = 2 + 0.5 + 2 + 0.5 + 4.5 + 0.5 + 4.5 + 0.5$$

$$= 15$$