INFO I-421: Applications in Data Mining

Homework 3

80 points

Based on the table of data, compute the following:

GNI Indexes are calculated as follows:

The Gini index $Gini(D)$ for the dataset D is computed as:

$$Gini(D) = 1 - \sum_{k=1}^{K} (p_k)^2$$

Where:

- p_k is the proportion of data points in class k with respect to the total number of data points in D.

Information Gain is calculated as follows:

The entropy $H(D)$ of dataset D is calculated as:

$$H(D) = -\sum_{k=1}^{K} p_k \log_2(p_k)$$

where p_k is the proportion of data points in class k with respect to the total number of data points in D, and K is the total number of distinct classes.

The entropy $H(D_v)$ of the subset D_v with respect to attribute A is calculated similarly:

$$H(D_v) = -\sum_{k=1}^{K} p_k^{(v)} \log_2(p_k^{(v)})$$

where $p_k^{(v)}$ is the proportion of data points in class k within subset D_v.

Then, the Information Gain $IG(D, A)$ when splitting D on attribute A is calculated as:

$$IG(D, A) = H(D) - \sum_{v=1}^{V} \frac{|D_v|}{|D|} H(D_v)$$

where:

- $|D_v|$ is the number of data points in subset D_v,
- $|D|$ is the total number of data points in D.

1. Compute the GINI Index for the gender attribute (10 pts)

```
> # Create a dataframe with your dataset
> data <- data.frame(
+   Customer = 1:20,
+   Gender = c("M", "M", "M", "M", "M", "M", "F", "F", "F", "F", " ..." ... [TRUNCATED]

> # Count occurrences of each class
> gender_counts <- table(data$Gender)

> # Calculate probabilities
> gender_probabilities <- gender_counts / sum(gender_counts)

> # Compute Gini Index
> gini_index <- 1 - sum(gender_probabilities^2)

> # Print the result
> print(gini_index)
[1] 0.5
```

2. Compute the GINI Index for car type attribute (10 pts)

```
> # Create a dataframe with your dataset
> data <- data.frame(
+   Customer = 1:20,
+   Car_Type = c("FAMILY", "SPORTS", "SPORTS", "SPORTS", "SPORTS", .... [TRUNCATED]

> # Count occurrences of each class
> car_type_counts <- table(data$Car_Type)

> # Calculate probabilities
> car_type_probabilities <- car_type_counts / sum(car_type_counts)

> # Compute Gini Index
> gini_index <- 1 - sum(car_type_probabilities^2)

> # Print the result
> print(gini_index)
[1] 0.64
```

3. Compute the GINI Index for shirt size attribute (10 pts)

```
> # Create a dataframe with your dataset
> data <- data.frame(
+   Customer = 1:20,
+   Shirt_Size = c("SMALL", "MEDIUM", "MEDIUM", "LARGE", "EXTRA LA ..." ... [TRUNCATED]

> # Count occurrences of each class
> shirt_size_counts <- table(data$Shirt_Size)

> # Calculate probabilities
> shirt_size_probabilities <- shirt_size_counts / sum(shirt_size_counts)

> # Compute Gini Index
> gini_index <- 1 - sum(shirt_size_probabilities^2)

> # Print the result
> print(gini_index)
[1] 0.735
```

4. Based on the Gini Index calculations, which attribute should be your root node for your decision tree. (10 pts) = **Gender**

   Gender: Gini Index = 0.5

   Car Type: Gini Index = 0.64

   Shirt Size: Gini Index = 0.735

   So Gender will be selected as root node as it has the lowest Gini index.

5. Compute the Information gain for the gender attribute (10 pts)

   Information Gain for gender =0.02

```
> # Create the dataset
> data <- data.frame(
+   Customer = 1:20,
+   Gender = c("M", "M", "M", "M", "M", "M", "F", "F", "F", "F",
+              "M" .... [TRUNCATED]

> # Function to calculate Gini Index
> calculate_gini <- function(labels) {
+   probabilities <- table(labels) / length(labels)
+   gini_index <- 1 -  .... [TRUNCATED]

> # Calculate Gini Index for entire dataset
> gini_dataset <- calculate_gini(data$Class)

> # Split dataset by gender
> male_data <- data[data$Gender == "M", ]

> female_data <- data[data$Gender == "F", ]

> # Calculate Gini Index for each gender category
> gini_male <- calculate_gini(male_data$Class)

> gini_female <- calculate_gini(female_data$Class)

> # Calculate weighted average of Gini Index for each gender category
> weighted_average_gini <- (nrow(male_data) / nrow(data)) * gini_male +
+         .... [TRUNCATED]

> # Compute Information Gain
> information_gain <- gini_dataset - weighted_average_gini

> # Print the result
> print(information_gain)
[1] 0.02
```

6. Compute the Information gain for car type attribute (10 pts)

   Information Gain for car types = 0.3375

```
> # Create the dataset
> data <- data.frame(
+   Customer = 1:20,
+   Gender = c("M", "M", "M", "M", "M", "M", "F", "F", "F", "F",
+              "M" .... [TRUNCATED]

> # Function to calculate Gini Index
> calculate_gini <- function(labels) {
+   probabilities <- table(labels) / length(labels)
+   gini_index <- 1 -  .... [TRUNCATED]

> # Calculate Gini Index for entire dataset
> gini_dataset <- calculate_gini(data$Class)

> # Split dataset by car type
> family_data <- data[data$Car_Type == "FAMILY", ]

> sports_data <- data[data$Car_Type == "SPORTS", ]

> luxury_data <- data[data$Car_Type == "LUXURY", ]

> # Calculate Gini Index for each car type category
> gini_family <- calculate_gini(family_data$Class)

> gini_sports <- calculate_gini(sports_data$Class)

> gini_luxury <- calculate_gini(luxury_data$Class)

> # Calculate weighted average of Gini Index for each car type category
> weighted_average_gini <- (nrow(family_data) / nrow(data)) * gini_family +
+   .... [TRUNCATED]

> # Compute Information Gain
> information_gain <- gini_dataset - weighted_average_gini

> # Print the result
> print(information_gain)
[1] 0.3375
```

7. Compute the Information gain for shirt size attribute (10 pts)

Information Gain for shirt size = 0.01

```
> # Create the dataset
> data <- data.frame(
+   Customer = 1:20,
+   Gender = c("M", "M", "M", "M", "M", "M", "F", "F", "F", "F",
+              "M" .... [TRUNCATED]

> # Function to calculate Gini Index
> calculate_gini <- function(labels) {
+   probabilities <- table(labels) / length(labels)
+   gini_index <- 1 -  .... [TRUNCATED]

> # Calculate Gini Index for entire dataset
> gini_dataset <- calculate_gini(data$Class)

> # Split dataset by shirt size
> small_data <- data[data$Shirt_Size == "SMALL", ]

> medium_data <- data[data$Shirt_Size == "MEDIUM", ]

> large_data <- data[data$Shirt_Size == "LARGE", ]

> extra_large_data <- data[data$Shirt_Size == "EXTRA LARGE", ]

> # Calculate Gini Index for each shirt size category
> gini_small <- calculate_gini(small_data$Class)

> gini_medium <- calculate_gini(medium_data$Class)

> gini_large <- calculate_gini(large_data$Class)

> gini_extra_large <- calculate_gini(extra_large_data$Class)

> # Calculate weighted average of Gini Index for each shirt size category
> weighted_average_gini <- (nrow(small_data) / nrow(data)) * gini_small +
+  .... [TRUNCATED]

> # Compute Information Gain
> information_gain <- gini_dataset - weighted_average_gini

> # Print the result
> print(information_gain)
[1] 0.008571429
```

8. Based on the Information gain, which attributes should be your root node for your decision tree? (10 pts) = **CAR TYPES**
   Gender: Information Gain = 0.02
   Car Type: Information Gain = 0.3375
   Shirt Size: Information Gain = 0.008571429
   Information gain of car type is the highest, so selecting it as root node will be suitable option.

Customer	Gender	Car Type	Shirt Size	Class
1	M	FAMILY	SMALL	C0
2	M	SPORTS	MEDIUM	C0
3	M	SPORTS	MEDIUM	C0
4	M	SPORTS	LARGE	C0

5	M	SPORTS	EXTRA LARGE	C0
6	M	SPORTS	EXTRA LARGE	C0
7	F	SPORTS	SMALL	C0
8	F	SPORTS	SMALL	C0
9	F	SPORTS	MEDIUM	C0
10	F	LUXURY	LARGE	C0
11	M	FAMILY	LARGE	C1
12	M	FAMILY	EXTRA LARGE	C1
13	M	FAMILY	MEDIUM	C1
14	M	LUXURY	EXTRA LARGE	C1
15	F	LUXURY	SMALL	C1
16	F	LUXURY	SMALL	C1
17	F	LUXURY	MEDIUM	C1
18	F	LUXURY	MEDIUM	C1
19	F	LUXURY	MEDIUM	C1
20	F	LUXURY	LARGE	C1