DSC 40A

Theoretical Foundations of Data Science I

Agenda

- Law of total probability.
- Bayes theorem.

Question

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!
If the direct link doesn't work, click the "Lecture Questions" link in the top right corner of dsc40a.com.

Getting to Campus

- You conduct a survey:
  - How did you get to campus today? Walk, bike, or drive?
  - Were you late?

	Late	Not Late
Walk	6%	24%
Bike	3%	7%
Drive	36%	24%

Getting to Campus

	Late	Not Late
Walk	6%	24%
Bike	3%	7%
Drive	36%	24%

What is the probability that a randomly selected person is late?

A. 24%
B. 30%
C. 45%
D. 50%

Getting to Campus

	Late	Not Late
Walk	6%	24%
Bike	3%	7%
Drive	36%	24%

- Since everyone either walks, bikes, or drives,

P(Late) = P(Late AND Walk) + P(Late AND Bike) + P(Late AND Drive)

- This is called the **Law of Total Probability**.

Getting to Campus

	Late	Not Late
Walk	6%	24%
Bike	3%	7%
Drive	36%	24%

Suppose someone tells you that they walked. What is the probability that they were late?

A. 6%
B. 20%
C. 25%
D. 45%

Getting to Campus

	Late	Not Late
Walk	6%	24%
Bike	3%	7%
Drive	36%	24%

- Since everyone either walks, bikes, or drives,

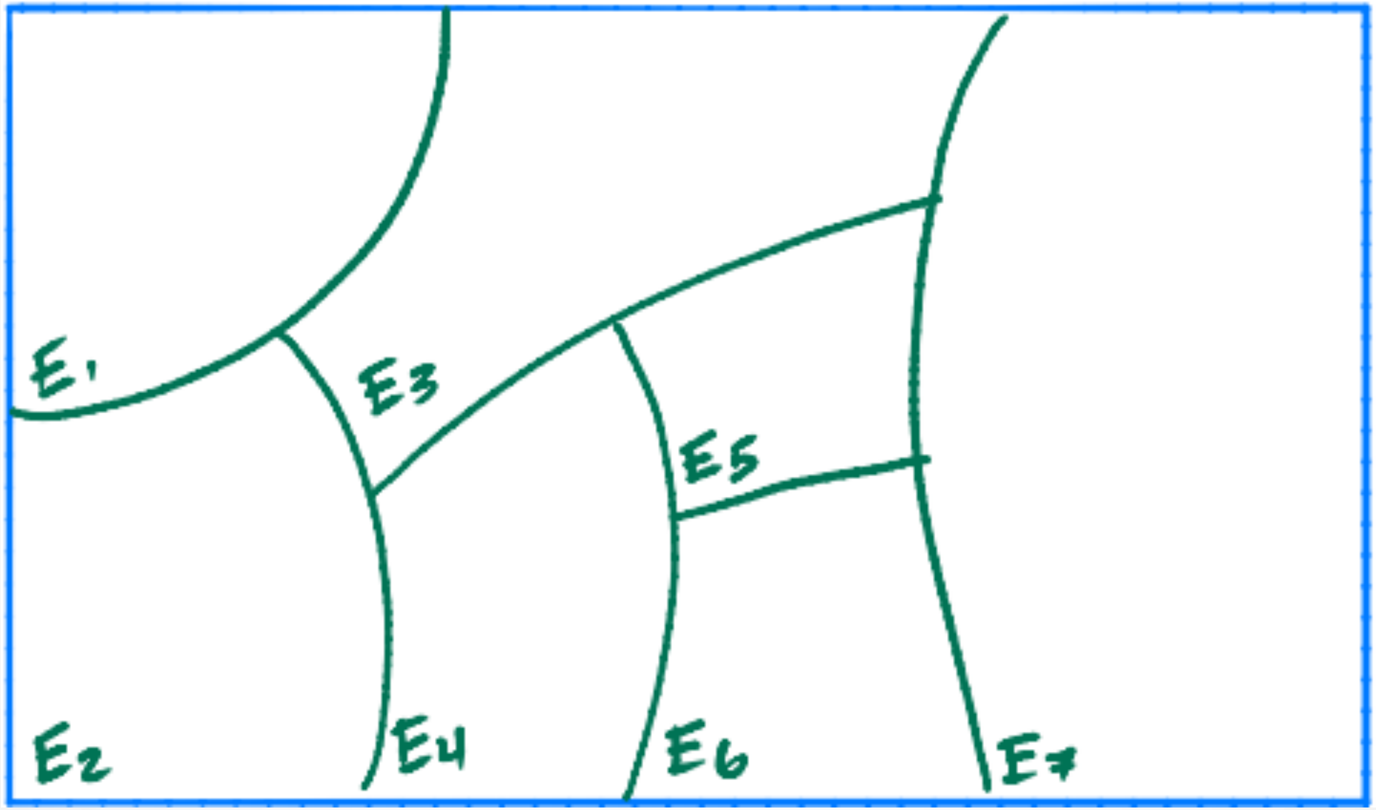
P(Late) = P(Late AND Walk) + P(Late AND Bike) + P(Late AND Drive)

P(Late) = P(Late|Walk)*P(Walk) + P(Late|Bike)*P(Bike)
+P(Late|Drive)*P(Drive)

Partitions

- A set of events $E_1, E_2, \ldots, E_k$ is a **partition** of S if
  - $P(E_i \cap E_j) = 0$ for all i,j
  - $P(E_1) + P(E_2) + \ldots + P(E_k) = 1$

Partitions



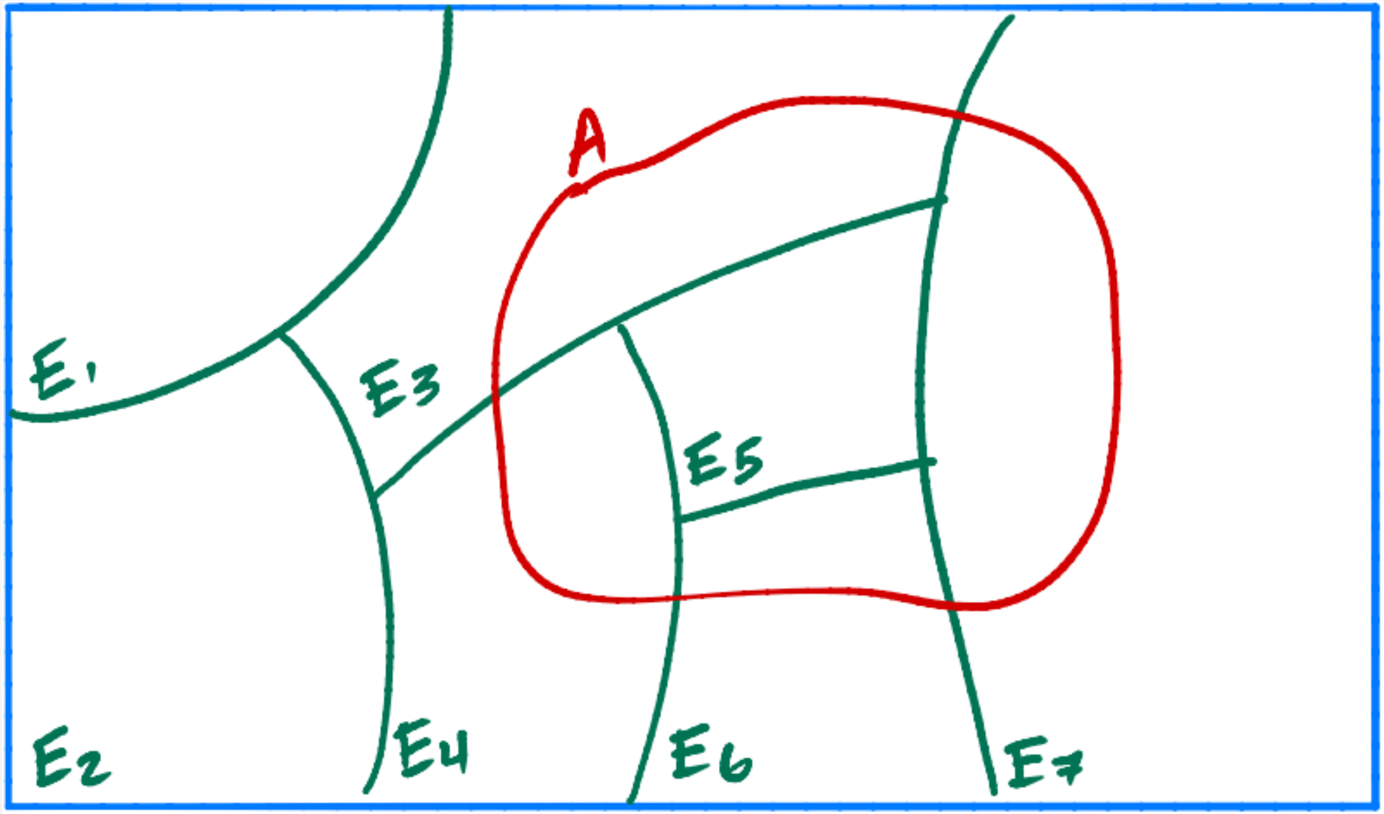
Law of Total Probability

- If A is an event and $E_1, E_2, \ldots, E_k$ is a **partition** of S, then

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + \ldots + P(A \cap E_k)$$

$$= \sum_{i=1}^{k} P(A \cap E_i)$$

Law of Total Probability



Law of Total Probability

- If A is an event and E_1, E_2, …, E_k is a **partition** of S, then

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + \ldots + P(A \cap E_k)$$

$$= \sum_{i=1}^{k} P(A \cap E_i)$$

- Written another way,

$$P(A) = P(A \mid E_1) \cdot P(E_1) + \ldots + P(A \mid E_k) \cdot P(E_k)$$

$$= \sum_{i=1}^{k} P(A \mid E_i) \cdot P(E_i)$$

Getting to Campus

	Late	Not Late
Walk	6%	24%
Bike	3%	7%
Drive	36%	24%

Suppose someone is late. What is the probability that they walked? Choose the best answer.

A. Close to 5%
B. Close to 15%
C. Close to 30%
D. Close to 40%

Getting to Campus

- Suppose all you know is
  - $P(Late) = 45\%$
  - $P(Walk) = 30\%$
  - $P(Late|Walk) = 20\%$
- Can you still find $P(Walk|Late)$?

Bayes' Theorem

Bayes' Theorem follows from the multiplication rule, or conditional probability.

$$P(A) * P(B|A) = P(A \text{ and } B) = P(B) * P(A|B)$$

Bayes' Theorem:

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

Bayes' Theorem

Bayes’ Theorem follows from the multiplication rule, or conditional probability.

$$P(A) * P(B|A) = P(A \text{ and } B) = P(B) * P(A|B)$$

Bayes’ Theorem:

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

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

not B

Bayes' Theorem

For hypothesis H and evidence (data) E

$$P(H \mid E) = \frac{P(E|H)}{P(E)}$$

- $P(H)$ - prior, initial probability before E is observed
- $P(H|E)$ - posterior, probability of H after E is observed
- $P(E|H)$ - likelihood, probability of E if the hypothesis is true
- $P(E)$ - marginal, probability of E regardless of H

The likelihood function is a function of E, while the posterior probability is a function of H.

Bayes' Theorem: Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\sim H)P(\sim H)}$$

A manufacturer claims that its drug test will **detect steroid use 95% of the time**. What the company does not tell you is that 15% of all steroid-free individuals also test positive (the false positive rate). 10% of the Tour de France bike racers use steroids. Your favorite cyclist just tested positive. What’s the probability that he used steroids?

What is your first guess?

A. Close to 95%
B. Close to 85%
C. Close to 40%
D. Close to 15%

Bayes' Theorem: Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\sim H)P(\sim H)}$$

A manufacturer claims that its drug test will **detect steroid use 95% of the time**. What the company does not tell you is that 15% of all steroid-free individuals also test positive (the false positive rate). 10% of the Tour de France bike racers use steroids. Your favorite cyclist just tested positive. What's the probability that he used steroids?

Now, calculate it and choose the best answer.

A. Close to 95%
B. Close to 85%
C. Close to 40%
D. Close to 15%

Bayes' Theorem: Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\sim H)P(\sim H)}$$

A manufacturer claims that its drug test will **detect steroid use 95% of the time.** What the company does not tell you is that 15% of all steroid-free individuals also test positive (the false positive rate). 10% of the Tour de France bike racers use steroids. Your favorite cyclist just tested positive. What's the probability that he used steroids?

Solution:

H: used steroids

E: tested positive

Bayes' Theorem: Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\sim H)P(\sim H)}$$

A manufacturer claims that its drug test will **detect steroid use 95% of the time.** What the company does not tell you is that 15% of all steroid-free individuals also test positive (the false positive rate). 10% of the Tour de France bike racers use steroids. Your favorite cyclist just tested positive. What's the probability that he used steroids?

Solution:

H: used steroids

E: tested positive

Despite manufacturer's claims, only **41% chance** that cyclist used steroids.

Bayes' Theorem: Example

Example

- 1% of people have a certain genetic defect
- 90% of tests accurately detect the gene (true positives).
- 7% of the tests are false positives.

If Olaf gets a positive test result, what are the odds he actually has the genetic defect?

Bayes' Theorem: Example

- Hypothesis: Olaf has the gene, $P(H) =$
- Evidence: Olaf got a positive test result, $P(E)$
- True positive: Probability of positive test result if someone has the gene $P(E|H) =$
- False positive: Probability of positive test result if someone doesn't have the gene $P(E|\overline{H}) =$

Bayes' Theorem: Example

Calculate

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

The probability that Olaf has the gene is only ________ despite the positive test result!

Bayes' Theorem: Example

What happens if there are less false positives?
Consider $P(E|\overline{H}) = 0.02$:

The probability that Olaf has the gene is now ___________.

Bayes' Theorem: Example

What happens if there are more true positives?
Consider $P(E|H) = 0.95$:

Improving the accuracy of true positives raised the probability that Olaf has the gene to ____.

Preview: Bayes' Theorem for Classification

Bayes' Theorem is very useful for classification problems, where we want to predict a class based on some features.

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

ain class

A = having certain features

$$P(\text{class}|\text{features}) = \frac{P(\text{features}|\text{class}) * P(\text{class})}{P(\text{features})}$$

Summary

- When a set of events partitions the sample space, the law of total probability applies.

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + \ldots + P(A \cap E_k)$$
$$= \sum_{i=1}^{k} P(A \cap E_i)$$

- Bayes Theorem says how to express P(B|A) in terms of P(A|B).
- **Next time:** independence and conditional independence