Probability for Learning

- Probability for classification and modeling concepts.
- Bayesian probability
  - Notion of probability interpreted as partial belief
- Bayesian Estimation
  - Calculate the validity of a proposition
    - Based on prior estimate of its probability
    - and New relevant evidence

Bayes Theorem

- **Goal:** To determine the most probable hypothesis, given the data *D* plus any initial knowledge about the prior probabilities of the various hypotheses in *H*.

Bayes Theorem

Bayes Rule: $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h | D) = probability of h given D (posterior density)
- P(D | h) = probability of D given h (likelihood of D given h)

Dilemma at the movies

This person dropped their ticket in the hallway.

Do you call out

"Excuse me, ma'am!"

or

"Excuse me, sir!"

You have to make a guess.

Dilemma at the movies

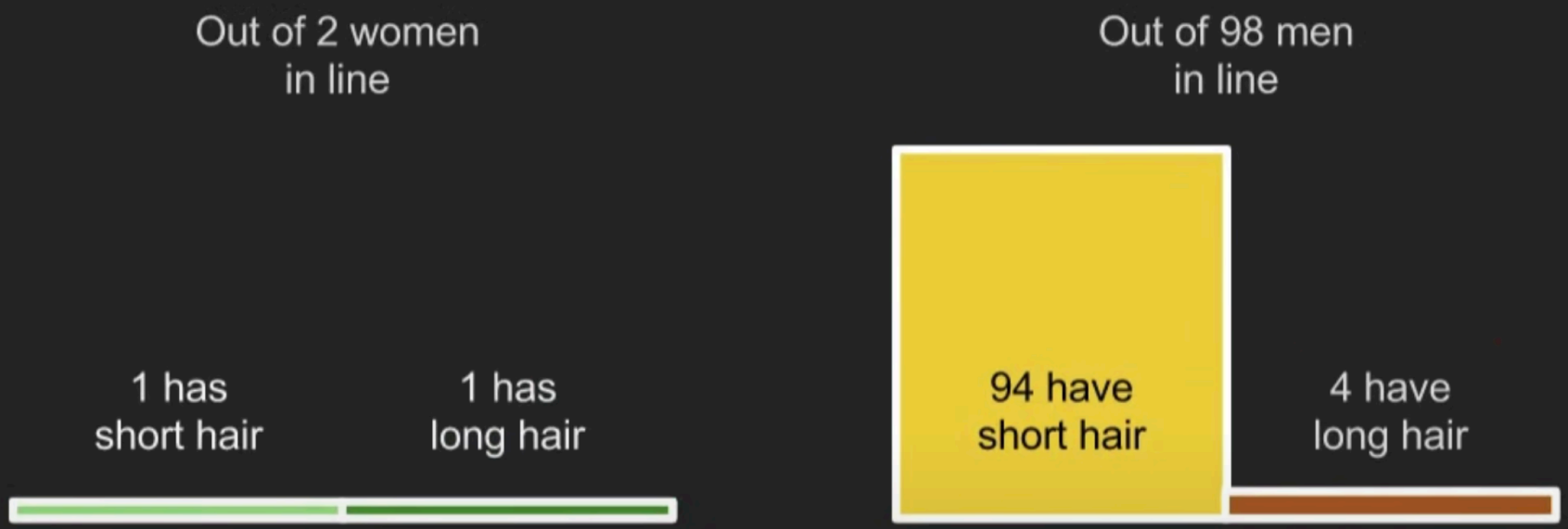
What if they're standing in line for the men's restroom?

Bayesian inference is a way to capture common sense.

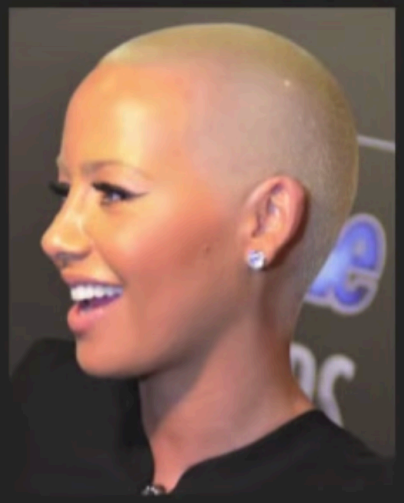
It helps you use what you know to make better guesses.

Put numbers to our dilemma

Put numbers to our dilemma



But there are 98 men and 2 women in line for the men's restroom.



Out of 100 people
at the movies

50 are women

50 are men

2 men have long hair

Out of 100 people
In line for the men's
restroom

2 are
women

98 are men

One
woman
has
short
hair

94 men
have
short hair

One
woman
has
long
hair

4 men have long hair

Translate to math

P(something) = # something / # everything

P(woman) = Probability that a person is a woman

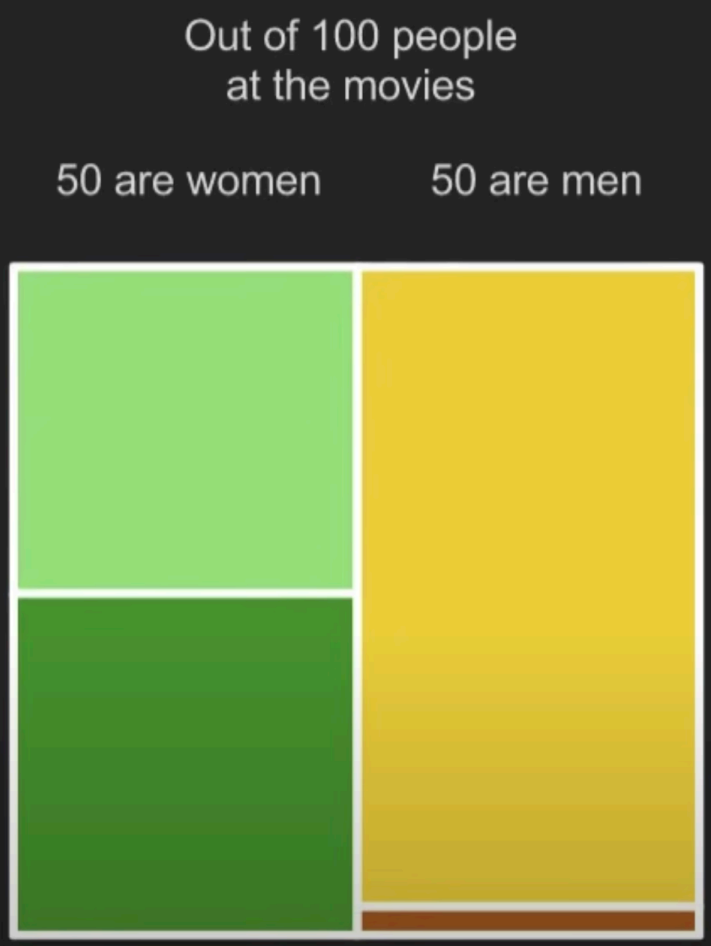
= # women / # people

= 50 / 100 = **.5**

P(man) = Probability that a person is a man

= # men / # people

= 50 / 100 = **.5**



Translate to math

P(something) = # something / # everything

P(woman) = Probability that a person is a woman

= # women / # people

= 2 / 100 = **.02**

P(man) = Probability that a person is a man

= # men / # people

= 98 / 100 = **.98**

Conditional probabilities

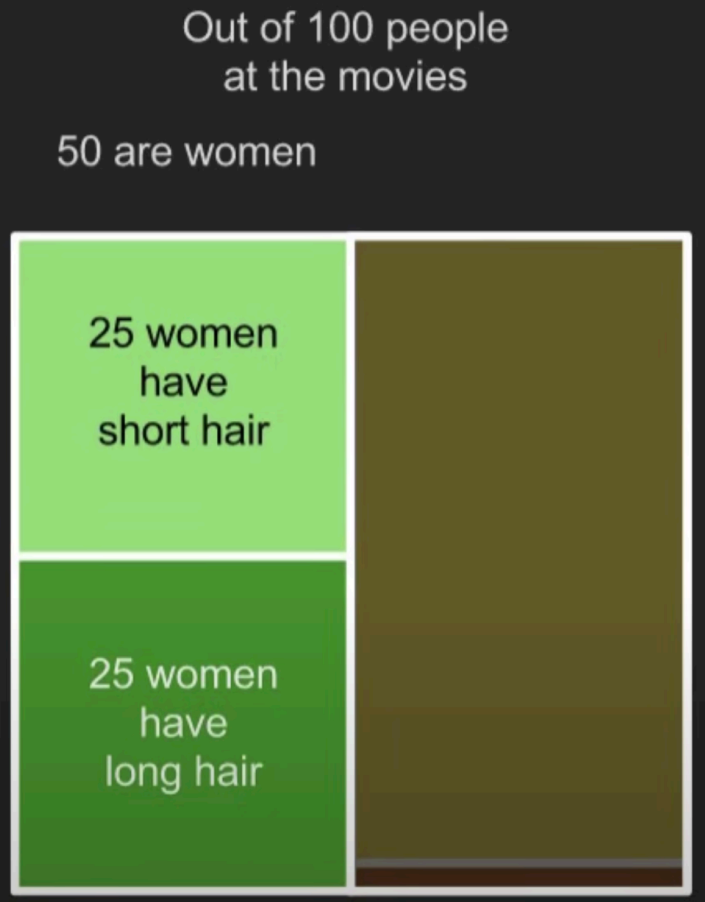
P(long hair | woman)

If I know that a person is a woman, what is the probability that person has long hair?

P(long hair | woman)

= # women with long hair / # women

= 25 / 50 = **.5**



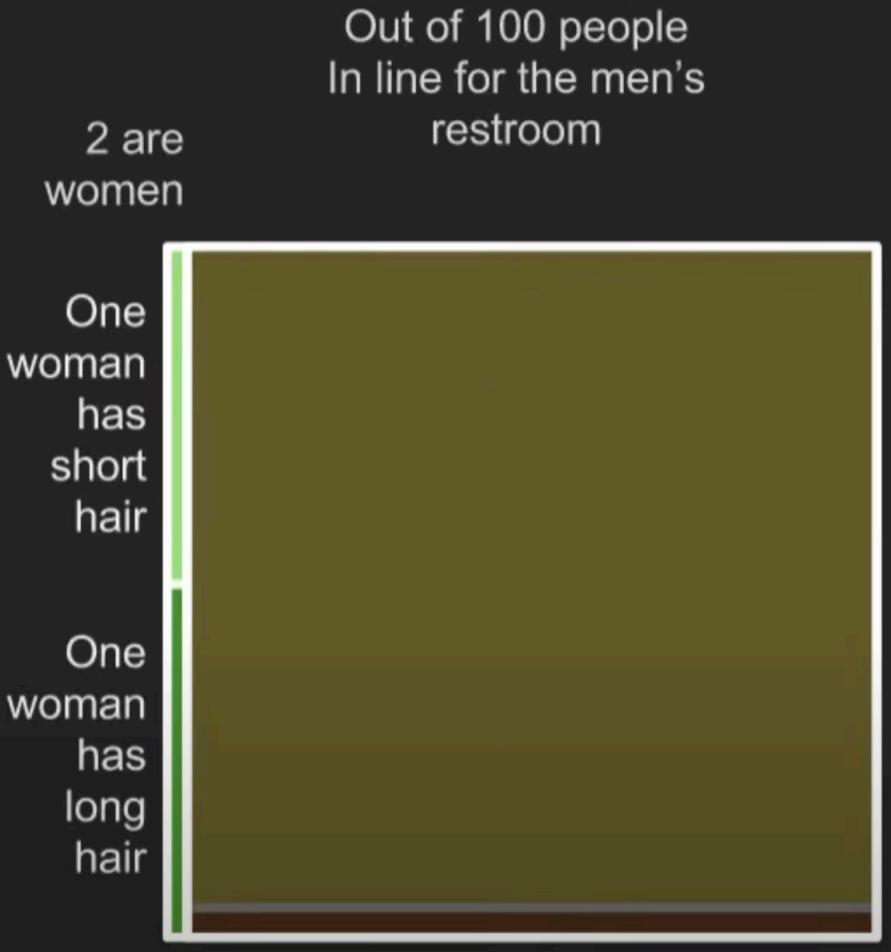
Conditional probabilities

This doesn't change when we consider people in line.

P(long hair | woman)

= # women with long hair / # women

= 1 / 2 = **.5**



Conditional probabilities

If I know that a person is a man, what is the probability that person has long hair?

P(long hair | man)

= # men with long hair / # men

= 2 / 50 = **.04**

Whether in line or not.

Conditional probabilities

P(A | B) is the probability of A, given B.

“If I know B is the case, what is the probability that A is also the case?”

P(A | B) is not the same as P(B | A).

P(cute | puppy) is not the same as P(puppy | cute)

If I know the thing I’m holding is a puppy, what is the probability that it is cute?

If I know the the thing I’m holding is cute, what is the probability that it is a puppy?

Joint probabilities

What is the probability that a person is both a woman and has short hair?

P(woman with short hair)

= P(woman) * P(short hair | woman)

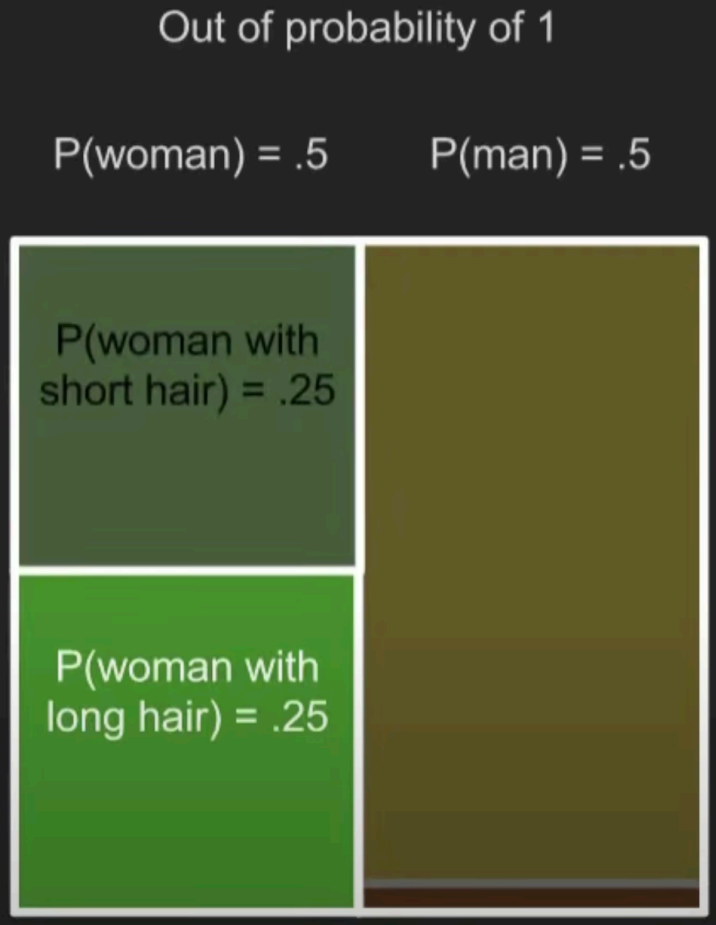
= .5 * .5 = **.25**

Joint probabilities

P(woman with long hair)

= P(woman) * P(long hair | woman)

= .5 * .5 = **.25**

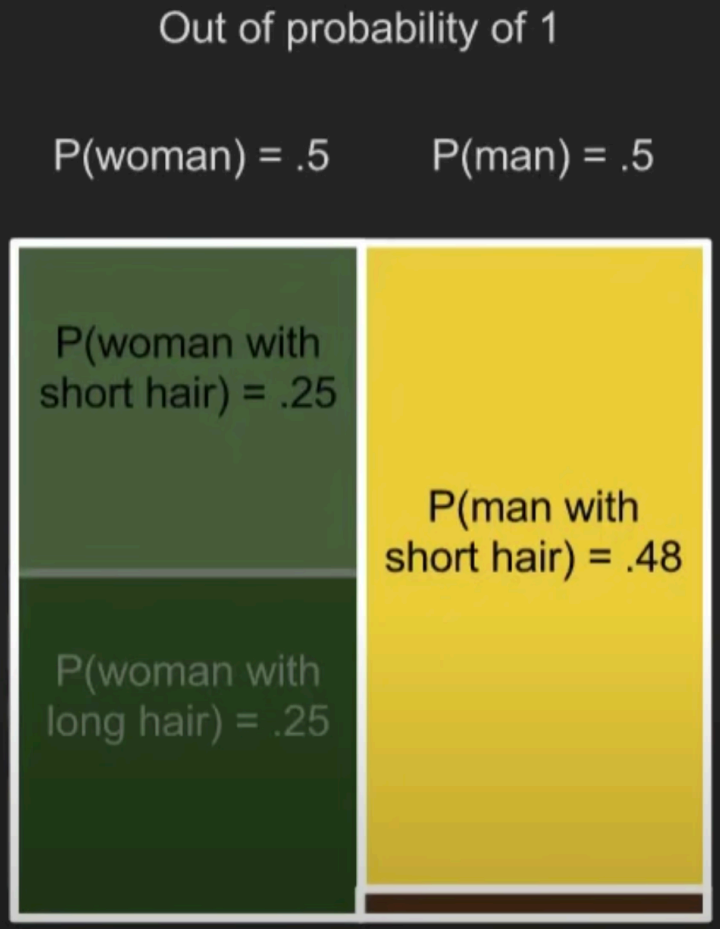


Joint probabilities

P(man with short hair)

= P(man) * P(short hair | man)

= .5 * .96 = **.48**

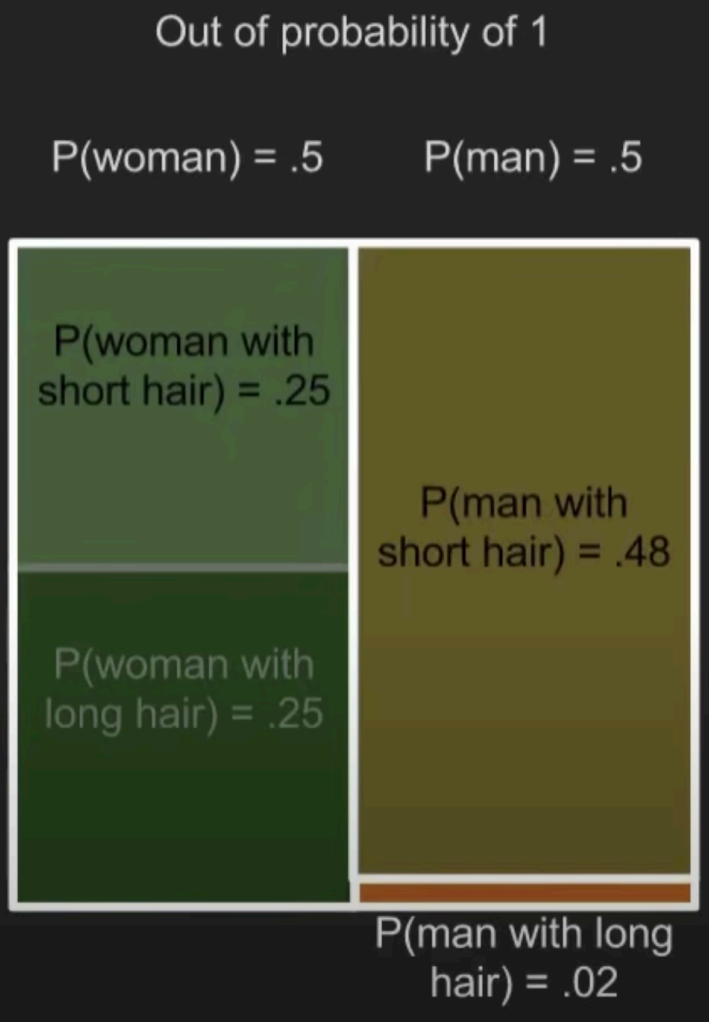


Joint probabilities

P(man with long hair)

= P(man) * P(long hair | man)

= .5 * .04 = **.02**



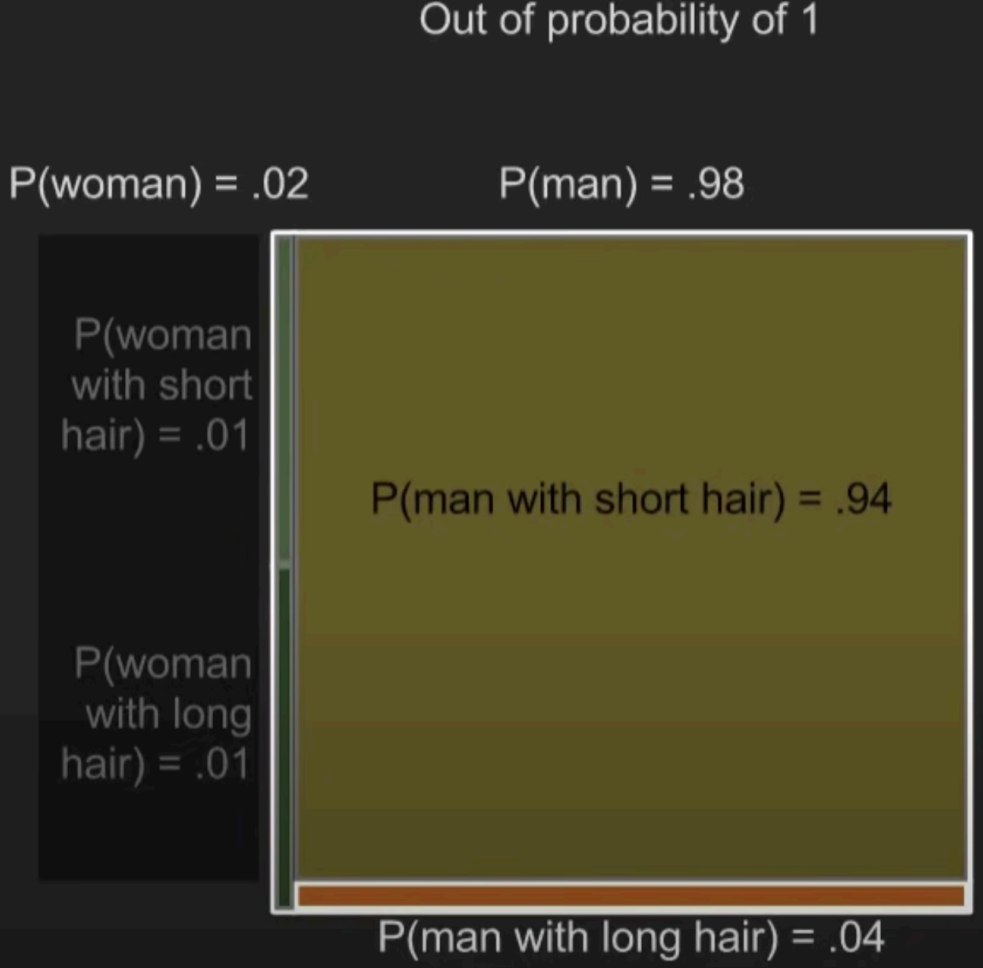
Joint probabilities

If P(man) = .98 and P(woman) = .02, then the answers change.

P(man with long hair)

= P(man) * P(long hair | man)

= .96 * .04 = **.04**



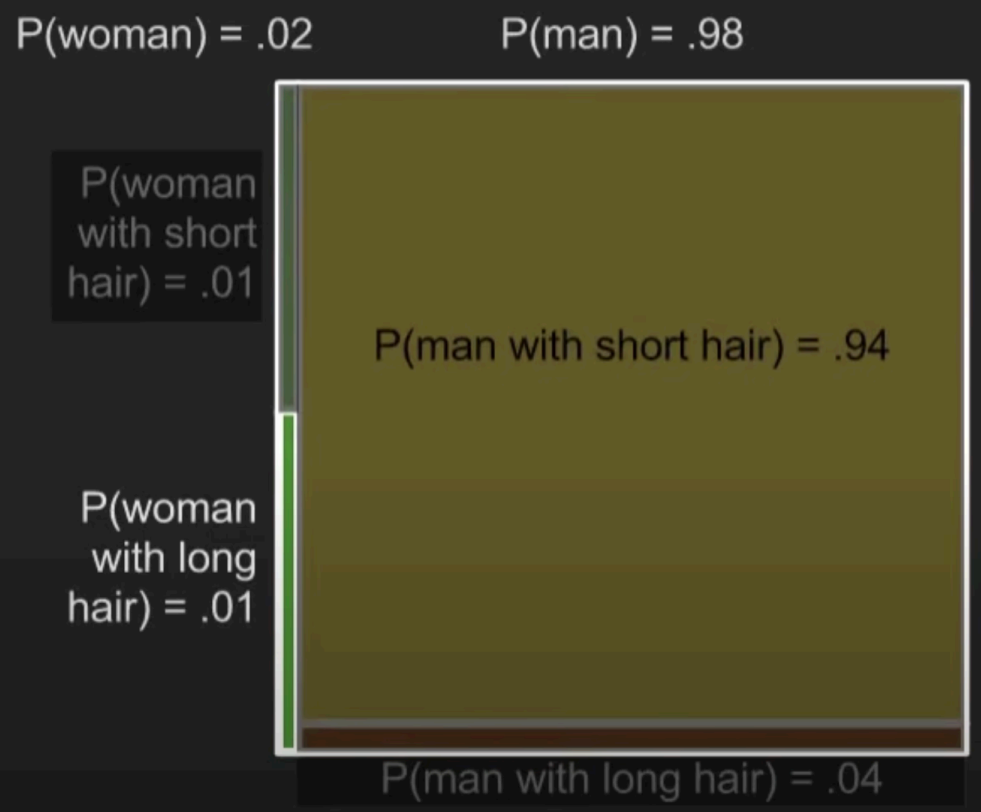
Joint probabilities

P(woman with long hair)

= P(woman) * P(long hair | woman)

= .02 * .5 = **.01**

Out of probability of 1



Joint probabilities

P(A and B) is the probability that both A and B are the case.

Also written P(A, B) or $P(A \cap B)$

P(A and B) is the same as P(B and A)

The probability that I am having a jelly donut with my milk is the same as the probability that I am having milk with my jelly donut.

P(donut and milk) = P(milk and donut)

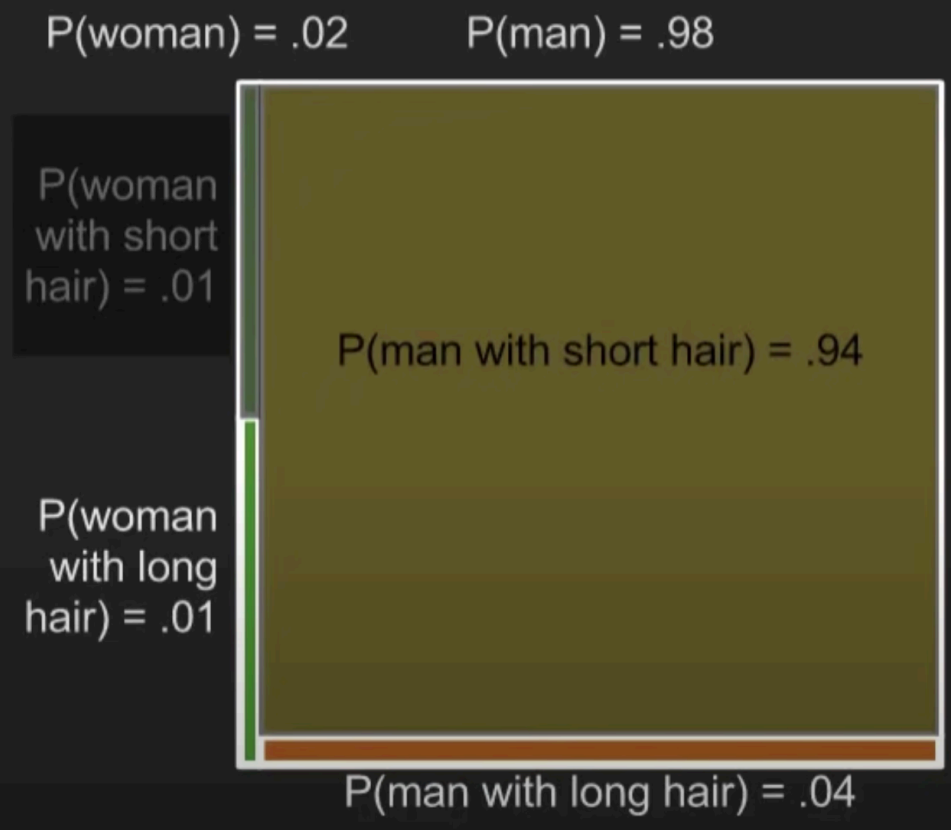
Marginal probabilities

Out of probability of 1

P(long hair) = P(woman with long hair) +

P(man with long hair)

= .01 + .04 = **.05**



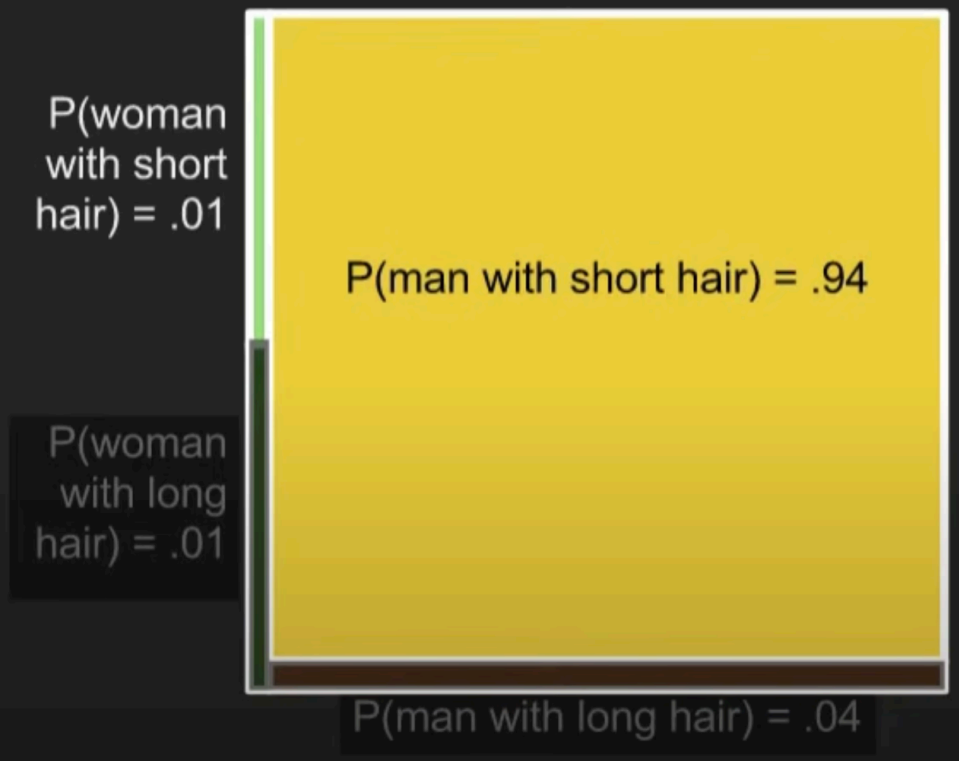
Marginal probabilities

Out of probability of 1

P(short hair) = P(woman with short hair) +

P(man with short hair)

= .01 + .94 = **.95**



What we really care about

We know the person has long hair.
Are they a man or a woman?

P(man | long hair)

We don’t know this answer yet.

Thomas Bayes noticed something cool

P(man with long hair) = P(long hair) * P(man | long hair)

P(long hair and man) = P(man) * P(long hair | man)

Because P(man and long hair) = P(long hair and man)

Thomas Bayes noticed something cool

P(man with long hair) = P(long hair) * P(man | long hair)

P(long hair and man) = P(man) * P(long hair | man)

Because P(man and long hair) = P(long hair and man)

P(long hair) * P(man | long hair) = P(man) * P(long hair | man)

P(man | long hair) = P(man) * P(long hair | man) / P(long hair)

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

Back to the movie theater, this time with Bayes

$$P(\text{man} \mid \text{long hair}) = \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{long hair})}$$

$$= \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{woman with long hair}) + P(\text{man with long hair})}$$

$$P(\text{man} \mid \text{long hair}) = \frac{.5 * .04}{.25 + .02} = .02 / .27 = .07$$

Back to the movie theater, this time with Bayes

$$P(\text{man} \mid \text{long hair}) = \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{long hair})}$$

$$= \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{woman with long hair}) + P(\text{man with long hair})}$$

$$P(\text{man} \mid \text{long hair}) = \frac{.98 * .04}{.01 + .04} = .04 / .05 = .80$$

Probability distributions

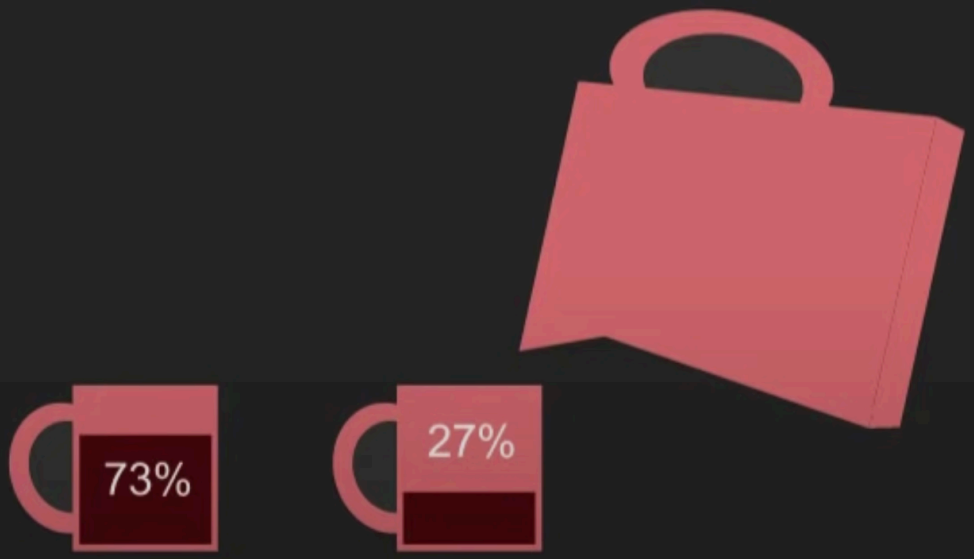
Probability is like a pot with just one cup of coffee left in it.

Probability distributions

If you only have one cup, you can fill it completely.

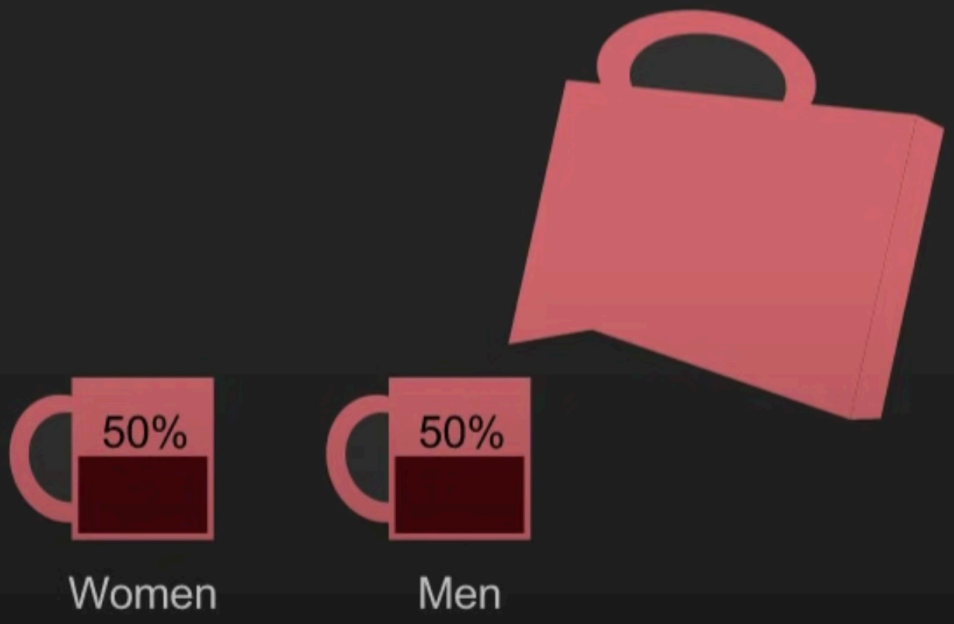
Probability distributions

If you have two cups, you have to decide how to share (distribute) it.



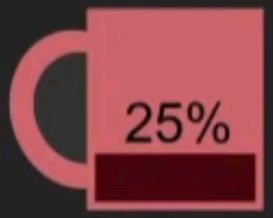
Probability distributions

Our people are distributed between two groups, women and men.

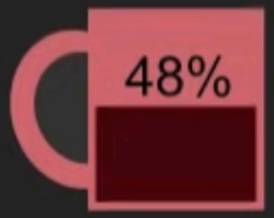


Probability distributions

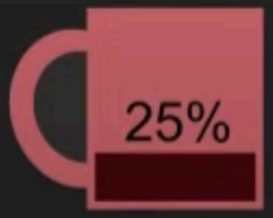
We can distribute them more.



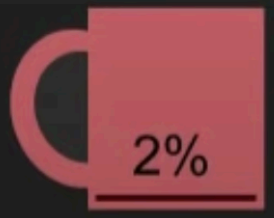
Women with short hair



Men with short hair



Women with long hair

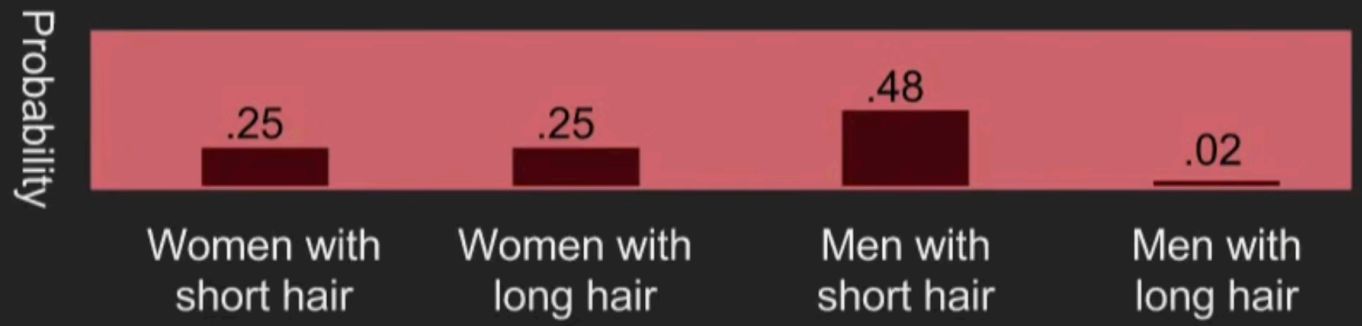


Men with long hair

Probability distributions

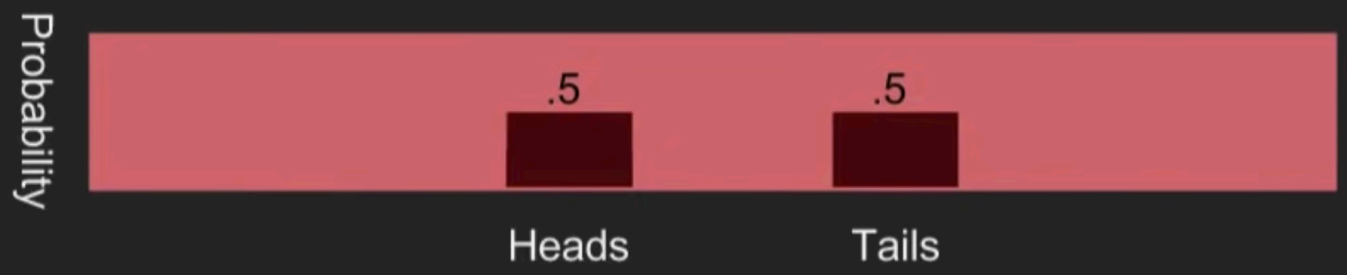
Probability distributions

It's helpful to think of probabilities as beliefs



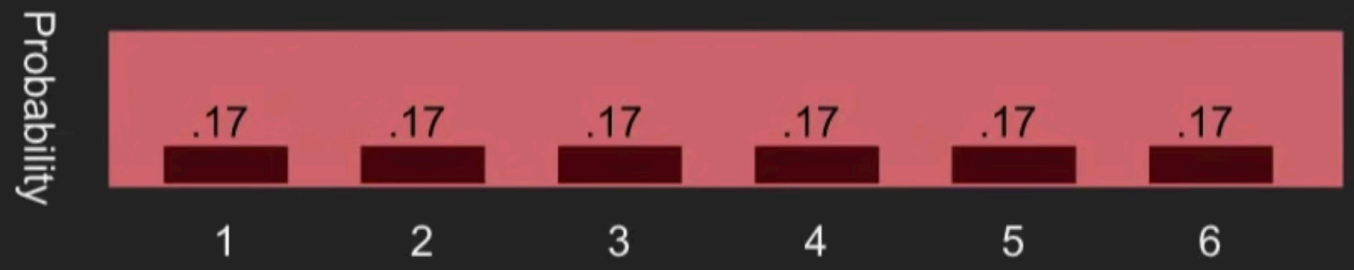
Probability distributions

Flipping a fair coin



Probability distributions

Rolling a fair die



Bayes Theorem

Bayes Rule: $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h | D) = probability of h given D (posterior density)
- P(D | h) = probability of D given h (likelihood of D given h)

Maximum A Posteriori (MAP) Hypothesis

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

The Goal of Bayesian Learning: the most probable hypothesis given the training data (Maximum A Posteriori hypothesis)

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

$$= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D \mid h)P(h)$$

Compute ML Hypo

$$h_{ML} = \arg\max_{h \in H} p(D \mid h)$$

$$= \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2}$$

$$= \arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$$

$$= \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

Bayes Optimal Classifier

Question: Given new instance x, what is its most probable classification?

- $h_{MAP}(x)$ is not the most probable classification!

Example: Let P(h1|D) = .4, P(h2|D) = .3, P(h3 |D) =.3

Given new data x, we have h1(x)=+, h2(x) = -, h3(x) = -

What is the most probable classification of x ?

Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

where V is the set of all the values a classification can take and v_j is one possible such classification.

Example:

P(h1| D) =.4,	P(-|h1)=0,	P(+|h1)=1	$\sum_{hi \in H} P(+ \mid h_i) P(h_i \mid D) = .4$
P(h2|D) =.3,	P(-|h2)=1,	P(+|h2)=0	
P(h3|D)=.3,	P(-|h3)=1,	P(+|h3)=0	$\sum_{hi \in H} P(- \mid h_i) P(h_i \mid D) = .6$

features: $X = (X_1, X_2, ..., X_n)$

label: Y

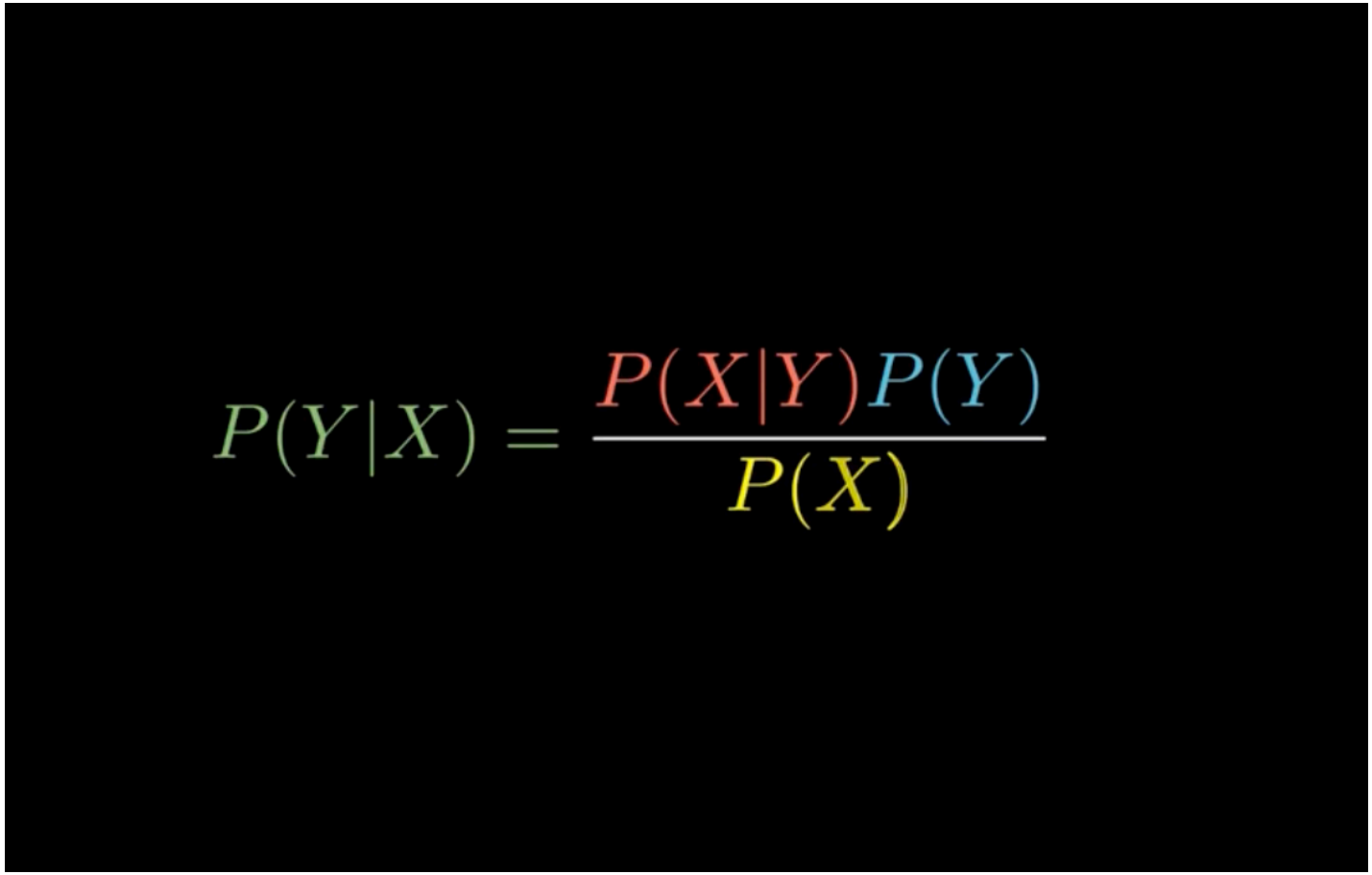
$$P(Y = y | X = (x_1, x_2, ..., x_n))$$

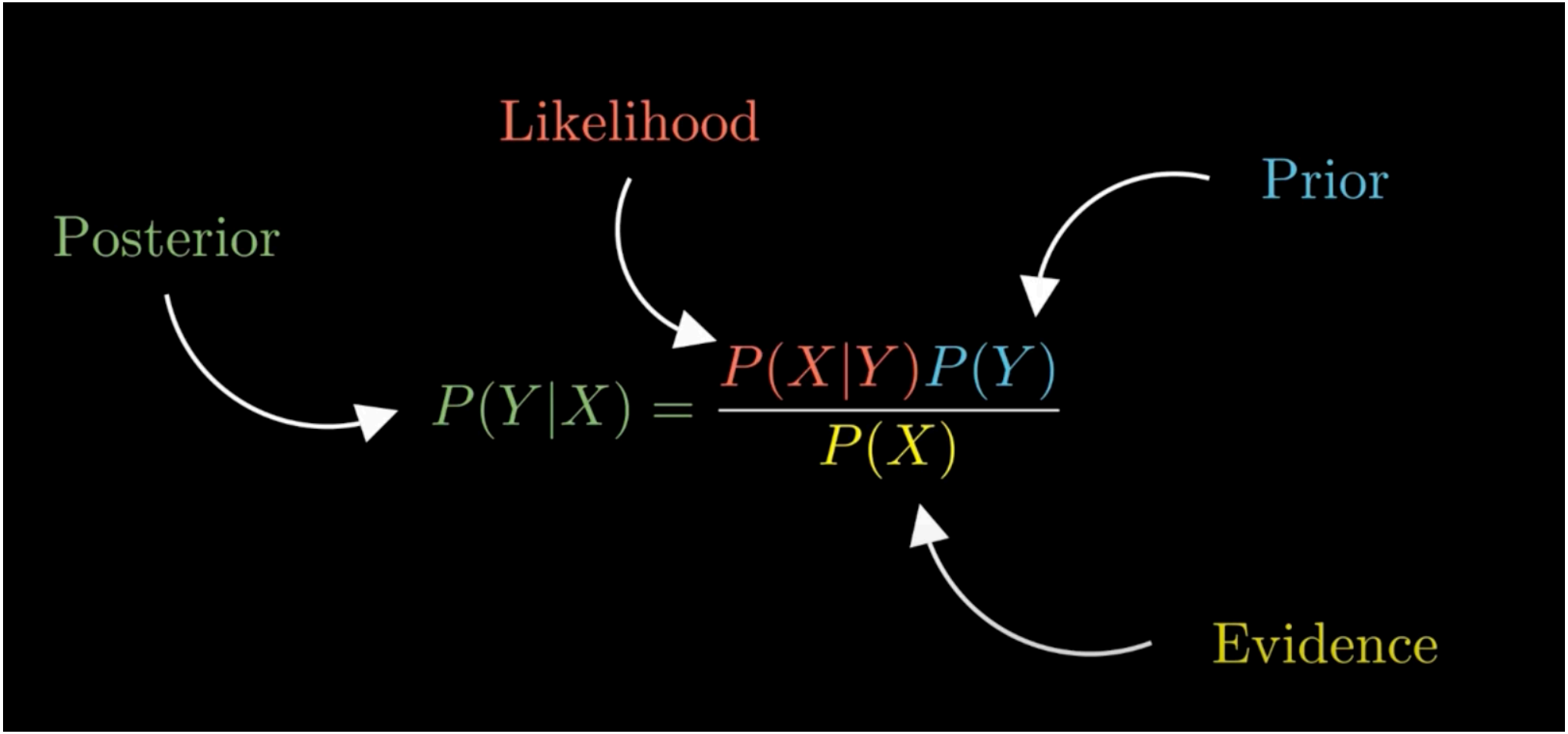
for what value of y

$$P(Y = y | X = (x_1, x_2, ..., x_n))$$

is maximum

but...$P(Y|X)$ is hard to find!


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


Likelihood
Prior
Posterior
$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$
Evidence

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

$$X_1, X_2 \in \{0, 1, 2\}$$
$$Y \in \{0, 1\}$$

Estimate the value of Y given that $X = (0, 2)$

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Let's compute $P(Y = 0|X = (0, 2))$ and $P(Y = 1|X = (0, 2))$ ...

$$P(Y = 0) = \frac{\#Y = 0}{\#Y = 0 + \#Y = 1} = \frac{6}{10}$$

$$P(Y = 1) = \frac{\#Y = 1}{\#Y = 0 + \#Y = 1} = \frac{4}{10}$$

$$P(X = (0, 2)|Y = 1) = \frac{1}{4}$$

$$P(X = (0, 2)|Y = 0) = 0$$

$$P(X = (0, 2)|Y = 0) * P(Y = 0) = 0 * \frac{6}{10} = 0$$

$$P(X = (0, 2)|Y = 1) * P(Y = 1) = \frac{1}{4} * \frac{4}{10} = \frac{1}{10}$$

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Let's compute $P(Y = 0|X = (0, 2))$ and $P(Y = 1|X = (0, 2))$ ...

$$P(Y = 0) = \frac{\#Y = 0}{\#Y = 0 + \#Y = 1} = \frac{6}{10}$$

$$P(Y = 1) = \frac{\#Y = 1}{\#Y = 0 + \#Y = 1} = \frac{4}{10}$$

$$P(X = (0, 2)|Y = 1) = \frac{1}{4}$$

$$P(X = (0, 2)|Y = 0) = 0$$

Estimated value of Y is 1 given $X = (0, 2)$

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Let's compute $P(Y = 0|X = (0, 2))$ and $P(Y = 1|X = (0, 2))$ ...

$$P(Y = 0) = \frac{\#Y = 0}{\#Y = 0 + \#Y = 1} = \frac{6}{10}$$

$$P(Y = 1) = \frac{\#Y = 1}{\#Y = 0 + \#Y = 1} = \frac{4}{10}$$

$$P(X = (0, 2)|Y = 1) = \frac{1}{4}$$

$$P(X = (0, 2)|Y = 0) = 0$$

Let's assume X_1, X_2 are independant!

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Let's compute $P(Y = 0|X = (0, 2))$ and $P(Y = 1|X = (0, 2))$ ...

$$P(Y = 0) = \frac{\#Y = 0}{\#Y = 0 + \#Y = 1} = \frac{6}{10}$$

$$P(Y = 1) = \frac{\#Y = 1}{\#Y = 0 + \#Y = 1} = \frac{4}{10}$$

$$P(X = (0, 2)|Y = 1) = P(X_1 = 0|Y = 1) * P(X_2 = 2|Y = 1)$$

$$P(X = (0, 2)|Y = 0) = P(X_1 = 0|Y = 0) * P(X_2 = 2|Y = 0)$$

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Let's compute $P(Y=0|X=(0,2))$ and $P(Y=1|X=(0,2))$ ...

$$P(Y=0) = \frac{\#Y=0}{\#Y=0+\#Y=1} = \frac{6}{10}$$

$$P(Y=1) = \frac{\#Y=1}{\#Y=0+\#Y=1} = \frac{4}{10}$$

$$P(X=(0,2)|Y=1) = P(X_1=0|Y=1) * P(X_2=2|Y=1) = \frac{3}{4}\ \frac{2}{4}$$

$$P(X=(0,2)|Y=0) = P(X_1=0|Y=0) * P(X_2=2|Y=0) = \frac{1}{6}\ \frac{1}{6}$$

$$4/10 * \frac{3}{4} * \frac{2}{4} > \frac{1}{6} * \frac{1}{6} * 6/10$$

don't forget to multiply the priors
i.e. P(Y=1) and P(Y=0)

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Naïve Bayes

Bayes rule:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)}$$

Assuming conditional independence among X_i's:

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, classification rule for $X^{new} = < X_1, \cdots, X_n >$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 parameters...