

LING 473: Day 5

START THE RECORDING
Bayes Theorem

Announcements

- I will not be physically here August 8 & 10
- Lectures will be made available right before I go to sleep in Oslo
 - So, something like 2:30-3:00pm here. I'll send out an email when it goes up.
 - This means that Assignment 2 will have to be reviewed on August 10.
- There is a grader for the class
 - He'll be helping me with grades, but I'm the final arbiter.
 - Let me know if you have a question about your grade

Projects Generally

- Read instructions carefully
- Modeling language data vs full linguistic analysis
 - Your requested implementation may not be fully linguistically correct (that's ok!)
- Must run on Patas with all requested files turned in
- Make sure to log out of Patas when you're done by typing "logout"
 - You may encounter strange state problems if you are disconnected without logging out

Writing assignment

- Due September 5th , 2017
<http://courses.washington.edu/ling473/writing-assignment.html>
- Short Critical review of a paper from the computational linguistics literature
- Formatted according to ACL-2017 guidelines
 - <http://acl2017.org/calls/papers/>
- Any published journal or peer-reviewed paper on a comp. ling. topic is acceptable

Review: Conditional Probability

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

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

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

joint probability = conditional probability × marginal probability
(or “prior” probability)

Independent random variables

Random variables A and B are independent *iff*

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

Recall conditional probability:

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

This means that, if A and B are independent,

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

$P(A \cap B)$ $P(A, B)$ $P(AB)$
Reminder: these are three notations for the same thing:
the **joint probability** of A and B . That is, that both events occur in a single trial

Conditional independence

A and B are **independent** *iff*

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

A and B are **conditionally independent given K** *iff*

$$P(A \cap B|K) = P(A|K)P(B|K)$$



Just as with conditional probability, K constrains the sample space. Conditional independence means that A and B are independent if we know that K has occurred.

Conditional independence

A and B are **conditionally independent given K** iff

$$P(A \cap B|K) = P(A|K)P(B|K)$$



Given that K has occurred, knowing that B has occurred gives us no additional information about the probability of A (and vice-versa)

Q: Does this imply that A and B are independent?

A: No. A and B could be either independent or dependent in the absence of knowledge about K

Conditional independence

$$P(A \cap B|K) = P(A|K)P(B|K)$$

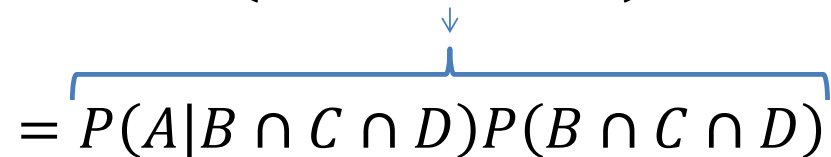
Two events (A and B) are **conditionally independent** given a third event (K) if their probabilities conditioned on K are independent. The following will also be true:

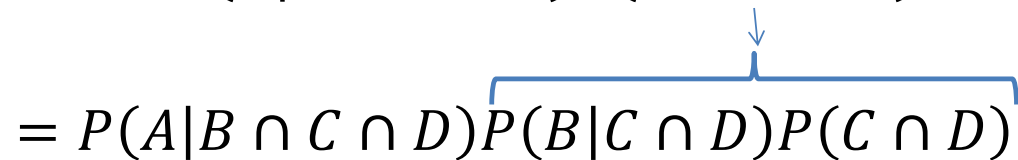
$$\begin{aligned}P(A|B \cap K) &= P(A|K) \\ P(B|A \cap K) &= P(B|K)\end{aligned}$$

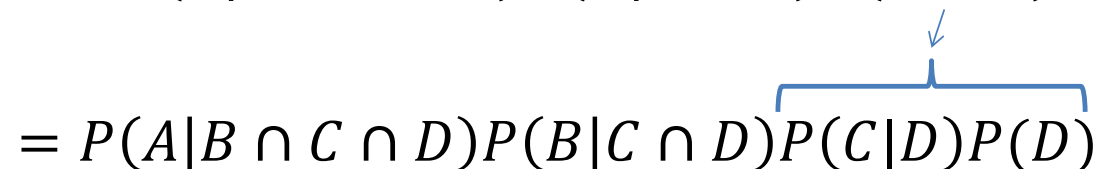
Chain rule

- This can be extended $P(A \cap B) = P(A|B)P(B)$

$$P(A \cap B \cap C \cap D)$$


$$= P(A|B \cap C \cap D)P(B \cap C \cap D)$$


$$= P(A|B \cap C \cap D)P(B|C \cap D)P(C \cap D)$$


$$= P(A|B \cap C \cap D)P(B|C \cap D)P(C|D)P(D)$$

etc... This is called the **chain rule**

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

$$\begin{aligned} & P(ABCDE) \\ & \quad \underbrace{\hspace{1.5cm}} \\ &= P(A|BCDE)P(BCDE) \\ & \quad \underbrace{\hspace{1.5cm}} \\ &= P(A|BCDE)P(B|CDE)P(CDE) \\ & \quad \underbrace{\hspace{1.5cm}} \\ &= P(A|BCDE)P(B|CDE)P(C|DE)P(DE) \\ & \quad \underbrace{\hspace{1.5cm}} \\ &= P(A|BCDE)P(B|CDE)P(C|DE)P(D|E)P(E) \end{aligned}$$

Chain rule

$$P(X_1 = x_1, \dots, X_n = x_n) = \\ P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_1 = x_1) \times P(X_{n-1} = x_{n-1}, \dots, X_1 = x_1)$$

$$P(A, B, C, D) = P(A \cap B \cap C \cap D) \\ = P(A|B, C, D)P(B|C, D)P(C|D)P(D)$$

(The “given” notation ‘|’ has lowest precedence)

Chain Rule and Bayes' Rule

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

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

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

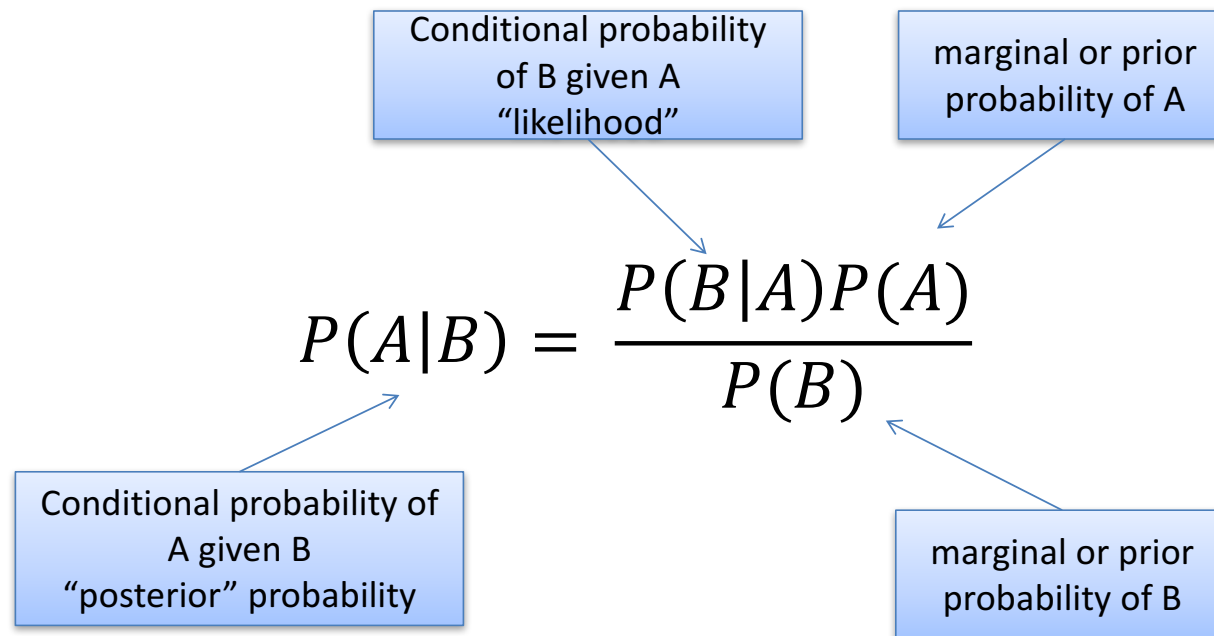
$$P(B|A)P(A) = P(B \cap A)$$

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

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

Bayes Theorem

Rev. Thomas Bayes (1701-1761)



Bayes Theorem

- Relates hypothesis to observation (evidence or prior knowledge)

The diagram illustrates Bayes Theorem using two light blue circles. The left circle is labeled 'hypothesis' and contains the expression $P(H|E)$. The right circle is labeled 'observation' and contains the fraction $\frac{P(E|H)P(H)}{P(E)}$. An equals sign is placed between the two circles, indicating that the posterior probability of the hypothesis given the evidence is equal to the product of the likelihood and the prior probability, divided by the total probability of the evidence.

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

Bayes Theorem

- Expresses one probability in terms of another
- $P(A|B)$ depends on B , but also $P(A)$ and $P(B)$ in the general population.
- When might Bayes theorem be useful?

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

- Medical test vs condition! $P(\text{condition}|\text{test})$ is hard to know, but $P(\text{test}|\text{condition})$ is easier!

Bayes Theorem

		actual condition A	
		yes $P(A)$	no
test result B	positive $P(B)$	true positive	false positive
	negative	false negative	true negative

- We can empirically discover $P(B|A=true)$ given a population of people with a condition. Same for $P(B|A=false)$.
- We can use a sample population to get $P(B=true)$ and $P(B=false)$.
- We can use other sources to estimate $P(A)$ in the general population.
- Now we have enough to generate $P(A|B)$, the probability of an actual condition given a test result.

Example 1

A gambler has two coins in his pocket, one fair coin and one two-headed one.

a. He selects one at random and flips it. It comes up heads. What is the probability that is the fair coin?



Example 1

- Assuming an equal chance of picking from the pocket:

$$P(F) = P(F^C) = \frac{1}{2}$$

- Probability of obtaining heads from the fair coin:

$$P(H | F) = \frac{1}{2}$$

- Probability of obtaining heads from the two-headed coin:

$$P(H | F^C) = 1$$

Example 1

- Overall prior probability of flipping heads:

$$\begin{aligned} P(H) &= P(H|F)P(F) + P(H|F^C)P(F^C) \\ &= \frac{1}{2} \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{3}{4} \end{aligned}$$

- Probability of having the fair coin given heads:

$$P(F|H) = \frac{P(H|F)P(F)}{P(H)} = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{3}{4}} = \frac{1}{3}$$

Example 1

b. He now flips the same coin a second time, and it again comes up heads. What is the probability that it is the fair coin?



Example 1

- Probability of two heads given the fair coin

$$P(H,H|F) = \frac{1}{4} \text{ (it is one of 4 outcomes)}$$

- Probability of two heads given the double-headed coin

$$P(H,H|F^C) = 1$$

- Overall probability of two heads:

$$\begin{aligned} & P(H,H|F)P(F) + P(H,H|F^C)P(F^C) \\ &= \frac{1}{4} \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{5}{8} \end{aligned}$$

Example 1

- Probability of having selected the fair coin given the observation {H,H}

$$P(F|H, H) = \frac{P(H, H|F)P(F)}{P(H)} = \frac{\frac{1}{4} \times \frac{1}{2}}{\frac{5}{8}} = \frac{1}{5}$$

Example 1

c. Suppose he flips the coin a third time, and it comes up tails.
What is the probability that it is the fair coin?



Example 1

- 1.0
- The double headed coin can't come up tails.

Example 2

The Monty Hall problem

- There are 3 doors: A, B, & C. One of them has a prize behind it.
- You choose door A. The host knows where the prize is and reveals that door B does not have the prize. The host asks if you want to switch.
- Should you switch?



Example 2

- Original chance of choosing the prize: $\frac{1}{3}$
- Event B: {door B is revealed}
- Random variable Z: { door with the prize }
- $$P(Z = a|B) = \frac{P(B|Z = a)P(Z=a)}{P(B)} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$
- $$P(Z = b|B) = \frac{P(B|Z = b)P(Z=b)}{P(B)} = \frac{0 \times \frac{1}{3}}{\frac{1}{2}} = 0$$
- $$P(Z = c|B) = \frac{P(B|Z = c)P(Z=c)}{P(B)} = \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Example 2

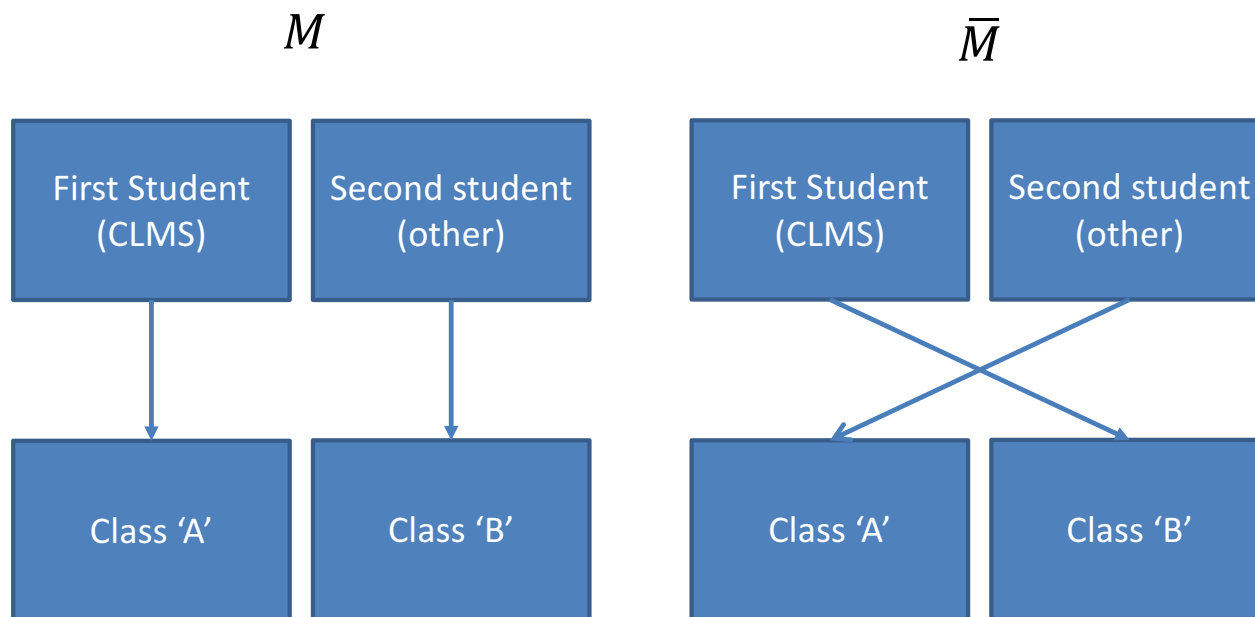
- Assuming the host always selects at random when he can, it is always better to switch your choice.



Example 3

- Class A has 15 PhD students, 10 CLMS students, and 5 students from other majors.
- Class B has 5 PhD students, 10 CLMS students, and 15 students from other majors.
- One student from each class is chosen at random. The first is a CLMS student and the second is from another major. What is the probability the CLMS student is from Class A?

2 possibilities for the actual matchup



Example 3

- Prior probabilities:


$$P(M) = 0.5$$

$$P(\bar{M}) = 1 - P(M) = 0.5$$


(either match-up is equally likely)

We observe the sequence: (CLMS, other)

Probability of seeing this given M :


$$P((\text{CLMS, other}) | M) = \frac{10}{30} \times \frac{15}{30} = \frac{1}{6}$$

Probability of seeing this given \bar{M} :


$$P((\text{CLMS, other}) | \bar{M}) = \frac{10}{30} \times \frac{5}{30} = \frac{1}{18}$$

Example 3

- Overall prior probability of seeing (CLMS, other):

$$P(M) \times P(\text{(CLMS, other)} | M) + P(\bar{M}) \times P(\text{(CLMS, other)} | \bar{M})$$

$$= \frac{1}{2} \times \frac{1}{6} + \frac{1}{2} \times \frac{1}{18}$$

$$P(\text{(CLMS, other)}) = \frac{1}{9}$$

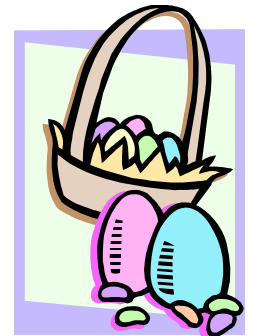
Example 3

- $$P(M|(CLMS, other)) = \frac{P((CLMS, other)|M) \times P(M)}{P((CLMS, other))}$$
- $$= \frac{\frac{1}{6} \times \frac{1}{2}}{\frac{1}{9}} = \frac{3}{4}$$

Example 4

A basket contains many small plastic eggs, some painted red and some are painted blue.

40% of the eggs in the bin contain pearls



30% of eggs containing pearls are painted blue, and 10% of eggs containing nothing are painted blue.

What is the probability that a blue egg contains a pearl?

Example 4

- $P(\text{pearl}|\text{blue}) = \frac{P(\text{blue}|\text{pearl})P(\text{pearl})}{P(\text{blue})}$
- $P(\text{pearl}|\text{blue}) = \frac{0.3 \times 0.4}{P(\text{blue})}$
- $P(\text{blue}) = P(\text{blue}|\text{pearl})P(\text{pearl}) + P(\text{blue}|\overline{\text{pearl}})P(\overline{\text{pearl}})$
- $P(\text{blue}) = 0.3 \times 0.4 + 0.1 \times 0.6 = 0.18$
- $P(\text{pearl}|\text{blue}) = \frac{0.12}{0.18} = \frac{2}{3}$

Earthquakes & Burglaries

- You own a house in California with an alarm system. If your alarm goes off, one of your neighbors will call you.
- The alarm could go off because of an earthquake, or a burglary.
- $P(\text{burglary}) = 0.001$
- $P(\text{earthquake}) = 0.002$

Earthquakes & Burglaries

- $P(\text{alarm, burglary, earthquake}) =$

Burglary	Earthquake	$P(\text{Alarm} B, E)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

Earthquakes & Burglaries

- The alarm is going off. What is the most likely reason for it?
- $P(burglary|alarm) = \frac{P(alarm|burglary)P(burglary)}{P(alarm)}$
- $P(earthquake|alarm) = \frac{P(alarm|earthquake)P(earthquake)}{P(alarm)}$
- $P(nothing|alarm) = \frac{P(alarm|nothing)P(nothing)}{P(alarm)}$

Earthquakes & Burglaries

- $P(A) = P(A|B, E)P(B)P(E) + P(A|B, \bar{E})P(B)P(\bar{E}) + P(A|\bar{B}, E)P(\bar{B})P(E) + P(A|\bar{B}, \bar{E})P(\bar{B})P(\bar{E})$
- $= 0.95 \times 0.001 \times 0.002 + 0.94 \times 0.001 \times 0.998 + 0.29 \times 0.9999 \times 0.002 + 0.001 \times 0.9999 \times 0.998$
- $P(A) = 0.002516964$

Earthquakes & Burglaries

- $$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{(P(A|B,\bar{E})P(\bar{E}) + P(A|B,E)P(E))P(B)}{P(A)}$$
- $$\frac{(0.94 \times 0.998 + 0.95 \times 0.002)(0.001)}{0.002516964} = 0.3734737$$
- $$P(E|A) = \frac{P(A|E)P(E)}{P(A)} = \frac{(P(A|E,\bar{B})P(\bar{B}) + P(A|E,B)P(B))P(E)}{P(A)}$$
- $$\frac{(0.29 \times 0.999 + 0.95 \times 0.001)(0.002)}{0.002516964} = 0.2309607$$

Earthquakes & Burglaries

- $P(\bar{E}, \bar{B} | A) = \frac{P(A|\bar{E}, \bar{B})P(\bar{E})P(\bar{B})}{P(A)}$
- $\frac{0.001 \times 0.998 \times 0.999}{0.002516964} = 0.3937314$
- It's almost a toss-up between nothing and a burglary. Ask your neighbor if they can see anyone (and if the ground is shaking).

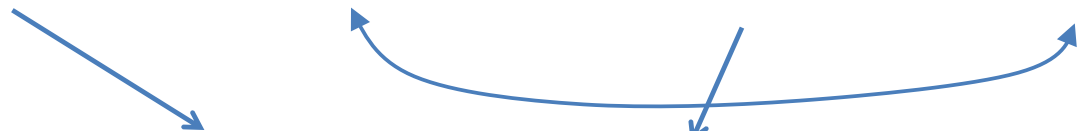
Review: Derivation of Bayes Theorem

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

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

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

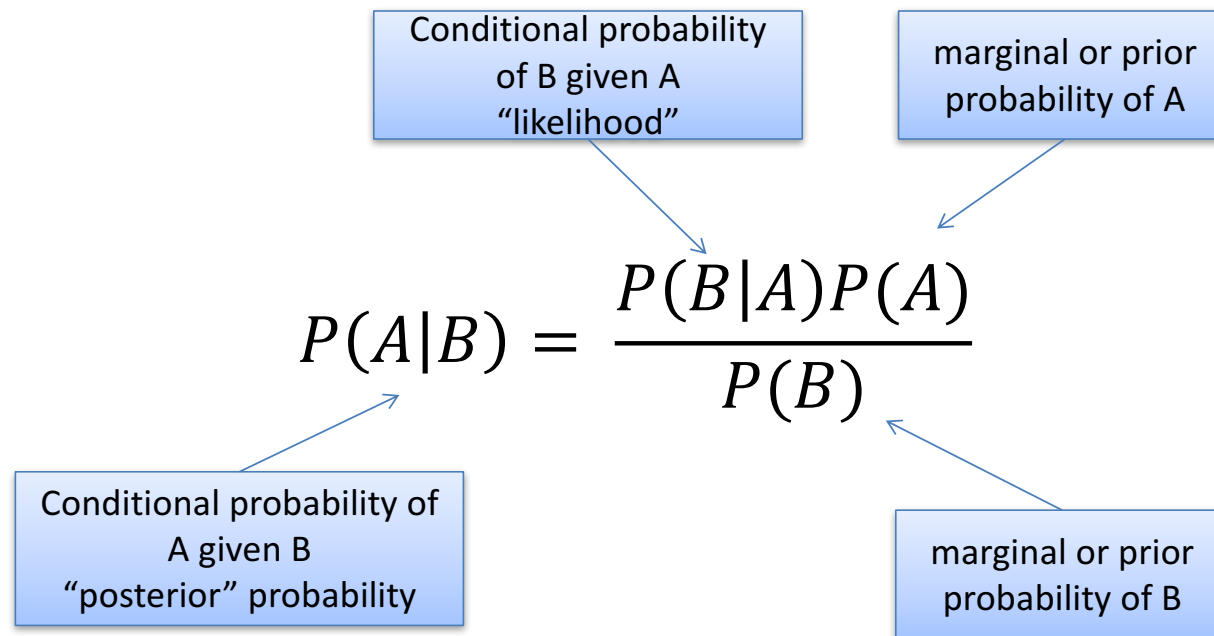
$$P(B|A)P(A) = P(B \cap A)$$


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

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

Bayes Theorem

Rev. Thomas Bayes (1701-1761)



Bayes Theorem

- Expresses one conditional probability in terms of its inverse
- $P(A|B)$ depends not only on B , but also on $P(A)$ and $P(B)$ in the general population

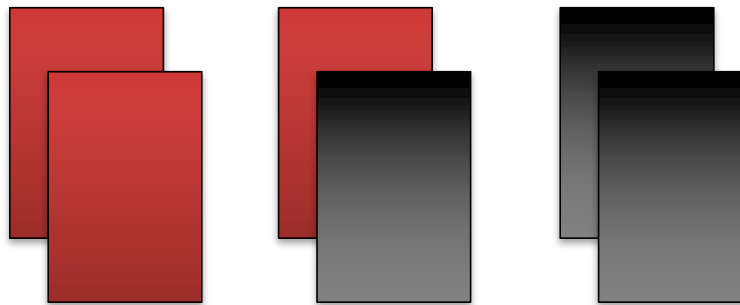
		actual condition A	
		yes $P(A)$	no
test result B	positive $P(B)$	true positive	false positive
	negative	false negative	true negative

Recipe for Bayes Theorem

- What you need:
 1. The probability of **actually satisfying** the criteria (regardless of 2)
 2. The probability of **testing positive** for the criteria (regardless of 1)
 3. And either:
 - a. the probability of **testing positive** given the **criteria is satisfied**
 - b. the probability of **satisfying the criteria** given the **test is positive**

Bayes theorem

- 3 cards:



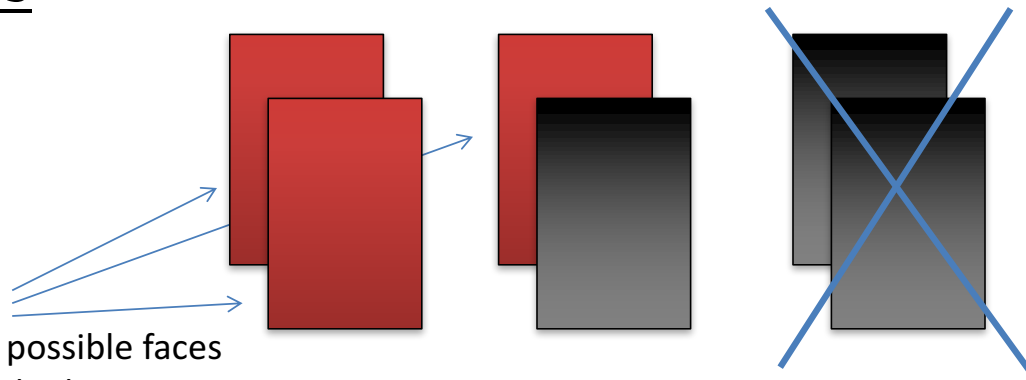
- We select a card at random and note that one side is red. What is the chance that it's the red-red card?

Bayes theorem

$$P((red, red)|R) = \frac{P(R|(red, red))P((red, red))}{P(R)}$$

$$= \frac{1 \times \frac{1}{3}}{\frac{3}{6}}$$

$$= \frac{2}{3}$$



For each of the 3 possible faces
that you could be looking at,
how many of them have red on
its *other* side?

Probability distributions

Assuming that a **random variable** exhibits a fixed, characteristic **probability distribution**, e.g.

$$\Omega = \{ a, b, c \}$$
$$P(X = x) = \begin{cases} 1/3, & \text{if } x = \{a\}; \\ 1/3, & \text{if } x = \{b\}; \\ 1/3, & \text{if } x = \{c\}; \end{cases}$$

allows us justify our intuition about events from last week:

$$A = \{a\}$$

$$A^c = \{ b, c \}$$

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

$$P(A^c) = \frac{|A^c|}{|\Omega|}$$

$$P(A) + P(A^c) = \frac{|A|}{|\Omega|} + \frac{|A^c|}{|\Omega|} = \frac{|A| + |A^c|}{|\Omega|} = \frac{|\Omega|}{|\Omega|} = 1$$

Probability distributions

- A random variable's probability distribution encapsulates both:
 - a characteristic type of “spread” or “shape” (distribution)
 - uniform
 - normal
 - etc.
 - the scaling and normalization factors that map between probabilities [0.0, 1.0] and the range of measurement values

This is why the capital letter subscript is (supposed to be) used: $P_X(X = x)$

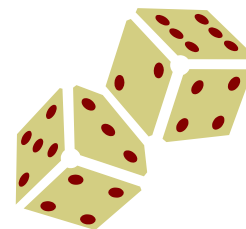
Uniform distribution

- Dividing the probability mass evenly between the values of a discrete random variable creates a uniform distribution

$$a = 1, b = 6$$

the mean μ is the average value

$$\mu = \frac{a + b}{2} = 3.5$$



Non-uniform distribution

(the, cat, in, the, hat)

$X = \{ \text{the word which is selected} \}$

$$P_X(X = \text{the}) = 0.4$$

$$P_X(X = \text{cat}) = 0.2$$

$$P_X(X = \text{in}) = 0.2$$

$$P_X(X = \text{hat}) = 0.2$$



$Y = \{ \text{the number of times } X=\text{the in 3 trials, with replacement} \}$

$$P_Y(Y = 0) = .6 \times .6 \times .6 = .216$$

$$P_Y(Y = 3) = .4 \times .4 \times .4 = .064$$

$$P_Y(Y = 1) = .4 \times .6 \times .6 \times \binom{3}{1} = .432$$

$$P_Y(Y = 2) = .4 \times .4 \times .6 \times \binom{3}{1} = .288$$

$$P_Y(Y \geq 2) = .352$$