CSCI-GA.3033-061 Predictive Analytics August 19, 2022

Solutions to Problem 2 of Final Exam Summer 2022 (45 Points)

Name: Anav Prasad (ap7152; N18439284) *Due: 11:59pm on Saturday, August 20*

Collaborators:

Apply a decision Tree algorithm to derive the decision tree learned from the following dataset (25pts)

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W4	Rainy	Yes	Rich	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema

Show all steps.

Explain how you would you apply Random Forests to the same dataset (briefly describe the algorithm being applied (20pts)), you do not need to apply the algorithm.

Solution:
First of all, let's compute the entropy of *Decision*.

$$\begin{aligned} Entropy(Decision) &= -\frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) - \frac{1}{5} \cdot \log_2\left(\frac{1}{5}\right) - \frac{1}{5} \cdot \log_2\left(\frac{1}{5}\right) \\ &\approx 1.371 \end{aligned}$$

- **Step 1:**
  Now, let's compute the entropy and information gains of the various columns:

- Weather:

$$
\begin{aligned}
Entropy(Sunny) &= -\frac{1}{2} \cdot \log\left(\frac{1}{2}\right) - \frac{1}{2} \cdot \log\left(\frac{1}{2}\right) \\
&= 1 \\
Entropy(Rainy) &= -\frac{2}{3} \cdot \log\left(\frac{2}{3}\right) - \frac{1}{3} \cdot \log\left(\frac{1}{3}\right) \\
&\approx 0.918 \\
Entropy(Weather) &= \frac{2}{5} \cdot 1 + \frac{3}{5} \cdot 0.918 \\
&= 0.9508 \\
IG(Weather, Decision) &= Entropy(Decision) - Entropy(Weather) \\
&= 1.371 - 0.9508 \\
&= 0.4202
\end{aligned}
$$

- Parents:

$$
\begin{aligned}
Entropy(Yes) &= -\frac{3}{3} \cdot \log\left(\frac{3}{3}\right) \\
&= 0 \\
Entropy(No) &= -\frac{1}{2} \cdot \log\left(\frac{1}{2}\right) - \frac{1}{2} \cdot \log\left(\frac{1}{2}\right) \\
&= 1 \\
Entropy(Parents) &= \frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 1 \\
&= 0.4 \\
IG(Parents, Decision) &= Entropy(Decision) - Entropy(Parents) \\
&= 1.371 - 0.4 \\
&= 0.971
\end{aligned}
$$

- Money:

$$\begin{aligned}
Entropy(Rich) &= -\frac{2}{4}\cdot\log\left(\frac{2}{4}\right) - \frac{1}{4}\cdot\log\left(\frac{1}{4}\right) - \frac{1}{4}\cdot\log\left(\frac{1}{4}\right) \\
&= 1.5 \\
Entropy(Poor) &= -\frac{1}{1}\cdot\log\left(\frac{1}{1}\right) \\
&= 0 \\
Entropy(Money) &= \frac{4}{5}\cdot 1.5 + \frac{1}{5}\cdot 0 \\
&= 1.2 \\
IG(Money, Decision) &= Entropy(Decision) - Entropy(Money) \\
&= 1.371 - 1.2 \\
&= 0.171
\end{aligned}$$

Since the Information Gain of *Parents* is the highest, I'm adding it as the first node in the decision tree.

$$\begin{array}{c} \text{Parents} \\ \swarrow \quad \searrow \\ \text{Yes} \qquad \text{No} \end{array}$$

- **Step 2:**
First, let's compute entropy of $Decision|Yes$ and $Decision|No$

$$\begin{aligned}
Entropy(Decision|Yes) &= -\frac{3}{3}\cdot\log\left(\frac{3}{3}\right) \\
&= 0 \\
Entropy(Decision|No) &= -\frac{1}{2}\cdot\log\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\log\left(\frac{1}{2}\right) \\
&= 1
\end{aligned}$$

Now, let's compute the entropy and information gains of the various columns:

- $Weather|Yes$:

$$\begin{aligned}
Entropy(Sunny|Yes) &= -\frac{1}{1} \cdot \log\left(\frac{1}{1}\right) \\
&= 0 \\
Entropy(Rainy|Yes) &= -\frac{2}{2} \cdot \log\left(\frac{2}{2}\right) \\
&= 0 \\
Entropy(Weather|Yes) &= \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 0 \\
&= 0 \\
IG(Weather, Decision|Yes) &= Entropy(Decision|Yes) - Entropy(Weather|Yes) \\
&= 0 - 0 \\
&= 0
\end{aligned}$$

- $Money|Yes$:

$$\begin{aligned}
Entropy(Rich|Yes) &= -\frac{2}{2} \cdot \log\left(\frac{2}{2}\right) \\
&= 0 \\
Entropy(Poor|Yes) &= -\frac{1}{1} \cdot \log\left(\frac{1}{1}\right) \\
&= 0 \\
Entropy(Money|Yes) &= \frac{2}{3} \cdot 0 + \frac{1}{3} \cdot 0 \\
&= 0 \\
IG(Money, Decision|Yes) &= Entropy(Decision|Yes) - Entropy(Money|Yes) \\
&= 0 - 0 \\
&= 0
\end{aligned}$$

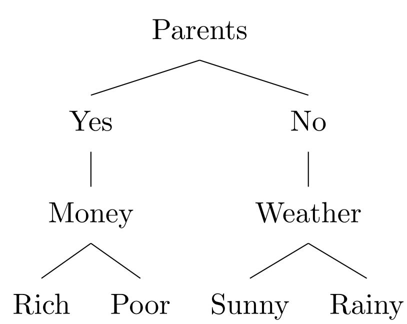
- $Weather|No$:

$$\begin{aligned}
Entropy(Sunny|No) &= -\frac{1}{1}\cdot\log\left(\frac{1}{1}\right)\\
&= 0\\
Entropy(Rainy|Yes) &= -\frac{1}{1}\cdot\log\left(\frac{1}{1}\right)\\
&= 0\\
Entropy(Weather|No) &= \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0\\
&= 0\\
IG(Weather, Decision|No) &= Entropy(Decision|No) - Entropy(Weather|No)\\
&= 1 - 0\\
&= 1
\end{aligned}$$

- $Money|No$:

$$\begin{aligned}
Entropy(Rich|No) &= -\frac{1}{2}\cdot\log\left(\frac{1}{2}\right)\\
&= 1\\
Entropy(Money|No) &= \frac{2}{2}\cdot 1\\
&= 1\\
IG(Money, Decision|No) &= Entropy(Decision|No) - Entropy(Money|No)\\
&= 1 - 1\\
&= 0
\end{aligned}$$

So, we can arbitrarily pick *Money* to go under **Yes** (because the information gain for it is 0 in both cases).
We must pick *Weather* to go under *No* because of its higher information gain.

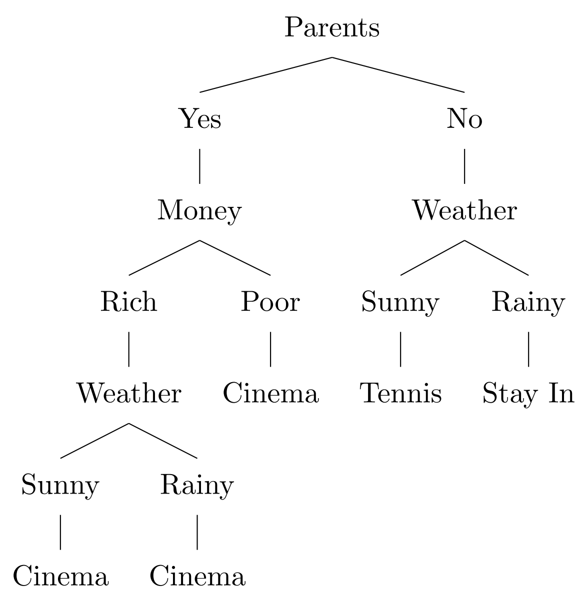


- **Step 3:**
At this point, we only face one node choice at each branch; so, we can easily complete the

tree.

$$
\begin{aligned}
(Decision|Yes, Rich, Weather = Sunny) &\rightarrow Cinema \\
(Decision|Yes, Rich, Weather = Rainy) &\rightarrow Cinema \\
(Decision|Yes, Poor, Weather = Sunny) &\rightarrow NA \\
(Decision|Yes, Poor, Weather = Rainy) &\rightarrow Cinema \\
(Decision|No, Sunny, Money = Rich) &\rightarrow Tennis \\
(Decision|No, Sunny, Money = Poor) &\rightarrow NA \\
(Decision|No, Rainy, Money = Rich) &\rightarrow StayIn \\
(Decision|No, Sunny, Money = Poor) &\rightarrow NA
\end{aligned}
$$

So, the final decision tree would be like as follows:



□