Decision Trees Homework
CS 4499/5599

For this homework you will work through the steps of the ID3 algorithm for building a Decision Tree that decides the following dataset:

Meat N,Y	Crust D,S,T	Veg N,Y	Quality B,G,Gr
Y	Thin	N	Great
N	Deep	N	Bad
N	Stuffed	Y	Good
Y	Stuffed	Y	Great
Y	Deep	N	Good
Y	Deep	Y	Great
N	Thin	Y	Good
Y	Deep	N	Good
N	Thin	N	Bad

The information entropy of this dataset (which has 2 “Bad” instances, 4 “Good” instances, and 3 “Great” instances) is computed as:

$$Info(S) = -\sum_{i=1}^{|C|} p_i \log_2(p_i)$$

$$\text{Info(S)} = -\ 2/9 \cdot \log_2 2/9 - 4/9 \cdot \log_2 4/9 - 3/9 \cdot \log_2 3/9 = 1.53$$

Starting with all instances, calculate the information gain for each attribute (i.e., how much is entropy reduced if we split on each attribute?). For example, for the Meat attribute (which splits the dataset into two subsets based on the values of the Meat attribute), the information *entropy* is computed as

$$Info_A(S) = \sum_{i=j}^{|A|} \frac{|S_j|}{|S|} Info(S_j) = -\sum_{j=1}^{|A|} \frac{|S_j|}{|S|} \cdot \sum_{i=1}^{|C|} p_i \log_2(p_i)$$

$$\text{InfoMeat(S)} = 4/9 \cdot (-2/4 \log_2 2/4 - 2/4 \cdot \log_2 2/4 - 0 \cdot \log_2 0/4) + 5/9 \cdot (-0/5 \cdot \log_2 0/5 - 2/5 \cdot \log_2 2/5 - 3/5 \cdot \log_2 3/5) = .98$$

Thus the Information Gain for the Meat attribute is 1.53 - .98 = .55.

1. Compute the Information Gain for all other features at this level.
2. Find best attribute (i.e., highest Information *Gain*—not Entropy) and split
3. Find the best attribute for at least the left most node at the next level **assuming sub-nodes are sorted alphabetically left to right by attribute/value**