4a. For the transactions shown in the table compute the following.

(i) Entropy of the collection of transaction records of the table with respect to classification.

(ii) what are the information gain of a_1 and a_2 relative to the transactions of the table?

Instance	1	2	3	4	5	6	7	8	9
a_1	T	T	T	F	F	F	F	T	F
a_2	T	T	F	F	T	T	F	F	T
Target class	+	+	−	+	−	−	−	+	−

Solution:

* $$\text{Entropy}(S) \equiv -p_{\oplus}\log_2 p_{\oplus} - p_{\ominus}\log_2 p_{\ominus}$$

* $$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$$

(i) Here, there are 9 instances out of which

* 4 are positive instances
* 5 are Negative instances

$$\text{Entropy}(4+, 5-) = -\left(\frac{4}{9}\right)\log_2\left(\frac{4}{9}\right) - \left(\frac{5}{9}\right)\log_2\left(\frac{5}{9}\right)$$

$$= \underline{0.9910}$$

So, $\boxed{\text{Entropy}(S) = 0.9910}$

(ii) * **For attribute a_1,**

Target class

Values

a_1	+	−
T	3	1
F	1	4

find the entropy for a_1. ie., $\frac{|S_v|}{|S|} \text{Entropy}(S_v)$

$$= \frac{4}{9}\left[-\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right)\right] +$$

$$\frac{5}{9}\left[-\left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right) - \left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right)\right]$$

$= 0.3605 + 0.4010$

$= \underline{0.7615}$ substitute in gain formula.

ie,

$$\text{Gain}(S, a_1) = \text{Entropy}(s) - \sum_{v \in \text{Values}(a_1)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$= 0.9910 - 0.7615$$

$$\boxed{\text{Gain}(S, a_1) = 0.2295}$$

* **For attribute a_2**

Target class

Values

a_2	+	−
T	2	3
F	2	2

find the entropy for a_2 ie., $\frac{|S_v|}{|S|} \text{Entropy}(S_v)$

$$= \frac{5}{9}\left[-\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)\right] + \frac{4}{9}\left[1\right]$$

∵ Equal no. of +ve & -ve intances

$$= 0.5394 + 0.4444$$

$= \underline{0.9838}$ Substitute in Information Gain formula

ie., $$\text{Gain}(S, a_2) = \text{Entropy}(S) - \sum_{v \in \text{Values}(a_2)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$= 0.9910 - 0.9838$$

$$\boxed{\text{Gain}(S, a_2) = 0.0072}$$

7. c. Consider a football game between two rival teams: Team0 and Team1. Suppose Team0 wins 95% of the time and Team1 wins the remaining matches. Among the games won by Team0, only 30% of them come from playing on Team1's football field. On the otherhand, 75% of the victories for Team1 are obtained while playing at home. If Team1 is to host the next match between the two teams, which team will most likely emerge as the winner?

Solution:

- Probability that Team0 wins is $P(Y_0) = \underline{0.95}$
- Probability that Team1 wins is $P(Y_1) = 1 - P(Y_A)$

$$= 1 - 0.95$$

$$= \underline{0.05}$$

- Probability that Team1 hosted the match it had won is $P(X_1 | Y_1) = \underline{0.75}$
- probability that Team1 hosted the match won by Team0 is $P(X_1 | Y_0) = \underline{0.30}$

The problem can be solved by computing $P(Y_1 | X_1)$, which is the conditional probability that Team1 wins the next match it hosts.

Using Bayes theorem,

$$P(Y_1|X_1) = \frac{P(X_1|Y_1) * P(Y_1)}{P(X_1)}$$

$$= \frac{P(X_1|Y_1) * P(Y_1)}{P(X_1|Y_1) * P(Y_1) + P(X_1|Y_0) P(Y_0)}$$

$$= \frac{0.75 * 0.05}{(0.75 * 0.05) + (0.30 * 0.95)}$$

$$P(Y_1|X_1) = \underline{\underline{0.1162}}$$

$$P(Y_0|X_1) = 1 - P(Y_1|X_1) = 1 - 0.1162$$

$$= \underline{\underline{0.8838}}$$

Since $P(Y_1|X_1) < P(Y_0|X_1)$

Team0 has a better probability of winning than Team1

Note: Here, there are two events - win & host

① win ← denoted by Y

② Host ← denoted by X

③ Team0 ← denoted by 0

④ Team1 ← denoted by 1

8. c The following table gives data set about stolen vehicles. Using Naïve Bayes classifier classifythe new data {Color = Red, Type = SUV, Origin =Domestic}

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

$$y' = \underset{y \in Y}{\operatorname{argmax}} \, P(y) \prod_{i=1}^{m} P(x_i | y)$$

(Attributes, Values) ⇒ (Color | Red, yellow)
(Type | Sports, SUV)
(Origin | Domestic, Imported)

Target class

Values

Color	Yes	No
Red	3	2
Yellow	2	3

$P(\text{Color} = \text{Red} | \text{Stolen} = \text{yes}) = \frac{3}{5} = 0.6$

$P(\text{color} = \text{Red} | \text{Stolen} = \text{No}) = \frac{2}{5} = 0.4$

$P(\text{color} = \text{yellow} | \text{Stolen} = \text{yes}) = \frac{2}{5} = 0.4$

$P(\text{Color} = \text{yellow} | \text{Stolen} = \text{No}) = \frac{3}{5} = 0.6$

Type	Yes	No
Sport	4	2
SUV	1	3

$P(\text{Type} = \text{Sport} | \text{Stolen} = \text{yes}) = 4/5 = 0.8$

$P(\text{Type} = \text{Sport} | \text{Stolen} = \text{No}) = 2/5 = 0.4$

$P(\text{Type} = \text{SUV} | \text{stolen} = \text{yes}) = 1/5 = 0.2$

$P(\text{Type} = \text{SUV} | \text{Stolen} = \text{No}) = 3/5 = 0.6$

Target

Value	Origin	yes	No
	Domestic	2	3
	Imported	3	2

P(Origin = Domestic | Stolen = yes) = 2/5 = 0.4

P(origin = Domestic | stolen = No) = 3/5 = 0.6

P(origin = Imported | Stolen = yes) = 3/5 = 0.6

P(origin = Imported | Stolen = no) = 2/5 = 0.4

Classify the new data = (Red, SUV, Domestic)

* For Stolen = yes :

⇒ (color = Red | stolen = yes) * (Type = SUV | stolen = yes) * (origin = Domestic | Stolem = yes) * P(yes)

⇒ 0.6 * 0.2 * 0.4 * 0.5

⇒ 0.024

* For Stolen = No :

⇒ (color = Red | Stolen = No) * (Type = SUV | stolen = No) * (origin = Domestic | stolen = No) * P(NO)

⇒ 0.4 * 0.6 * 0.6 * 0.5

⇒ 0.072

So, we would classify the new data as not Stolen