SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CS508 - BIG DATA ANALYTICS

III YEAR / V SEMESTER

Unit 2- CLUSTERING AND CLASSIFICATION

Topic 1 : The General Algorithm and Decision Tree Algorithms

How does the Decision Tree algorithm Work?

•**Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.

•**Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**

•**Step-3:** Divide the S into subsets that contains possible values for the best attributes.

•**Step-4:** Generate the decision tree node, which contains the best attribute.

•**Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Algorithm

4.6 Pseudo-code for the decision tree learner algorithm

Algorithm 1 Decision Tree Learner (examples, features)

```
1: if all examples are in the same class then
2:     return the class label.
3: else if no features left then
4:     return the majority decision.
5: else if no examples left then
6:     return the majority decision at the parent node.
7: else
8:     choose a feature f.
9:     for each value v of feature f do
10:        build edge with label v.
11:        build sub-tree using examples where the value of f is v.
```

Problem

Consider whether a dataset based on which we will determine whether to play football or not.

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Problem Here There are for independent variables to determine the dependent variable. The independent variables are Outlook, Temperature, Humidity, and Wind. The dependent variable is whether to play football or not.

https://www.shiksha.com/online-courses/articles/understanding-decision-tree-algorithm-in-machine-learning/

Problem and Solution

1.Entropy

In machine learning, entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Find the entropy of the class variable.

E(S) = -[(9/14)log(9/14) + (5/14)log(5/14)] = 0.94

Problem and Solution

1.Entropy

In machine learning, entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Find the entropy of the class variable.

$E(S) = -[(9/14)\log(9/14) + (5/14)\log(5/14)] = 0.94$

Problem and Solution

Information Gain

Information gain can be defined as the amount of information gained about a random variable or signal from observing another random variable.It can be considered as the difference between the entropy of parent node and weighted average entropy of child nodes.

$$IG(S, A) = H(S) - H(S, A)$$

Alternatively,

$$IG(S, A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

To find the information gain. It is the difference between parent entropy and average weighted entropy we found above.

IG(S, outlook) = 0.94 - 0.693 = 0.247

Problem and Solution

Gini Impurity

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

The next step is to find the information gain. It is the difference between parent entropy and average weighted entropy we found above.

IG(S, outlook) = 0.94 - 0.693 = 0.247

Similarly find Information gain for Temperature, Humidity, and Windy.

IG(S, Temperature) = 0.940 - 0.911 = 0.029

IG(S, Humidity) = 0.940 - 0.788 = 0.152

IG(S, Windy) = 0.940 - 0.8932 = 0.048

- **Types of Decision Tree Algorithms**
- The different decision tree algorithms are listed below:
- ID3(Iterative Dichotomiser 3)
- C4.5
- CART(Classification and Regression Trees)
- CHAID (Chi-Square Automatic Interaction Detection)
- MARS(Multivariate Adaptive Regression Splines)

Activity

- **Advantages of Decision Trees**
- Decision trees are super interpretable
- Require little data preprocessing
- Suitable for low latency applications

- **Disadvantages of Decision Trees**
- More likely to overfit noisy data. The probability of overfitting on noise increases as a tree gets deeper.
- A solution for it is **pruning**. You can read more about pruning from my ***Kaggle notebook***.
- Another way to avoid overfitting is to use bagging techniques like Random Forest. You can read more about Random Forest from an article from ***neptune.ai***.
- **References:**

- **Applications of Decision Trees**
- **Business Decision Making**: Used in strategic planning and resource allocation.
- **Healthcare**: Assists in diagnosing diseases and suggesting treatment plans.
- **Finance**: Helps in credit scoring and risk assessment.
- **Marketing**: Used to segment customers and predict customer behavior.

Assessment 1

1. List out the advantages of Decision Trees
   a)_____
   b)_____
   c)_____
   d)_____

2. Identify the disadvantages of Decision Trees
   a)_____
   b)_____
   c)_____
   d)_____

REFERENCES

1. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012.
2. David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", Morgan Kaufmann/Elsevier Publishers, 2013
3. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", Wiley publishers, 2015.

THANK YOU