**NAME: SHIVA SAI PRANEETH CHAKINALA**

**ID: 801147603**

**Cloud Computing for Data Analysis**

**Exercise 09: Decision Tree**

**Part 1**

Consider the training examples shown in Table below for a binary classification

problem.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer ID | Gender | Car Type | 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 |

(a)Compute the Gini index for the overall collection of training examples.

Ans: Gini(t)=1-p((C0/Class)^2)-p((C1/Class)^2)

=1-((10/20)^2)-((10/20)^2)

=0.5

(b) Compute the Gini index for the Customer ID attribute.

Ans: If we take each of the customer ID’s the Gini value is 0, so the customer ID is for 20 Gini values,

Gini(1)=1-((1/1)^2)-((0/1)^2)

=0

(c) Compute the Gini index for the Gender attribute

Ans: Gini(Male)= 1- p((C0/M)^2)-p((C1/M)^2)

=1-((6/10)^2)-((4/10)^2)

=0.48

Gini(Female)=1-p((C0/F)^2)-p((C1/F)^2)

=1-((6/10)^2)-((4/10)^2)=0.48

Gini(Gender)=[(T(Male/T(Male+Female)]\*Gini(Male)+[(T(Female/T(Male+Female)]\*Gini(Female)

=((10/20)\*0.48)+((10/20)\*48)=0.48

(d) Compute the Gini index for the Car Type attribute using multiway

split.

Ans: Gini(Family)=1-p((C0/Family)^2)-p((C1/Family)^2)

=1-((1/4)^2)-((3/4)^2)=0.375

Gini(Sports)=1-p((C0/Sports)^2)-p((C1/Sports)^2)

=1-((8/8)^2)-((0/8)^2))=0

Gini(Luxury)=1-p((C0/Luxury)^2)-p((C1/Luxury)^2)

=1-((1/8)^2)-((7/8)^2))=0.2188

Gini(Car Type)= [[T(Family)/T(Family+ Sports+ Luxury)]\*Gini(Family)]+[[T(Sports)/T(Family + Sports+ Luxury)]\*Gini(Sports)]+[[T(Luxury)/T(Family+ Sports+ Luxury)]\*Gini(Luxury)]

= [(4/20)\*0.375]+0+[(8/20)\*0.2188]

=0.1625

(e) Compute the Gini index for the Shirt Size attribute using multiway

split.

Ans:

Gini(Small)=1-p((C0/Small)^2)-p((C1/Small)^2)

=1-((3/5)^2)-((2/5)^2)=0.48

Gini(Medium) = 1 - p((C0/Medium)^2)- p((C1/Medium)^2)

= 1 - ((3/7)^2) - ((4/7)^2) = 0.4898

Gini(Large) = 1 - p((C0/Large)^2) – p((C1/Large)^2)

= 1 - ((2/4)^2) – ((2/4)^2) = 0.5

Gini(Extra Large) = 1 - p((C0/Extra Large)^2) – p((C1/Extra Large)^2)

= 1 – ((2/4)^2) – ((2/4)^2) = 0.5

Gini(Shirt Size) =

[[(T(Small)/T(Small+ Medium+ Large+ Extra Large)]\*Gini(Small)]+ [(T(Medium)/T(Small+ Medium+ Large+ Extra Large)]\*Gini(Medium)]+ [[(T(Large)/T(Small+Medium+Large+ExtraLarge)]\*Gini(Large)]+[[(T(Extra Large)/T(Small+ Medium+ Large+ Extra Large)]\*Gini(Extra Large)]

= [(5/20)\*0.48] + [(7/20)\*0.4898] + [(4/20)\*0.5]+ [(4/20)\*0.5]

= 0.4914

(f) Which attribute is better, Gender, Car Type, or Shirt Size?

Ans: The Gini index for Gender is 0.48, Gini index for car type is 0.1625 and Gini index for shirt size is 0.4914.

* The attribute with lowest Gini index is better.
* So, Car Type is better.

(g) Explain why Customer ID should not be used as the attribute test

condition even though it has the lowest Gini.

Ans: The customer id is unique. So, it cannot be used as a predictive attribute.

- It should not be used even if it has lowest Gini because no matter the number of divisions, it will not be useful without the need of any predictive behaviour.