Grant Pemberton — SID: 3034347047
Data X
Homework 7
October 18, 2018

Problem 1

Using this dataset identify the best feature to do the first split in a binary decision tree, so as to maximize the information gain in the next split, show your calculations

1 Entropy of a DataSet

HasJob	HasFamily	IsAbove30years	Defaulter
1	1	1	0
1	1	1	0
1	0	1	0
0	1	0	0
0	0	1	1
0	1	0	1
1	0	1	1
1	0	1	1

The entropy of the Defaulter Column is as follows:

H(Defaulter) = P(0) + P(1)

P(0) = -(4/8)log_2 (4/8) = .5 , P(1) = -(4/8)log_2 (4/8) = .5 → H(Defaulter) = .5 + .5 = **1**

1.1 Calculating Entropy of each feature

1.1.1 HasJob

For the 'HasJob' feature:

When 'HasJob' is equal to 0 (3 times out of 8), Defaulter is equal to 0 for 1/3 instances, and 1 for 2/3 instances. When 'HasJob' is equal to 1 (5 times out of 8), Defaulter is equal to 0 for 3/5 instances, and 1 for 2/5 instances. E('HasJob',0) = -(1/3)log_2 (1/3) = .91829

E('HasJob',1) = -(2/3)log_2 (2/3) = .9709505

So E('HasJob') = (.91829+.9709505)/2 = **.94462025**

1.1.2 HasFamily

For the 'HasFamily' feature:

When 'HasFamily' is equal to 0 (4 times out of 8), Defaulter is equal to 0 for 1/4 instances, and 1 for 3/4 instances. When 'HasFamily' is equal to 1 (4 times out of 8), Defaulter is equal to 0 for 3/4 instances, and 1 for 1/4 instances. E('HasFamily',0) = -(1/4)log_2 (1/4) = .811278

E('HasFamily',1) = -(3/4)log_2 (3/4) = .811278

So E('HasFamily') = (.811278+.811278)/2 = **.811278**

1.1.3 IsAbove30years

For the 'IsAbove30years' feature:

When 'IsAbove30years' is equal to 0 (2 times out of 8), Defaulter is equal to 0 for 1/2 instances, and 1 for 1/2 instances. When 'IsAbove30years' is equal to 1 (6 times out of 8),

Defaulter is equal to 0 for 3/6 instances, and 1 for 3/6 instances. E('IsAbove30years',0) =

-(1/2)log_2 (1/2) = 1 E('IsAbove30years',1) = -(3/6)log_2 (3/6) = 1

So E('IsAbove30years') = (1+1)/2 = **1**

To find the best feature to split on, we will calculate the information attained by each feature.

'HasJob': H(Y—'HasJob') = H(Defaulter) - H('HasJob') = 1-.94462025 = **0.05537975**

'HasFamily': H(Y—'HasFamily') = H(Defaulter) - H('HasFamily') = 1-.811278 = **0.188722**

'IsAbove30years': H(Y—'IsAbove30years') = H(Defaulter)-H('IsAbove30years') = 1-1 = **0**

Since 'HasFamily' gives the highest amount of information, it is the feature on which we should split first.

Problem 2

Given a signal of three symbols S=(A,B,C) and P(A) =0.7, P(B)=0.2, P(C)=0.1
What is the entropy of S? What does it mean according to the Source coding Theorem?

The Entropy of S is given by the following equation

Entropy = (-(p(A))log_2 (p(A)))+(-(p(B))log_2 (p(B)))+(-(p(C))log_2 (p(C)))

Entropy = (-(.7)log_2 (.7)+(-(.2)log_2 (.2)+(-(.1)log_2 (.1)) = **1.156779**

In the source coding theorem, this means that given a sequence of letters containing 70% A, 20% B, and 10% C, you would need at least 1.156779 bits to encode the information of that string of letters. Alternatively if you have the same string of letters and pick a letter at random, one would need to ask on average 1.156779 yes or no questions to find out which letter you picked.