Solution: SAMPLE SOLUTIONS

Contents

1 Learning Goals

By the end of the exercise, you should be able to

- Construct a decision tree given an order of testing the features.
- Determine the prediction accuracy of a decision tree on a test set.
- Compute the entropy of a probability distribution.
- Compute the expected information gain for testing a feature.
- Trace the execution of and implement the ID3 algorithm. Construct a decision tree by selecting a feature for each node using the expected information gain metric.

2 Jeeves the valet - the data set

Jeeves the valet – the training set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Jeeves the valet – the test set

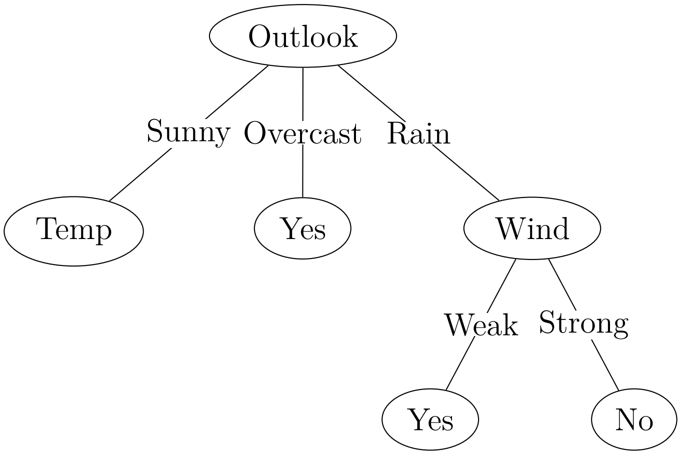
Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Mild	High	Strong	No
2	Rain	Hot	Normal	Strong	No
3	Rain	Cool	High	Strong	No
4	Overcast	Hot	High	Strong	Yes
5	Overcast	Cool	Normal	Weak	Yes
6	Rain	Hot	High	Weak	Yes
7	Overcast	Mild	Normal	Weak	Yes
8	Overcast	Cool	High	Weak	Yes
9	Rain	Cool	High	Weak	Yes
10	Rain	Mild	Normal	Strong	No
11	Overcast	Mild	High	Weak	Yes
12	Sunny	Mild	Normal	Weak	Yes
13	Sunny	Cool	High	Strong	No
14	Sunny	Cool	High	Weak	No

3 Practice Questions

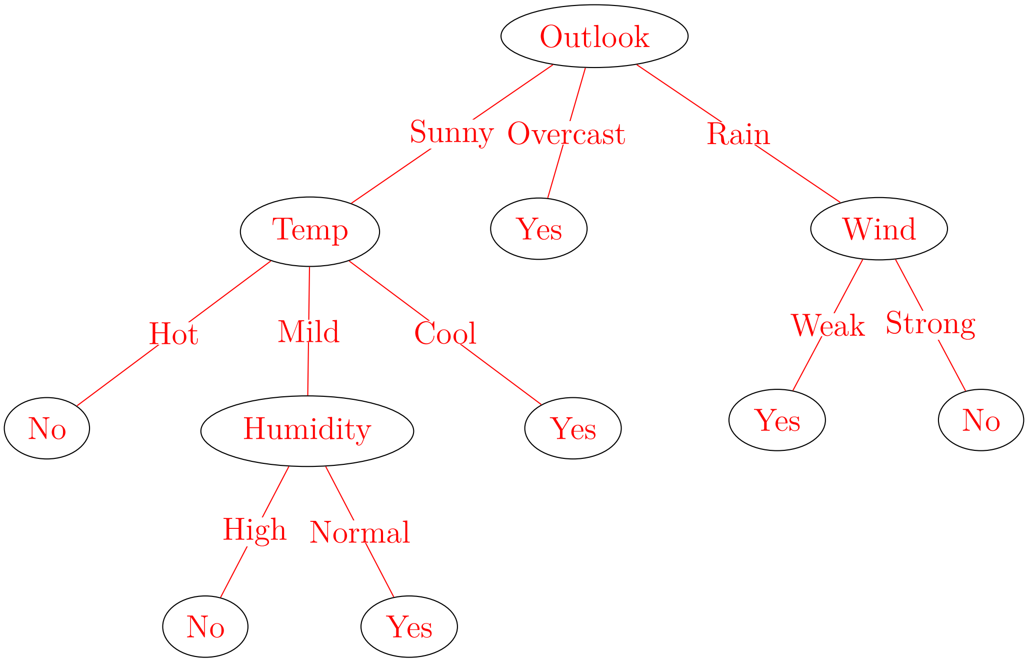
3.1 Construct a decision tree given an order of testing the features

Question 1:

Considering the following partial decision tree generated using the training set. Assume that after we test *Outlook* = *Sunny* we are going to test *Temp* next. Also, assume that we always test Humidity before testing Wind. Generate the full decision tree using the training set.



Solution:



3.2 The ID-3 Algorithm

Algorithm 1 ID3 Algorithm (Features, Examples)

```
If all examples are positive, return a leaf node with decision yes.
If all examples are negative, return a leaf node with decision no.
If no features left, return a leaf node with the most common decision of the examples.
If no examples left, return a leaf node with the most common decision of the examples in the parent.
else
    choose the most important feature $f$
    for each value $v$ of feature $f$ do
        add arc with label $v$
        add subtree $ID3(F - f, s \in S | f(s) = v)$
    end for
```

3.3 Trace the execution of ID-3 Algorithm

Question 2:

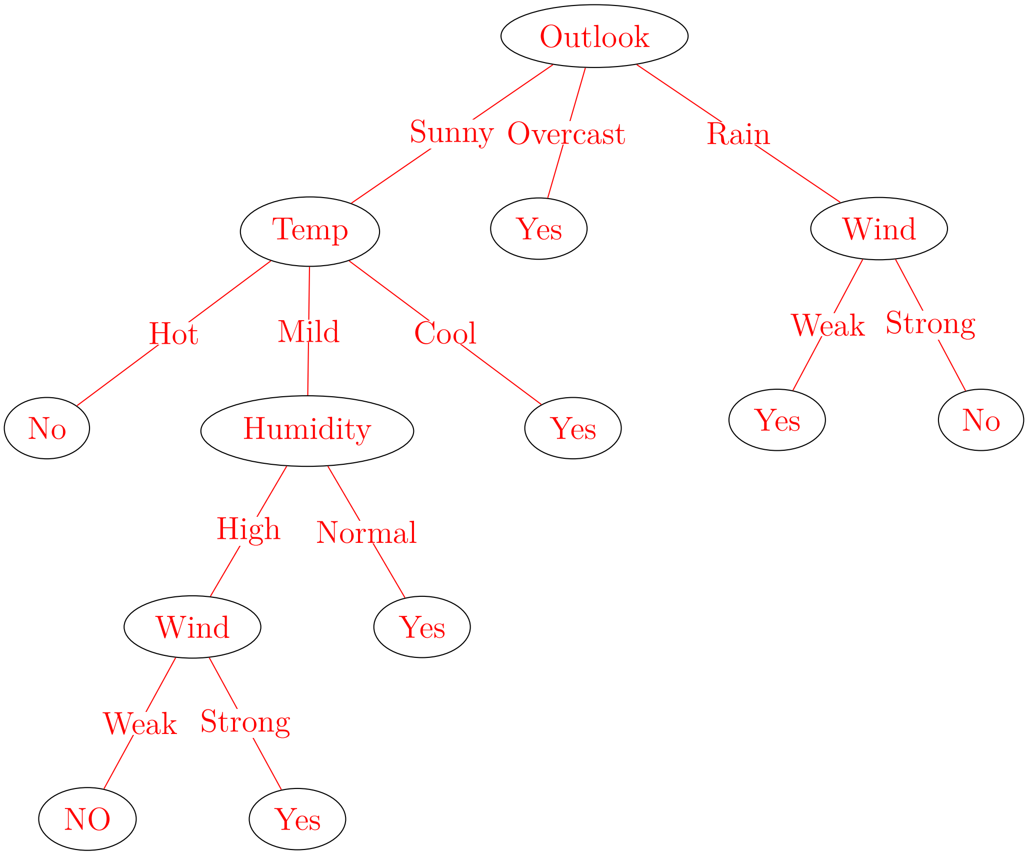
Suppose we add 3 extra data points into the training dataset. The training set becomes the following:

Jeeves the valet – the training set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No
15	**Sunny**	**Mild**	**High**	**Weak**	**No**
16	**Sunny**	**Mild**	**High**	**Weak**	**Yes**
17	**Sunny**	**Mild**	**High**	**Strong**	**Yes**

Consider the partial decision tree generated from the above training set, generate the full decision tree using ID-3 Algorithm.

Solution: Note that there are multiple different decisions for the same set of feature values of For the set of feature values $\{Outlook = Sunny, Temp = Mild, Humidity = High, Wind = Weak\}$, there are multiple different decisions (2 No's and 1 Yes). There is no feature left to classify these examples. Therefore, we choose the majority decision of the examples, which is NO.



Question 3:

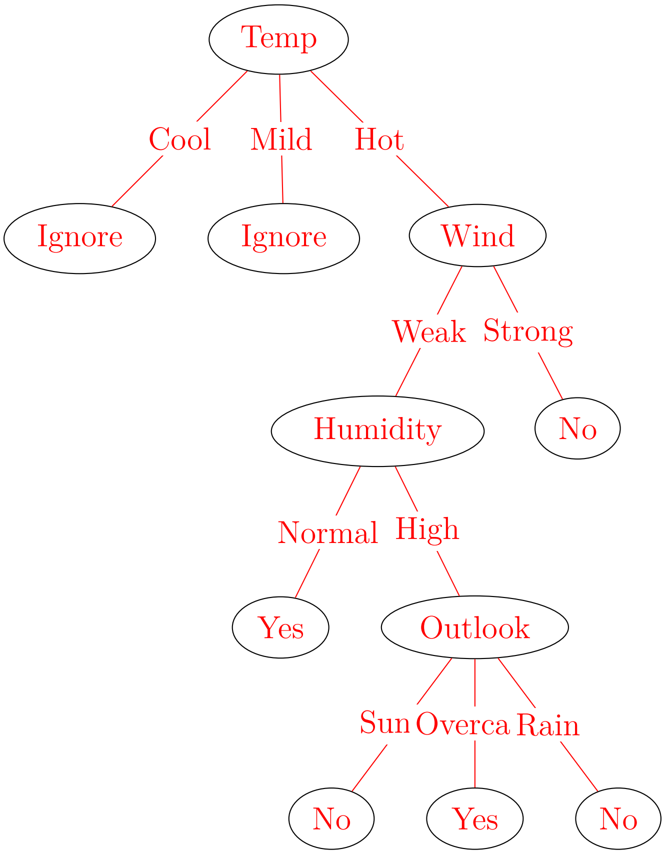
Suppose we add 1 extra data point to the original training set. The training set becomes the following:

Jeeves the valet – the training set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No
15	**Sunny**	**Hot**	**High**	**Weak**	**No**

Consider the following partial decision tree generated from the above training set. Complete the branch of $Temp = Hot$.

Solution: For $\{Temp = Hot, Wind = Weak, Humidity = High, Outlook = Rain\}$, there are no examples. Check the parent node $Humidity = High$, The parent node "Humidity" has 2 No's and 1 Yes. Therefore, we choose No as the decision.



Question 4:

What is the entropy of the distribution $(0.5, 0.5)$?

What is the entropy of the distribution $(0.01, 0.99)$?

Which distribution has more uncertainty?

Solution: $I(0.5, 0.5) = -1/2log_2(1/2) - 1/2log_2(1/2) = 1$.

$I(0.99, 0.01) = -0.99log_2(0.99) - 0.01log_2(0.01) = 0.08$.

The first distribution contains more uncertainty.

Question 5:

Consider the original training set.

Jeeves the valet – the training set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

To construct a decision tree, we need to determine which feature to test first by Information Gain metric.

The following are the Information Gain of testing Temp and Wind at the first step. Calculate the Information Gain for the remaining features and determine which feature we should test first.

- $Gain(Temp) = 0.029$

- $Gain(Wind) = 0.048$

Solution: If we split on Outlook, we would get

- Outlook = Sunny. +: 9, 11. -: 1, 2, 8.
- Outlook = Overcast. +: 3, 7, 12, 13. -: none.
- Outlook = Rain. +: 4, 5, 10. -: 6, 14.

$$
\begin{aligned}
& Gain(Outlook) && (1)\\
& = 0.940 - \left(\frac{5}{14} I\left(\frac{2}{5}, \frac{3}{5}\right) + \frac{4}{14} I\left(\frac{4}{4}, \frac{0}{4}\right) + \frac{5}{14} I\left(\frac{3}{5}, \frac{2}{5}\right)\right) && (2)\\
& = 0.940 - \left(\frac{5}{14} 0.971 + \frac{4}{14} 0 + \frac{5}{14} 0.971\right) && (3)\\
& = 0.940 - 0.694 && (4)\\
& = 0.247. && (5)
\end{aligned}
$$

If we split on Humidity, we would get

- Humidity = Normal. +: 5, 7, 9, 10, 11, 13. -: 6.
- Humidity = High. +: 3, 4, 12. -: 1, 2, 8, 14.

$$
\begin{aligned}
& Gain(Humidity) && (6)\\
& = 0.940 - \left(\frac{7}{14} I\left(\frac{3}{7}, \frac{4}{7}\right) + \frac{7}{14} I\left(\frac{6}{7}, \frac{1}{7}\right)\right) && (7)\\
& = 0.940 - \left(\frac{7}{14} 0.985 + \frac{7}{14} 0.592\right) && (8)\\
& = 0.940 - 0.789 && (9)\\
& = 0.151 && (10)
\end{aligned}
$$

The expected information gain is the largest if we test Outlook. Thus, we will choose Outlook as the root of the decision tree.

3.4 Handling Continuous Values

Question 6:

The following dataset is sorted according to the values of Temp.

Day	Outlook	Temp	Humidity	Wind	Tennis?
7	Overcast	17.7	Normal	Strong	Yes
6	Rain	18.3	Normal	Strong	No
5	Rain	20.0	Normal	Weak	Yes
9	Sunny	20.6	Normal	Weak	Yes
4	Rain	21.1	High	Weak	Yes
14	Rain	21.7	High	Strong	No
8	Sunny	22.2	High	Weak	No
12	Overcast	22.2	High	Strong	Yes
10	Rain	23.9	Normal	Weak	Yes
11	Sunny	23.9	Normal	Strong	Yes
2	Sunny	26.6	High	Strong	No
13	Overcast	27.2	Normal	Weak	Yes
3	Overcast	28.3	High	Weak	Yes
1	Sunny	29.4	High	Weak	No

Determine whether we would consider the following values as possible split points

- $Temp = 18$
- $Temp = 20.3$
- $Temp = 23.05$
- $Temp = 26.6$

Solution:

- The classification for 17.7 is Yes, whereas the classification for 18.3 is No. Thus, we will consider (17.7 + 18.3) / 2 = 18 as a possible split point.
- The classification for 20.0 and the classification for 20.6 are both Yes. Therefore, we will NOT consider the midway value(20.3) between these two as a possible split point.
- The classification for 2 data points with 22.2 are No and Yes, whereas the classification for the two data points with 23.9 are both Yes. We will consider (22.2 + 23.9) / 2 = 23.05 as a possible split point (because No for 22.2 is different from Yes for 23.9.)

- Since we only consider the midpoint of 2 values as a possible split point, 26.6 is not a split point.