DSC 40A

Theoretical Foundations of Data Science I

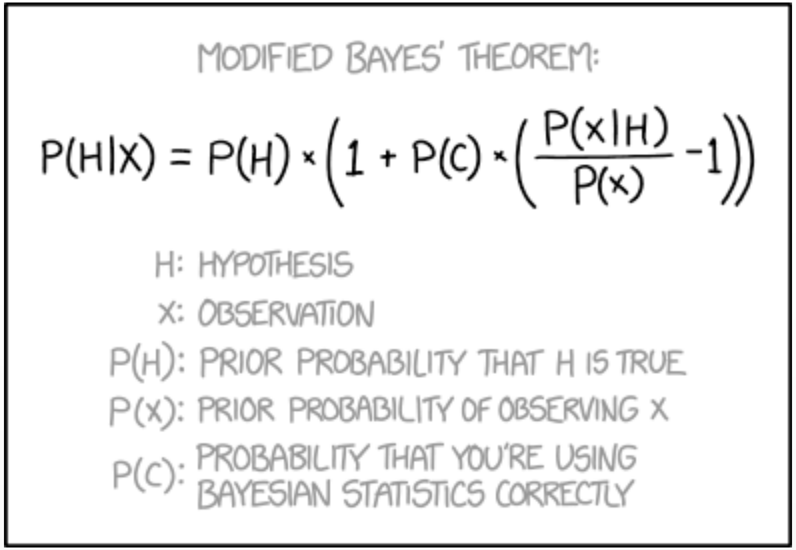
Question

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.

Agenda

- Bayes Theorem
- Naïve Bayes Classifier



Source: xkcd

Bayes Theorem

Last Week

- We defined Bayes' Theorem:

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

- Bayes' Theorem describes how to update the probability of one event given that another has occurred.

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:

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

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

law of total probability

not B

Bayes' Theorem

For hypothesis H and evidence (data) E

$$P(H \mid E) = \frac{P(E|H)\,P(H)}{P(E)} = \frac{P(E|H)\cdot P(H)}{P(E|H)\,P(H) + P(E|\bar{H})\cdot P(\bar{H})}$$

- $P(H)$ - 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

$$P(E|\bar{H}) \neq P(\bar{E}|H)$$

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%
70% B. Close to 85%
C. Close to 40%
D. Close to 15%

Bayes' Theorem: Example

$\bar{H}$, not H

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

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%

E - positive drug test

H - uses steroids

$\bar{H}$ - doesn't use steroids

Bayes' Theorem: Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\sim H)P(\sim H)} = \frac{0.95 \cdot 0.1}{0.95 \cdot 0.1 + 0.5 \cdot 0.9} \approx 0.41$$

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:

$P(E|H) = 95\%$ (TP)

H: used steroids

$P(E|\bar{H}) = 15\%$ (FP)

$\Longrightarrow P(H|E) = 41\%$

E: tested positive

$P(H) = 10\%$ $\quad$ $P(\bar{H}) = 90\%$

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

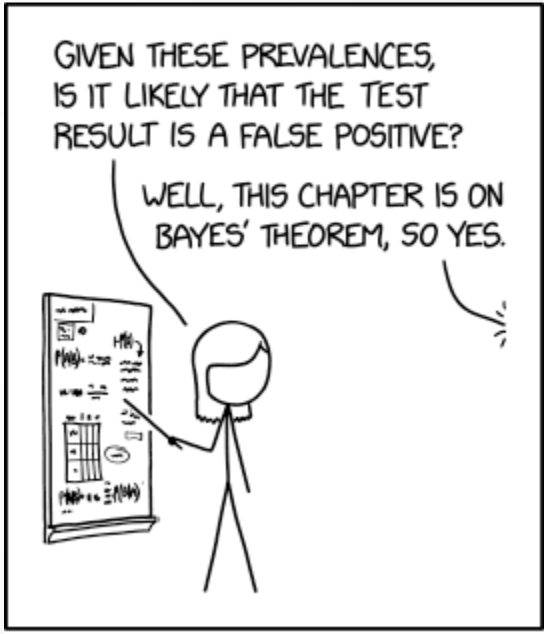
Example

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

H – has genetic disorder

E – positive test result

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



SOMETIMES, IF YOU UNDERSTAND BAYES' THEOREM WELL ENOUGH, YOU DON'T NEED IT.

Bayes' Theorem: Example

- Hypothesis: Olaf has the gene, $P(H) = 0.01$
- Evidence: Olaf got a positive test result, $P(E)$
- True positive: Probability of positive test result if someone has the gene $P(E|H) = 0.9$
- False positive: Probability of positive test result if someone doesn't have the gene $P(E|\bar{H}) = 0.07$

Bayes' Theorem: Example

Calculate

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

$$= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.07 \cdot \underbrace{0.99}_{1-0.01}} = \ldots = 0.115$$

The probability that Olaf has the gene is only 11.5% despite the positive test result!

Bayes' Theorem: Example

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

$$P(H|E) = \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + \boxed{0.02 \cdot 0.99}} = \ldots \approx 31\%$$

The probability that Olaf has the gene is now 31 %.

Bayes' Theorem: Example

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

$$P(H|E) = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.11 + 0.07 \cdot 0.99} = \ldots = 0.12$$

Improving the accuracy of true positives raised the probability that Olaf has the gene to 12%.

The Monty Hall Problem

The Monty Hall Problem

You're in a game show.
Behind one door is a car.
Behind the others, goats.

You pick one of three doors,
say #1.

The host, Monty Hall, opens one door, revealing…a goat!

The Monty Hall Problem

He then gives you the option of switching doors or sticking with your original choice.

The question is: **should you switch?**

The Monty Hall Problem

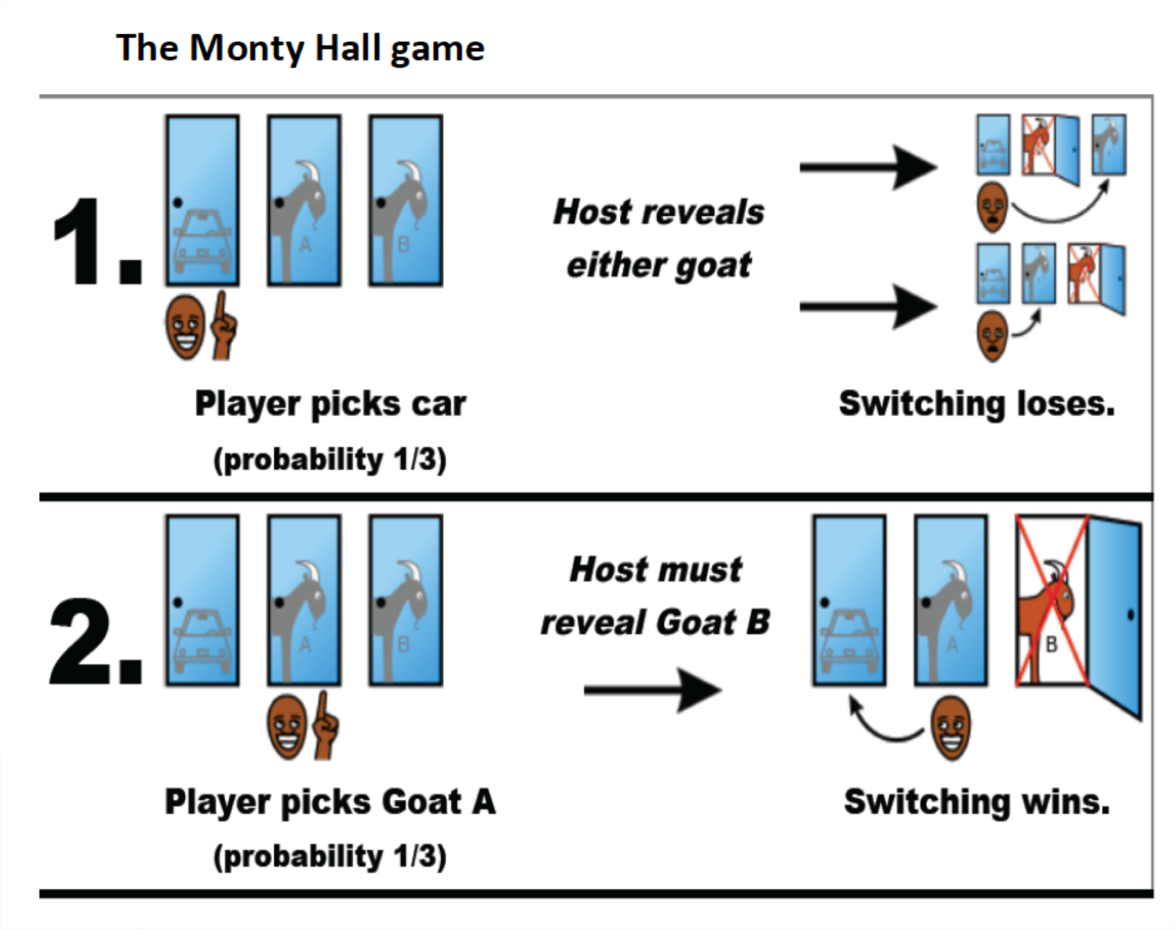
$\{1, 2, 3\}$ - door numbers

C - door number with a car

S - door number player selected

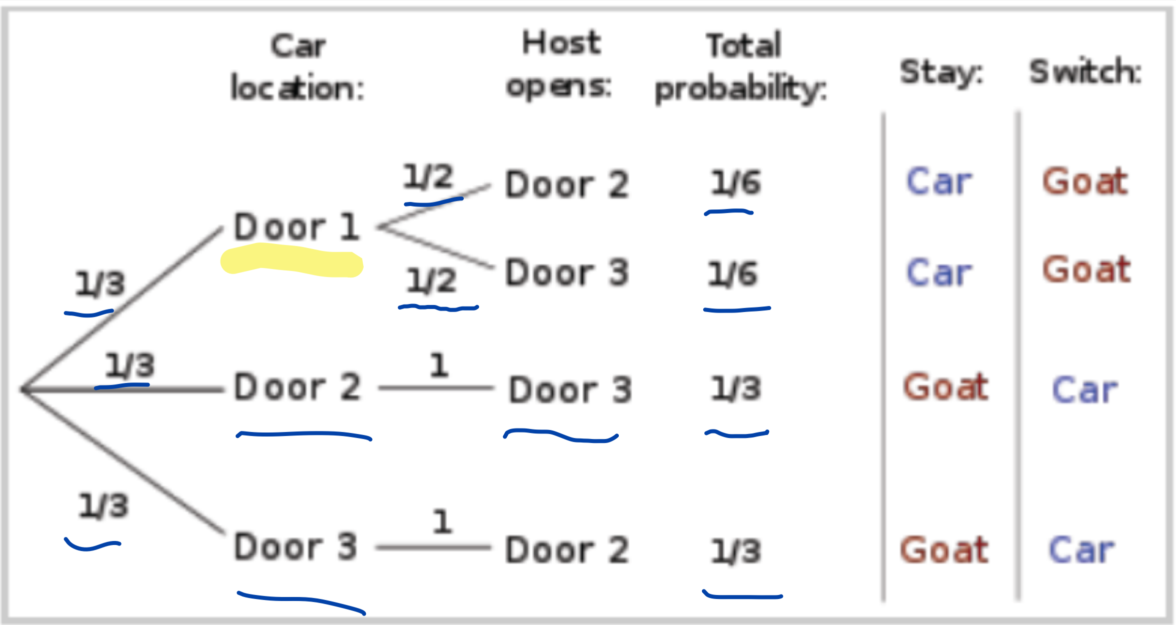
H - door number host opened

The Monty Hall Problem



The Monty Hall Problem

selected door #1



Tree showing the probability of every possible outcome if the player initially picks Door 1

The Monty Hall Problem

$S=1 \qquad H=2$

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

$$\rightarrow P(A \cap B) = P(A|B)\,P(B) = P(B|A) \cdot P(A)$$

$$P(C=1 \mid H=2, S=1)$$

$$P(C=3 \mid H=2, S=1)$$

$$P(C=1 \mid H=2) = \frac{P(C=1 \cap H=2)}{P(H=2)}$$

can open 2 or 3

$$P(C=1, H=2) = P(H=2 \mid C=1) \cdot P(C=1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

The Monty Hall Problem

$$P(C=3 \mid H=2) = \frac{P(C=3 \cap H=2)}{P(H=2)}$$

has to open door 2

$$P(C=3 \cap H=2) = P(H=2 \mid C=3)\, P(C=3) = 1 \cdot \frac{1}{3} = \frac{1}{3}$$

$$P(H=2) = P(H=2, C=1) + \underline{P(H=2, C=2)} + P(H=2, C=3)$$

$$= \frac{1}{6} + 0 + \frac{1}{3} = \frac{1}{2}$$

conditioned on $S=1$

$$P(C=1 \mid H=2) = \frac{1/6}{1/2} = \frac{1}{3} < P(C=3 \mid H=2) = \frac{1/3}{1/2} = \frac{2}{3}$$