CSE 422: Artificial Intelligence
Brac University
Quiz 02 (SEC 8)

Total Marks: 10 **Time: 20 mins**

Name: **ID:** **Section:**

An email filtering system uses a Naive Bayes classifier to distinguish between spam and non-spam ("ham") emails. Based on the training data, the system has derived the following probabilities:

- The probability of any email being spam is 20%.
- The probability of any email being non-spam is 80%.
- The word "free" appears in 30% of spam emails.
- The word "free" appears in 5% of non-spam emails.

A new email arrives in your inbox, and it contains the word "free".

1. Calculate the probability of any email containing the word “free”. [4]
2. Calculate the Probability that the New Email is Spam. [4]
3. In the context of Naive Bayes, discuss the potential impact on classification accuracy if the word "free" frequently co-occurs with other specific words in spam emails, considering the classifier's assumption about feature independence. [2]

Ans:

Calculate P ("free") which is the probability of the word "free" appearing in any email.

$$P("free") = P("free"|Spam) \times P(Spam) + P("free"|spam') \times P(spam')$$

Use Bayes' theorem to calculate

$$P(Spam|"free") = \frac{P("free"|Spam) \times P(Spam)}{P("free")}$$

The Naive Bayes classifier assumes that features (words in this case) are independent of each other given the class label (spam or non-spam). This assumption can lead to inaccuracies in the classification if, in reality, certain words like "free" tend to co-occur with other indicative words in spam emails. Such correlations can provide additional context that Naive Bayes fails to capture, potentially reducing the effectiveness of spam detection in complex real-world data where word interdependencies are common.

CSE 422: Artificial Intelligence
Brac University
Quiz Optional (SEC 8)

Total Marks: 10 **Time: 20 mins**

Name: **ID:** **Section:**

An online retailer wants to use decision tree modeling to predict customer purchasing behavior. The dataset consists of customers categorized by whether they bought a particular product ("Buy" or "Not Buy") and divided based on two attributes: "Age Group" (Youth, Adult, Senior) and "Income Level" (Low, Medium, High). The data distribution is as follows:

800 customers in total: 400 "Buy". 400 "Not Buy".
Breakdown by Age Group:
Youth: 300 customers. 150 "Buy". 150 "Not Buy"
Adult: 300 customers. 225 "Buy". 75 "Not Buy"
Senior: 200 customers. 25 "Buy", 175 "Not Buy"
Breakdown by Income Level:
Low: 200 customers. 50 "Buy". 150 "Not Buy"
Medium: 300 customers. 200 "Buy". 100 "Not Buy"
High: 300 customers. 150 "Buy". 150 "Not Buy"

1. Calculate the initial entropy of the dataset based on customer purchasing behavior. [2]
2. Calculate the entropy for subsets of the data based on age group (Youth, Adult, Senior). [6]
3. Calculate the information gain from splitting the data based on age group. [1]
4. How would the decision tree model's accuracy and complexity be affected if an additional attribute with low information gain were included in the model? [1]

Ans:

Overall Entropy Calculation

Formula: $E(S) = -p_+ log_2(p_+) - p_- log_2(p_-)$

Step 1: Calculate the probabilities:

$p_+(probability\ of\ "Buy"): 400/800 = 0.5$

$p_-\ (probability\ of\ "Will\ not\ buy"): 400/800 = 0.5$

Step 2: Plug into the entropy formula.

Entropy of Subsets by Age Group

For the subset Youth

$Total\ number\ of\ Youths\ =\ 300,\ Will\ Buy\ =\ 150,\ Will\ Not\ Buy\ =\ 150$

$$p_{-} = \frac{Will\ Not\ Buy}{Total\ Number\ of\ Youths}$$

$$p_{+} = \frac{Will\ Buy}{Total\ Number\ of\ Youths}$$

Calculate entropy for this subset using

$$E(Young) = -\ p_{+} log_2(p_{+})\ -\ p_{-} log_2(p_{-})$$

Similarly the entropy for the subset Adult and Senior can be calculated

Total number of adults = 300, Buy = 225, Not Buy = 75

$$p_{-} = \frac{Not\ Buy}{Total\ number\ of\ Adults}$$

$$p_{+} = \frac{Will\ Buy}{Total\ number\ of\ Adults}$$

Calculate entropy for this subset using

$$E(Adults)\ = -\ p_{+} log_2(p_{+})\ -\ p_{-} log_2(p_{-})$$

Similarly entropy for other seniors can be calculated.

Total number of seniors = 200, Buy = 25, Not Buy = 175

$$p_{-} = \frac{Not\ Buy}{Total\ number\ of\ Seniors}$$

$$p_{+} = \frac{Will\ Buy}{Total\ number\ of\ Seniors}$$

Information Gain Calculation

Formula:

$$IG(S,\ Age\ Group)\ =\ E(S)\ -\ (\frac{Total\ Youth}{Total}\ E(Youth)\ +\ \frac{Total\ Adult}{Total} E(Adult)\ +\ \frac{Total\ Senior}{Total} E(Senior))$$

Including an attribute with low information gain in the decision tree model could lead to increased complexity and overfitting without significant improvement in accuracy. Low information gain indicates that the attribute does little to reduce uncertainty or impurity in the classification of the target variable, thus potentially complicating the model with minimal benefit. This can make the model less generalizable and more sensitive to noise in the training data.

CSE 422: Artificial Intelligence
Brac University
Quiz 02 (SEC 7)

Total Marks: 10 **Time: 20 mins**

Name: **ID:** **Section:**

Suppose a small clinic has records of patients suffering from a rare disease. The disease can be classified into two types: Type A and Type B. Based on historical data, the following probabilities are observed:

- 30% of the clinic's patients have Type A of the disease.
- 70% of the clinic's patients have Type B of the disease.
- Of the patients with Type A, 40% are smokers.
- Of the patients with Type B, 20% are smokers.

1. A new patient arrives at the clinic and is diagnosed with the disease. You also learn that this patient is a smoker. Calculate the Probability that the New Patient Has Type A or Type B Given They Are a Smoker. [8]

2. What assumption does the Naive Bayes classifier make about the features in a dataset, and how might this impact its performance in real-world scenarios? [2]

Ans:

Calculate *P*(Smoker) which is the probability of a randomly chosen patient from the clinic being a smoker.

$$P(Smoker) = P(Smoker|Type\ A)P(Type\ A) + P(Smoker|Type\ B)P(Type\ B)$$

Use Bayes' theorem to calculate,

$$P(Type\ A|Smoker) = \frac{P(Smoker|Type\ A)\times P(Type\ A)}{P(Smoker)}$$

and similarly, $P(Type\ B|Smoker) = \frac{P(Smoker|Type\ B)\times P(Type\ B)}{P(Smoker)}$

Naive Bayes theorem assumes that all features in a dataset are mutually independent given the class label. This assumption simplifies the computation significantly but is often practically violated because features can be correlated. For instance, in medical diagnosis, symptoms may be related to each other, which this model ignores.

CSE 422: Artificial Intelligence
Brac University
Quiz Optional (SEC 7)

Total Marks: 10 **Time: 20 mins**

Name: **ID:** **Section:**

A company is analyzing employee data to predict which employees are likely to leave the company within the next year. The dataset consists of 1000 employees categorized into two classes: "Will Leave" and "Will Stay." The employees are also divided based on their job satisfaction level: "High" or "Low".

The data is distributed as follows:

- 600 employees are classified as "Will Stay".
- Out of these, 400 have "High" job satisfaction.
- 200 have "Low" job satisfaction.
- 400 employees are classified as "Will Leave".
- Out of these, 100 have "High" job satisfaction.
- 300 have "Low" job satisfaction.

1. Calculate the Overall Entropy of the System. [2]
2. Calculate the Entropy for each subset of the data based on job satisfaction (High and Low). [4]
3. Calculate the information gain from splitting the data based on job satisfaction. [2]
4. Discuss how the choice of splitting criteria (e.g., job satisfaction) impacts the effectiveness of a decision tree in classifying new instances. [2]

Ans:

Overall Entropy Calculation

Formula: $E(S) = -p_+ log_2(p_+) - p_- log_2(p_-)$

Step 1: Calculate the probabilities:

$p_+(probability\ of\ "Will\ Stay"): \ 600/1000 \ = \ 0.6$

$p_- \ (probability\ of\ "Will\ not\ buy"): \ 400/1000 \ = \ 0.4$

Step 2: Plug into the entropy formula.

Entropy of Subsets by Job Satisfactions

For the subset High Job Satisfaction

$Total\ number\ of\ High\ Job\ Satisfaction\ =\ 500,\ Will\ Stay\ =\ 400,\ Will\ Leave\ =\ 100$

$$p_{-} = \frac{Will\ leave}{Total\ number\ of\ High\ Job\ Satisfaction}$$

$$p_{+} = \frac{Will\ Stay}{Total\ Number\ Low\ Job\ Satisfaction}$$

Calculate entropy for this subset using

$$E(High) = -\ p_{+} log_2(p_{+})\ -\ p_{-} log_2(p_{-})$$

Similarly the entropy for the subset Low Job Satisfaction can be calculated

Total number of Low Job Satisfaction= 500, Stay = 200, Leave = 300

$$p_{-} = \frac{Leave}{Total\ number\ of\ Low\ Job\ Satisfaction}$$

$$p_{+} = \frac{Stay}{Total\ number\ of\ Low\ Job\ Satisfaction}$$

Calculate entropy for this subset using

$$E(Low\ Job\ Satisfaction)\ = -\ p_{+} log_2(p_{+})\ -\ p_{-} log_2(p_{-})$$

Information Gain Calculation

Formula:

$$IG(S,\ Job\ Satisfaction)\ =\ E(S)\ -\ (\frac{Total\ number\ of\ Low\ Job\ Satisfaction}{Total}\ E(Low)\ +\ \frac{Total\ number\ of\ High\ Job\ Satisdaction}{Total} E(Hig$$

Conceptual understanding

The choice of splitting criteria significantly influences the decision tree's performance. Ideally, a split should create subsets that are as pure as possible, meaning they lean strongly towards a single class. The effectiveness of the tree in classifying new instances hinges on selecting features that best reduce uncertainty in node impurity, leading to a more accurate and generalizable model.