Quick review of probability theory
9.S918
Spring 2024

Roger Levy

MIT Course 9 (Brain & Cognitive Sciences)

2 April 2024

Core introductory concepts in probability theory

- Foundations of probability theory
- Joint, marginal, and conditional probability
- Bayes' Rule
- Conditional Independence
- Discrete and continuous random variables
- Mean, variance, covariance, and correlation

Probability spaces

Traditionally, probability spaces are defined in terms of **sets**. An event E is a subset of a **sample space** Ω: $E \subseteq \Omega$.

Probability spaces

Traditionally, probability spaces are defined in terms of **sets**. An event E is a subset of a **sample space** Ω: $E \subseteq \Omega$.

A **probability space** P on a sample space Ω is a function from events E in Ω to real numbers such that the following three axioms hold:

1. $P(E) \geq 0$ for all $E \subseteq \Omega$ (non-negativity).
2. If E_1 and E_2 are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).

Probability spaces

Traditionally, probability spaces are defined in terms of **sets**. An event E is a subset of a **sample space** Ω: $E \subseteq \Omega$.

A **probability space** P on a sample space Ω is a function from events E in Ω to real numbers such that the following three axioms hold:

1. $P(E) \geq 0$ for all $E \subseteq \Omega$ (non-negativity).
2. If E_1 and E_2 are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).

Note that the set-theoretic characterization of events can also be translated into fundamental operations in Boolean logic:

	Sets	Boolean logic
Subset	$A \subseteq B$	$A \rightarrow B$
Disjointness	$E_1 \cap E_2 = \emptyset$	$\neg(E_1 \wedge E_2)$
Union	$E_1 \cup E_2$	$E_1 \vee E_2$

A simple example

In historical English, object NPs could be *preverbal* or *postverbal*.

There is a broad cross-linguistic tendency for *pronominal* objects to occur earlier on average than *non-pronominal* objects.

So, hypothetical probabilities from historical English:

		Y:	
		Pronoun	**Not Pronoun**
X:	Object **Preverbal**	0.224	0.655
	Object **Postverbal**	0.014	0.107

A simple example

In historical English, object NPs could be *preverbal* or *postverbal*.

There is a broad cross-linguistic tendency for *pronominal* objects to occur earlier on average than *non-pronominal* objects.

So, hypothetical probabilities from historical English:

		Y:	
		Pronoun	**Not Pronoun**
X:	Object **Preverbal**	0.224	0.655
	Object **Postverbal**	0.014	0.107

We will sometimes call this the **joint distribution** $P(X, Y)$ over two **random variables**—here, verb-object word order X and object pronominality Y.

Checking the axioms of probability

1. $P(E) \geq 0$ for all $E \subset \Omega$ (non-negativity).
2. If E_1 and E_2 are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).

	Object	
	Pronoun	**Not Pronoun**
Object **Preverbal**	0.224	0.655
Object **Postverbal**	0.014	0.107

- We can consider the sample space to be

$$\begin{aligned}\Omega = \{&\textbf{Preverbal+Pronoun}, \textbf{Preverbal+Not Pronoun},\\ &\textbf{Postverbal+Pronoun}, \textbf{Postverbal+Not Pronoun}\}\end{aligned}$$

Checking the axioms of probability

1. $P(E) \geq 0$ for all $E \subset \Omega$ (non-negativity).
2. If E_1 and E_2 are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).

	Object	
	Pronoun	**Not Pronoun**
Object **Preverbal**	0.224	0.655
Object **Postverbal**	0.014	0.107

- We can consider the sample space to be

$$\Omega = \{\textbf{Preverbal+Pronoun}, \textbf{Preverbal+Not Pronoun}, \textbf{Postverbal+Pronoun}, \textbf{Postverbal+Not Pronoun}\}$$

- Disjoint union tells us the probabilities of non-atomic events:

Checking the axioms of probability

1. $P(E) \geq 0$ for all $E \subset \Omega$ (non-negativity).
2. If E_1 and E_2 are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).

	Object	
	Pronoun	**Not Pronoun**
Object **Preverbal**	0.224	0.655
Object **Postverbal**	0.014	0.107

- We can consider the sample space to be

$$\begin{aligned}\Omega =\{&\textbf{Preverbal+Pronoun}, \textbf{Preverbal+Not Pronoun},\\ &\textbf{Postverbal+Pronoun}, \textbf{Postverbal+Not Pronoun}\}\end{aligned}$$

- Disjoint union tells us the probabilities of non-atomic events:
  - If we define $E_1 = \{\textbf{Preverbal+Pronoun}, \textbf{Postverbal+Not Pronoun}\}$, then $P(E_1) = 0.224 + 0.107 = 0.331$.

Checking the axioms of probability

1. $P(E) \geq 0$ for all $E \subset \Omega$ (non-negativity).
2. If E_1 and E_2 are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).

	Object	
	Pronoun	**Not Pronoun**
Object **Preverbal**	0.224	0.655
Object **Postverbal**	0.014	0.107

- We can consider the sample space to be

$$\Omega = \{\textbf{Preverbal+Pronoun}, \textbf{Preverbal+Not Pronoun}, \textbf{Postverbal+Pronoun}, \textbf{Postverbal+Not Pronoun}\}$$

- Disjoint union tells us the probabilities of non-atomic events:
  - If we define $E_1 = \{\textbf{Preverbal+Pronoun}, \textbf{Postverbal+Not Pronoun}\}$, then $P(E_1) = 0.224 + 0.107 = 0.331$.
- Check for properness: $P(\Omega) = 0.224 + 0.655 + 0.014 + 0.107 = 1$

Marginal probability

- Sometimes we have a joint distribution $P(X, Y)$ over random variables X and Y, but we're interested in the distribution implied over one of them (here, without loss of generality, X)

Marginal probability

- Sometimes we have a joint distribution $P(X, Y)$ over random variables X and Y, but we're interested in the distribution implied over one of them (here, without loss of generality, X)
- The **marginal probability distribution** $P(X)$ is

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

Marginal probability

- Sometimes we have a joint distribution $P(X, Y)$ over random variables X and Y, but we're interested in the distribution implied over one of them (here, without loss of generality, X)
- The **marginal probability distribution** $P(X)$ is

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

- This is sometimes known as the **law of total probability**.

Marginal probability: an example

		Y: Pronoun	Not Pronoun
X:	Object **Preverbal**	0.224	0.655
	Object **Postverbal**	0.014	0.107

Finding the marginal distribution on X:

$$\begin{aligned} P(X = \textbf{Preverbal}) &= P(X = \textbf{Preverbal}, Y = \textbf{Pronoun}) \\ &\quad + P(X = \textbf{Preverbal}, Y = \textbf{Not Pronoun}) \\ &= 0.224 + 0.655 \\ &= 0.879 \end{aligned}$$

$$\begin{aligned} P(X = \textbf{Postverbal}) &= P(X = \textbf{Postverbal}, Y = \textbf{Pronoun}) \\ &\quad + P(X = \textbf{Postverbal}, Y = \textbf{Not Pronoun}) \\ &= 0.014 + 0.107 \\ &= 0.121 \end{aligned}$$

Marginal probability: an example

		Y: Pronoun	Not Pronoun
X:	Object **Preverbal**	0.224	0.655
	Object **Postverbal**	0.014	0.107

So, the marginal distribution on X is

	$P(X)$
Preverbal	0.879
Postverbal	0.121

Likewise, the marginal distribution on Y is

	$P(Y)$
Pronoun	0.238
Not Pronoun	0.762

Conditional probability

The conditional probability of event B given that A has occurred/is known is defined as follows:

$$P(B|A) \equiv \frac{P(A, B)}{P(A)}$$

Conditional Probability: an example

		Y: Pronoun	Not Pronoun
X:	Object **Preverbal**	0.224	0.655
	Object **Postverbal**	0.014	0.107

	$P(X)$
Preverbal	0.879
Postverbal	0.121

	$P(Y)$
Pronoun	0.238
Not Pronoun	0.762

Conditional Probability: an example

		Y: Pronoun	Not Pronoun
X:	Object **Preverbal**	0.224	0.655
	Object **Postverbal**	0.014	0.107

	$P(X)$
Preverbal	0.879
Postverbal	0.121

	$P(Y)$
Pronoun	0.238
Not Pronoun	0.762

How do we calculate the following?

$$P(Y = \textbf{Pronoun} | X = \textbf{Postverbal}) = \frac{P(X = \textbf{Postverbal}, Y = \textbf{Pronoun})}{P(X = \textbf{Postverbal})}$$

$$= \frac{0.014}{0.121} = 0.116$$

Conditional Probability: an example

		Y:	
		Pronoun	**Not Pronoun**
X:	Object **Preverbal**	0.224	0.655
	Object **Postverbal**	0.014	0.107

	$P(X)$
Preverbal	0.879
Postverbal	0.121

	$P(Y)$
Pronoun	0.238
Not Pronoun	0.762

How do we calculate the following?

$$P(Y = \textbf{Pronoun} | X = \textbf{Postverbal})$$

$$= \frac{0.014}{0.121} = 0.116$$

Conditional Probability: an example

		Y: Pronoun	Not Pronoun
X:	Object **Preverbal**	0.224	0.655
	Object **Postverbal**	0.014	0.107

	$P(X)$
Preverbal	0.879
Postverbal	0.121

	$P(Y)$
Pronoun	0.238
Not Pronoun	0.762

How do we calculate the following?

$$P(Y = \textbf{Pronoun}|X = \textbf{Postverbal}) = \frac{P(X = \textbf{Postverbal}, Y = \textbf{Pronoun})}{P(X = \textbf{Postverbal})}$$
$$= \frac{0.014}{0.121} = 0.116$$

The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

And this generalizes to more than two variables:

The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

And this generalizes to more than two variables:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

And this generalizes to more than two variables:

$$\begin{aligned} P(E_1, E_2) &= P(E_2|E_1)P(E_1) \\ P(E_1, E_2, E_3) &= P(E_3|E_1, E_2)P(E_2|E_1)P(E_1) \end{aligned}$$

The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

And this generalizes to more than two variables:

$$\begin{aligned}
P(E_1, E_2) &= P(E_2|E_1)P(E_1) \\
P(E_1, E_2, E_3) &= P(E_3|E_1, E_2)P(E_2|E_1)P(E_1) \\
\vdots \qquad & \qquad \vdots \\
P(E_1, E_2, \ldots, E_n) &= P(E_n|E_1, E_2, \ldots, E_{n-1}) \ldots P(E_2|E_1)P(E_1)
\end{aligned}$$

The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

And this generalizes to more than two variables:

$$\begin{aligned}
P(E_1, E_2) &= P(E_2|E_1)P(E_1) \\
P(E_1, E_2, E_3) &= P(E_3|E_1, E_2)P(E_2|E_1)P(E_1) \\
\vdots \quad & \quad \quad \vdots \\
P(E_1, E_2, \ldots, E_n) &= P(E_n|E_1, E_2, \ldots, E_{n-1}) \ldots P(E_2|E_1)P(E_1)
\end{aligned}$$

The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

And this generalizes to more than two variables:

$$\begin{aligned}
P(E_1, E_2) &= P(E_2|E_1)P(E_1) \\
P(E_1, E_2, E_3) &= P(E_3|E_1, E_2)P(E_2|E_1)P(E_1) \\
\vdots \qquad & \qquad \vdots \\
P(E_1, E_2, \ldots, E_n) &= P(E_n|E_1, E_2, \ldots, E_{n-1}) \ldots P(E_2|E_1)P(E_1)
\end{aligned}$$

The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

$$P(E_1, E_2) = P(E_2|E_1)P(E_1)$$

And this generalizes to more than two variables:

$$\begin{aligned}
P(E_1, E_2) &= P(E_2|E_1)P(E_1) \\
P(E_1, E_2, E_3) &= P(E_3|E_1, E_2)P(E_2|E_1)P(E_1) \\
\vdots \qquad & \qquad \vdots \\
P(E_1, E_2, \ldots, E_n) &= P(E_n|E_1, E_2, \ldots, E_{n-1}) \ldots P(E_2|E_1)P(E_1)
\end{aligned}$$

Breaking a joint probability down into the product of a marginal probability and several conditional probabilities this way is called **chain rule decomposition**.

Bayes' Rule (Bayes' Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Rule (Bayes' Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

With extra "background" random variables I:

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)}$$

Bayes' Rule (Bayes' Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

With extra "background" random variables I:

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)}$$

This "theorem" follows directly from def'n of conditional probability:

$$P(A, B) = P(B|A)P(A)$$

Bayes' Rule (Bayes' Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

With extra "background" random variables I:

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)}$$

This "theorem" follows directly from def'n of conditional probability:

$$P(A, B) = P(B|A)P(A)$$
$$P(A, B) = P(A|B)P(B)$$

Bayes' Rule (Bayes' Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

With extra "background" random variables I:

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)}$$

This "theorem" follows directly from def'n of conditional probability:

$$P(A, B) = P(B|A)P(A)$$
$$P(A, B) = P(A|B)P(B)$$

So

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

Bayes' Rule (Bayes' Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

With extra "background" random variables I:

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)}$$

This "theorem" follows directly from def'n of conditional probability:

$$P(A, B) = P(B|A)P(A)$$
$$P(A, B) = P(A|B)P(B)$$

So

$$P(A|B)P(B) = P(B|A)P(A)$$
$$\frac{P(A|B)P(B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Rule (Bayes' Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

With extra "background" random variables I:

$$P(A|B,I) = \frac{P(B|A,I)P(A|I)}{P(B|I)}$$

This "theorem" follows directly from def'n of conditional probability:

$$P(A,B) = P(B|A)P(A)$$
$$P(A,B) = P(A|B)P(B)$$

So

$$P(A|B)P(B) = P(B|A)P(A)$$
$$\frac{P(A|B)\cancel{P(B)}}{\cancel{P(B)}} = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Rule, more closely inspected

$$\overbrace{P(A|B)}^{\text{Posterior}} = \frac{\overbrace{P(B|A)}^{\text{Likelihood}}\,\overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Normalizing constant}}}$$

Bayes' Rule in action

Let me give you the same information you had before:

$$P(Y = \textbf{Pronoun}) = 0.238$$
$$P(X = \textbf{Preverbal} | Y = \textbf{Pronoun}) = 0.941$$
$$P(X = \textbf{Preverbal} | Y = \textbf{Not Pronoun}) = 0.860$$

[1] A "transitive" verb is one that requires an object.

Bayes' Rule in action

Let me give you the same information you had before:

$$
\begin{aligned}
P(Y = \textbf{Pronoun}) &= 0.238 \\
P(X = \textbf{Preverbal} | Y = \textbf{Pronoun}) &= 0.941 \\
P(X = \textbf{Preverbal} | Y = \textbf{Not Pronoun}\ ) &= 0.860
\end{aligned}
$$

Imagine you're an incremental sentence processor. You encounter a transitive verb[1] but haven't encountered the object yet. **Inference under uncertainty:** How likely is it that the object is a pronoun?

[1] A "transitive" verb is one that requires an object.

Bayes Rule in Action

$$P(Y = \textbf{Pronoun}) = 0.238$$
$$P(X = \textbf{Preverbal}|Y = \textbf{Pronoun}) = 0.941$$
$$P(X = \textbf{Preverbal}|Y = \textbf{Not Pronoun}) = 0.860$$

$$P(Y = \textbf{Pron}|X = \textbf{PostV})$$

Bayes Rule in Action

$$P(Y = \textbf{Pronoun}) = 0.238$$
$$P(X = \textbf{Preverbal} | Y = \textbf{Pronoun}) = 0.941$$
$$P(X = \textbf{Preverbal} | Y = \textbf{Not Pronoun}) = 0.860$$

$$P(Y = \textbf{Pron} | X = \textbf{PostV}) = \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{P(X = \textbf{PostV})}$$

Bayes Rule in Action

$$P(Y = \textbf{Pronoun}) = 0.238$$
$$P(X = \textbf{Preverbal} | Y = \textbf{Pronoun}) = 0.941$$
$$P(X = \textbf{Preverbal} | Y = \textbf{Not Pronoun}\ ) = 0.860$$

$$\begin{aligned} P(Y = \textbf{Pron} | X = \textbf{PostV}) &= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{P(X = \textbf{PostV})} \\ &= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV}, Y = y)} \end{aligned}$$

Bayes Rule in Action

$$P(Y = \textbf{Pronoun}) = 0.238$$
$$P(X = \textbf{Preverbal} | Y = \textbf{Pronoun}) = 0.941$$
$$P(X = \textbf{Preverbal} | Y = \textbf{Not Pronoun}) = 0.860$$

$$\begin{aligned}
P(Y = \textbf{Pron} | X = \textbf{PostV}) &= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{P(X = \textbf{PostV})} \\
&= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV}, Y = y)} \\
&= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV} | Y = y) P(Y = y)}
\end{aligned}$$

Bayes Rule in Action

$$\begin{aligned} P(Y = \textbf{Pronoun}) &= 0.238 \\ P(X = \textbf{Preverbal}|Y = \textbf{Pronoun}) &= 0.941 \\ P(X = \textbf{Preverbal}|Y = \textbf{Not Pronoun}\ ) &= 0.860 \end{aligned}$$

$$\begin{aligned} P(Y = \textbf{Pron}|X = \textbf{PostV}) &= \frac{P(X = \textbf{PostV}|Y = \textbf{Pron})P(Y = \textbf{Pron})}{P(X = \textbf{PostV})} \\ &= \frac{P(X = \textbf{PostV}|Y = \textbf{Pron})P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV}, Y = y)} \\ &= \frac{P(X = \textbf{PostV}|Y = \textbf{Pron})P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV}|Y = y)P(Y = y)} \\ &= \frac{P(X = \textbf{PostV}|Y = \textbf{Pron})P(Y = \textbf{Pron})}{\mathrm{P}(\textbf{PostV}|\textbf{Pron})\mathrm{P}(\textbf{Pron}) + \mathrm{P}(\textbf{PostV}|\textbf{NotPron})\mathrm{P}(\textbf{NotPron})} \end{aligned}$$

Bayes Rule in Action

$$\begin{aligned} P(Y = \textbf{Pronoun}) &= 0.238 \\ P(X = \textbf{Preverbal} | Y = \textbf{Pronoun}) &= 0.941 \\ P(X = \textbf{Preverbal} | Y = \textbf{Not Pronoun}) &= 0.860 \end{aligned}$$

$$\begin{aligned} P(Y = \textbf{Pron} | X = \textbf{PostV}) &= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{P(X = \textbf{PostV})} \\ &= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV}, Y = y)} \\ &= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV} | Y = y) P(Y = y)} \\ &= \frac{P(X = \textbf{PostV} | Y = \textbf{Pron}) P(Y = \textbf{Pron})}{\mathrm{P}(\textbf{PostV}|\textbf{Pron})\mathrm{P}(\textbf{Pron}) + \mathrm{P}(\textbf{PostV}|\textbf{NotPron})\mathrm{P}(\textbf{NotPron})} \\ &= \frac{(1 - 0.941) \times 0.238}{(1 - 0.941) \times 0.238 + (1 - 0.860) \times (1 - 0.238)} \end{aligned}$$

Bayes Rule in Action

$$
\begin{aligned}
P(Y = \textbf{Pronoun}) &= 0.238 \\
P(X = \textbf{Preverbal}|Y = \textbf{Pronoun}) &= 0.941 \\
P(X = \textbf{Preverbal}|Y = \textbf{Not Pronoun}\ ) &= 0.860
\end{aligned}
$$

$$
\begin{aligned}
P(Y = \textbf{Pron}|X = \textbf{PostV}) &= \frac{P(X = \textbf{PostV}|Y = \textbf{Pron})P(Y = \textbf{Pron})}{P(X = \textbf{PostV})} \\
&= \frac{P(X = \textbf{PostV}|Y = \textbf{Pron})P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV}, Y = y)} \\
&= \frac{P(X = \textbf{PostV}|Y = \textbf{Pron})P(Y = \textbf{Pron})}{\sum_y P(X = \textbf{PostV}|Y = y)P(Y = y)} \\
&= \frac{P(X = \textbf{PostV}|Y = \textbf{Pron})P(Y = \textbf{Pron})}{\mathrm{P}(\textbf{PostV}|\textbf{Pron})\mathrm{P}(\textbf{Pron}) + \mathrm{P}(\textbf{PostV}|\textbf{NotPron})\mathrm{P}(\textbf{NotPron})} \\
&= \frac{(1 - 0.941) \times 0.238}{(1 - 0.941) \times 0.238 + (1 - 0.860) \times (1 - 0.238)} \\
&= 0.116
\end{aligned}
$$

Other ways of writing Bayes' Rule

$$P(A|B) = \frac{\overbrace{P(B|A)}^{\text{Likelihood}}\,\overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Normalizing constant}}}$$

- The hardest part of using Bayes' Rule was calculating the normalizing constant (a.k.a. the **partition function**)

Other ways of writing Bayes' Rule

$$P(A|B) = \frac{\overbrace{P(B|A)}^{\text{Likelihood}}\ \overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Normalizing constant}}}$$

- The hardest part of using Bayes' Rule was calculating the normalizing constant (a.k.a. the **partition function**)
- Hence there are often two other ways we write Bayes' Rule:

Other ways of writing Bayes' Rule

$$P(A|B) = \frac{\overbrace{P(B|A)}^{\text{Likelihood}}\overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Normalizing constant}}}$$

- The hardest part of using Bayes' Rule was calculating the normalizing constant (a.k.a. the **partition function**)
- Hence there are often two other ways we write Bayes' Rule:
  1. Emphasizing explicit marginalization:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_a P(A=a,B)}$$

Other ways of writing Bayes' Rule

$$P(A|B) = \frac{\overbrace{P(B|A)}^{\text{Likelihood}}\overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Normalizing constant}}}$$

- The hardest part of using Bayes' Rule was calculating the normalizing constant (a.k.a. the **partition function**)
- Hence there are often two other ways we write Bayes' Rule:
  1. Emphasizing explicit marginalization:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_a P(A=a,B)}$$

  2. Ignoring the partition function:

$$P(A|B) \propto P(B|A)P(A)$$

(Conditional) Independence

Events A and B are said to be Conditionally Independent given information C if

$$P(A, B|C) = P(A|C)P(B|C)$$

Conditional independence of A and B given C is often expressed as

$$A \perp B|C$$

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
  - Each possible value has a probability **mass** $P(X = x)$ (or just $P(x)$ for short)

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
  - Each possible value has a probability **mass** $P(X = x)$ (or just $P(x)$ for short)
  - Properness is characterized in terms of a sum: $\sum_x P(X = x) = 1$

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
  - Each possible value has a probability **mass** $P(X = x)$ (or just $P(x)$ for short)
  - Properness is characterized in terms of a sum: $\sum_x P(X = x) = 1$
- A **continuous** random variable's support is a continuum (e.g., [0,1])

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
  - Each possible value has a probability **mass** $P(X = x)$ (or just $P(x)$ for short)
  - Properness is characterized in terms of a sum: $\sum_x P(X = x) = 1$
- A **continuous** random variable's support is a continuum (e.g., [0,1])
  - Each possible value has a probability **density** $p(X = x)$ (or just $p(x)$ for short)

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
  - Each possible value has a probability **mass** $P(X = x)$ (or just $P(x)$ for short)
  - Properness is characterized in terms of a sum: $\sum_x P(X = x) = 1$
- A **continuous** random variable's support is a continuum (e.g., [0,1])
  - Each possible value has a probability **density** $p(X = x)$ (or just $p(x)$ for short)
  - The probability *mass* of any value on the continuum is zero, regardless of the density!

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
  - Each possible value has a probability **mass** $P(X = x)$ (or just $P(x)$ for short)
  - Properness is characterized in terms of a sum: $\sum_x P(X = x) = 1$
- A **continuous** random variable's support is a continuum (e.g., [0,1])
  - Each possible value has a probability **density** $p(X = x)$ (or just $p(x)$ for short)
  - The probability *mass* of any value on the continuum is zero, regardless of the density!
  - Properness is characterized in terms of an integral: $\int_x P(X = x)\,\mathrm{d}x = 1$ (note the derivative!)

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
  - Each possible value has a probability **mass** $P(X = x)$ (or just $P(x)$ for short)
  - Properness is characterized in terms of a sum: $\sum_x P(X = x) = 1$
- A **continuous** random variable's support is a continuum (e.g., [0,1])
  - Each possible value has a probability **density** $p(X = x)$ (or just $p(x)$ for short)
  - The probability *mass* of any value on the continuum is zero, regardless of the density!
  - Properness is characterized in terms of an integral: $\int_x P(X = x)\, \mathrm{d}x = 1$ (note the derivative!)
  - Remember that probability densities have units (the inverse of the unit of the continuum), and the densities can exceed 1 per unit!

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
- A **continuous** random variable's support is a continuum (e.g., [0,1])

- (We are presently eliding over cases where a random variable can have both mass and density on different sets of values)

Discrete vs continuous random variables

- The **support** of a random variable is the set of values for which its probability is non-zero
- A **discrete** random variable's support is a finite or countably infinite number of values
- A **continuous** random variable's support is a continuum (e.g., [0,1])

- (We are presently eliding over cases where a random variable can have both mass and density on different sets of values)
- Unless I mention otherwise, things I say will hold for both discrete and continuous random variables, and I will freely use sums or integrals with the implicit understanding that what I say applies to both cases

Mean and variance

- (Population) **mean**, or **expected value**:

$$E[X] = \sum_x x\, P(X = x) \qquad \text{(discrete)}$$

$$E[X] = \int_x x\, p(X = x)\, \mathrm{d}x \qquad \text{(continuous)}$$

Mean and variance

- (Population) **mean**, or **expected value**:

$$E[X] = \sum_x x\, P(X = x) \qquad \text{(discrete)}$$

$$E[X] = \int_x x\, p(X = x)\, \mathrm{d}x \qquad \text{(continuous)}$$

- (Population) **variance**:

$$\mathrm{Var}[X] = \sum_x (x - E[X])^2\, P(X = x) \qquad \text{(discrete)}$$

$$\mathrm{Var}[X] = \int_x (x - E[X])^2\, p(X = x)\, \mathrm{d}x \qquad \text{(continuous)}$$

Covariance and correlation

- The **covariance** between two random variables is how much they vary together:

$$\mathrm{Cov}(X, Y) = \int_{x,y} (x - E[X])(y - E[Y]) \mathrm{d}x \mathrm{d}y$$