Data Mining; Assignemt 4

Jesse Annan | ID: 002708111

March 29, 2023

1. **Question 1**
First we need to determine the root node - the feature with most information gain.
This is obtained by calculating the weighted entropy of each attribute and taking it out of the entropy of the datasets (Grass).
Step 1: Determining Root Node:

$$\begin{aligned}\mathbf{H}(grass) &= -\sum_{i\in\{dry,wet\}} p(i)\cdot\log_2 p(i)\\ &= -((9/16)\times\log_2(9/16)+(7/16)\times\log_2(7/16)) = .9887\end{aligned}$$

$$\begin{aligned}\mathbf{H}(sprinkler=yes|grass) &= \sum_{i\in\{dry,wet\}} p(i)\cdot\log_2 p(i)\\ &= -((1/6)\times\log_2(1/6)+(5/6)\times\log_2(5/6)) = .6500\\ \mathbf{H}(sprinkler=no|grass) &= \sum_{i\in\{dry,wet\}} p(i)\cdot\log_2 p(i\\ &= -((8/10)\times\log_2(8/10)+(2/10)\times\log_2(2/10)) = .7219\\ \mathbf{InfoGain}(Sprinkler) &= \frac{\#sprinkler{=}yes}{n}\times\mathbf{H}(sprinkler=yes|grass)\\ &+\frac{\#sprinkler{=}no}{n}\times\mathbf{H}(sprinkler=no|grass)\\ &= \frac{6}{16}\times.6500+\frac{10}{16}\times.7219 = .2937\end{aligned}$$

$$\begin{aligned}\mathbf{H}(rain=yes|grass) &= \sum_{i\in\{dry,wet\}} p(i)\cdot\log_2 p(i)\\ &= -((1/4)\times\log_2(1/4)+(3/4)\times\log_2(3/4)) = .8113\\ \mathbf{H}(rain=no|grass) &= \sum_{i\in\{dry,wet\}} p(i)\cdot\log_2 p(i\\ &= -((8/12)\times\log_2(8/12)+(4/12)\times\log_2(4/12)) = .9183\\ \mathbf{InfoGain}(rain) &= \frac{\#rain{=}yes}{n}\times\mathbf{H}(rain=yes|grass)\\ &+\frac{\#rain{=}no}{n}\times\mathbf{H}(rain=no|grass)\\ &= \frac{4}{16}\times.8113+\frac{12}{16}\times.9183 = .0972\end{aligned}$$

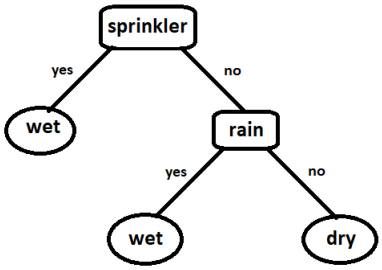
Since **InfoGain**(sprinkler) is greater than the **InfoGain**(rain), we select sprinkler as our parent node.

Step 2: Determining Sub Root Node:
Now, we determine the sub root for our parent root i.e given sprinkler what is the information gain

$$\mathbf{H}(rain = yes|grass, sprinkler = no) = \sum_{i \in \{dry,wet\}} p(i) \cdot \log_2 p(i)$$
$$= -((1/3) \times \log_2(1/3) + (2/3) \times \log_2(2/3)) = .9183$$
$$\mathbf{H}(rain = no|grass, sprinkler = no) = \sum_{i \in \{dry,wet\}} p(i) \cdot \log_2 p(i$$
$$= -(7/7) \times \log_2(7/7) = 0$$

$$\mathbf{H}(rain = yes|grass, sprinkler = yes) = \sum_{i \in \{dry,wet\}} p(i) \cdot \log_2 p(i)$$
$$= (1/1) \times \log_2(1/1) = 0$$
$$\mathbf{H}(rain = no|grass, sprinkler = yes) = \sum_{i \in \{dry,wet\}} p(i) \cdot \log_2 p(i)$$
$$= -((1/5) \times \log_2(1/5) + (4/5) \times \log_2(4/5)) = .7219$$

Whenever we find $\mathbf{H} = 0$, it implies that there is one distinct class in the particular sample. Therefore, from our calculations, $\mathbf{H}(rain = yes|grass, sprinkler = yes) = 0$ means that whenever the sprinkler is on and it rains then the grass is wet similarly $\mathbf{H}(rain = no|grass, sprinkler = no) = 0$ means that whenever the sprinkler is off and it isn't raining then the grass is dry.



Also with the calculations we can conclude that there's 4/5 chance that the grass is wet if sprinkler is off and it's raining and also there's 2/3 chance that the grass is wet if the sprinkler is on and it's raining. Hence the decision tree above is the best three that David can obtain using information gain on **Table 1**.

2. **Question 2**

Actual	Wet	Dry	Dry	Wet	Wet	Dry	Dry	Wet	Wet	Dry
Predicted	Dry	Dry	Wet	Wet	Wet	Wet	Dry	Wet	Wet	Dry

Confusion Matrix

		Predicted		
		Wet	**dry**	**Total**
Actual	**Wet**	4	1	5
	Dry	2	3	5
	Total	6	4	10

Accuracy Values

Classification Accuracy	$\frac{4+3}{10} = .7$
Error rate	$\frac{2+1}{10} = .3$
Sensitivity	$\frac{4}{5} = .8$
Precision	$\frac{4}{6} = .6667$
Recall	$\frac{4}{5} = .8$
F-score	$\frac{2 \times \frac{4}{6} \times \frac{4}{5}}{\frac{4}{6} + \frac{4}{5}} = .7273$

3. **Question 3**
Given $\mathbb{X} = \{Rain = No, Sprinkler = Yes\}$ we want to find
$max\ \mathbf{P}(\mathbb{C}_i|\mathbb{X}_j)\ \ \alpha\ \ \mathbf{P}(\mathbb{X}_j|\mathbb{C}_i) \cdot \mathbf{P}(\mathbb{C}_i)$
$\Rightarrow \mathbf{P}(grass|\mathbb{X}_j)\ \ \alpha\ \ \mathbf{P}(\mathbb{X}_j|grass) \cdot \mathbf{P}(grass)$

$$\mathbf{P}(grass = wet) = \frac{7}{16}$$
$$\mathbf{P}(grass = dry) = \frac{9}{16}$$
$$\mathbf{P}(rain = no|grass = wet) = \frac{4}{7}$$
$$\mathbf{P}(rain = no|grass = dry) = \frac{8}{9}$$
$$\mathbf{P}(sprinkler = yes|grass = wet) = \frac{5}{7}$$
$$\mathbf{P}(sprinkler = yes|grass = dry) = \frac{1}{7}$$
$$\mathbf{P}(\mathbb{X}|grass = wet) = \mathbf{P}(rain = no|grass = wet) \times \mathbf{P}(sprinkler = yes|grass = wet) = \frac{20}{49}$$
$$\mathbf{P}(\mathbb{X}|grass = dry) = \mathbf{P}(rain = no|grass = dry) \times \mathbf{P}(sprinkler = yes|grass = dry) = \frac{8}{81}$$
$$\mathbf{P}(grass = wet|\mathbb{X}) = \mathbf{P}(\mathbb{X}|grass = wet) \times \mathbf{P}(grass = wet) = \frac{5}{28}$$
$$\mathbf{P}(grass = dry|\mathbb{X}) = \mathbf{P}(\mathbb{X}|grass = dry) \times \mathbf{P}(grass = dry) = \frac{1}{18}$$

Since the value of $\mathbf{P}(grass = wet|\mathbb{X}) > \mathbf{P}(grass = dry|\mathbb{X})$ our Naïve Bayesian classifier will predict
Grass = wet for $\mathbb{X} = \{Rain = No, Sprinkler = Yes\}$

4. **Question 4**

P(Rain)

no	$\frac{12}{16}$
yes	$\frac{4}{16}$

P(Sprinkler | Rain)

		Sprinkler	
		no	**yes**
Rain	**no**	$\frac{7}{12}$	$\frac{5}{12}$
	yes	$\frac{3}{4}$	$\frac{1}{4}$

P(Grass |Rain, Sprinkler)

		Grass	
Rain	**Sprinkler**	**Wet**	**Dry**
No	**No**	0	1
No	**Yes**	$\frac{4}{5}$	$\frac{1}{5}$
Yes	**No**	$\frac{2}{3}$	$\frac{1}{3}$
Yes	**Yes**	1	0