CS 5350: Machine Learning HW1

Richard Child u0581030

February 11, 2019

1 Decision Tree [40 points + 10 bonus]

1. [7 points] Decision tree construction.

   (a) [5 points]

$$ID3(S, Attributes, Labels)$$
$$S = \text{full dataset}, Attributes = \{x_1, x_2, x_3, x_4\}, Labels = \{0, 0, 1, 1, 0, 0, 0\}$$

Create root node
A = attribute with maximum Information Gain

$$Entropy(S) = -\frac{2}{7}log_2(\frac{2}{7}) - \frac{5}{7}log_2(\frac{5}{7}) = 0.863$$
$$Gain(S, x_1) = 0.863 - \sum_{v \epsilon Values_{x_1}} \frac{\mid S_v \mid}{\mid S \mid} Entropy(S_v)$$
$$Entropy(x_1 = 0) = -\frac{1}{5}log_2(\frac{1}{5}) - \frac{4}{5}log_2(\frac{4}{5}) = 0.722; Entropy(x_1 = 1) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$$
$$Gain(S, x_1) = 0.863 - \left[\frac{5}{7}(0.722) + \frac{2}{7}(1)\right] = 0.062$$
$$Gain(S, x_2) = 0.863 - \sum_{v \epsilon Values_{x_2}} \frac{\mid S_v \mid}{\mid S \mid} Entropy(S_v)$$
$$Entropy(x_2 = 0) = -\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.918; Entropy(x_2 = 1) = -\frac{0}{4}log_2(\frac{0}{4}) - \frac{4}{4}log_2(\frac{4}{4}) = 0$$
$$Gain(S, x_2) = 0.863 - \left[\frac{5}{7}(0.722) + \frac{2}{7}(1)\right] = 0.470$$
$$Gain(S, x_3) = 0.863 - \sum_{v \epsilon Values_{x_3}} \frac{\mid S_v \mid}{\mid S \mid} Entropy(S_v)$$
$$Entpy(x_3 = 0) = -\frac{1}{4}log(\frac{1}{4}) - \frac{3}{4}log(\frac{3}{4}) = 0.811; Entpy(x_3 = 1) = -\frac{1}{3}log(\frac{1}{3}) - \frac{2}{3}log(\frac{2}{3}) = 0.918$$
$$Gain(S, x_3) = 0.863 - \left[\frac{4}{7}(0.811) + \frac{3}{7}(0.918)\right] = 0.006$$
$$Gain(S, x_4) = 0.863 - \sum_{v \epsilon Values_{x_4}} \frac{\mid S_v \mid}{\mid S \mid} Entropy(S_v)$$
$$Entropy(x_4 = 0) = -\frac{0}{4}log_2(\frac{0}{4}) - \frac{4}{4}log_2(\frac{4}{4}) = 1; Entropy(x_4 = 1) = -\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.918$$
$$Gain(S, x_4) = 0.863 - \left[\frac{4}{7}(0) + \frac{3}{7}(0.918)\right] = 0.470$$

There is a tie for maximum Information Gain $Gain(S, x_2) = Gain(S, x_4) = 0.470$. We will arbitrarily use x_4 to split on.
Add two new branches to tree corresponding to possible values of x_4 which are 0 and 1. Will again perform ID3 on each branch.

$$ID3(S_{x_4=0}, Attributes, Labels)$$
$$S_{x_4=0} = \text{examples where } x_4 = 0, Attributes = \{x_1, x_2, x_3\}, Labels = \{0, 0, 0, 0\}$$

All labels in this example subset are equal, therefore a leaf node is returned.
Now perform ID3 on $x_4 = 1$

$$ID3(S_{x_4=1}, Attributes, Labels)$$
$$S_{x_4=1} = \text{examples where } x_4 = 1, Attributes = \{x_1, x_2, x_3\}, Labels = \{1, 1, 0\}$$

Create a root node for the tree.
A = attribute with maximum Information Gain on $S_{x_4=1}$

$$Entropy(S_{x_4=1}) = -\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.918$$
$$Gain(S_{x_4=1}, x_1) = 0.918 - \sum_{v \epsilon Values_{x_1}} \frac{|S_v|}{|S_{x_4=1}|} Entropy(v)$$
$$Entropy(x_1 = 0) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1; Entropy(x_1 = 1) = -\frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$$
$$Gain(S_{x_4=1}, x_1) = 0.918 - \left[\frac{2}{3}(1) + \frac{1}{3}(0)\right] = 0.251$$
$$Gain(S_{x_4=1}, x_2) = 0.918 - \sum_{v \epsilon Values_{x_2}} \frac{|S_v|}{|S_{x_4=1}|} Entropy(v)$$
$$Entropy(x_2 = 0) = -\frac{2}{2}log_2(\frac{2}{2}) - \frac{0}{2}log_2(\frac{0}{2}) = 0; Entropy(x_2 = 1) = -\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1}) = 0$$
$$Gain(S_{x_4=1}, x_2) = 0.918 - \left[\frac{2}{3}(0) + \frac{1}{3}(0)\right] = 0.918$$
$$Gain(S_{x_4=1}, x_3) = 0.918 - \sum_{v \epsilon Values_{x_3}} \frac{|S_v|}{|S_{x_4=1}|} Entropy(v)$$
$$Entropy(x_3 = 0) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1; Entropy(x_3 = 1) = -\frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$$
$$Gain(S_{x_4=1}, x_2) = 0.918 - \left[\frac{2}{3}(1) + \frac{1}{3}(0)\right] = 0.251$$

A = x_2, so split on x_2 then add two branches to tree 0 and 1.
The subgroup $S_{x_4=1,x_2=0}$ is non-empty and the subgroup $S_{x_4=1,x_2=1}$ is also non-empty. Perform ID3 on both branches.
First we perform ID3 on $x_2 = 0$ given $x_4 = 1$.

$$ID3(S_{x_4=1,x_2=0}, Attributes, Labels)$$
$$S_{x_4=1,x_2=0} = \text{examples where } x_4 = 1 \text{ and } x_2 = 0, Attributes = x_1, x_3, Labels = \{1, 1\}$$

All labels for this subset are equal, therefore a leaf node is returned.
Now perform ID3 on $x_4 = 1$ and $x_2 = 1$

$$ID3(S_{x_4=1,x_2=1}, Attributes, Labels)$$
$$S_{x_4=1,x_2=0} = \text{examples where } x_4 = 1 \text{ and } x_2 = 1, Attributes = x_1, x_3, Labels = \{0\}$$

All labels for this subset are equal, therefore a leaf node is returned.
The tree is now completely constructed as each branch of the tree has a leaf node.

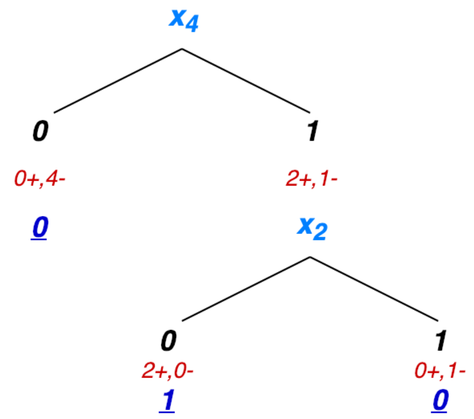

Figure 1: Tree structure

(b) [2 points] The boolean function for this tree is $x_2 \wedge x_4$. Truth table for this function is as follows:

x_1	x_2	x_3	x_4	y
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	0
0	1	0	0	0
0	1	0	1	1
0	1	1	0	0
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

Table 1: Truth table for $x_2 \wedge x_4$

2. [17 points] Decision tree construction using Majority Error and Gini Index

(a) [7 points] Use majority error (ME) to calculate the gain, and select the best feature to split the data in ID3 framework.

$$ID3(S, Attributes, Labels)$$
$$S = \text{full dataset}, Attributes = \{Outlook, Temperature, Humidity, Wind\}, Labels = \{Yes, No\}$$

Create root node.
A = attribute with maximum Gain using majority error

$$ME(S) = \frac{5}{14} = 0.357$$
$$Gain(S, Outlook) = 0.357 - \sum_{v \epsilon Values_O} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{O=sunny}) = \frac{2}{5} = 0.400, ME(S_{O=overcast}) = \frac{0}{4} = 0, ME(S_{O=rainy}) = \frac{2}{5} = 0.400$$
$$Gain(S, Outlook) = 0.357 - \left[\frac{5}{14}(0.4) + \frac{4}{15}(0) + \frac{5}{14}(0.4)\right] = 0.071$$
$$Gain(S, Temp) = 0.357 - \sum_{v \epsilon Values_T} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{T=hot}) = \frac{2}{4} = 0.5, ME(S_{T=med}) = \frac{2}{6} = 0.333, ME(S_{T=cool}) = \frac{1}{4} = 0.25$$
$$Gain(S, Temp) = 0.357 - \left[\frac{4}{14}(0.5) + \frac{6}{14}(0.333) + \frac{4}{14}(0.25)\right] = 0.0$$
$$Gain(S, Humidity) = 0.357 - \sum_{v \epsilon Values_H} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{H=high}) = \frac{3}{7} = 0.429; ME(S_{H=Normal}) = \frac{1}{7} = 0.143$$
$$Gain(S, Humidity) = 0.357 - \left[\frac{7}{14}(0.429) + \frac{7}{14}(0.143)\right] = 0.071$$
$$Gain(S, Wind) = 0.357 - \sum_{v \epsilon Values_W} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{W=strong}) = \frac{3}{6} = 0.5; ME(S_{W=weak}) = \frac{2}{8} = 0.25$$
$$Gain(S, Wind) = 0.357 - \left[\frac{6}{14}(0.5) + \frac{8}{14}(0.25)\right] = 0.0$$

Both *Outlook* and *Humidity* have the same majority error equal to 0.071, we will arbitrarily pick *Outlook* as the attribute to split on the first level. So now we need to add 3 branches to the node: $\{Overcast, Sunny, Rainy\}$. Now the subgroups $S_{O=overcast}$, $S_{O=sunny}$, and $S_{O=rainy}$ are all non-empty, so we must perform $ID3$ on each of these branches. First we consider $S_{O=overcast}$.

$$ID3(S_{O=overcast}, Attributes, Labels)$$
$$S_{O=overcast} = \text{examples where Outlook is Overcast}$$
$$Attributes = \{Temp, Humidity, Wind\}, Labels = \{yes, no\}$$

All labels for this subset are equal, therefore a leaf node is returned. Now we will consider $S_{O=rainy}$.

$$ID3(S_{O=rainy}, Attributes, Labels)$$
$$S_{O=rainy} = \text{examples where Outlook is Rainy}$$
$$Attributes = \{Temp, Humidity, Wind\}, Labels = \{yes, no\}$$

Create root node.
A = attribute with the highest Gain using majority error

$$ME(S_{O=rainy}) = \frac{2}{5} = 0.4$$
$$Gain(S_{O=rainy}, Temp) = 0.4 - \sum_{v \epsilon Values} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{O=rainy,Temp=hot}) = 0.0; ME(S_{O=rainy,Temp=med}) = \frac{1}{3} = 0.333; ME(S_{O=cold}) = \frac{1}{2} = 0.5$$
$$Gain(S_{O=rainy}, Temp) = 0.4 - \left[\frac{3}{5}(0.333) + \frac{2}{5}(0.5)\right] = 0.0$$
$$Gain(S_{O=rainy}, Humid) = 0.4 - \sum_{v \epsilon Values} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{O=rainy,Humid=high}) = \frac{1}{2} = 0.5; ME(S_{O=rainy,Humid=norm}) = \frac{1}{3} = 0.333$$
$$Gain(S_{O=rainy}, Humid) = 0.4 - \left[\frac{2}{5}(0.5) + \frac{3}{5}(0.333)\right] = Gain(S_{O=rainy}, Wind) = 0.4 - \sum_{v \epsilon Values} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{O=rainy,Wind=strong}) = 0.0; ME(S_{O=rainy,Wind=weak}) = 0.0$$
$$Gain(S_{O=rainy}, Wind) = 0.4$$

Wind has the highest information gain via majority error 0.4. We will split on $Wind$ and create two branches under this node $\{Strong, Weak\}$. The subgroups $S_{O=r,W=s}$ and $S_{O=r,W=w}$ are both non-empty, so we must perfom $ID3$ on each branch. First the branch of $S_{O=r,W=s}$.

$$ID3(S_{O=r,W=s}, Attributes, Labels)$$
$$S_{O=r,W=s} = \text{examples where } Outlook = rainy \text{ and } Wind = strong$$
$$Attributes = \{Temp, Humid\}; Labels = \{yes, no\}$$

This subset contains labels that are all equal, therefore leaf node is returned. Now we consider branch $S_{O=r,W=w}$.

$$ID3(S_{O=r,W=w}, Attributes, Labels)$$
$$S_{O=r,W=w} = \text{examples where } Outlook = rainy \text{ and } Wind = weak$$
$$Attributes = \{Temp, Humid\}; Labels = \{yes, no\}$$

All labels in this subset are equal, therefore a leaf node is returned for this branch. Now we go back up the tree to consider the branch $S_{O=sunny}$. We perform $ID3$ as follows.

$$ID3(S_{O=sunny}, Attributes, Labels)$$
$$S_{O=sunny} = \text{examples where } Outlook = sunny$$
$$Attributes = \{Temp, Humid, Wind\}; Labels\{yes, no\}$$

Create root node.
A = attribute that maximizes information gain using Majority Error.

$$ME(S_{O=s}) = \frac{2}{5} = 0.4$$
$$Gain(S_{O=s,Temp}) = 0.4 - \sum_{v \epsilon Values_T} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{O=s,T=h}) = \frac{0}{2} = 0.0; ME(S_{O=s,T=m}) = \frac{1}{2} = 0.5; ME(S_{O=s,T=c}) = \frac{0}{1} = 0.0$$
$$Gain(S_{O=s,Temp}) = 0.4 - \left[\frac{2}{5}(0) + \frac{2}{5}(0.5) + \frac{1}{5}(0)\right] = 0.2$$
$$Gain(S_{O=sunny,Humid}) = 0.4 - \sum_{v \epsilon Values_H} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{O=s,H=h}) = \frac{0}{3} = 0.0; ME(S_{O=s,H=n}) = \frac{0}{2} = 0.0$$
$$Gain(S_{O=sunny,Humid}) = 0.4 - \left[\frac{3}{5}(0) + \frac{2}{5}(0)\right] = 0.4$$
$$Gain(S_{O=sunny,Wind}) = 0.4 - \sum_{v \epsilon Values_W} \frac{\mid S_v \mid}{\mid S \mid} ME(S_v)$$
$$ME(S_{O=s,W=s}) = \frac{1}{2} = 0.5; ME(S_{O=s,W=w} = \frac{1}{3} = 0.333$$
$$Gain(S_{O=sunny,Wind}) = 0.4 - \left[\frac{2}{5}(0.5) + \frac{3}{5}(0.333)\right] = 0.0$$

The attribute to split on is *Humidity* with gain of 0.4. $S_{O=s,H=h}$ and $S_{O=s,H=n}$ are both non-empty so we will add those branches to the node perfom $ID3$ on them, but $S_{O=s,H=l}$ is empty so that branch will not be added to the node. First we perform $ID3$ on $S_{O=s,H=h}$

$$ID3(S_{O=s,H=h}, Attributes, Labels)$$
$$S_{O=s,H=h} = \text{examples where } Outlook = sunny \text{ and } Humidity = high$$
$$Attributes = \{Temp, Wind\}; Labels = \{yes, no\}$$

All labels of $S_{O=s,H=h}$ are equal so we return a leaf node. Now we perform $ID3$ on $S_{O=s,H=n}$.

$$ID3(S_{O=s,H=n}, Attributes, Labels)$$
$$S_{O=s,H=n} = \text{examples where } Outlook = sunny \text{ and } Humidity = normal$$
$$Attributes = \{Temp, Wind\}; Labels = \{yes, no\}$$

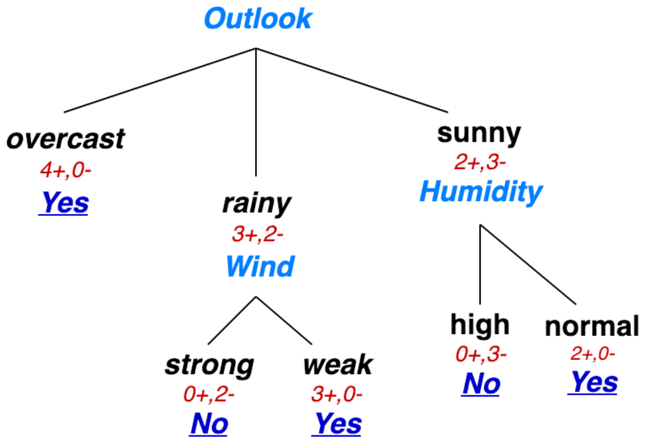

Figure 2: Tree structure for Tennis data using Majority Error

All labels of $S_{O=s,H=n}$ are equal so we return a leaf node.
At this point the entire tree has been built and it can be pictured as follows.

(b) [7 points] Use gini index (GI) to calculate the gain, and conduct tree learning with ID3 framework.

$$Gain(S, A) = GI(S) - \sum_{v\epsilon Values_A} \frac{\mid S_v \mid}{\mid S \mid} GI(S_v)$$

$$GI(S) = 1 - \sum_{t=1}^{k} p_t^2$$

$$ID3(S, Attributes, Labels)$$

$$S = \text{full dataset}, Attributes = \{Outlook, Temperature, Humidity, Wind\}, Labels = \{Yes, No\}$$

Create root node.
A = attribute with maximum Gain using gini index

$$GI(S) = 1 - \frac{9}{14}^2 - \frac{5}{14}^2 = 0.459$$

First calculate *Gain* for *Outlook*

$$GI(S_{O=s}) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48; GI(S_{O=r}) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48; GI(S_{O=o}) = 1 - (\frac{4}{4})^2 = 0.0$$

$$Gain(S, O) = 0.459 - \left[\frac{5}{14}(0.48) + \frac{5}{14}(0.48) + \frac{4}{14}(0)\right] = 0.116$$

Now calculate *Gain* for *Temp*

$$GI(S_{T=h}) = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5; GI(S_{T=m}) = 1 - (\frac{4}{6})^2 - (\frac{2}{6})^2 = 0.44; GI(S_{T=c}) = 1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = 0.375$$

$$Gain(S, T) = 0.459 - \left[\frac{4}{14}(0.5) + \frac{6}{14}(0.444) + \frac{4}{14}(0.375)\right] = 0.019$$

Now calculate *Gain* for *Humidity*

$$GI(S_{H=h}) = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.49; GI(S_{H=n}) = 1 - (\frac{6}{7})^2 - (\frac{1}{7})^2 = 0.245$$

$$Gain(S, H) = 0.459 - \left[\frac{7}{14}(0.49) + \frac{7}{14}(0.245)\right] = 0.092$$

$$\text{Finally calculate } Gain \text{ for } Wind$$

$$GI(S_{W=s}) = 1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5; GI(S_{W=w}) = 1 - (\frac{6}{8})^2 - (\frac{2}{8})^2 = 0.375$$

$$Gain(S, W) = 0.459 - \left[\frac{6}{14}(0.5) + \frac{8}{14}(0.375)\right] = 0.030$$

The dataset will be split on the attribute *Outlook* since it has the highest information gain via gini index. We add three branches to the node corresponding to $\{Overcast, Sunny, Rainy\}$. The subgroups $S_{O=o}$,$S_{O=s}$,and$S_{O=r}$ are all non-empty so we will perform the $ID3$ on each of them.

$$ID3(S_{O=o}, Attributes, Labels)$$

$$S_O = \text{examples where } Outlook = overcast, Attributes = \{Temp, Humid, Wind\}, Labels = \{yes\}$$

The labels of this subset are all equal so we return a leaf node with label *yes*. Next we will perform $ID3$ on $S_{O=s}$.

$$ID3(S_{O=s}, Attributes, Labels)$$

$$S_{O=s} = \text{examples where } Outlook = sunny, Attributes = \{Temp, Humid, Wind\}, Labels = \{yes, no\}$$

Create root node.
A = the attribute that maximizes information gain using gini index

$$GI(S_{O=s}) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$$

$$\text{First consider } Outlook = s \text{ and } Temp$$

$$GI(S_{O=s,T=h}) = 1 - (\frac{2}{2})^2 = 0; GI(S_{O=s,T=m}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5; GI(S_{O=s,T=c}) = 1 - (\frac{1}{1})^2 = 0$$

$$Gain(S_{O=s,T}) = 0.48 - \left[\frac{2}{5}(0) + \frac{2}{5}(0.5) + \frac{1}{5}(0)\right] = 0.28$$

$$\text{Now consider } Outlook = s \text{ and } Humidity$$

$$GI(S_{O=s,H=h}) = 1 - (\frac{3}{3})^2 - (\frac{0}{3})^2 = 0.0; GI(S_{O=s,H=n}) = 1 - (\frac{2}{2})^2 - (\frac{0}{2})^2 = 0.0$$

$$Gain(S_{O=s,H}) = 0.48$$

$$\text{Finally consider } Outlook = s \text{ and } Wind$$

$$GI(S_{O=s,W=s}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5; GI(S_{O=s,W=w}) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = 0.444$$

$$Gain(S_{O=s,W}) = 0.48 - \left[\frac{2}{5}(0.5) + \frac{3}{5}(0.444)\right] = 0.014$$

The attribute to split on is *Humidity* with an information gain of 0.48 using gini index. We will add the branches for *Humidity* values *High* and *Normal* since those subgroups are non-empty, but the subgroup for *Low* is empty so it is not added. Now we perform $ID3$ on the branches of *Humidity*.

$$ID3(S_{O=s,H=h}, Attributes, Labels)$$

$$S_{O=s,H=h} = \text{examples where } Outlook = sunny \text{ and } Humidity = high$$

This subset has all the same label of *no* so a leaf node is returned. Now we will consider the other branch from $Humidity$.

$$ID3(S_{O=s,H=n}, Attributes, Labels)$$
$$S_{O=s,H=n} = \text{examples where } Outlook = sunny \text{ and } Humidity = normal$$

All the examples in this subset have the same label of *yes* so a leaf node is returned. We have considered all *Humidity* branches, now we move back up to finish the last branch of *Outlook* which is the *Rainy* branch.

$$ID3(S_{O=r}, Attributes, Labels)$$
$$S_{O=r} = \text{examples where } Outlook = rainy, Attributes = \{Temp, Humid, Wind\}, Label = \{yes, no\}$$

Create a root node.
A = the attribute that best splits data, maximizes information gain using gini index.

$$GI(S_{O=r}) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$$
First we consider data set $Outook = r$ and $Temp$
$$GI(S_{O=r,T=m}) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = 0.44; GI(S_{O=r,T=c}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$$
$$Gain(S_{O=r,T}) = 0.48 - \left[\frac{3}{5}(0.44) + \frac{2}{5}(0.5)\right] = 0.016$$
Next we consider data set $Outlook = r$ and $Humidity$
$$GI(S_{O=r,H=h}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5; GI(S_{O=r,H=n}) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 = 0.44$$
$$Gain(S_{O=r,H}) = 0.48 - \left[\frac{2}{5}(0.5) + \frac{3}{5}(0.44)\right] = 0.016$$
Finally we consider data set $Outlook = r$ and $Wind$
$$GI(S_{O=r,W=s}) = 1 - (\frac{2}{2})^2 = 0; GI(S_{O=r,W=w}) = 1 - (\frac{3}{3})^2 = 0$$
$$Gain(S_{O=r,W}) = 0.48$$

The attribute is $Wind$ to split with gain equals 0.48. We will add the two branches to this node *Strong* and $Weak$. Then we perform $ID3$ on each subgroup. Both $S_{O=r,W=s}$ and $S_{O=r,W=w}$ have uniform labels among their examples. Therefore we add a leaf node for each branch and return. $S_{O=r,W=s}$ has a *no* leaf node and $S_{O=r,W=w}$ has a *yes* leaf node. These are all the remaining branches for the algorithm. A tree structure is as follows.

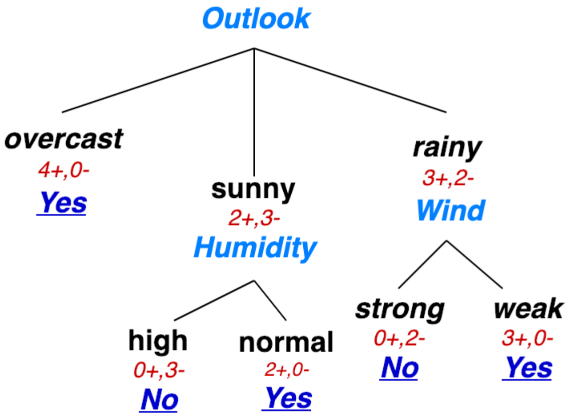

Figure 3: Tree structure for Tennis data using Gini Index

(c) [3 points] The two trees we have just created using Majority Error and Gini Index are equivalent to the tree we created in class. But, they could have been different. There were ties for which attribute to split on initially in the Majority Error tree. So instead of choosing to split on *Outlook* we could have initially split on *Humidity* and it could have been different.

3. [16 points] Continue with the same training data in Problem 2. Suppose before the tree construction, we receive one more training instance where Outlook's value is missing: {Outlook: Missing, Temperature: Mild, Humidity: Normal, Wind: Weak, Play: Yes}.

(a) [3 points] Use the most common value in the training data as the missing value, and calculate the information gains of the four features.
We will assign the missing feature value to $Outlook = Sunny$ since it is the most common.

$$Entropy(S) = -\frac{10}{15}log_2(\frac{10}{15}) - \frac{5}{15}log_2(\frac{5}{15}) = 0.918$$
$$Gain(S,O) = 0.918 - \left[\frac{6}{15}(1) + \frac{4}{15}(0) + \frac{5}{15}(0.971)\right] = 0.194$$
$$Gain(S,T) = 0.918 - \left[\frac{4}{15}(1) + \frac{7}{15}(0.863) + \frac{4}{15}(0.811)\right] = 0.032$$
$$Gain(S,H) = 0.918 - \left[\frac{7}{15}(0.985) + \frac{8}{15}(0.544)\right] = 0.168$$
$$Gain(S,W) = 0.918 - \left[\frac{9}{15}(0.764) + \frac{6}{15}(1)\right] = 0.111$$

The best attribute to split on is *Outlook*.

(b) [3 points] Use the most common value among the training instances with the same label.
The most common value among "Yes" labels for *Outlook* attribute is *Overcast*.

$$Entropy(S) = -\frac{10}{15}log_2(\frac{10}{15}) - \frac{5}{15}log_2(\frac{5}{15}) = 0.918$$
$$Gain(S,O) = 0.918 - \left[\frac{5}{15}(0.971) + \frac{5}{15}(0) + \frac{5}{15}(0.971)\right] = 0.271$$
$$Gain(S,T) = 0.918 - \left[\frac{4}{15}(1) + \frac{7}{15}(0.863) + \frac{4}{15}(0.811)\right] = 0.032$$

$$Gain(S, H) = 0.918 - \left[\frac{7}{15}(0.985) + \frac{8}{15}(0.544)\right] = 0.168$$
$$Gain(S, W) = 0.918 - \left[\frac{9}{15}(0.764) + \frac{6}{15}(1)\right] = 0.111$$

Again the best attribute to split on is *Outlook*.

(c) [3 points] Use the fractional counts to infer the feature values, and then calculate the information gains of the four features. Indicate the best feature.

(d) [7 points] Continue with the fractional examples, and build the whole free with information gain. List every step and the final tree structure.

4. [**Bonus question 1**] [5 points]. Prove that the information gain is always non-negative. That means, as long as we split the data, the purity will never get worse! (Hint: use convexity)

5. [**Bonus question 2**] [5 points]. We have discussed how to use decision tree for regression (i.e., predict numerical values) — on the leaf node, we simply use the average of the (numerical) labels as the prediction. Now, to construct a regression tree, can you invent a gain to select the best attribute to split data in ID3 framework?

2 Decision Tree Practice [60 points]

1. [5 Points] The Github repository is `https://github.com/richardlynnchild/machine-learning.git`.

2. [30 points]

(a) [15 points] The ID3 algorithm was implemented in the *DecisionTree/DecisionTree.py* file.

(b) [10 points] The implementation of ID3 was used with varying tree depths from 1 to 6. The six different trees were used to predict values in the training and testing data sets. The following table lists the prediction errors for each information gain heuristic and tree depth level for each dataset.

Depth	ME_{Train}	$Gini_{Train}$	$Entropy_{Train}$	ME_{Test}	$Gini_{Test}$	$Entropy_{Test}$
1	0.302	0.302	0.302	0.297	0.297	0.297
2	0.301	0.222	0.222	0.316	0.223	0.223
3	0.189	0.176	0.181	0.224	0.184	0.196
4	0.097	0.089	0.082	0.166	0.137	0.151
5	0.029	0.027	0.027	0.091	0.084	0.084
6	0.000	0.000	0.000	0.091	0.084	0.084

Table 2: Prediction errors for Train and Test data sets

(c) [5 points] We can conclude that the training errors decrease at a higher rate as the tree depth increases. The training errors are able to get to zero while the testing errors cannot quite reach zero.

3. [25 points]

(a) [10 points] The following table contains prediction error rates for all three heuristics. Both the training and testing data sets were used with tree depth from 1 to 16.

Depth	ME_{Train}	$Gini_{Train}$	$Entropy_{Train}$	ME_{Test}	$Gini_{Test}$	$Entropy_{Test}$
1	0.109	0.109	0.119	0.117	0.117	0.125
2	0.104	0.104	0.106	0.109	0.109	0.111
3	0.096	0.093	0.101	0.114	0.115	0.107
4	0.083	0.075	0.079	0.117	0.121	0.120
5	0.071	0.060	0.061	0.120	0.134	0.128
6	0.065	0.047	0.047	0.122	0.149	0.138
7	0.062	0.035	0.035	0.122	0.157	0.147
8	0.056	0.027	0.029	0.128	0.164	0.150
9	0.049	0.021	0.023	0.132	0.172	0.159
10	0.042	0.017	0.017	0.139	0.174	0.161
11	0.037	0.015	0.014	0.146	0.175	0.162
12	0.030	0.014	0.014	0.153	0.176	0.164
13	0.025	0.014	0.014	0.161	0.176	0.165
14	0.020	0.014	0.014	0.167	0.176	0.165
15	0.016	0.014	0.014	0.168	0.176	0.165
16	0.014	0.014	0.014	0.170	0.176	0.165

Table 3: Prediction errors for Bank train and test data sets, keeping unknowns

(b) [10 points] The following table are prediction error rates for all three heuristics, ranging from tree depth 1 to 16, but the 'unknown' values were substituted with majority values.

Depth	ME_{Train}	$Gini_{Train}$	$Entropy_{Train}$	ME_{Test}	$Gini_{Test}$	$Entropy_{Test}$
1	0.109	0.109	0.119	0.117	0.117	0.125
2	0.105	0.105	0.106	0.110	0.110	0.111
3	0.098	0.101	0.102	0.114	0.108	0.109
4	0.086	0.088	0.087	0.116	0.115	0.116
5	0.078	0.074	0.071	0.117	0.127	0.127
6	0.072	0.057	0.057	0.122	0.135	0.135
7	0.069	0.045	0.045	0.123	0.149	0.143
8	0.066	0.037	0.039	0.125	0.150	0.147
9	0.061	0.029	0.032	0.129	0.160	0.155
10	0.056	0.025	0.026	0.133	0.166	0.164
11	0.051	0.022	0.023	0.138	0.164	0.161
12	0.045	0.022	0.022	0.141	0.165	0.162
13	0.037	0.022	0.022	0.147	0.165	0.161
14	0.031	0.022	0.022	0.151	0.165	0.161
15	0.026	0.022	0.022	0.157	0.165	0.161
16	0.022	0.022	0.022	0.157	0.165	0.161

Table 4: Prediction errors for Bank train and test data sets, replacing unknowns

(c) [5 points] We can conclude that the test data predictions become more accurate with more depth. The difference between keeping 'unknown' values and replacing them is slight, but it is enough to see it makes an improvement in the testing data set. It also seems that there is a point of diminishing returns for testing data as tree depth increases. There doesn't appear to be any improvement past level 3.