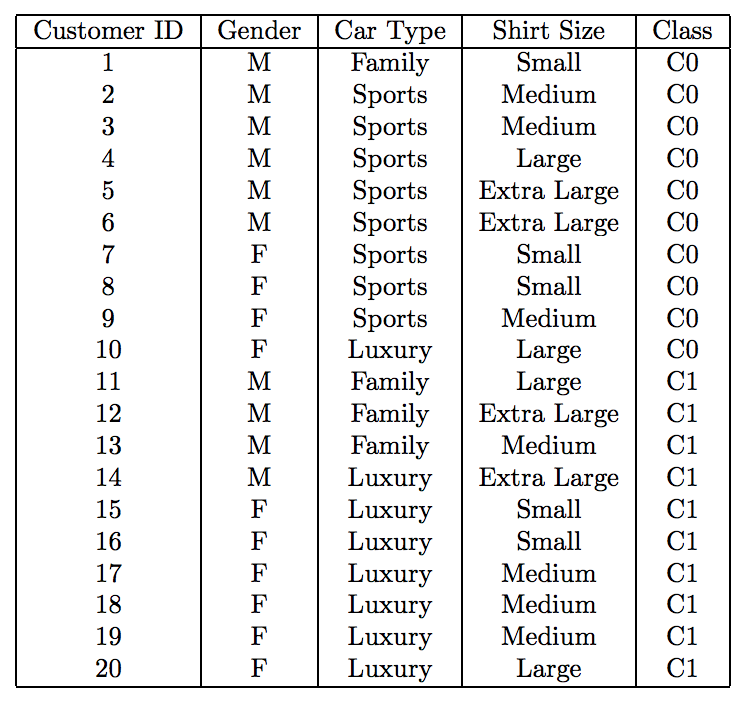
**Consider the training examples shown in attached Table for a binary classification problem.**



1. **Compute the Gini index for the overall collection of training examples.**

G index (Class) = 1 - ((10/20)^2 + (10/20)^2) = 0.5

1. **Compute the Gini index for the Customer ID attribute.**

|  |  |
| --- | --- |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |
| 6 | 1 |
| 7 | 1 |
| 8 | 1 |
| 9 | 1 |
| 10 | 1 |
| 11 | 1 |
| 12 | 1 |
| 13 | 1 |
| 14 | 1 |
| 15 | 1 |
| 16 | 1 |
| 17 | 1 |
| 18 | 1 |
| 19 | 1 |
| 20 | 1 |

G index (Customer ID) = 1 – ((1/1)^2 + (0/1)^2) = 1 – 1 = 0

1. **Compute the Gini index for the Gender attribute.**

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| C0 | 4 | 6 |
| C1 | 6 | 4 |
| Total | 10 | 10 |

Gini index (Gender) = 1 – [ P(F)^2 +P(M)^2] = 1 – (10/20)^2 – (10/20)^2 = 0.5

Using multiway split:

Gini female = 1 – (4/10)^2 – (6/10)^2 = 1 – 0.16 – 0.36 = 0.48

Gini male = 1 – (6/10)^2 – (4/10)^2 = 1 – 0.36 – 0.16 = 0.48

Gini index (Gender Type) = 10/20 \* Gini F + 10/20 \* Gini M = 0.48

1. **Compute the Gini index for the Car Type attribute using multiway split.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Family | Sports | Luxury |
| C0 | 1 | 8 | 1 |
| C1 | 3 | 0 | 7 |
| Total | 4 | 8 | 8 |

Gini Family = 1 – ((1/4)^2 + (3/4)^2) = 1 – 5/8 = 1 – 0.625 = 0.375

Gini Sports = 1 – ((8/8)^2 + 0) = 0

Gini Luxury = 1 – ((1/8)^2 + (7/8)^2) = 1 – 25/32 = 1 – 0.78125 = 0.218

Gini Car Type = 4/20 \* Gini Family + 8/20 \* Gini Family + 8/20 \* Gini Family =

= 4/20\*0.375 + 8/20\*0 + 8/20 \* 0.218 = 9/50 + 18/125 + 1/5= 13/80 = 0.1625

1. **Compute the Gini index for the Shirt Size attribute using multiway split.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Small | Medium | Large | Extra-Large |
| C0 | 3 | 3 | 2 | 2 |
| C1 | 2 | 4 | 2 | 2 |
| Total | 5 | 7 | 4 | 4 |

Gini Small = 1 – ((3/5)^2 + (2/5)^2) = 1 – 9/25 – 4/25 = 1 – 0.52 = 0.48

Gini Medium = 1 – ((3/7)^2 + (4/7)^2) = 1 – 9/49 – 16/49 = 0.489

Gini Large = 1 – ((2/4)^2 + (2/4)^2) = 1 – 1/4 – 1/4 = 0.5

Gini Extra - Large = 1 – ((2/4)^2 + (2/4)^2) = 0.5

Gini Shirt Size = 5/20 \* Gini S + 7/20 \* Gini M + 4/20 \* Gini L + 4/20 \* Gini E-L = 5/20 \* 0.48 + 7/20 \* 0.489 + 4/20 \* 0.5 + 4/20 \* 0.5 =

= 3/25 + 6/35 + 1/10 + 1/10 = 86/175 = 0.4914

1. **Which attribute is better, Gender, Car Type, or Shirt Size?**

Car Type, because is by far the best attribute with the lowest Gini index.

1. **Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.**

Because it doesn’t give you any information about de customer. Every customer has a different identifier.