Notation and Math

Probability

- If A and B are independent then $p(A \text{ and } B) = p(A)p(B)$
- If A, B, C are independent then $p(A, B, C) = p(A)p(B)p(C)$
- $\log\big(p(A)p(B)p(C)\big) = \log p(A) + \log p(B) + \log p(C)$

Conditional probability

- Suppose you have 3 coins:
  - The first coin c_1 has a 0.4 prob of landing on heads
  - The second coin c_2 has a 0.7 prob of landing on heads
  - The third coin c_3 has a .2 prob of landing on heads
- If you tossed the first coin twice, the second coin once and the third coin once (in that order). What is the probability of seeing HTHT?

$$p(H \mid c_1)p(T \mid c_1)p(H \mid c_2)p(T \mid c_3)$$

Pair share

$$(0.4)(1 - 0.4)(0.7)(1 - 0.2)$$

$$p(H \mid c_1)\big(1 - p(H \mid c_1)\big)\, p(H \mid c_2)\big(1 - p(H \mid c_3)\big)$$

argmax

Which argument maximizes the function

- Suppose you have 3 coins.
  - c_1 has a $p(H \mid c_1) = \theta_1 = 0.4$
  - c_2 has a $p(H \mid c_2) = \theta_2 = 0.7$
  - c_3 has a $p(H \mid c_3) = \theta_3 = 0.2$
- If you randomly chose one of the three coins and tossed it 100 times, receiving 71 heads and 29 tails what do you think is the probability of heads?

$$\theta^* = \arg \max_{\theta \in \{\theta_1, \theta_2, \theta_3\}} \theta^{71}(1-\theta)^{29}$$

MLE!

If I didn’t know it was one of 3 coins… how would I estimate the prob of heads for the coin?

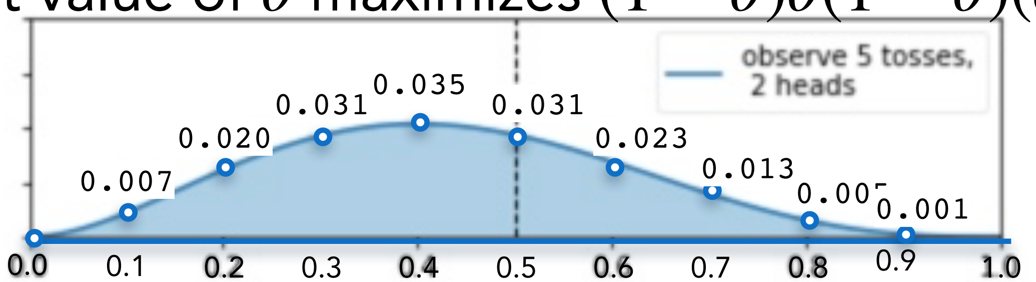
Data → Estimation

How does the likelihood of seeing the data change as we change θ?

Which θ makes observing the data $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}$ most likely?

- T, H, T, H, T,
- If we want to predict θ (the probability of heads), how can we estimate (learn) θ
- One measure of goodness is the θ that most *likely* generated the data

What value of θ maximizes $(1-\theta)\theta(1-\theta)(\theta)(1-\theta)$?



0	$0^2(1-0)^3$
0.1	$0.1^2(1-0.1)^3$
0.2	$0.2^2(1-0.2)^3$
0.3	$0.3^2(1-0.3)^3$
0.4	$0.4^2(1-0.4)^3$
0.5	$0.5^2(1-0.5)^3$
0.6	$0.6^2(1-0.6)^3$
0.7	$0.7^2(1-0.7)^3$
0.8	$0.8^2(1-0.8)^3$
0.9	$0.9^2(1-0.9)^3$
1	$1^2(1-1)^3$

$$L(\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

$$L(\theta; D) = p(D; \theta) = \theta^{N_H}(1-\theta)^{N_T}$$

Data → Estimation

- T, H, T, H, T,
- If we want to predict θ (the probability of heads), how can we estimate (learn) θ
- One measure of goodness is the θ that most *likely* ge[nerates] the data

How does the likelihood of seeing the data change as we change θ?

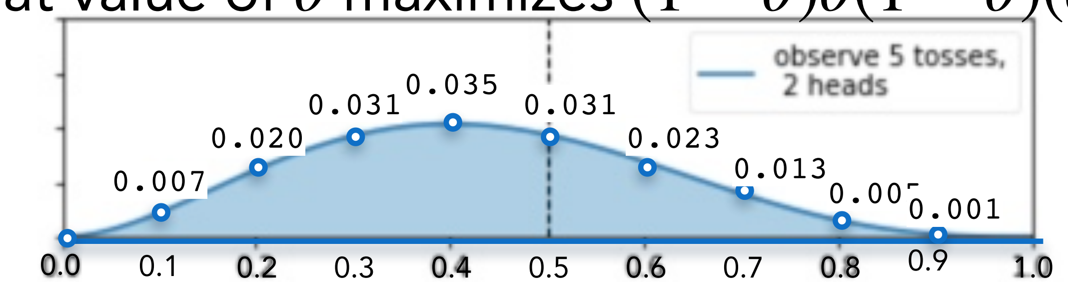
Which θ makes observing the data $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}$ most likely?

Likelihood is always relative to some model - In this example the model is Bernoulli

0.1	$0.1^2(1-0.1)^3$
0.2	$0.2^2(1-0.2)^3$
1	$1^2(1-1)^3$

Typically we view the distribution θ as fixed, and the examples as parameters. We are turning this idea "on its head". Here the examples are fixed and we are considering different choices for the parameter values

What value of θ maximizes $(1-\theta)\theta(1-\theta)(\theta)(1-\theta)$?



$$L(\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

$$L(\theta; D) = p(D;\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

Maximum likelihood estimation (MLE)


https://upload.wikimedia.org/wikipedia/commons/3/3b/

- Flip a (unfair?) coin 100 times and if N_H=55 and N_T
- What is p(H)?
- **Likelihood function** $L(\theta)$ is the probability of the observed data as a function of θ.
  For this example: $L(\theta) = p(D \mid \theta) = \theta^{N_H}(1-\theta)^{N_T}$
- **Log-likelihood function** $\ell(\theta) = \log L(\theta)$
- **Maximum likelihood criterion** the most likely parameter is the one that maximizes $\ell(\theta)$
- How to maximize $\ell(\theta)$?

Extremely small value

If θ was 0.5, then
$L(0.5) = 0.5^{100} \approx 7.9 \times 10^{-31}$

If θ was 0.5, then
$\ell(0.5) = \log 0.5^{100} = 100 \log 0.5 = -69.31$

Maximum likelihood estimation (MLE)

https://upload.wikimedia.org/wikipedia/commons/3/3b/

Coin flips are conditionally independently p(Heads)=θ and identically distributed (i.i.d.)

- Flip a (unfair?) coin 100 times and if N_H=55 and N_T = 45
- What is p(H)?
- **Likelihood function** $L(\theta)$ is the probability of the observed data as a function of θ.
  For this example: $L(\theta) = p(D \mid \theta) = \theta^{N_H}(1-\theta)^{N_T}$
- **Log-like** ... $= \log L(\theta)$
- **Maximum likelihood criterion**: the most likely parameter is the one that maximizes $\ell(\theta)$
- How to maximize $\ell(\theta)$?

L is a function of the model parameters, not the data

Maximizing $\ell(\theta)$ is the same as maximizing $L(\theta)$. Why?

In computer science log is always base 2…. In Machine learning log is always base e

Extremely small value

If θ was 0.5, then
$L(0.5) = 0.5^{100} \approx 7.9 \times 10^{-31}$

If θ was 0.5, then
$\ell(0.5) = \log 0.5^{100} = 100 \log 0.5 = -69.31$

What if we had 100 coin tosses, 40 heads and 60 tails

Which is the right likelihood function?
θ will be your estimated probability of flipping a coin and getting heads.

A) $L(\theta) = (0.4)^{40}(1 - 0.4)^{60}$

B) $L(\theta) = (\theta)^{40}(1 - \theta)^{60}$

C) $L(\theta) = (0.4)^{\theta}(1 - 0.4)^{1-\theta}$

D) $L(\theta) = (0.8)^{60}(1 - 0.8)^{100}$

What if we had 100 coin tosses, 40 he
and 60 tails

Which is the right likelihood function?
θ will be your estimated probability of flipping a coin and getting heads.

A) $L(\theta) = (0.4)^{40}(1 - 0.4)^{60}$

B) $L(\theta) = (\theta)^{40}(1 - \theta)^{60}$

C) $L(\theta) = (0.4)^{\theta}(1 - 0.4)^{1-\theta}$

D) $L(\theta) = (0.8)^{60}(1 - 0.8)^{100}$