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

Solutions to Problem 1 of Final Exam Summer 2022 (125 Points)

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

Collaborators:

Unsupervised Learning Algorithms

Consider the following supervised learning algorithms (most of which has been adopted in your project and some were discussed in class)

- Support Vector Machines (25pts)
- Decision Trees (25pts)
- Random Forest (25pts)
- Neural Nets (25pts)
- KNN (25pts)

For each algorithm answer the following points, feel free to cite your sources if you used any.

1. Briefly describe the algorithm
2. Advantages of the algorithm
3. Disadvantage of the algorithm

Solution:

- **Support Vector Machines** (25pts)
  - **Description:**
    Support Vector Machines (SVMs) are supervised machine learning algorithms used for classification and regression problems. Given a labelled training dataset, they work by training and perfecting a hyperplane that tries to perfectly classify the training dataset into two classes. The trained hyperplane can then be used to classify the test dataset by just plotting the points and seeing on which side of the hyperplane they fall.
  - **Advantages:**
    * More effective in high dimensional spaces (that is when number of features are high).
    * Subsequently, also more effective when the number of dimensions (or features) are greater than the number of test samples.
    * Works well when the classes' data points, when plotted, land far from each other (making the drawing of a hyperplane easier).
    * Memory efficient as it uses a subset of training points (called support vectors) in the decision function to classify test points.

    - Different kernel functions (to increase to the number of dimensions so as to make the SVM more effective) can be specified for the decision functions and, its possible to specify custom kernels.
    - Also, SVMs are largely quite stable since small tweaks and minor changes to the data points does not greatly affect the hyperplane and, thus, its accuracy.
  - **Disadvantages:**
    - Not suitable for large datasets.
    - Since SVMs work by drawing hyperplanes through the data to classify it, it does not work well in cases when the data has more noise and, the different classes' data points are intermixed so thoroughly that even the use of custom kernels cannot properly separate them.
    - It might be apparent from the last point but the choosing and finding of an appropriate kernel is also quite difficult.
    - SVMs tend to under-perform when the number of features end up actually exceeding the number of training data points.
    -
  - **Sources:** Geeks For Geeks

- **Decision Trees** (25pts)
  - **Description:**
  - **Advantages:**
  - **Disadvantages:**

- **Random Forest** (25pts)
  - **Description:**
  - **Advantages:**
  - **Disadvantages:**

- **Neural Nets** (25pts)
  - **Description:**
  - **Advantages:**
  - **Disadvantages:**

- **kNN** (25pts)
  - **Description:**
  - **Advantages:**
  - **Disadvantages:**

□

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:

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

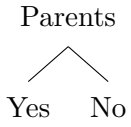
– Parents:

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

– 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.



- **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:

– $Weather|Yes$:

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

– $Money|Yes$:

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

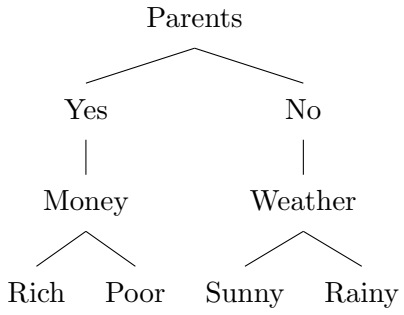
– $Weather|No$:

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

– $Money|No$:

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

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.

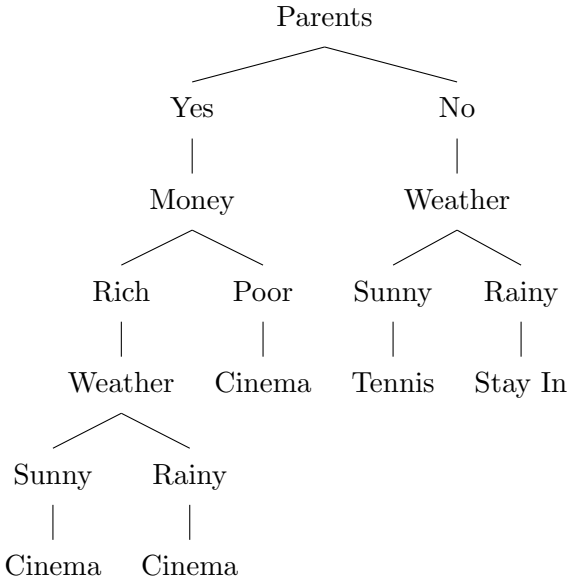


- **Step 3:**
  At this point, we only face one node choice at each branch; so, we can easily complete the

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:



□