****

**Assignment-01**

**on**

**“Decision trees.”**

**Course Title: Machine Learning.**

**Course Code: CSE475**

**Section: 02**

**Submitted by:**

**Md Tariqul Islam**

**ID:2021-1-60-054**

**Submitted to:**

**Dr.Raihan ul Islam(DRUI)**

**Associate Professor**

**Department of Computer Science and Engineering**

**East West University**

**Submission Date: 10/9/2024**

**ANSWER TO THE QUESTION NO:1**

In decision trees, entropy and information gain help us figure out the best way to split data at each step.

**Entropy**

Entropy measures the uncertainty or randomness in a dataset. Imagine you have a big jar of mixed candies: if you can't tell which candy you'll pick next, the entropy is high. If all the candies are the same, the entropy is low.

**Information Gain**

Information Gain is about finding the best way to reduce that uncertainty. When you split the data based on a feature (like color or size of the candies), you want to make the groups as different from each other as possible. The more you can reduce uncertainty (entropy) by making these splits, the higher the information gain.

So, in a decision tree, the goal is to split the data in a way that maximizes information gain, effectively making the resulting groups as distinct and informative as possible.

**ANSWER TO THE QUESTION NO:2**

To construct a decision tree using the ID3 algorithm, we will use the given dataset to perform calculations for entropy and information gain at each step to decide the best attribute to split on.

The formula for entropy is:

Entropy(S) = - p\_+ \* log2(p\_+) - p\_- \* log2(p\_-)

where:

p\_+ is the proportion of positive examples (Loan Default = Yes).

p\_- is the proportion of negative examples (Loan Default = No).

In the dataset:

Number of "Yes" = 6

Number of "No" = 9

Total = 15

p\_+ = 6 / 15 = 0.4

p\_- = 9 / 15 = 0.6

Calculating entropy:

Entropy(S) = - (0.4 \* log2(0.4) + 0.6 \* log2(0.6))

Substitute the values for log2:

log2(0.4) ≈ -1.3219

log2(0.6) ≈ -0.7369

Entropy(S) = - (0.4 \* -1.3219 + 0.6 \* -0.7369)

Entropy(S) = - (-0.5288 - 0.4421)

Entropy(S) ≈ 0.9709

Now, we compute the entropy for each attribute (Age, Income, Education, Credit Score).

Entropy(Low):

Entropy(Low) = - (3/4) \* log2(3/4) - (1/4) \* log2(1/4)

Entropy(Low) = - (0.75 \* -0.4150 + 0.25 \* -2)

Entropy(Low) = - (-0.3113 - 0.5)

Entropy(Low) ≈ 0.8113

Entropy(Medium):

Entropy(Medium) = - (2/5) \* log2(2/5) - (3/5) \* log2(3/5)

Entropy(Medium) = - (0.4 \* -1.3219 + 0.6 \* -0.7369)

Entropy(Medium) ≈ 0.9709

Entropy(High):

Entropy(High) = - (1/6) \* log2(1/6) - (5/6) \* log2(5/6)

Entropy(High) = - (0.1667 \* -2.5850 + 0.8333 \* -0.2630)

Entropy(High) ≈ 0.6500

Overall Entropy for "Income":

Entropy(Income) = (4/15) \* Entropy(Low) + (5/15) \* Entropy(Medium) + (6/15) \* Entropy(High)

Entropy(Income) = (4/15) \* 0.8113 + (5/15) \* 0.9709 + (6/15) \* 0.6500

Entropy(Income) ≈ 0.7999

Information Gain for "Income":

IG(Income) = Entropy(S) - Entropy(Income)

IG(Income) = 0.9709 - 0.7999

IG(Income) ≈ 0.1710

Entropy for "Education"

Entropy(High School):

Entropy(High School) = - (3/5) \* log2(3/5) - (2/5) \* log2(2/5)

Entropy(High School) = - (0.6 \* -0.7369 + 0.4 \* -1.3219)

Entropy(High School) ≈ 0.9709

Entropy(Bachelor's):

Entropy(Bachelor's) = - (3/7) \* log2(3/7) - (4/7) \* log2(4/7)

Entropy(Bachelor's) = - (0.4286 \* -1.2224 + 0.5714 \* -0.8074)

Entropy(Bachelor's) ≈ 0.9852

Entropy(Master's):

Entropy(Master's) = - (0/3) \* log2(0/3) - (3/3) \* log2(3/3)

Entropy(Master's) = 0

Overall Entropy for "Education":

Entropy(Education) = (5/15) \* Entropy(High School) + (7/15) \* Entropy(Bachelor's) + (3/15) \* Entropy(Master's)

Entropy(Education) = (5/15) \* 0.9709 + (7/15) \* 0.9852 + (3/15) \* 0

Entropy(Education) ≈ 0.7834

Information Gain for "Education":

IG(Education) = Entropy(S) - Entropy(Education)

IG(Education) = 0.9709 - 0.7834

IG(Education) ≈ 0.1875

Entropy for "Credit Score"

Entropy(Fair):

Entropy(Fair) = - (3/4) \* log2(3/4) - (1/4) \* log2(1/4)

Entropy(Fair) = - (0.75 \* -0.4150 + 0.25 \* -2)

Entropy(Fair) = - (-0.3113 - 0.5)

Entropy(Fair) ≈ 0.8113

Entropy(Good):

Entropy(Good) = - (1/5) \* log2(1/5) - (4/5) \* log2(4/5)

Entropy(Good) = - (0.2 \* -2.3219 + 0.8 \* -0.3219)

Entropy(Good) ≈ 0.7219

Entropy(Excellent):

Entropy(Excellent) = 0 (since all are "No")

Entropy(Poor):

Entropy(Poor) = 0 (since all are "Yes")

Overall Entropy for "Credit Score":

Entropy(Credit Score) = (4/15) \* Entropy(Fair) + (5/15) \* Entropy(Good) + (5/15) \* Entropy(Excellent) + (2/15) \* Entropy(Poor)

Entropy(Credit Score) = (4/15) \* 0.8113 + (5/15) \* 0.7219 + (3/15) \* 0 + (3/15) \* 0

Entropy(Credit Score) = 0.2150 + 0.2406

Entropy(Credit Score) ≈ 0.4556

Information Gain for "Credit Score"

IG(Credit Score) = Entropy(S) - Entropy(Credit Score)

IG(Credit Score) = 0.9709 - 0.4556

IG(Credit Score) ≈ 0.5153

So,Information Gain

**IG(Income) =** 0.1710

**IG(Education) =** 0.1875

**IG(Credit Score) =** 0.5153

IMG_20240911_002505.png

**ANSWER TO THE QUESTION NO:3**

Overfitting happens when a decision tree becomes too complex, capturing not just the useful patterns in the data but also the random noise. This makes the model perform well on the data it was trained on but poorly on new data. In critical areas like medical diagnosis or financial forecasting, overfitting can have serious and harmful effects.

**1. Impact on Model Reliability**

**Poor Generalization**: An overfitted model may fail to recognize important patterns in new data, leading to poor performance. For example, in medical diagnosis, it might correctly identify diseases in the training data but miss them in new patients. In finance, it may predict past trends accurately but fail when market conditions change.

**Inconsistent Predictions**: Overfitting makes the model overly sensitive to small changes in input data. In healthcare, slight changes in test results could lead to different diagnoses. In finance, small fluctuations in data could cause unreliable predictions, leading to bad decisions.

**2. Implications in Medical Diagnosis**

**False Positives and Negatives**: Overfitting can increase incorrect diagnoses (false positives) or fail to detect a condition (false negatives). False positives can lead to unnecessary treatments, while false negatives might delay critical care, putting lives at risk.

**Loss of Trust**: Frequent mistakes can cause doctors and other professionals to lose trust in the model, leading them to avoid using it, even if it could be helpful.

**Legal and Ethical Issues**: Incorrect predictions could result in legal issues like malpractice suits and raise ethical concerns about patient safety and consent.

**3. Implications in Financial Forecasting**

**Poor Investment Decisions**: Overfitting can cause bad predictions, leading to poor investment choices, financial losses, and even economic instability.

**Reduced Confidence in AI Systems**: Repeated errors can make people less willing to rely on AI models, slowing down the adoption of automated tools in finance.

**4. Decision-Making Impact**

**Suboptimal Decisions**: Overfitting can lead to poor decisions. In healthcare, this could mean unnecessary treatments, while in finance, it could mean inefficient use of money and resources.

**Costly Mistakes**: Mistakes from overfitting can be expensive, both financially and in human terms, especially in critical applications where even small errors can have big consequences.

**5. Mitigating Overfitting**

To avoid overfitting in decision trees:

**Prune the Tree**: Simplify the model by cutting unnecessary branches.

**Use Cross-Validation**: Test the model on different parts of the data to make sure it works well in general.

**Apply Regularization**: Penalize overly complex models to prevent them from fitting noise.

**Limit Depth and Splits**: Control the tree’s depth and the number of splits to prevent it from becoming too complex.

Overfitting can make decision tree models unreliable in crucial areas like healthcare and finance, leading to poor decisions, increased risks, and loss of trust. To prevent these problems, it's important to carefully design, validate, and simplify models to keep them accurate and dependable in real-world situations.