Assignment 5

CS 4783/5783

Shafi-Al-Salman Romeo

Consider the following data about the possible taste of a particular fruit based on some visual characteristics.

Taste	Farm climate	Visual defects	Size
Meh	Warm	Some	Small
Meh	Cold	None	Large
Meh	Cold	None	Large
Yummy	Cold	Many	Small
Yummy	Warm	Many	Small
Meh	Warm	Some	Large
Yummy	Warm	Many	Large
Yummy	Cold	None	Small
Yummy	Cold	None	Small
Meh	Warm	Some	Large

Answer the following questions based on your understanding of decision trees and Naïve Bayes.

[Question 1]

Support you want to build a decision tree. What is the initial entropy of the target variable taste?

[Answer 1]

As we can see from the table, there are 5 Meh and 5 Yummy out of 10. Therefore, initial entropy of the target variable taste is:

$$H(y) = -\sum_{i=1}^{k} P(y = y_i)\, log_2 P(y = y_i)$$

Here,

$$P(taste = Meh) = \frac{5}{10}$$

$$P(taste = Yummy) = \frac{5}{10}$$

So,

$$H(taste) = -\frac{5}{10} log_2\left(\frac{5}{10}\right) - \frac{5}{10} log_2\left(\frac{5}{10}\right) = 1$$

[Question 2]

Consider that the variable Visual defects is chosen as the root of the decision tree. What is the information gain (IG) of the decision tree?

[Answer 2]

We consider that the variable Visual defects is chosen as the root of the decision tree. To compute the information gain of the decision tree, entropy of the target variable is needed.

Visual defects		Taste		
		Meh	Yummy	Total
	Some	3	0	3
	None	2	2	4
	Many	0	3	3
Total				10

Therefore, entropy of Visual defects is:

$$H(Taste\ \mathrm{l}\ Visual\ defects) = \frac{3}{10}H(0{,}3) + \frac{4}{10}\left\{-\frac{2}{4}log_2\left(\frac{2}{4}\right) - \frac{2}{4}log_2\left(\frac{2}{4}\right)\right\} - \frac{3}{10}H(3{,}0)$$

$$H(Taste\ \mathrm{l}\ Visual\ defects) = 0 + 0.4 \times 1 + 0\ = 0.4$$

$$IG(Visual\ defects) = H(Taste) - H(Taste\ \mathrm{l}\ Visual\ defects) = 1 - 0.4 = 0.6$$

<u>Check:</u>

Entropy of Farm climate is:

Farm climate		Taste		
		Meh	Yummy	Total
	Warm	3	2	5
	Cold	2	3	5
Total				10

$$H(Taste\ \mathrm{l}\ Farm\ climate) = \frac{5}{10}\left\{-\frac{3}{5}log_2\left(\frac{3}{5}\right) - \frac{2}{5}log_2\left(\frac{2}{5}\right)\right\} + \frac{5}{10}\left\{-\frac{2}{5}log_2\left(\frac{2}{5}\right) - \frac{3}{5}log_2\left(\frac{3}{5}\right)\right\}$$

$$H(Taste\ \mathrm{l}\ Farm\ climate) = 0.5 \times (0.52 + 0.44) + 0.5 \times (0.44 + 0.52) = 0.96$$

$$IG(Farm\ climate) = H(Taste) - H(Taste\ \mathrm{l}\ Farm\ climate) = 1 - 0.96 = 0.04$$

Entropy of Size is:

Size		Taste		
		Meh	Yummy	Total
	Small	1	4	5
	Large	4	1	5
Total				10

$$H(Taste\ \mathrm{l\ Size}) = \frac{5}{10}\left\{-\frac{1}{5}log_2\left(\frac{1}{5}\right)-\frac{4}{5}log_2\left(\frac{4}{5}\right)\right\}+\frac{5}{10}\left\{-\frac{4}{5}log_2\left(\frac{4}{5}\right)-\frac{1}{5}log_2\left(\frac{1}{5}\right)\right\}$$

$$H(Taste\ \mathrm{l\ Size}) = 0.5 \times (0.46 + 0.25) + 0.5 \times (0.25 + 0.46) = 0.71$$

$$IG(Size) = H(Taste) - H(Taste\ \mathrm{l}\ Size) = 1 - 0.71 = 0.29$$

So, we can see that Visual defects has the largest information gain of the decision tree. So, we can choose it as the root of the decision tree. Information gain of the decision tree of root node is 0.6.

[Question 3]

What is entropy H(Taste|Visual Defect == Some) and the entropy H(Taste|Visual Defect == None)?

[Answer 3]

		Taste		
		Meh	Yummy	Total
Visual defects	Some	3	0	3
	None	2	2	4
	Many	0	3	3
Total				10

Entropy of $H(Taste\ \mathrm{l\ Visual\ Defect\ \ ==\ \ Some})$ is:

$$\frac{3}{10}\left\{-\frac{3}{3}log_2\left(\frac{3}{3}\right)-\frac{0}{3}log_2\left(\frac{0}{3}\right)\right\} = 0$$

Entropy of $H(Taste\ \mathrm{l\ Visual\ Defect\ \ ==\ \ None})$ is:

$$\frac{4}{10}\left\{-\frac{2}{4}log_2\left(\frac{2}{4}\right)-\frac{2}{4}log_2\left(\frac{2}{4}\right)\right\} = 0.4$$