GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

SESSION 13 SUPPLEMENT: PROBABILITY

Rob Hall

DAT13 SF // April 20, 2015

I. PROBABILITY SUPPLEMENT

Q: What is a **probability***?*

Q: What is a **probability***?*

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

Q: What is a **probability**?

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

The probability of event A is denoted $P(A)$.

Q: What is the set of all possible events called?

Q: What is the set of all possible events called?

A: This set is called the **sample space** Ω. *Event* A *is a member of the sample space, as is every other event.*

Q: What is the set of all possible events called?
A: This set is called the **sample space** Ω*. Event* A *is a member of the sample space, as is every other event.*

The probability of the sample space $P(\Omega)$ *is 1.*

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

A: With the **joint probability** *of A and B, written $P(AB)$.*

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

A: The intersection of A & B divided by region B.

Q: Suppose event B has occurred. What quantity represents the probability of A ***given*** *this information about B?*

A: The intersection of A & B divided by region B.

NOTE

This information about B transforms the sample space.

Take a moment to convince yourself of this!

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

A: The intersection of A & B divided by region B.

This is called the **conditional probability** *of A given B, written $P(A|B) = P(AB) / P(B)$.*

NOTE

This information about B transforms the sample space.

Take a moment to convince yourself of this!

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

A: The intersection of A & B divided by region B.

This is called the **conditional probability** *of A given B, written $P(A|B) = P(AB) / P(B)$.*

Notice, with this we can also write $P(AB) = P(A|B) * P(B)$.

NOTE

This information about B transforms the sample space.

Take a moment to convince yourself of this!

Q: What does it mean for two events to be **independent***?*

Q: What does it mean for two events to be **independent***?*

A: Information about one does not affect the probability of the other.

Q: What does it mean for two events to be **independent**?

A: Information about one does not affect the probability of the other.

This can be written as $P(A|B) = P(A)$.

*Q: What does it mean for two events to be **independent**?*

A: Information about one does not affect the probability of the other.

This can be written as $P(A|B) = P(A)$.

Using the definition of the conditional probability, we can also write:

$$P(A|B) = P(AB) / P(B) = P(A) \rightarrow \quad P(AB) = P(A) * P(B)$$

A motivating example: COOKIES!

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

In other words, we want: P(Bowl 1 | vanilla) This is a conditional probability.

Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

In other words, we want: P(Bowl 1 | vanilla) This is a conditional probability.

How can we compute this?

Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

In other words, we want: P(Bowl 1 | vanilla)

What about P(vanilla | Bowl1) ?

Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

In other words, we want: P(Bowl 1 | vanilla)

What about P(vanilla | Bowl1) ? That's easy! P(vanilla | Bowl1) = 30/40 = 3/4

Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

In other words, we want: P(Bowl 1 | vanilla)

But P(Bowl1 | vanilla) is NOT equal to P(vanilla | Bowl1) = 3/4

The way we get from P(Bowl1 | vanilla) to P(vanilla | Bowl1) is as follows:

The way we get from P(Bowl1 | vanilla) to P(vanilla | Bowl1) is as follows:

$P(AB) = P(A|B) * P(B)$ *from earlier slide*

The way we get from P(Bowl1 | vanilla) to P(vanilla | Bowl1) is as follows:

$P(AB) = P(A|B) * P(B)$ *from earlier slide*

$P(BA) = P(B|A) * P(A)$ *by substitution*

CHECK THIS OUT

The way we get from P(Bowl1 | vanilla) to P(vanilla | Bowl1) is as follows:

$P(AB) = P(A|B) * P(B)$ *from earlier slide*

$P(BA) = P(B|A) * P(A)$ *by substitution*

But $P(AB) = P(BA)$ *since event* AB = *event* BA

The way we get from P(Bowl1 | vanilla) to P(vanilla | Bowl1) is as follows:

$P(AB) = P(A|B) * P(B)$ *from earlier slide*

$P(BA) = P(B|A) * P(A)$ *by substitution*

But $P(AB) = P(BA)$ *since event* AB = *event* BA

→ $P(A|B) * P(B) = P(B|A) * P(A)$ *by combining the above*

The way we get from P(Bowl1 | vanilla) to P(vanilla | Bowl1) is as follows:

$P(AB) = P(A|B) * P(B)$	*from earlier slide*
$P(BA) = P(B|A) * P(A)$	*by substitution*
But $P(AB) = P(BA)$	*since event* AB = *event* BA
→ $P(A|B) * P(B) = P(B|A) * P(A)$	*by combining the above*
→ $P(A|B) = P(B|A) * P(A) / P(B)$	*by rearranging last step*

This result is called **Bayes' theorem**.

$$P(A|B) = P(A) * P(B|A) / P(B)$$

We want: P(Bowl 1 | vanilla)

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

P(A|B) = P(A) * P(B|A) / P(B)

What is P(A)?
What is P(B)?
What is P(B|A)?

We want: P(Bowl 1 | vanilla)

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

$P(A|B) = P(A) * P(B|A) / P(B)$

$P(A) = 0.5$

$P(B) = 50 / 80 = 5/8$

$P(B|A) = 30/40 = 3/4$

We want: P(Bowl 1 | vanilla)

Bowl 1 contains:
30 vanilla cookies
10 chocolate chip cookies

Bowl 2 contains:
20 vanilla cookies
20 chocolate chip cookies

$P(A|B) = P(A) * P(B|A) / P(B) = 0.5 * 6/8 / 5/8 = \mathbf{3/5}$

$P(A) = 0.5$

$P(B) = 50 / 80 = 5/8$

$P(B|A) = 30/40 = 3/4$

This result is called **Bayes' theorem**. *Here it is again:*

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

This result is called **Bayes' theorem**. *Here it is again:*

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

- It's interesting because it's kind of a "wormhole" between two different "interpretations" of probability.

This result is called **Bayes' theorem**. *Here it is again:*

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

- It's interesting because it's kind of a "wormhole" between two different "interpretations" of probability.

- It's a very powerful computational tool.