CSIS734-01 Data Mining & Predictive Analytics
Garth Mortensen, mort0052@stthomas.edu

Graduate Program in Software
CSIS 734-01: Data Mining & Predictive Analytics
Assignment #5 (100 points)
Due Date: March 24th, 2018

1. Assume we are going to use the "Traffic Violation" from the following table as our target attribute in the classification analysis.

1.1 What is the system entropy before we begin building a decision tree.

$Entropy(System) = Entropy(p_1 \ldots p_n) = -p_1 log_2 p_1 - p_2 log_2 p_2 \ldots - p_n log_2 p_n$

1.2 Which attribute are you going to select as the first level node in the decision tree and why?

Because Information Gain for Entropy(TV, Seat Belt) is not as high as Information Gain Entropy(TV, Driving Condition), we should select the latter as the first level node. It results in a more pure split.

Driving Condition	Traffic Violation	Seat Belt
Alcohol-impaired	Speeding	No
Sober	None	Yes
Sober	No stop sign	Yes
Sober	Speeding	Yes
Sober	No traffic signal	No
Alcohol-impaired	No stop sign	Yes
Alcohol-impaired	None	Yes
Sober	No traffic signal	Yes
Alcohol-impaired	None	No
Sober	No traffic signal	No
Alcohol-impaired	Speeding	Yes
Sober	No stop sign	Yes
	Target	

CSIS734-01 Data Mining & Predictive Analytics
Garth Mortensen, mort0052@stthomas.edu
Graduate Program in Software
CSIS 734-01: Data Mining & Predictive Analytics
Assignment #5 (100 points)
Due Date: March 24th, 2018

Driving Condition	Traffic Violation	Seat Belt
Sober	No stop sign	Yes
Alcohol-impaired	No stop sign	Yes
Sober	No stop sign	Yes
Sober	No traffic signal	No
Sober	No traffic signal	Yes
Sober	No traffic signal	No
Sober	None	Yes
Alcohol-impaired	None	Yes
Alcohol-impaired	None	No
Alcohol-impaired	Speeding	No
Sober	Speeding	Yes
Alcohol-impaired	Speeding	Yes
	Target	

Question 1.1

What is the system entropy before we begin building a decision tree.

Classification 1 slides, page 31

Entropy in the Playing Golf Example

- $\text{Entropy}(S) = -p_1\log_2 p_1 - p_2\log_2 p_2 \ldots - p_n\log_2 p_n$
  - p_i is the percentage of S belonging to class i
    - p_1 is 9/14 (64.3%) of S is class **Y**
    - p_2 is 5/14 (35.7%) of S is class **N**
  - $\text{Entropy}(S) = -0.643\ \log_2 0.643 - 0.357\ \log_2 0.357 = 0.94$
  - System entropy BEFORE building DT process

Outlook	Temp.	Humid	Windy	Play
S	W	H	F	N
S	W	H	T	N
R	C	L	T	N
S	M	H	F	N
R	M	H	T	N
O	W	H	F	Y
R	M	H	F	Y
R	C	L	F	Y
O	C	L	T	Y
S	C	L	F	Y
R	M	L	F	Y
S	M	L	T	Y
O	M	H	T	Y
O	W	L	F	Y

Entropy(System) is calculated per attribute. Slide 31

Note - Max entropy for N classes is log2N. That is, if N = 4, log2(4) = 2, or log(4, 2)

Traffic Violations	Count	
No stop sign	3	25%
No traffic signal	3	25%
None	3	25%
Speeding	3	25%
Total	12	

Entropy(Traffic Violations) = Entropy($p_1 \ldots p_n$) = $-p_1 \log_2 p_1 - p_2 \log_2 p_2 \ldots - p_n \log_2 p_n$

= -(25%*log(2)*25%)-(25%*log(2)*25%)-(25%*log(2)*25%)-(25%*log(2)*25%)

= -(0.25*LOG(0.25,2))-(0.25*LOG(0.25,2))-(0.25*LOG(0.25,2))-(0.25*LOG(0.25,2))

Entropy(TV) = **2.00**

Seat Belt	Count	
Yes	8	67%
No	4	33%
Total	12	

Entropy(Seat Belt) = Entropy($p_1 \ldots p_n$) = $-p_1 \log_2 p_1 - p_2 \log_2 p_2 \ldots - p_n \log_2 p_n$

=-(67%*log(2)*67%)-(33%*log(2)*33%)

=-(0.666*LOG(0.666,2))-(0.333*LOG(0.333,2))

Entropy(SB) = **0.92**

Driving Condition	Count	
Alcohol-impaired	5	42%
Sober	7	58%
Total	12	

Entropy(Driving Condition) = Entropy($p_1 \ldots p_n$) = $-p_1 \log_2 p_1 - p_2 \log_2 p_2 \ldots - p_n \log_2 p_n$

= -(42%*log(2)*42%)-(58%*log(2)*58%)

= -(0.42*LOG(0.42,2))-(0.58*LOG(0.58,2))

Entropy(DC) = **0.98**

Question 1.2

Which attribute are you going to select as the first level node in the decision tree and why?

Entropy characterizes the purity of samples

Entropy (uncertainty) can be further reduced IF we begin by divide-and-conquer by selecting attributes in the RIGHT order
Information gain = (entropy before splitting) – (entropy after splitting on an attribute)

Step 1: Calculate the target entropy.

Traffic Violations	Count	%
No stop sign	3	0.25
No traffic signal	3	0.25
None	3	0.25
Speeding	3	0.25
Total	12	

Entropy(Traffic Violations) = Entropy($p_1 \ldots p_n$) = $-p_1 \log_2 p_1 - p_2 \log_2 p_2 \ldots - p_n \log_2 p_n$
= -(25%*log(2)*25%)-(25%*log(2)*25%)-(25%*log(2)*25%)-(25%*log(2)*25%)
= -(0.25*LOG(0.25,2))-(0.25*LOG(0.25,2))-(0.25*LOG(0.25,2))-(0.25*LOG(0.25,2))
Entropy(TV) = **2**

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated.
Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split.
The result is the Information Gain, or decrease in entropy.

	Traffic Violation										
Driving Condition	**None**	**No stop sign**	**No traffic signal**	**Speeding**	**Total**		**% of total**				
Sober	1	2	3	1	7	58%	14%	29%	43%	14%	
Alcohol-impaired	2	1	0	2	5	42%	40%	20%	0%	40%	
					12						

				% of total	**Weighted Entropy**
= -(0.14*LOG(0.14,2))-(0.29*LOG(0.29,2))-(0.43*LOG(0.43,2))-(0.14*LOG(0.14,2))	=	E(Sober)=	1.84	58%	1.07
= -(0.4*LOG(0.4,2))-(0.2*LOG(0.2,2))-(0.4*LOG(0.4,2))	=	E(Alcohol)=	1.52	42%	0.63
				Sum Entropy=	**1.70**

	Traffic Violation										
Seat Belt	**None**	**No stop sign**	**No traffic signal**	**Speeding**	**Total**		**% of total**				
Yes	2	3	1	2	8	67%	25%	38%	13%	25%	
No	1	0	2	1	4	33%	25%	0%	50%	25%	
					12						

				% of total	**Weighted Entropy**
= -(0.25*LOG(0.25,2))-(0.38*LOG(0.38,2))-(0.13*LOG(0.13,2))-(0.25*LOG(0.25,2))	=	E(Yes)=	1.91	67%	1.28
= -(0.25*LOG(0.25,2))-(0.5*LOG(0.5,2))-(0.25*LOG(0.25,2))	=	E(No)=	1.50	33%	0.50
				Sum Entropy=	**1.78**

Slide 31

Information Gain =	*Entropy(TV)*	- Entropy(TV, Driving Condition)	=	2-1.7 =	**0.30**	**=Gain(TV, DC)**	
Information Gain =	*Entropy(TV)*	- Entropy(TV, Seat Belt)	=	2-1.78 =	**0.22**	**=Gain(TV, SB)**	*Lower entropy, more pure split.*

Because Information Gain for Entropy(TV, Seat Belt) is not as high as Information Gain Entropy(TV, Driving Condition), we should select the latter as the first level node. It results in a more pure split.