Discussion 2
Machine Learning,Spring 2019

1. Decision Trees

Concepts

(a) False. You can't have more splits than training examples.

(b) True

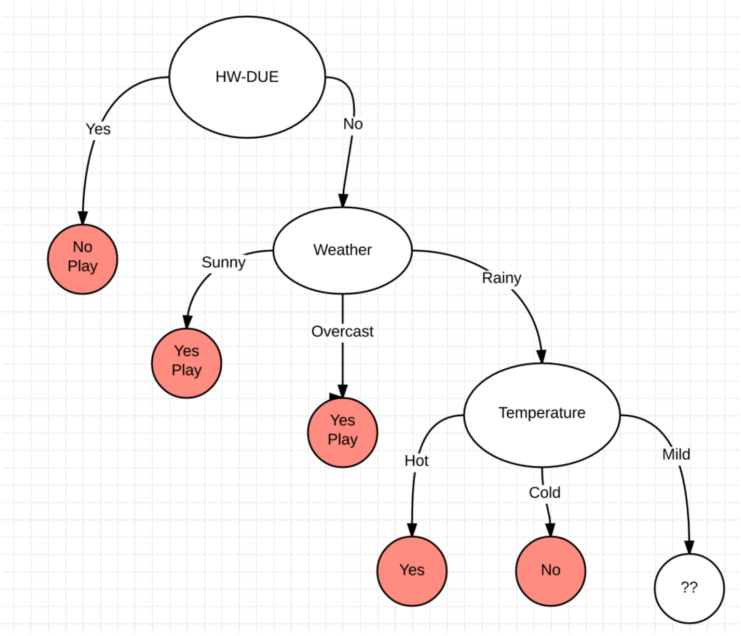
(c) True

(d) False

(e) False

Practice

Consider the following dataset for a problem where we try to predict the last column (can play). We're gonna be using the **ID3 algorithm** for construction of the decision tree.



Training

Figure 1: Decision Trees.

(a) What is the first feature that you would split on ?

Solution First attribute we split on is HW-Due.

(b) What is the number of levels of the decision tree constructed (leaf nodes included)?

Solution The number of levels here is 4 (since we count the nodes at the bottom as well).

(c) What's the training error of this decision tree?

Solution 0.00, since we were able to classify all points to the right classes.

(d) Given the test dataset, how many testing points do we classify INCORRECTLY? Enter as an integer.

Solution Only one testing point is predicted incorrectly.

2. Entropy

The entropy H of a discrete random variable X with n possible values is defined by the formula

$$H(X) = -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i).$$

Similarly, the joint entropy of two random variables X with n outcomes and Y with m outcomes is

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ log\ p(x_i,y_j).$$

2.1

As $p(x)$ is the probability of outcome x, $0 \leq p(x) \leq 1$ for any outcome. Use this fact to prove that $H(X) \geq 0$ for any random variable X.

Answer

$0 \leq p(x) \leq 1$ implies that $log\ p(x)$ is always negative. Since $p(x)$ is always positive, $p(x)\ log\ p(x)$ must be negative and therefore the sum $\sum_{i=1}^{n} p(x_i)\ log\ p(x_i)$ must in turn be negative. This implies that $-\sum_{i=1}^{n} p(x_i)\ log\ p(x_i)$ is positive.

2.2

Consider a discrete random variable Y with a uniform probability distribution over n outcomes, i.e. $p(y) = 1/n$. Is $H(Y)$ bounded as $n \to \infty$?

Answer

$$H(Y) = -\sum_{i=1}^{n} \frac{1}{n} \log\ \frac{1}{n}$$

$$= -n\frac{1}{n}(\log 1 - \log n) = \log n$$

$\log n$ is unbounded as $n \to \infty$.

2.3

The conditional entropy $H(Y|X)$ is defined

$$H(Y|X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\ log\ p(y_j|x_i).$$

Show that $H(X,Y) = H(X) + H(Y|X)$. Here are some hints:

- $log\ \frac{a}{b} = log\ a - log\ b$
- $p(y|x) = \frac{p(x,y)}{p(x)}$
- $\sum_{j=1}^{m} p(x_i, y_j) = p(x_i)$.

Answer

$$\begin{aligned}
H(X) + H(Y|X) &= -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\ log\ p(y_j|x_i) \\
&= -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\ log\ \frac{p(x_i, y_j)}{p(x_i)} \\
&= -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} [p(x_i, y_j)\ (log\ p(x_i, y_j) - log\ p(x_i))] \\
&= -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\ \ log\ p(x_i, y_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\ log\ p(x_i)) \\
&= -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\ \ log\ p(x_i, y_j) + \sum_{i=1}^{n} p(x_i)\ log\ p(x_i) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\ log\ p(x_i, y_j) \\
&= H(X,Y)
\end{aligned} \tag{1}$$

Information Gain

The term "information gain" has been used in our discussion of decision trees to describe the extra information we gain about some variable X by including a new variable Y that may lend extra predictive power in a model. Formally, the information gain (also called the mutual information) is defined as:

$$I(X;Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\ log\ \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

2.4

Provide an intuitive explanation of why the *information gain criterion* helps us choose a good split in a decision tree.

Answer

It makes the leaves purer (closer to certain about the correct class) on average.

2.5

When two random variables X and Y are independent, then their joint distribution $P(X,Y)$ factorizes into $P(X,Y) = P(X)P(Y)$. What is the information gain for two independent random variables? What does this say about the ability of one variable to explain the other?

Answer

$$\begin{aligned} I(X;Y) =& \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j) log\ \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \\ =& \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j) log\ \frac{p(x_i)p(y_j)}{p(x_i)p(y_j)} \\ =& \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j) log\ 1 \\ =& 0 \end{aligned} \tag{2}$$

The information gain is zero. This means that no new information is gained when adding a conditionally independent variable.

2.6

Show that $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

Answer

$$\begin{aligned} I(X;Y) =& \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ log\ \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \\ =& \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ log\ p(x_i,y_j) - log\ p(x_i)p(y_j) \\ =& \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ log\ p(x_i,y_j) - log\ p(x_i) - log\ p(y_j) \\ =& \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ log\ p(x_i,y_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ log\ p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ log\ p(y_j) \\ =& \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ log\ p(x_i,y_j) - \sum_{i=1}^{n} p(x_i)\ log\ p(x_i) - \sum_{j=1}^{m} p(y_j)\ log\ p(y_j) \\ =& - H(X,Y) + H(X) + H(Y) \end{aligned} \tag{3}$$