Lecture 7: Decision Trees

Aru Gyani

Based on Prof. Rishabh Iyer's slides.

Part I

Probability, Random Variables, and Entropy

1 Probability

The following content is sourced from the following:

- seeing-theory
- Khan Academy
- Professor Rishabh Iyer's class notes
- BYJU'S
- ThoughtCo

1.1 Discrete Probability

Definition 1.1 Sample space

Sample space specifies the set of possible outcomes.

For example, $\Omega = \{H, T\}$ would be the set of possible outcomes of a coin flip. Simply put, we are saying a coin flip could be either *heads* or *tails*.

Definition 1.2 Probability

Probability is simply how likely something is to happen

For each element ω inside of our sample space Ω, there is a number $p(\omega) \epsilon [0, 1]$ called a probability. This represents how likely the event ω is to happen.

The total probability of all possible outcomes (each outcome being an element ω) within the sample space Ω adds up to 1. This means that one of the outcomes in the sample space is certain to occur.

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

For exaxmple, a biased coin might have $p(H) = .6$ and $p(T) = .4$.

Note: $[0, 1]$ is a range, not a set. This number is always between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.

Definition 1.3 Event

An **event** is a subset of the sample space Ω, e.g.

- Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ be the 6 possible outcomes of a dice roll.
- $A = \{1, 5, 6\} \subseteq \Omega$ would be the event that the dice roll comes up as a one, five, or six.

The probability of an event is just the sum of all the outcomes that it contains.

$$p(A) = p(1) + p(5) + p(6)$$

Note: $A \subseteq \Omega$ means that A is a subset of Ω. Below, is a brief recap of what a subset is.

Definition 1.4 Subset

A set A is a **subset** of another set B if all of elements of the set A are elements of the set B. In other words, the set A is contained inside the set B.

1.2 Conditional Probability

1.2.1 Independence

We say two events are independent if knowing one event occurred doesn't change the probability of the other event. For example, the probability that a fair coin shows "heads" after being flipped is $\frac{1}{2}$. What if we knew the day was Tuesday? Does this change the probability of getting "heads"? Of course not. The probability of getting "heads", given that it's a Tuesday, is still $\frac{1}{2}$. Thus, the result of a coin flip and the day being Tuesday are independent events. More reading .

Definition 1.5 Independence

Two events A and B are independent if

$$p(A \cap B) = p(A)p(B)$$

Informally, $p(A \cap B)$ indicates the probability of A and B (aka probability of A intersection B). This means that it indicates the likelihood of both events happening.

Now, when we multiply $p(A)p(B)$, we are calculating the probability that A occurs and then, independently, that B occurs.

The probability that both events occur simultaneously (the intersection) is the product of their individual probabilities.

Definition 1.6 Multiplication Rule

The **multiplication rule** for independent events relates the probabilities of two events to the probability that they both occur.

For example, suppose that we roll a six-sided die and then flip a coin. These two events are independent. The probability of rolling a 1 is $\frac{1}{6}$. The probability of a head is $\frac{1}{2}$. The probability of rolling a 1 *and* getting a head is $\frac{1}{6} * \frac{1}{2} = \frac{1}{12}$

Definition 1.7 Intersection

The intersection of sets A and B is the set of all elements which are common to both A and B.

Suppose A is the set of even numbers less than 10 and B is the set of the first five multiples of 4, then the intersection of these two can be identified as:

$$A = \{2, 4, 6, 8\} \tag{1}$$
$$B = \{4, 8, 12, 16, 20\} \tag{2}$$

The elements common to A and B are 4 and 8. Therefore, the set of elements in the intersection A and B is:

$$P(A \cap B) = \{4, 8\}$$

Independence Example 1 Let's suppose that we have a fair die: $p(1) = \cdots = p(6) = \frac{1}{6}$ (the probability of rolling any number on the dice is equal). If $A = \{1, 2, 5\}$ and

$B = \{3, 4, 6\}$ are A and B independent?

No, they aren't independent. We know this because the intersection of A and B is $P(A \cap B) = \{\}$ or $\emptyset$. This means that their intersection has a probability of 0, meaning it will never happen. Compare this to $P(A)P(B)$ and we find that:

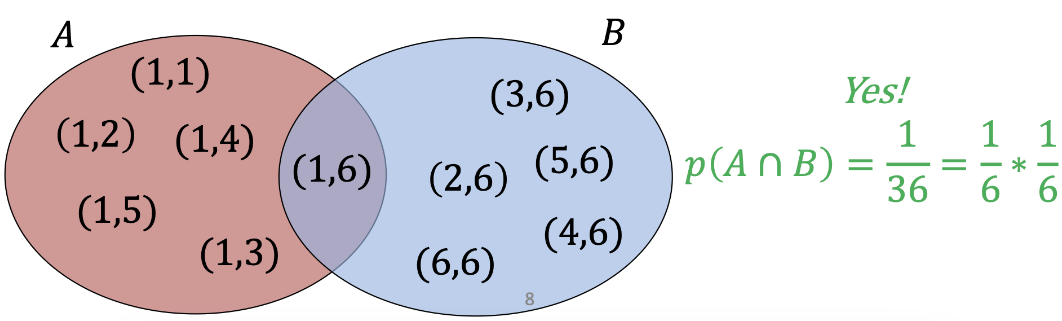
$$P(A) = P(1) + P(2) + P(5) = \frac{3}{6}$$
$$P(B) = P(3) + P(4) + P(6) = \frac{3}{6}$$
$$P(A)P(B) = \frac{3}{6} * \frac{3}{6} = \frac{1}{4} = 0.25$$

Thus, given that $P(A)P(B)$ represents the probability of both A and B occurring independently (see Definition 1.6: Multiplication Rule), we find that the intersection does not equal their independent occurrence. This means that A and B are not independent.

Independence Example 2 Now, suppose that $\Omega = \{(1,1), (1,2), \cdot, (6,6)\}$ is the set of all possible rolls of two **unbiased** dice.

- Let $A = \{(1,1), (1,2), (1,3), \cdot, (1,6)\}$ be the event that the first die rolls a 1.
- Let $B = \{(1,6), (2,6), \cdot, (6,6)\}$ be the event that the second die rolls a 6.

Are A and B independent?



1.2.2 Conditional Probability

Definition 1.8 Conditional Probability

The **conditional probability** of an event A given an event B with B with $p(B) > 0$ is defined to be:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

This formula tells us that to find the probability of A given B, you divide the probability of both A and B occurring together by the probability of B occurring.

Note: The vertical bar "|" means "given". So, $P(A|B)$ can be read as "the probability that Event A occurs <u>given</u> Event B has occurred".

This is the probability of the event $A \cap B$ happening over the sample space $\Omega' = B$.

Properties of Conditional Probability:

- $\sum_{\omega \in \Omega'} p(\omega|B) = 1$
- If A and B are independent, then $p(A|B) = p(A)$.

$$\begin{aligned} p(A|B) &= \frac{p(A \cap B)}{p(B)} \\ &= \frac{p(A)p(B)}{p(B)} \\ &= p(A) \end{aligned}$$

Conditional Probability Example 1: Dice Rolls To demonstrate the calculations of conditional probability, let's take two events.

- $A = \{\text{Dice Roll is even}\} = \{2, 4, 6\}$
- $B = \{\text{Dice Roll} > 4\} = \{5, 6\}$

Conditional probability is about looking at what the chances are of one thing happening if we know that something else already happened.

$P(A|B)$: "If we know that the dice is greater than 4, what's the chance it's also even"?
To figure this out let's go through the following steps:

1. Look at the potential outcomes in B, which are $\{5, 6\}$.

2. From those, find the ones that are also in A. Only the number 6 is both greater than 4 and even.

3. Since there are two numbers in B and only one of them is also in A, the conditional probability $P(A|B)$ is $\frac{1}{2}$.

Since we know B has happened, we know we are only looking at a world where the dice roll is 5 or 6 (> 4). In that world, there's only one even number, which is 6. **Conditional probability adjusts your focus**. Instead of looking at all possible dice rolls, you're only looking at the ones where B is true. Then you check, out of those, how often A is true.

Question: How is the probability $\frac{1}{2}$ when we aren't guaranteed a 6 when we look A (it could also be 5).

When we calculate the conditional probability $P(A|B)$, we are not talking about rolling the dice again. We're talking about a scenario where we already know that a certain event (in this case, event B: rolling a number greater than 4) has occurred. This knowledge "shrinks" our sample space — the set of possible outcomes we consider — to only those that fit this condition.

So, when we're looking at $P(A|B)$, we're asking "Given that we have rolled a number greater than 4, what is the probability that this number is also even"?

Since we know that B has happened, we ignore the rest of our original sample space of the dice $(1, 2, 3, 4)$ and only consider $\{5, 6\}$. Out of these two possible outcomes, what's the probability the number we rolled is even? Thus, we reach at $\frac{1}{2}$.

To show this same process done using the formula…

$$
\begin{aligned}
p(A|B) &= \frac{p(A \cap B)}{p(B)} \\
p(A \cap B) &= p(6) = \frac{1}{6} \\
p(B) &= p(5) + p(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \\
\frac{p(A \cap B)}{p(B)} &= \frac{\frac{1}{6}}{\frac{1}{3}} \\
&= \frac{1}{6} * \frac{3}{1} \\
&= \frac{1}{2}
\end{aligned}
$$

2 Discrete Random Variables

Definition 2.1 Discrete Random Variables

A **discrete random variable** X, is a function from the state (sample) space Ω into a discrete space D.

1. For each $x \in D$:

$$p(X = x) \equiv p(\{\omega \in \Omega : X(\omega) = x\})$$

   is the probability that X takes the value of x.

2. $p(X)$ defines a probability distribution:

$$\sum_{x \in D} p(X = x) = 1$$

Random variables partition the state space into disjoint events.

Definition 2.2 Discrete Space

A **discrete** space D refers to a set of distinct, separate values that a random value X can take on. In the context of probability, *discrete* refers to a space that is countable (if you can find the first element in the space, then you can find the second, then the third, etc.)

Notation:

- ω stands for an individual outcome of the random process you're considering. It's like a placeholder for any one of the specific results that could happen.

- Ω represents the sample space, which is the set of all possible outcomes of the random process.

  So, when you see $\omega \in \Omega$, it means that w is one of the possible outcomes within the sample space Ω.

Let's break down the properties of discrete random variables we see in the definition with an example.

Suppose you roll a six-sided die, and you're intereseted in whether the result is odd or even.

1. Your state space Ω is the possible die results: $\{1, 2, 3, 4, 5, 6\}$.

2. You define a random variable X where $X(\omega)$ is 0 if ω (the die result) is odd,

and 1 if ω is even.

3. The discrete space D that X maps to is $\{0, 1\}$ because those are the only two values X can take.

4. Now, X is a function because it takes each outcome of the die roll and gives you either a 0 or a 1 based on the rule you defined.

Let's try to model a fair coin flip using the formal notation:

$$\Omega = \{H, T\}$$
$$D = \{0, 1\}$$
$$X(\omega) = \Omega \rightarrow D = \begin{cases} \omega = H : 1 \\ \omega = T : 0 \end{cases}$$

We can now use this in a probability distribution, $p(X)$.

3 Entropy

3.1 Conditional Entropy

3.2 Information Gain

Part II
Decision Trees

Definitions