Homework 2
Karsen Diepholz
23 Feb 20

Recitation Exercises

2.

a. $1 - [(\frac{10}{20})^2 + (\frac{10}{20})^2] = 1 - [\frac{1}{4} + \frac{1}{4}] = \frac{1}{2}$

b. $1 - [(\frac{0}{1})^2 + 1^2] = 0$, weighted average is 0

c. Female: $1 - [(\frac{6}{20})^2 + (\frac{4}{20})^2] \; = \; .48$
Male: $1 - [(\frac{6}{20})^2 + (\frac{4}{20})^2] \; = \; .48$
Since they are the same, the weighted average is .48

d. Family: $1 - [(\frac{1}{4})^2 + (\frac{3}{4})^2] \; = \; .375$
Luxury: $1 - [(\frac{1}{8})^2 + (\frac{7}{8})^2] \; = \; .21875$
Sports: $1 - [(1^2 - 0^2)] \; = \; 0$
Weighted average = $[(\frac{4}{20}) * .375] + [(\frac{8}{20}) * .21875] \; + \; [(\frac{8}{20}) * 0] \; = \; .163$

e. Small: 1-[.36+.16] = .48
Medium: 1-[.184+.327] = .49
Large: 1-[.25+.25] = .5
Extra Large: 1-[.25+.25] = .5

Weighted average: .4915

f. Car type would be the better attribute, since it has the lowest Gini Index.

g. Because each attribute is unique.

3.

a. Entropy with respect to class attributes:

P(C1) = 4/9
P(C2) = 5/9

Entropy = $-(4/9) * log2(4/9) \; - \; (5/9) * log2(5/9) \; = \; .9911$

b. Information gain of a_1 = P(t)*Entropy(3,1) + P(F) * Entropy(1,4)

$(4/9)[-(3/4) * log(3/4) - (\frac{1}{4}) * log(\frac{1}{4})] \; + \; \frac{5}{9}[-\frac{1}{5} * log(\frac{1}{5}) - \frac{4}{5} * log(\frac{4}{5})] \; = \; .7616$

Gain of a_1=.9911-.7616 = **.23**

Information gain of a_2 = P(T)*Entropy(2,2) + P(F)*Entropy(3,2)

$$\frac{4}{9}(-\frac{2}{4} * log(\frac{2}{4}) - \frac{2}{4} * log(\frac{2}{4})) \ + \ \frac{5}{9}(-\frac{3}{5} * log(\frac{3}{5}) - \frac{2}{5} * log(\frac{2}{5})) \ = \ .9839$$

Gain of a_2= .9911-.9839 = **.0072**

c.

	1	3	4	5	6	7	8	
Split	0.5	2	3.5	4.5	5.5	6.5	7.5	8.5
	<=, >	<=,>	<=,>	<=,>	<=,>	<=,>	<=,>	<=,>
P(t)	0,4	1,3	1,3	2,2	2,2	3,1	4,0	4,0
P(-)	0,5	0,5	1,4	1,4	3,2	3,2	4,1	5,0

I only included gains, not every equation, except for the first one listed below:

$$P(\leq .5) * Entropy(0,0) \ + \ P(> .5) * Entropy(4,5)$$

Gain(1)=.9911-.9911= **0**
Gain(3)=.9911-.8484=**.1427**
Gain(4)=.9911-.9885=**.0026**
Gain(5)=.9911-.9183=**.0728**
Gain(6)=.9911-.9839=**.0072**
Gain(7)=.9911-.9728=**.0183**
Gain(8)=.9911-.8889=**.1022**

d. Best gain of a_3 is Gain(3)=.1427 since it is the largest.
Gain(a_2) = .0072, Gain(a_1) = .2296

Gain(a_1) has the highest value so it provides the best split of the 3.

5.

a. $[-(4/7) * log(4/7) - (\frac{3}{7}) * log(\frac{3}{7})] = .9852$

$[-(3/3) * log(3/3) - (\frac{0}{3}) * log(\frac{0}{3})] \ = \ 0$

Original - .9852 = .2813

Attribute B will be used to split the mode.

b. $G_{original}$ = 1-$(.4)^2$-$(.6)^2$ = .48

$$1 - (\tfrac{4}{7})^2 - (\tfrac{3}{7})^2 \; = \; .4898$$
$$1 - (\tfrac{3}{3})^2 - (\tfrac{0}{3})^2 \; = \; 0$$

$G_{original}$ - .4898 = .1371

$G_{original}$ - .1371-.4898 = .1633

Attribute B will be chosen to split the mode

c. Yes, even though some measures have same range behavior, the respective gain does not behave in the same mannar, as shown in parts a and b.

12.

a. Choose T10 on unseen data because it has better accuracy on unseen dataset. Additionally, it does not fit to noise of training dataset unlike T100.

b. This basically means that on an unseen dataset we prefer a model which performs better on the unseen portion, so we would choose T10 again.