Data Mining, CSCI 347, Fall 2019
Homework 2, Trees and evaluation, due Sept. 30

1. Calculate the bits of information for the following.

   Given 5 instances with two class values.

   Info([5,0]) = entropy(5/5,0/5)
   = -5/5 * $\log_2$(5/5) -0/5 * $\log_2$(0/5)
   = -1 * 0 - 0 * undefined
   = 0

   Info([4,1]) = 0.722 bits

   Info([3,2]) = 0.971 bits

   Info([2,3]) = 0.971 bits

   Info([1,4]) = 0.722 bits

   Info([0,5]) = 0 bits

2. Calculate the bits of information for the following.

   Given 4 instances with two class values.

   Info([4,0]) = 0 bits

   Info([3,1]) = 0.811 bits

   Info([2,2]) = 1.0 bits

   Info([1,3]) = 0.811 bits

   Info([0,4]) = 0 bits

3. Calculate the bits of information for the following splits with 3 class values.

Info([5,5,5])

Info([5,5,5]) = entropy(5/15,5/15,5/15)

$= -1/3 * \log_2(1/3) - 1/3 * \log_2(1/3) - 1/3 * \log_2(1/3)$

$= 3*[-1/3 * \log_2(1/3)]$

$= - \quad -1 \;\; * -1.585$

= 1.585 bits

Info([6,6,6])

Info([6,6,6]) = entropy(6/18,6/18,6/18)

$= -1/3 * \log_2(1/3) - 1/3 * \log_2(1/3) - 1/3 * \log_2(1/3)$

$= 3*[-1/3 * \log_2(1/3)]$

$= - \quad -1 \;\; * -1.585$

= 1.585 bits

4. Calculate the bits of information for the following splits with 4 class values.

Info([5,5,5,5])

Info([5,5,5,5]) = entropy(5/20,5/20,5/20,5/20)

$= -1/4 * \log_2(1/4) - 1/4 * \log_2(1/4) - 1/4 * \log_2(1/4) - 1/4 * \log_2(1/4)$

$= 4*[-1/4 * \log_2(1/4)]$

$= - \quad -1 \;\; * -2$

= 2 bits

5. Calculate the bits of information for the following splits with 8 class values.

Info([5,5,5,5,5,5,5,5])

Info([5,5,5,5,5,5,5,5]) = entropy(5/40,5/40,5/40,5/40, 5/40,5/40,5/40,5/40,))

$= 8* [-1/8 * \log_2(1/8)]$

$= \quad -1 \;\; * \quad -3$

= 3 bits

6. Look over your answers to questions 1-5. State patterns that you notice.

Info([a,b]) = Info([b,a])
Info([n,0]) is 0.
Info([n,n)] is 1

The number of bits is at its minimum when the node is pure (i.e. only one class value is represented). It is at its maximum when the class values are evenly split. The bits of information increases as the split becomes closer and closer to even.

The more class values being split between, the higher the entropy. More specifically, every time you increase the number of splits by a power of 2, the entropy goes up 1.

7. Use the measure of gain ratio to determine the best attribute to use as the root of a tree that predicts the contact-lens recommendation (none, soft or hard) when trained on the following dataset.

   This is the contact-lenses dataset with the tear_prod-rate attribute removed. Note that the first column (the numbers) is there for your convenience. It is not part of the data.

   Show all work.

Relation: contact-lenses-weka.filters.unsupervised.attribute.Remove-R4

No.	1: age Nominal	2: spectacle-prescrip Nominal	3: astigmatism Nominal	4: contact-lenses Nominal
1	young	myope	no	none
2	young	myope	no	soft
3	young	myope	yes	none
4	young	myope	yes	hard
5	young	hypermetrope	no	none
6	young	hypermetrope	no	soft
7	young	hypermetrope	yes	none
8	young	hypermetrope	yes	hard
9	pre-presbyopic	myope	no	none
10	pre-presbyopic	myope	no	soft
11	pre-presbyopic	myope	yes	none
12	pre-presbyopic	myope	yes	hard
13	pre-presbyopic	hypermetrope	no	none
14	pre-presbyopic	hypermetrope	no	soft
15	pre-presbyopic	hypermetrope	yes	none
16	pre-presbyopic	hypermetrope	yes	none
17	presbyopic	myope	no	none
18	presbyopic	myope	no	none
19	presbyopic	myope	yes	none
20	presbyopic	myope	yes	hard
21	presbyopic	hypermetrope	no	none
22	presbyopic	hypermetrope	no	soft
23	presbyopic	hypermetrope	yes	none
24	presbyopic	hypermetrope	yes	none

Use the order: [# of none, # of soft, # of hard]

Top node Info([15, 5, 4]) = 1.326 bits

Using the ‘age’ attribute as the root produces a 3-way split:

info([4, 2, 2]) = 1.5 bits
info([5, 2, 1]) = 1.299 bits and
info([6, 1, 1]) = 1.061 bits

Intrinsic value is info([8, 8, 8]) = 1.585 bits

$$\text{Overall gain ratio} = \frac{1.326 - (8/24*1.5 + 8/24 * 1.299 + 8/24 * 1.061)}{1.585}$$

$$= \frac{1.326 - \quad 1.287}{1.585}$$

$$= \frac{0.39}{1.585}$$

$$= 0.025$$

Using the ‘spectacle-prescrip’ attribute as the root produces a 2-way split:

info([7, 2, 3]) = 1.385 bits and
info([8, 3, 1]) = 1.189 bits

Intrinsic value is info([12, 12]) = 1.0 bits

$$\text{Overall gain ratio} = \frac{1.326 - (12/24*1.385 + 12/24 * 1.189)}{1.0}$$

$$= \frac{1.326 - \quad 1.287}{1.0}$$

$$= \frac{0.039}{1.0}$$

$$= 0.039$$

Using the ‘astigmatism attribute as the root produces a 2-way split:

info([8, 0, 4]) = 0.918 bits and

info([7, 5, 0]) = 0.980

Intrinsic value is info([12, 12]) = 1.0 bits

$$\text{Overall gain ratio} = \frac{1.326 - (12/24 * 0.918 + 12/24 * 0.980)}{1.0} = 0.79$$

$$= \frac{1.326 - 0.949}{1.0}$$

$$= \frac{0.377}{1.0}$$

$$= 0.377$$

The attribute astigmatism would be used as the root