CS4445 Data Mining and Knowledge Discovery in Databases B Term 2014

Project 2. Part I. INDIVIDUAL HOMEWORK ASSIGNMENT

Solutions by Chiying Wang

Consider the following dataset.

```
@relation simple-weather
@attribute outlook {sunny,overcast,rainy}
@attribute humidity numeric
@attribute windy {TRUE,FALSE}
@attribute play {yes,no}
@data

sunny,    80, FALSE, no
sunny,    90, TRUE,  no
overcast, 80, FALSE, yes
rainy,    96, FALSE, yes
rainy,    80, FALSE, yes
rainy,    72, TRUE,  no
overcast, 72, TRUE,  yes
sunny,    96, FALSE, no
sunny,    72, FALSE, yes
rainy,    80, FALSE, yes
sunny,    72, TRUE,  yes
overcast, 90, TRUE,  yes
overcast, 80, TRUE,  yes
rainy,    96, TRUE,  no
```

where the *play* attribute is the classification target.

1. (30 points) Construct the full ID3 decision tree using entropy to rank the predicting attributes (*outlook, humidity, and windy*) with respect to the target/classification attribute (*play*). Keep *humidity* as a numeric attribute (do not discretize it). Show all the steps of the calculations. Make sure you compute log in base b (for the appropriate b) correctly as some calculators don't have a log_b primitive for all b's. Also, state explicitly in your tree what instances exactly belong to each tree node using the line numbers provided next to each data instance in the dataset above.

Solution:

(0) Calculating the entropy of the entire dataset:

The dataset is sorted by the attribute "Humidity" as shown in the following table.

No.	Outlook	Humidity	Windy	Play
1	rainy	72	TRUE	no
2	overcast	72	TRUE	yes
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes
5	sunny	80	FALSE	no
6	overcast	80	FALSE	yes
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
9	overcast	80	TRUE	yes
10	sunny	90	TRUE	no
11	overcast	90	TRUE	yes
12	rainy	96	FALSE	yes
13	sunny	96	FALSE	no
14	rainy	96	TRUE	no

The general entropy of the dataset is:

$$\text{entropy(Dataset)} = \text{entropy}([9,5]) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} \approx 0.940$$

(1) <u>Picking the attribute for the root node of the decision tree</u>:

There are three attributes: outlook, humidity, and windy to choose from as the root node of the decision tree. For each attribute, we compute the information gain given the attribute value and pick the attribute with the largest information gain as the root node in the decision tree.

The candidate conditions that can be used to split the dataset instances in the root node are: Outlook, Humidity≤76, Humidity≤85, Humidity≤93, and Windy.

(1.1) <u>Entropy and information gain for Outlook</u>:

When outlook = "sunny", the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes
5	sunny	80	FALSE	no
10	sunny	90	TRUE	no
13	sunny	96	FALSE	no

$$\text{entropy(outlook} = \text{sunny)} = \text{entropy}([2,3]) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} \approx 0.971$$

When outlook = “rainy”, the instances are:

No.	Outlook	Humidity	Windy	Play
1	rainy	72	TRUE	no
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
12	rainy	96	FALSE	yes
14	rainy	96	TRUE	no

$$\text{entropy}(\text{outlook} = \text{rainy}) = \text{entropy}([3,2]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} \approx 0.971$$

When outlook = “overcast”, the instances are:

No.	Outlook	Humidity	Windy	Play
2	overcast	72	TRUE	yes
6	overcast	80	FALSE	yes
9	overcast	80	TRUE	yes
11	overcast	90	TRUE	yes

$$\text{entropy}(\text{outlook} = \text{overcast}) = \text{entropy}([4,0]) = -\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4} = 0$$

Therefore,

$$\text{entropy}(\text{outlook}) = \text{entropy}([2,3],[3,2],[4,0]) = \frac{5}{14} * 0.971 + \frac{5}{14} * 0.971 + \frac{4}{14} * 0$$

Then the information gain:

$$\text{gain}(\text{outlook}) = \text{entropy}(\text{Dataset}) - \text{entropy}([2,3],[3,2],[4,0]) \approx 0.940 - 0.694 \approx 0.246$$

(1.2) <u>Entropy and information gain for Humidity</u>:

Each possible split point for the Humidity attribute is considered. A split point is the midpoint between a pair of different values in the sorted values of humidity. The split points at the root node are: (72+80)/2=76, (80+90)/2=85, and (90+96)/2=93.

For split point (72 + 80) / 2 = 76, when humidity ≤ 76, the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes
1	rainy	72	TRUE	no
2	overcast	72	TRUE	yes

$$\text{entropy}(\text{humidity} \leq 76) = \text{entropy}([3,1]) = -\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4} \approx 0.811$$

When humidity is greater than 76, the instances are:

No.	Outlook	Humidity	Windy	Play
5	sunny	80	FALSE	no
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
6	overcast	80	FALSE	yes
9	overcast	80	TRUE	yes
10	sunny	90	TRUE	no
11	overcast	90	TRUE	yes
13	sunny	96	FALSE	no
12	rainy	96	FALSE	yes
14	rainy	96	TRUE	no

$$\text{entropy}(\text{humidity} > 76) = \text{entropy}([6,4]) = -\frac{6}{10}log_2\frac{6}{10} - \frac{4}{10}log_2\frac{4}{10} \approx 0.971$$

Therefore,

$$\text{entropy}(\text{humidity "}\leq\text{or}>\text{" } 76) = \text{entropy}([3,1],[6,4]) = \frac{4}{14} * 0.811 + \frac{10}{14} * 0.971$$

Then the information gain:

$$\text{gain}(\text{humidity"}\leq\text{or}>\text{"}76) = \text{entropy}(\text{Dataset}) - \text{entropy}([3,1],[6,4])$$
$$\approx 0.940 - 0.925 \approx 0.015$$

For split point = (80 + 90) / 2 = 85, when humidity is less than or equal to 85, the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes
1	rainy	72	TRUE	no
2	overcast	72	TRUE	yes
5	sunny	80	FALSE	no
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
6	overcast	80	FALSE	yes
9	overcast	80	TRUE	yes

$$\text{entropy}(\text{humidity} \leq 85) = \text{entropy}([7,2]) = -\frac{7}{9}log_2\frac{7}{9} - \frac{2}{9}log_2\frac{2}{9} \approx 0.764$$

When humidity is greater than 85, the instances are:

No.	Outlook	Humidity	Windy	Play
10	sunny	90	TRUE	no
11	overcast	90	TRUE	yes
13	sunny	96	FALSE	no
12	rainy	96	FALSE	yes
14	rainy	96	TRUE	no

$$\text{entropy}(\text{humidity} > 85) = \text{entropy}([2,3]) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} \approx 0.971$$

Therefore,

$$\text{entropy}(\text{humidity"}\leq\text{or}>\text{"}\ 85) = \ \text{entropy}([7,2],[2,3]) = \frac{9}{14} * 0.764 + \frac{5}{14} * 0.971$$

Then the information gain:

$$\text{gain}(\text{humidity "}\leq\text{or}>\text{" }85) = \ \text{entropy}(\text{Dataset}) - \text{entropy}([7,2],[2,3])$$
$$\approx 0.940 - 0.838 \approx 0.102$$

For split point = (90 + 96) / 2 = 93, when humidity is less than or equal to 93, the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes
1	rainy	72	TRUE	no
2	overcast	72	TRUE	yes
5	sunny	80	FALSE	no
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
6	overcast	80	FALSE	yes
9	overcast	80	TRUE	yes
10	sunny	90	TRUE	no
11	overcast	90	TRUE	yes

$$\text{entropy}(\text{humidity} \leq 93) = \text{entropy}([8,3]) = -\frac{8}{11}log_2\frac{8}{11} - \frac{3}{11}log_2\frac{3}{11} \approx 0.845$$

When humidity is greater than 93, the instances are:

No.	Outlook	Humidity	Windy	Play
13	sunny	96	FALSE	no
12	rainy	96	FALSE	yes
14	rainy	96	TRUE	no

$$\text{entropy}(\text{humidity} > 93) = \text{entropy}([1,2]) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} \approx 0.918$$

Therefore,

$$\text{entropy}(\text{humidity"}\leq\text{or}>\text{"}93) = \text{entropy}([8,3],[1,2]) = \frac{11}{14} * 0.845 + \frac{3}{14} * 0.918$$

Then the information gain:

$$\text{gain}(\text{humidity"}\leq\text{or}>\text{"}93) = \text{entropy}(\text{Dataset}) - \text{entropy}([8,3],[1,2])$$
$$\approx 0.940 - 0.861 \approx 0.079$$

For humidity, pick the split point with the highest information gain "85" which has information gain is 0.102.

(1.3) <u>Entropy and information gain for Windy</u>:

When windy = "true", the instances are:

No.	Outlook	Humidity	Windy	Play
4	sunny	72	TRUE	yes
1	rainy	72	TRUE	no
2	overcast	72	TRUE	yes
9	overcast	80	TRUE	yes
10	sunny	90	TRUE	no
11	overcast	90	TRUE	yes
14	rainy	96	TRUE	no

$$\text{entropy}(\text{windy} = \text{TRUE}) = \text{entropy}([4,3]) = -\frac{4}{7}log_2\frac{4}{7} - \frac{3}{7}log_2\frac{3}{7} \approx 0.985$$

When windy = "false", the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
5	sunny	80	FALSE	no
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
6	overcast	80	FALSE	yes
13	sunny	96	FALSE	no
12	rainy	96	FALSE	yes

$$\text{entropy}(\text{windy} = \text{FALSE}) = \text{entropy}([5,2]) = -\frac{5}{7}log_2\frac{5}{7} - \frac{2}{7}log_2\frac{2}{7} \approx 0.863$$

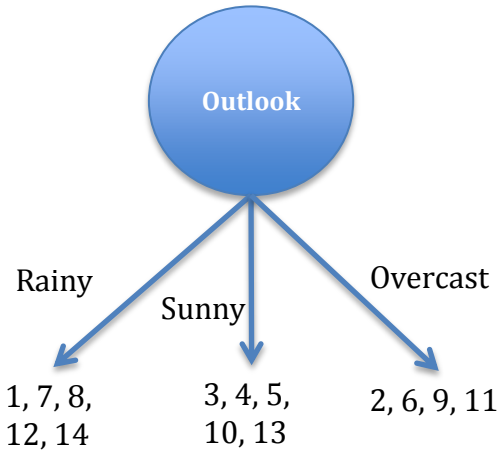
Therefore,

$$\text{entropy}(\text{windy}) = \text{entropy}([4,3],[5,2]) = \frac{7}{14} * 0.985 + \frac{7}{14} * 0.863$$

Then the information gain:

$$\text{gain(windy)} = \text{entropy(Dataset)} - \text{entropy}([4,3],[5,2]) \approx 0.940 - 0.924 \approx 0.016$$

The picked attribute is "Outlook" with the highest information gain and the corresponding tree is as follows:



(2) <u>Second Level Nodes:</u>

(2.1) <u>For the left-most node in the 2nd level "Rainy"</u>: the instances are:

No.	Outlook	Humidity	Windy	Play
1	rainy	72	TRUE	no
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
12	rainy	96	FALSE	yes
14	rainy	96	TRUE	no

The general entropy of the instances in this node is:

$$\text{entropy(Dataset in this node)} = \text{entropy}([3,2]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} \approx 0.971$$

(2.1.1) <u>Entropy and information gain for Humidity</u>:

For the split point = (72 + 80) / 2 = 76, when humidity is less than or equal to 76, the instances are:

No.	Outlook	Humidity	Windy	Play
1	rainy	72	TRUE	no

$$\text{entropy(humidity} \leq 76) = \text{entropy}([0,1]) = 0 - \frac{1}{1}log_2\frac{1}{1} \approx 0$$

When humidity is greater than 76, the instances are:

No.	Outlook	Humidity	Windy	Play
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
12	rainy	96	FALSE	yes
14	rainy	96	TRUE	no

$$\text{entropy}(\text{humidity} > 76) = \text{entropy}([3,1]) = -\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4} \approx 0.811$$

Therefore, $\text{entropy}(\text{humidity}<\text{or}\geq 76) = \text{entropy}([0,1],[3,1]) = \frac{1}{5} * 0 + \frac{4}{5} * 0.811$

Then the information gain:

$$\text{gain}(\text{humidity} <\text{or}\geq 76) = \text{entropy}(\text{Dataset}) - \text{entropy}([0,1],[3,1])$$
$$\approx 0.971 - 0.649 \approx 0.322$$

For split point = (80 + 96) / 2 = 88, when humidity is less than or equal to 88, the instances are:

No.	Outlook	Humidity	Windy	Play
1	rainy	72	TRUE	no
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes

$$\text{entropy}(\text{humidity} \leq 88) = \text{entropy}([2,1]) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} \approx 0.918$$

When humidity is greater than 88, the instances are:

No.	Outlook	Humidity	Windy	Play
12	rainy	96	FALSE	yes
14	rainy	96	TRUE	no

$$\text{entropy}(\text{humidity} > 88) = \text{entropy}([1,1]) = -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} \approx 1$$

Therefore, $\text{entropy}(\text{humidity} < \text{or} \geq 88) = \text{entropy}([2,1],[1,1]) = \frac{3}{5} * 0.918 + \frac{2}{5} * 1$

Then the information gain:

$$\text{gain}(\text{humidity} < \text{or} \geq 88) = \text{entropy}(\text{Dataset}) - \text{entropy}([2,1],[1,1])$$
$$\approx 0.971 - 0.951 \approx 0.020$$

For humidity, pick the split point with the highest information gain “76” for which the information gain is 0.322.

(2.1.2) Entropy and information gain for Windy:

When windy = "TRUE", the instances are:

No.	Outlook	Humidity	Windy	Play
1	rainy	72	TRUE	no
14	rainy	96	TRUE	no

$$\text{entropy}(\text{windy} = \text{TRUE}) = \text{entropy}([0,2]) = 0 - \frac{2}{2} log_2 \frac{2}{2} \approx 0$$

When windy = "FALSE", the instances are:

No.	Outlook	Humidity	Windy	Play
7	rainy	80	FALSE	yes
8	rainy	80	FALSE	yes
12	rainy	96	FALSE	yes

$$\text{entropy}(\text{windy} = \text{FALSE}) = \text{entropy}([3,0]) = -\frac{3}{3} log_2 \frac{3}{3} - 0 \approx 0$$

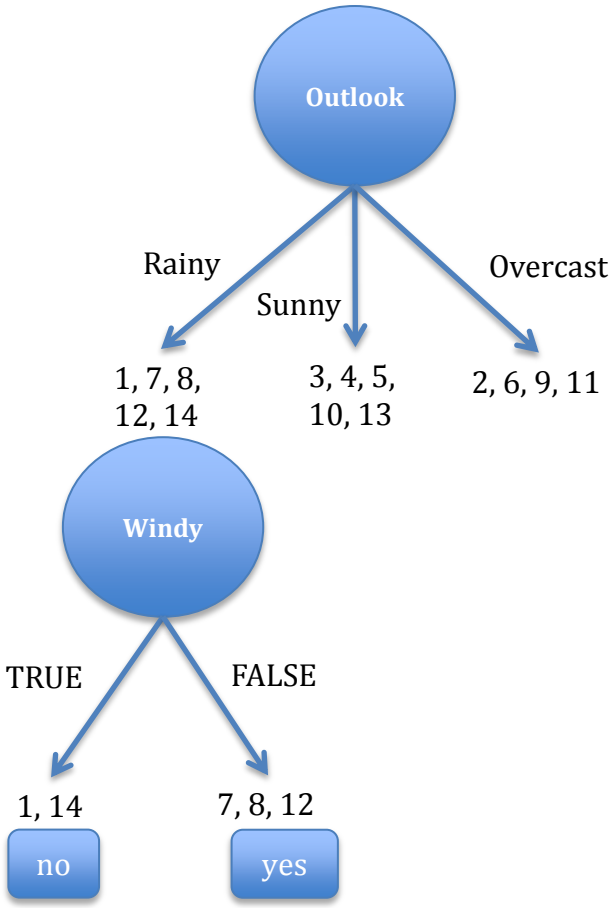
Therefore,

$$\text{entropy}(\text{windy}) = \text{entropy}([0,2],[3,0]) = \frac{3}{5} * 0 + \frac{2}{5} * 0$$

Then the information gain:

$$\text{gain}(\text{windy}) = \text{entropy}(\text{Dataset at this node}) - \text{entropy}([0,2],[3,0]) \approx 0.971 - 0 \approx 0.971$$

For the left-most node of level 2, the selected attribute is "Windy" since it has the highest information gain for this node. The corresponding tree is as follows:



Since the instances 1 and 14 have the same class value "no" and the instances 7, 8, 12 have the same class value "yes", this subtree has been completed.

(2.2) <u>For the middle node in the 2nd level "Sunny"</u>: the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes
5	sunny	80	FALSE	no
10	sunny	90	TRUE	no
13	sunny	96	FALSE	no

The general entropy of the dataset in this node is:

$$\text{entropy(Dataset)} = \text{entropy([2,3])} = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} \approx 0.971$$

(2.2.1) <u>Entropy and information gain for Humidity</u>:

For split point = (72 + 80) / 2 = 76, when humidity is less than or equal to 76, the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes

$$\text{entropy}(\text{humidity} \leq 76) = \text{entropy}([2,0]) = -\frac{2}{2}log_2\frac{2}{2} - 0 \approx 0$$

When humidity is greater than 76, the instances are:

No.	Outlook	Humidity	Windy	Play
5	sunny	80	FALSE	no
10	sunny	90	TRUE	no
13	sunny	96	FALSE	no

$$\text{entropy}(\text{humidity} > 76) = \text{entropy}([0,3]) = 0 - \frac{3}{3}log_2\frac{3}{3} \approx 0$$

Therefore,

$$\text{entropy}(\text{humidity} \leq \text{or} > 76) = \text{entropy}([2,0],[0,3]) = \frac{2}{5} * 0 + \frac{3}{5} * 0$$

Then the information gain:

$$\text{gain}(\text{humidity} \leq \text{or} > 76) = \text{entropy}(\text{Dataset at this node}) - \text{entropy}([2,0],[0,3]) \approx 0.971 - 0 \approx 0.971$$

Note: *Since this split point provides the lowest possible entropy (or equivalently, the largest information gain), we can stop our calculations here and use this split point as the chosen condition for the tree node under consideration. However, we continue to show the entropy and information gain calculations for the other candidates below, just for illustration purposes.*

For split point = (80 + 90) / 2 = 85, when humidity is less than or equal to 85, the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes
5	sunny	80	FALSE	no

$$\text{entropy}(\text{humidity} < 85) = \text{entropy}([2,1]) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} \approx 0.918$$

When humidity is greater than 85, the instances are:

No.	Outlook	Humidity	Windy	Play
10	sunny	90	TRUE	no
13	sunny	96	FALSE	no

$$\text{entropy}(\text{humidity} > 85) = \text{entropy}([0,2]) = 0 - \frac{2}{2} log_2 \frac{2}{2} \approx 0$$

Therefore,

$$\text{entropy}(\text{humidity} \leq \text{or} > 85) = \text{entropy}([2,1],[0,2]) = \frac{3}{5} * 0.918 + \frac{2}{5} * 0$$

Then the information gain:

$$\text{gain}(\text{humidity} \leq \text{or} > 85) = \text{entropy}(\text{Dataseta at this node}) - \text{entropy}([2,0],[0,3]) \approx 0.971 - 0.551 \approx 0.420$$

For split point = (90 + 96) / 2 = 93, when humidity is less than or equal to 93, the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
4	sunny	72	TRUE	yes
5	sunny	80	FALSE	no
10	sunny	90	TRUE	no

$$\text{entropy}(\text{humidity} \leq 93) = \text{entropy}([2,2]) = -\frac{2}{4} log_2 \frac{2}{4} - \frac{2}{4} log_2 \frac{2}{4} \approx 1$$

When humidity is greater than 93, the instances are:

No.	Outlook	Humidity	Windy	Play
13	sunny	96	FALSE	no

$$\text{entropy}(\text{humidity} > 93) = \text{entropy}([0,1]) = 0 - \frac{1}{1} log_2 \frac{1}{1} \approx 0$$

Therefore,

$$\text{entropy}(\text{humidity} \leq \text{or} > 93) = \text{entropy}([2,2],[0,1]) = \frac{4}{5} * 1 + \frac{1}{5} * 0$$

Then the information gain:

$$\text{gain}(\text{humidity} \leq \text{or} > 93) = \text{entropy}(\text{Dataset at this node}) - \text{entropy}([2,2],[0,1]) \approx 0.971 - 0.8 \approx 0.171$$

For humidity, pick the split point with the highest information gain “76” which has information gain is 0.971.

(2.2.2) <u>Entropy and information gain for Windy</u>:
When windy = "TRUE", the instances are:

No.	Outlook	Humidity	Windy	Play
4	sunny	72	TRUE	yes
10	sunny	90	TRUE	no

$$\text{entropy}(\text{windy} = \text{TRUE}) = \text{entropy}([1,1]) = -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} \approx 1$$

When windy = "FALSE", the instances are:

No.	Outlook	Humidity	Windy	Play
3	sunny	72	FALSE	yes
5	sunny	80	FALSE	no
13	sunny	96	FALSE	no

$$\text{entropy}(\text{windy} = \text{FALSE}) = \text{entropy}([1,2]) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} \approx 0.918$$

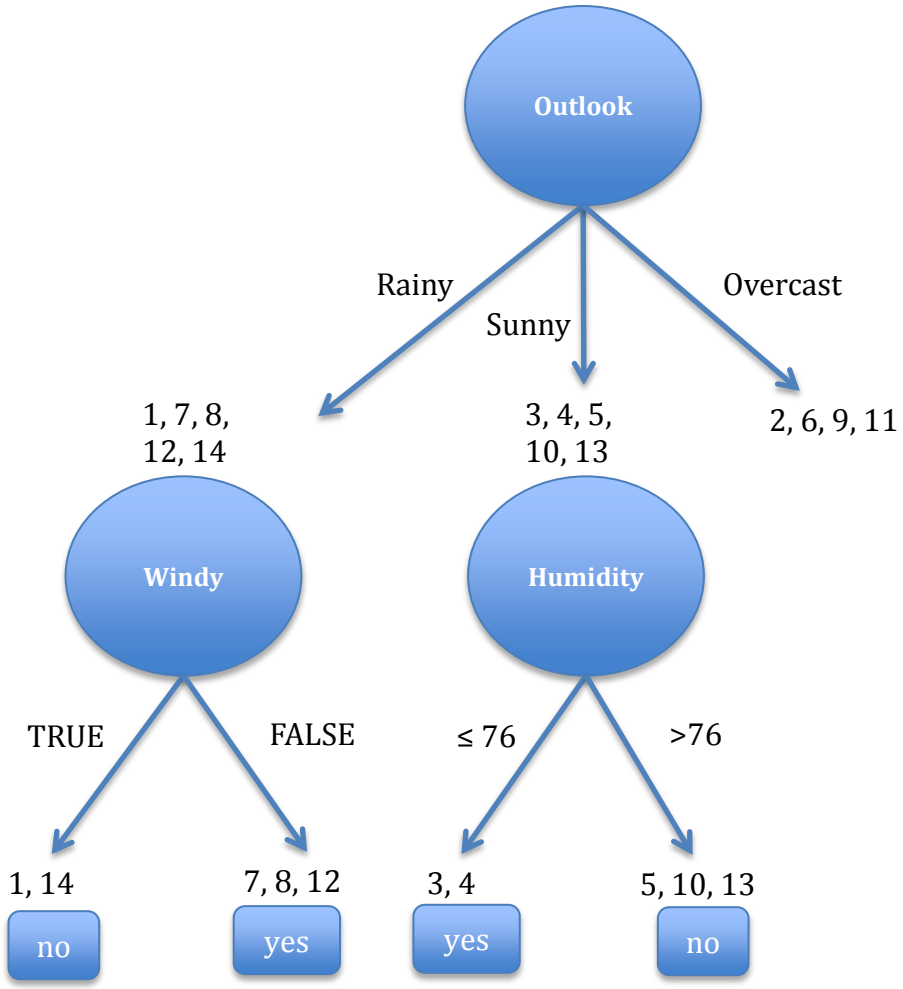
Therefore,

$$\text{entropy}(\text{windy}) = \text{entropy}([1,1],[1,2]) = \frac{2}{5} * 1 + \frac{3}{5} * 0.918$$

Then the information gain:

$$\text{gain}(\text{windy}) = \text{entropy}(\text{Dataset at this node}) - \text{entropy}([1,1],[1,2])$$
$$\approx 0.971 - 0.951 \approx 0.020$$

The picked condition is the split point "76" for humidity as it has the highest information gain among all candidates for this node. The corresponding tree is as follows:

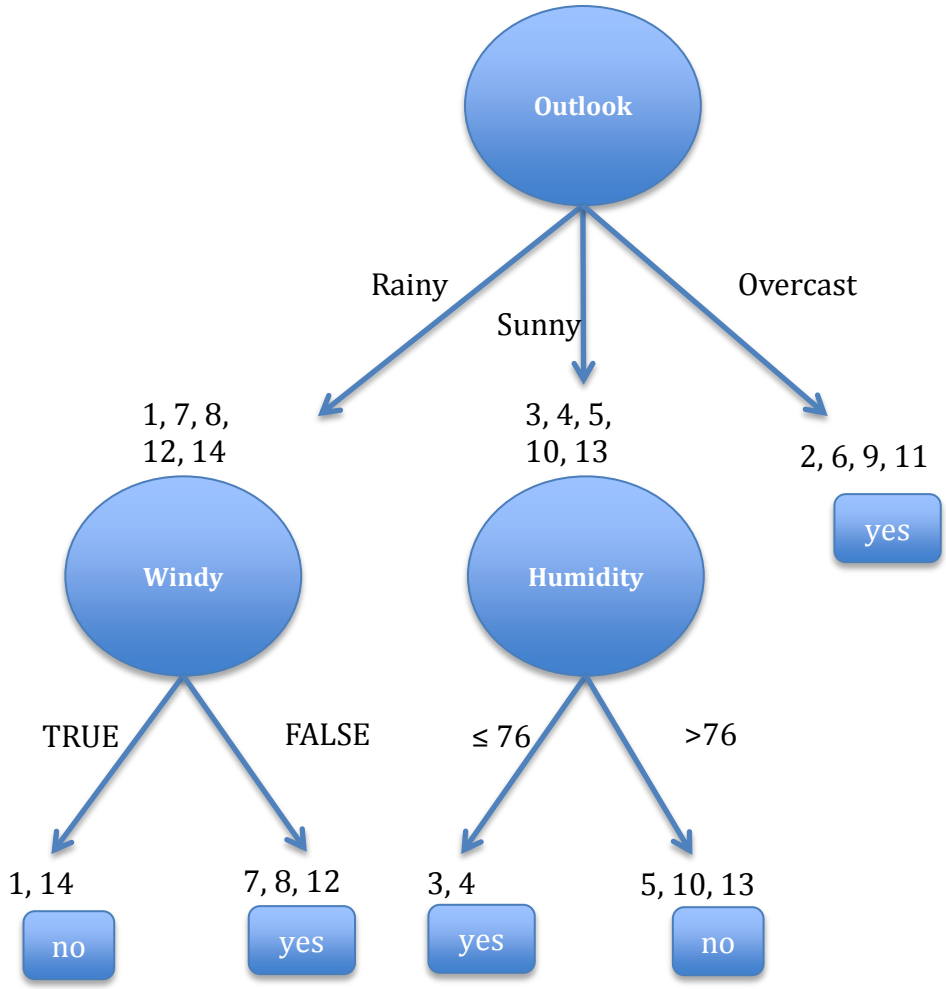


Since the instances 3 and 4 have the same class value "yes" and the instances 5, 10, 13 have the same class value "no", this subtree has been completed.

(2.1) <u>For the right-most node in the 2nd level "Overcast":</u> the instances are:

No.	Outlook	Humidity	Windy	Play
2	overcast	72	TRUE	yes
6	overcast	80	FALSE	yes
9	overcast	80	TRUE	yes
11	overcast	90	TRUE	yes

Since the target attribute has the same value in all the instances at this node, there is no need to further split this node. Therefore the entire decision tree is as follows:



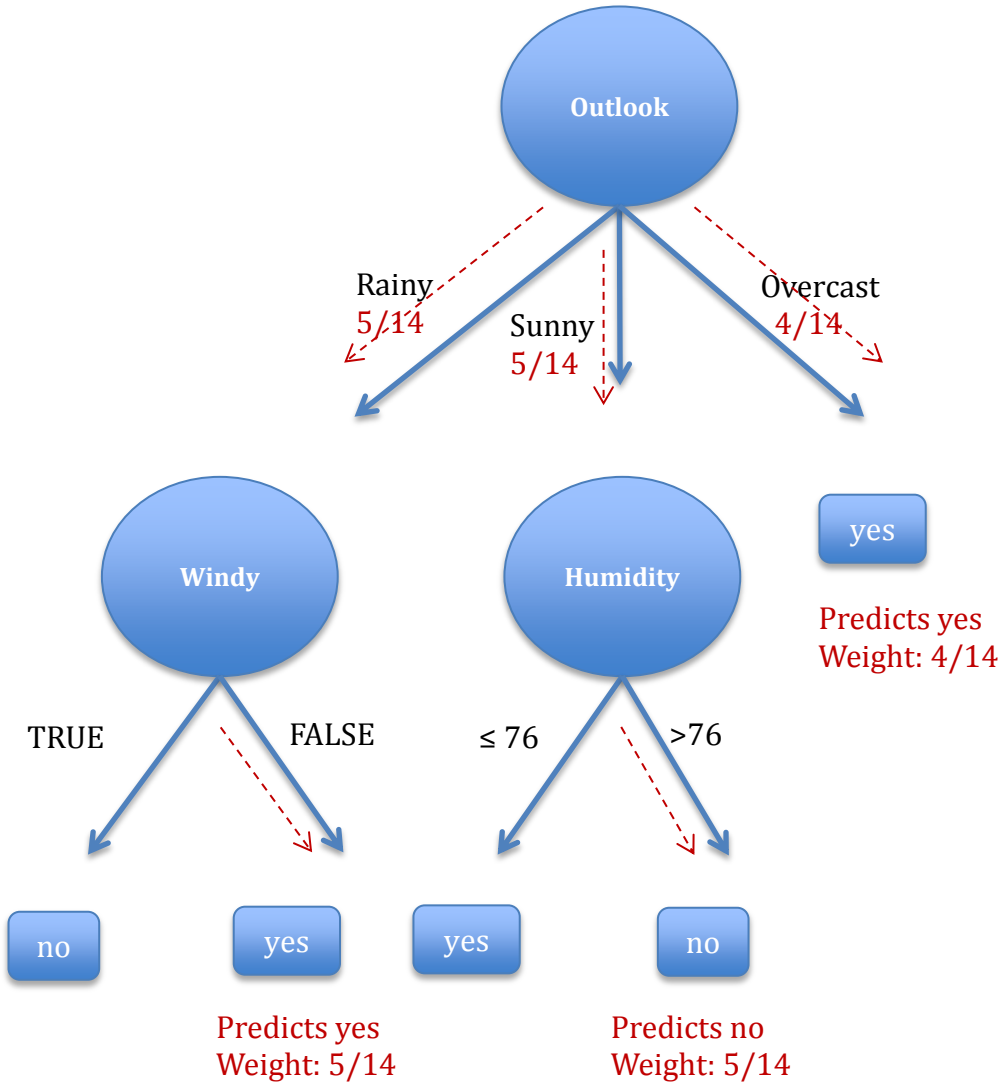
2. (5 points) Propose approaches to using your decision tree above to classify instances that contain missing values. Use the following instance to illustrate your ideas.

```
outlook = ?, humidy = 80, and windy = FALSE.
```

Solution:
Since this instance has a missing value for the attribute outlook, we could split this instance along the three children of the root node according to the probability of each possible value of outlook on the training dataset.

Outlook	Probability
sunny	5/14
Rainy	5/14
Overcast	4/14

Then we follow each branch on the tree that applies to this test instance. Those branches are shown in red (dotted lines) in the constructed tree:



Finally, the reached leaves vote (i.e., provide their predictions), and those votes are weighed by the weights associated with the corresponding branches:

Target value	**Weighted vote of the leaf nodes reached**
yes	5/14 + 4/14
no	5/14

Therefore, we could select "yes" as the predicted value for this test instance, as this target value received the largest weighted vote.

3. Study how J4.8 performs post-pruning by reading in detail:

Some useful comments:
J4.8 performs post-pruning by using sub-tree replacement and sub-tree raising.

Sub-tree replacement: for each subtree in the decision tree (in a bottom up fashion), replace the subtree with a leaf node. Assign to that leaf the majority class value of the data instances in the subtree. Then use the training dataset to examine the accuracy of the resulting tree. If the accuracy when removing the subtree is at least the same as the accuracy without replacement, the subtree should be replaced by the leaf node. This process could continue until the accuracy after post-pruning decreases.

Sub-tree raising: for each subtree in the decision tree (in a bottom up fashion), raise it to its parent node and incorporate in this subtree the data instances of its siblings (who will become "orphans"). Then use the training dataset to examine the accuracy of the resulting tree. If the accuracy doesn't decrease, the raising operation is adopted.