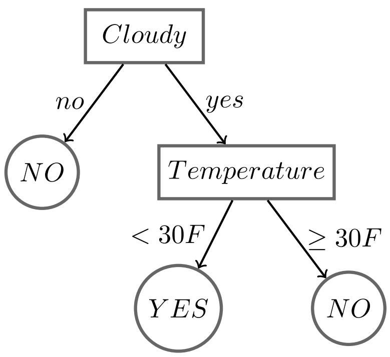
1 Definitions Oh My!

(a) **Decision Trees**: Popular representation and technique for classification and prediction, in which each node represents an attribute that is tested (or a decision), each branch represents an outcome, and each leaf node represents a resulting label

Example: Will it rainy today?



(b) **Entropy**: A measurement of the uncertainty in a random variable, hence with high entropy we are less confident about the outcome

$$H(Y) = -\sum_{y} P(Y = y) \log_2 P(Y = y)$$

(c) **Conditional-Entropy**: The expected value of the entropy of Y given X, over all values of X. This lets us quantify the entropy of Y given that we know X.

$$H(Y \mid X) = \sum_{x} P(X = x) H(Y \mid X = x)$$

(d) **Mutual Information**: Measurement of how much uncertainty about variable Y decreases if we know and have observed X.

$$I(Y; X) = H(Y) - H(Y \mid X = x)$$

2 Decision Trees

2.1 Mutual Information and Entropy Practice

1. Calculate the entropy of tossing a fair coin.

$$\begin{aligned} H(Y) &= -P(Y = \text{heads})\log_2 P(\text{Y} = \text{heads}) - P(\text{Y} = \text{tails})\log_2 P(\text{Y} = \text{tails}) \\ &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} \\ &= 1 \end{aligned}$$

2. Calculate the entropy of tossing a coin that lands only on tails. *note* : $0 \cdot \log_2 0 = 0$.

$$\begin{aligned} H(Y) &= -P(\text{Y} = \text{heads})\log_2 P(\text{Y} = \text{heads}) - P(\text{Y} = \text{tails})\log_2 P(Y = \text{tails}) \\ &= -0\log_2 0 - 1\log_2 1 \\ &= 0 \end{aligned}$$

3. Calculate the entropy of a fair dice roll.

$$\begin{aligned} H(Y) &= -\sum_{i=1}^{6}\frac{1}{6}\log_2\frac{1}{6} \\ &= -\log_2\frac{1}{6} \\ &= \log_2 6 \end{aligned}$$

4. When is the mutual information I(Y;X) = 0?

I(Y; X) = 0 if and only if X and Y are independent.

$$\begin{aligned} I(Y;X) &= H(Y) - H(Y \mid X) \\ 0 &= H(Y) - H(Y \mid X) \\ H(Y) &= H(Y \mid X) \end{aligned}$$

Intuitively, if X and Y are independent, then knowing X tells us nothing about Y, and vice versa, so their mutual information is zero.

2.2 Decision Tree *Run*through

Fun fact: Pat loves to go on runs! However, sometimes things come up which deter him from running outside. The following attributes have affected his running in the past:

- How the sky looks the day of his run (sunny, overcast, rainyy)
- The temperature outside (hot, mild, cool)
- The humidity 30 minutes before he plans on running (high, normal)

As a member of the 07-280 running club, you have gathered the following data and decided that you want to make a decision tree to predict whether Pat will go on a run.

Outlook (X_1)	Temperature (X_2)	Humidity (X_3)	Go on run? (Y)
sunny	hot	high	no
overcast	hot	high	yes
rainy	mild	high	yes
rainy	cool	normal	yes
sunny	mild	high	no
sunny	mild	normal	yes
rainy	mild	normal	yes
overcast	hot	normal	yes

To start, let's first determine the attribute for the root node of the decision tree. To do this, we need to determine the mutual information for each attribute.

1. Calculate $H(Y)$, the entropy of Pat going on a run.

$$\begin{aligned} H(Y) &= -P(Y = \text{yes}) \log_2 P(Y = \text{yes}) - P(Y = \text{no}) \log_2 P(Y = \text{no}) \\ &= -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} \\ &= 0.8113 \end{aligned}$$

2. Calculate $I(Y; X_1)$, the mutual information when splitting on the outlook.

$$\begin{aligned} I(Y; X_1) &= H(Y) - H(Y \mid X_1) \\ &= H(Y) - \left(\sum_x P(X_1 = x) H(Y \mid X_1 = x) \right) \\ &= H(Y) + \sum_x P(X_1 = x) \sum_y P(Y = y \mid X_1 = x) \log_2 P(Y = y \mid X_1 = x) \\ &= 0.8113 + \left(\frac{3}{8} (\frac{1}{3} \log_2 \frac{1}{3} + \frac{2}{3} \log_2 \frac{2}{3}) + \frac{2}{8} (\frac{2}{2} \log_2 \frac{2}{2} + \frac{0}{2} \log_2 \frac{0}{2}) + \frac{3}{8} (\frac{3}{3} \log_2 \frac{3}{3} + \frac{0}{3} \log_2 \frac{0}{3}) \right) \\ &= 0.8113 + \left(\frac{3}{8} (-0.9183) + \frac{2}{8} (0) + \frac{3}{8} (0) \right) \\ &= 0.4669 \end{aligned}$$

3. Calculate $I(Y; X_2)$, the mutual information when splitting on temperature.

$$
\begin{aligned}
I(Y; X_2) &= H(Y) + \sum_x P(X_2 = x) \sum_y P(Y = y \mid X_2 = x) \log_2 P(Y = y \mid X_2 = x) \\
&= 0.8113 + \left(\frac{3}{8}\left(\frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3}\right) + \frac{4}{8}\left(\frac{3}{4} \log_2 \frac{3}{4} + \frac{1}{4} \log_2 \frac{1}{4}\right) + \frac{1}{8}\left(\frac{1}{1} \log_2 \frac{1}{1} + \frac{0}{1} \log_2 \frac{0}{1}\right) \right) \\
&= 0.8113 + \left(\frac{3}{8}(-0.9183) + \frac{4}{8}(-0.8113) + \frac{1}{8}(0) \right) \\
&= 0.0613
\end{aligned}
$$

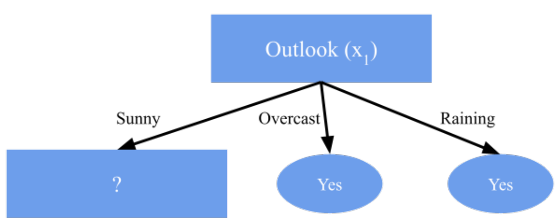
4. Calculate $I(Y; X_3)$, the mutual information when splitting on humidity.

$$
\begin{aligned}
I(Y; X_3) &= H(Y) + \sum_x P(X_3 = x) \sum_y P(Y = y \mid X_3 = x) \log_2 P(Y = y \mid X_3 = x) \\
&= 0.8113 + \left(\frac{4}{8}\left(\frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4}\right) + \frac{4}{8}\left(\frac{4}{4} \log_2 \frac{4}{4} + \frac{0}{4} \log_2 \frac{0}{4}\right) \right) \\
&= 0.8113 + \left(\frac{4}{8}(-1) + \frac{4}{8}(0) \right) \\
&= 0.3113
\end{aligned}
$$

5. Which attribute do we split on?

Outlook (X_1) because it has the highest mutual information.

6. What does the decision tree look like so far?



Now that we've split on the first attribute, the remaining data points which still cannot be classified are shown in the table below.

Outlook (X_1)	Temperature (X_2)	Humidity (X_3)	Go on run? (Y)
sunny	hot	high	no
sunny	mild	high	no
sunny	mild	normal	yes

From now on, we will be strictly considering the branch of the decision tree where the outlook is sunny.

7. Calculate $H(Y)$, the entropy of going on a run given that it is sunny.

$$\begin{aligned} H(Y) &= -\left(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right) \\ &= 0.9183 \end{aligned}$$

8. Calculate $I(Y;X_2)$.

$$\begin{aligned} I(Y;X_2) &= 0.9183+\left(\frac{1}{3}\left(\frac{0}{1}\log_2\frac{0}{1}+\frac{1}{1}\log_2\frac{1}{1}\right)+\frac{2}{3}\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)\right) \\ &= 0.2516 \end{aligned}$$

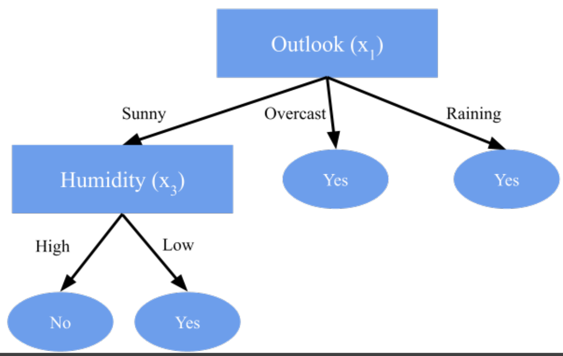
9. Calculate $I(Y;X_3)$.

$$\begin{aligned} I(Y;X_3) &= 0.9183+\left(\frac{2}{3}\left(\frac{0}{2}\log_2\frac{0}{2}+\frac{2}{2}\log_2\frac{2}{2}\right)+\frac{1}{3}\left(\frac{1}{1}\log_2\frac{1}{1}+\frac{0}{1}\log_2\frac{0}{1}\right)\right) \\ &= 0.9183 \end{aligned}$$

10. Which attribute do we split on now?

Humidity (X_3), which perfectly splits the data.

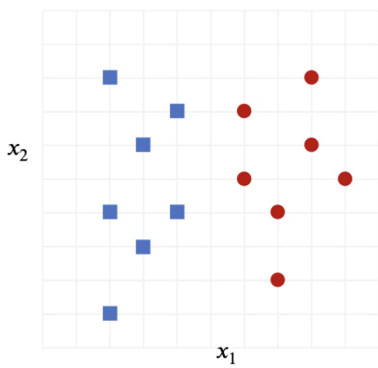
11. Draw the final decision tree.

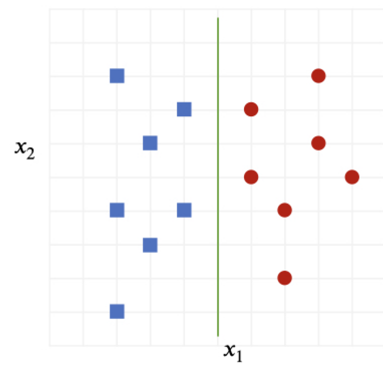


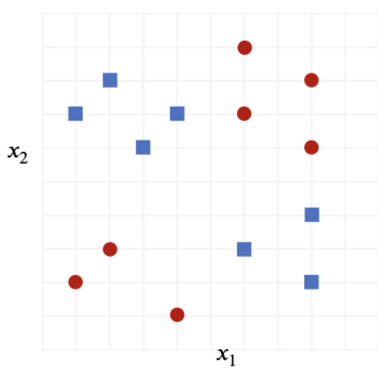
3 Decision Boundaries

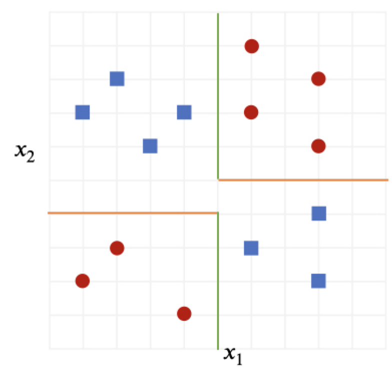
3.1 Decision Trees

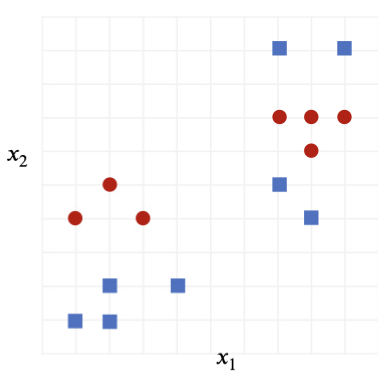
What could the decision boundaries look like for the following datatsets, assuming zero error?

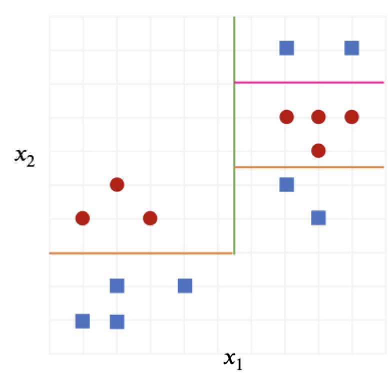












4 Decision Trees

Explore different values for max_depth by running through the DT Jupiter Notebook (see recitation materials).

4.1 Discussion Questions

1. What is the best value of max_depth for this dataset and why?

   Between 2 and 8, depending on the data split and training function. We are looking for the lowest validation error.

2. Why do you think the error is large with very small max_depth?

   When max_depth is small, we might not split enough. The leaf nodes may use a large dataset with still a mixture of different outputs to decide, resulting in high entropy and error. This leads to underfitting because it is insufficient to learn the underlying structure of the data.

3. How is the value of max_depth related to overfitting?

   Larger values of max_depth tend to overfit to the data because we might decide invidual or small groups of data points. This might fit to the noise of the training data, which results in low training error but higher test error due to its not learning the general trends of the data.

4. Why can we not display the decision boundaries like we saw in lecture?

   The Iris dataset has 4 features, so we cannot display the decision boundaries in a 2D space. In lecture, we picked 2 of the 4 features to display.

5. For which values of max_depth would the decision boundaries be the most complex?

   Decision boundaries become less smooth as max_depth increases since we split on more features. Therefore, the most complex decision boundaries occur with larger values of max_depth. If the boundaires become too complex, this might lead to overfitting.