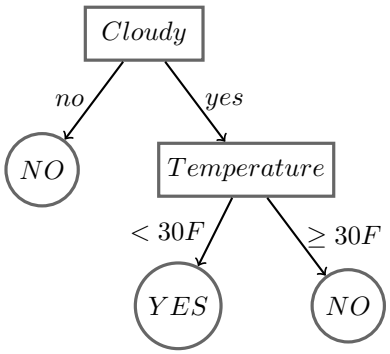
1 Definitions Oh My!

(a) **Decision Trees**: Popular representation and technique for classification and prediction. 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 be rainy today?



(b) **Entropy**: A measurement of the uncertainty in a random variable. Hence, high entropy means we are less confident about the outcome.
Note: A random variable assigns numerical values to outcomes of a random experiment or process. For example, a coin flip (F) is a random variable where $P(F = heads) = \frac{1}{2}$.

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

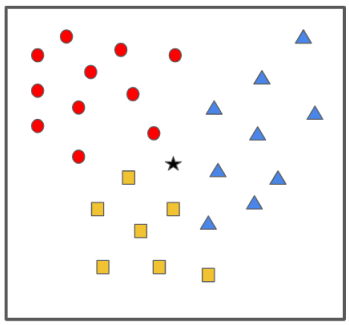
(c) **Conditional-Entropy**: The expected value of the entropy of Y given X, over all values of X. This lets us quantify the entropy of Y given that we know X.

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

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

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

(e) **K-Nearest Neighbors (KNN)**: Classification algorithm in which the predicted label for a particular data point is determined by the labels of the k nearest neighbors to that point.



2 Decision Trees

2.1 Mutual Information and Entropy Practice

1. Calculate the entropy of tossing a fair coin.

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

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

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

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, rainy)
- The temperature outside (hot, mild, cool)
- The humidity 30 minutes before he plans on running (high, normal)

As a member of the 10-315 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.

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

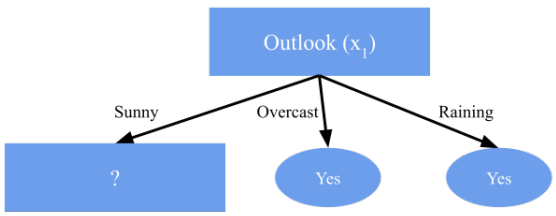
4. Calculate $I(Y; X_3)$, the mutual information when splitting on humidity.

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

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 $I(Y; X_2)$.

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

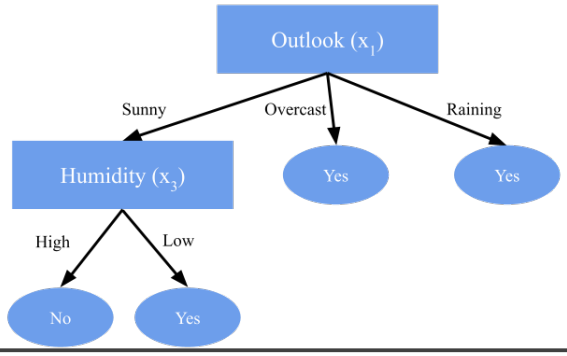
9. Calculate $I(Y; X_3)$.

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

10. Which attribute do we split on now?

Humidity (X_3), which perfectly splits the data.

11. Draw the final decision tree.



2.3 Discussion Questions

Explore different maximum depths of a decision tree by running through the DT Jupyter Notebook. Let d be the max-depth of the decision tree

1. What is the best value of d for this dataset?

   2 or 3

2. At which value of d do the training and validation errors begin to diverge?

   The errors start to diverge at depth 3, suggesting overfitting.

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

   Larger values of d tend to overfit to the data. The deeper the tree is developed, the more complex the tree becomes and the more it fits specifically to the training data. Hence it may not perform well to unseen and new data.

3 K Nearest Neighbor

Explore different values for "k" by running through the kNN Jupiter Notebook (see recitation materials).

3.1 Discussion Questions

1. What is the best value of k for this dataset?

   Between 5 and 20

2. Why do you think the error increases as k gets larger?

   As k increases, you use a larger set of points to determine the class of your current point. This leads to underfitting.

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

   Smaller values of k tend to overfit to 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 k would the decision boundaries be the most complex?

   Decision boundaries become more smooth as k increases. Therefore, the most complex decision boundaries occur with small values of k.

4 Putting it Together

Which methods can achieve zero-training error on this dataset?

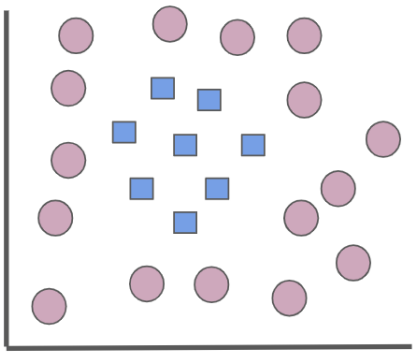
(a) Decision Trees

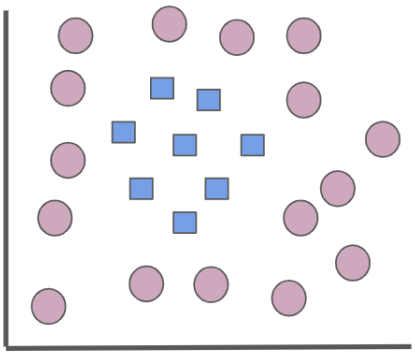
(b) 1-Nearest Neighbor

(c) Both

(d) Neither

Bonus: If zero error, what could the decision boundary look like?





(c) Both
Left - Decision Trees
Right - 1 Nearest Neighbor