Sohaib Syed
CS422
10/17/21
Dr. Gurbani

Homework 5

Problem 1.1:

Exercise 2:

A.

Probability of C0 is 10/20=1/2, probability of C1 is 10/20=1/2.

Thus Gini index: is 1- $\sum_{i=1}^{n} (P_i)^2$=1-(0.5^2+0.5^2)=**.5**

B.

By Gini index definition: give a zero impurity value if a node contains instances from a single class (Tan,Chapter 3). Thus the Gini index of Customer ID is 0.

C.

To find Gini index of gender we would first have to get gini index of each gender since both male and female can be classified as C0 or C1. We would then get the sum of the Gini indexes of each gender multiplied by their probability in general

Male Gini = 1-(0.6^2+0.4^2)=0.48

Female Gini= 1-(0.4^2+0.6^2)=0.48

Total gender Gini= 10/20(0.48)+10/20(0.48)=**0.48**

D.

Gini family car= 1- ((¼)^2+(¾)^2)=.375

Gini sports car = 1-((8/8)^2+0^2)=0

Gini luxury car= 1- ((⅛)^2+(⅞)^2)=.219

Gini of car = 4/20(.375)+8/20(0)+8/20(.219)=**.163**

E.

Gini small: 1-((⅖)^2+(⅖)^2)=.48

Gini medium: 1-((3/7)^2+(4/7)^2)=.490

Gini large: 1- ((2/4)^2+(2/4)^2)=.5

Gini extra large: 1-((2/4)^2+(2/4)^2)=.5

Gini shirt size= 5/20(.48)+7/20(.490)+4/20(.5)+4/20(.5)=**.4915**

F.

The better attribute is Car Type because it has the lowest Gini index. This means that it is the least impure.

G.

Customer ID has the lowest Gini index, but because for every new entry a new ID is necessary it would not help to predict the class.

Exercise 3:

A.

Entropy= $-\sum_{i=1}^{n} P_i * Log_2(P_i)$

-(((4/9)*$Log_2(4/9)$)+((5/9)*$Log_2(5/9)$))=**.991**

B.

This problem is sort of like a multiclass gini index problem so first we find entropy of T and then of F multiply to probability of T and F respectively. Sum them up.

A1:

(4/9)*-((¾*$Log_2(3/4)$)+(¼*$Log_2(1/4)$))=.361

(5/9)*-((⅕*$Log_2(1/5)$)+(⅘*$Log_2(4/5)$))=.401

.361+.401=.762

Gain = E(b)-E(a)=.991-.762=**.229**

A2:

(5/9)*-((2/5*$Log_2(2/5)$)+(3/5*$Log_2(3/5)$))=.539

(4/9)*-((2/4*$Log_2(2/4)$)+(2/4*$Log_2(2/4)$))=.444

.539+.444=.984

Gain = E(b)-E(a)=.991-.984=**.007**

C.

a3=1.0

a3 <= 1: (1/9)*-((1/1)*$Log_2(1/1)$+(0/1)* $Log_2(0/1)$)=0

a3 > 1: (8/9)*-((⅜)*$Log_2(3/8)$+(⅝)*$Log_2(5/8)$)=.848

gain= .991-.848=.143

a3=3.0

a3 <= 3: (2/9)*-((1/2)*$Log_2(1/2)$+(1/2)* $Log_2(1/2)$)=.222

a3 > 3: (7/9)*-((3/7)*$Log_2(3/7)$+(4/7)*$Log_2(4/7)$)=.766

gain= .991-(.222+.766)=.002

a3=4.0

a3<=4: (3/9)*-((2/3)*$Log_2(2/3)$+(1/3)* $Log_2(1/3)$)=.306

a3 > 4: (6/9)*-((2/6)*$Log_2(2/6)$+(4/6)*$Log_2(4/6)$)=.612

gain= .991-(.306+.612)=.073

a3=5.0

a3<=5: (5/9)*-((2/5)*$Log_2(2/5)$+(3/5)* $Log_2(3/5)$)=.539

a3 > 5: (4/9)*-((2/4)*$Log_2(2/4)$+(2/4)*$Log_2(2/4)$)=.444

gain= .991-(.539+.444)=.008

a3=6

a3<=6: (6/9)*-((3/6)*$Log_2(3/6)$+(3/6)* $Log_2(3/6)$)=.667

a3 > 6: (3/9)*-((1/3)*$Log_2(1/3)$+(2/3)*$Log_2(2/3)$)=.306

gain= .991-(.667+.306)=.018

a3=7

a3<=7: (8/9)*-((4/8)*$Log_2(4/8)$+(4/8)* $Log_2(4/8)$)=.889

a3 > 7: (1/9)*-((0/1)*$Log_2(0/1)$+(1/1)*$Log_2(1/1)$)=0

gain= .991-(.889+0)=.102

a3= 8

a3<=8: (9/9)*-((4/9)*$Log_2(4/9)$+(5/9)* $Log_2(5/9)$)=.991

a3 > 8: (0/9)*-((0/9)*$Log_2(0/9)$+(0/9)*$Log_2(0/9)$)=0

gain= .991-(.991+0)=0

The best split happens at a3=1.0, with gain of .143.

D.

According to the information gain a1 is the best split with gain of .229

E.

Error(a1)=¼(4/9)+⅕(5/9)=2/9

Error(a2)=⅝(5/9)+2/4(4/9)=5/9

Since a1 has the lower error rate, a1 is the best split.

F.

A1:

1-((¾)^2+(¼)^2)=.375

1-((⅘)^2+(⅕)^2)=.32

4/9(.375)+5/9(.32)=**.344**

A2:

1-((⅖)^2+(⅗)^2)=.48

1-((2/4)^2+(2/4)^2)=.5

5/9(.48)+4/9(.5)=**.489**

A1 has a lower gini index so a1 is the better split.

Exercise 5:

A.

Information gain E(b)-E(a)

E(b)= -((4/10)*$Log_2(4/10)$+(6/10)*$Log_2(6/10)$)=.971

A(T)= (7/10)*-((4/7)*$Log_2(4/7)$+(3/7)* $Log_2(3/7)$)=.690

A(F)=(3/10)*-((3/3*$Log_2(3/3)$+(0/3)$Log_2(0/3)$)=0

E(a)=.971-.690=**.281**

B(T)= (4/10)*-((3/4)*$Log_2(3/4)$+(1/4)* $Log_2(1/4)$)=.325

B(F)= (6/10)*-((5/6)*$Log_2(5/6)$+(1/6)* $Log_2(1/6)$)=.390

E(a)= .971-(.325+.390)=**.256**

Since splitting on A has the greater information gain that is where the induction algorithm would choose.

B.

Gini of entire class: 1-(.4^2+.6^2)=.48

Gini of A(T): 1-((4/7)^2+(3/7)^2)=.490

Gini of A(F):1-((3/3)^2+(0/3)^2)=

7/10(.490)+3/10(0)=**.343**

Gini of B(T): 1-((¾)^2+(¼)^2)=.375

Gini of B(F):1-((⅝)^2+(⅜)^2)=.278

4/10(.375)+6/10(.278)=**.317**

Gain a=.48-.343= .137

Gain b=.48-.317=**.163**

Gini index of B is lower, and the gain from splitting at B is greater.

C.

Yes, the two impurity measures can favor different attributes. This is because the two quantities have different equations and the gains of both are from different weights. The graph is misleading because it shows the impurity of the two measures, but not what the gains are of the two measures. An example of this is part a and part b of this problem.

Problem 1.2:

18

A.

If the data is half “+” and half “-”, and the classifier always predicts “+” then that means it is missing all the “-” which make up half the data. Thus the error should be 50%.

B.

Assuming data is still 50/50, the classifier would be 80% correct for the “+” and 20% correct “-” . This also means that 20% of the time a “+” would be classified as “-” , a false negative (FN) and 80% of the time a “-” is classified as “+”, a false positive (FP). Error is FP+FN/total, so (.8*.5+ .2*.5)/1=.5 or again 50%.

C.

In a similar situation to part a, if ⅔ data is “+” and ⅓ data is “-”, then always predicting “+” would mean that ⅓ of the data was misclassified, so the error would be ⅓ or 33%

D.

2/3 =”+” ; 1/3 = “-”

Prediction “+” = 2/3 ;False negatives= 1/3

Prediction “-” =1/3 ; False positives = 2/3

Error = (FP+FN)/Total= (⅔*⅓ + ⅓*⅔)/1=.444 or 44.4%