WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

Algorithmic Privacy, MLE

CSE 312 Su23
Lecture 19

https://xkcd.com/2169/

Application: Tail Bounds

Privacy Preservation

A real-world example (adapted from *The Ethical Algorithm* by Kearns and Roth; based on ‘randomized response’ protocol by Warner [1965]).

And gives a sense of how randomness is actually used to protect privacy.

Privacy Preservation with Randomness

You’re working with a social scientist. They want to get accurate data on the rate at which people cheat on their romantic partners.

Privacy Preservation with Randomness

You're working with a social scientist. They want to get accurate data on the rate at which people cheat on their romantic partners.

We know about polling accuracy!

Do a poll, call up a random sample of adults and ask them "have you ever cheated on your romantic partner?"

Use a tail-bound to estimate the needed number n get a guaranteed good estimate, right?

Privacy Preservation with Randomness

You're working with a social scientist. They want to get accurate data on the rate at which people cheat on their romantic partners.

We know about polling accuracy!

Do a poll, call up a random sample of adults and ask them "have you ever cheated on your romantic partner?"

Use a tail-bound to estimate the needed number n get a guaranteed good estimate, right?

You do that, and somehow, no one says they cheated.

What’s the problem?

People lie.

What’s the problem?

People lie.

Or they might be concerned about you keeping this data.

What’s the problem?

People lie.

Or they might be concerned about you keeping this data.

Databases can be leaked (or infiltrated. Or subpoenaed).

What’s the problem?

People lie.

Or they might be concerned about you keeping this data.

Databases can be leaked (or infiltrated. Or subpoenaed).

You don’t want to hold this data, and the people you’re calling don’t want you to hold this data.

Doing Better With Randomness

You don't really need to know **who** was cheating. Just how many people were.

Here's a protocol:

Please flip a coin.

If the coin is heads, or you have ever cheated, please tell me "heads"

If the coin is tails and you have not ever cheated, please tell me "tails"

		cheat	nc
coin	H	H	H
	T	H	T

Will it be private?

If you are someone who has cheated, and you report heads can that be used against you? Not substantially – just say "no the coin came up heads!"

You discover your partner said heads, what's the probability that they cheated?

$$\mathbb{P}(C \mid \text{said } H) = \frac{\mathbb{P}(\text{said } H \mid C)\cdot\mathbb{P}(C)}{\mathbb{P}(\text{said } H)} \qquad \text{[Bayes]}$$

$$= \frac{1 \cdot 0.15 \qquad\qquad \tfrac{1}{2}\,(1-0.15)}{\mathbb{P}(\text{said } H \mid C)\,\mathbb{P}(C) + \mathbb{P}(\text{said } H \mid \bar{C})\,\mathbb{P}(\bar{C})}$$

Will it be private?

If you are someone who has cheated on your spouse, and you report heads can that be used against you? Not substantially – just say "no the coin came up heads!"

15%

$$\mathbb{P}(C|H) = \frac{\mathbb{P}(H|C)\cdot\mathbb{P}(C)}{\mathbb{P}(H)} = \frac{1\cdot\mathbb{P}(C)}{\frac{1}{2}\mathbb{P}(\overline{C}) + 1\cdot\mathbb{P}(C)}$$

26%

Is this a substantial change?

No. For real world values (~15%) of $\mathbb{P}(C)$, the probability estimate would increase (to ~26%). But that isn't too damaging.

But will it be accurate?

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

But will it be accurate?

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

Suppose you poll n people, and let X be the number of people who said "heads". We'll find an estimate Y of the number of people who cheated in the sample, and let p be the true probability of cheating in the population.

What should Y be? Can we draw a margin of error around Y?

$$\underset{\text{said H}}{\mathbb{E}[X]} = \frac{n}{2} + \underset{\text{cheat}}{\frac{\mathbb{E}[Y]}{2}}$$

see H, say H

$$\mathbb{E}[X] = \frac{1}{2}(n + \mathbb{E}[Y])$$

$$X = \frac{1}{2}(n + Y)$$

$$2X - n = Y$$

But will it be accurate?

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

Suppose you poll n people, and let X be the number of people who said "heads" We'll find an estimate Y of the number of people who cheated in the sample, and let p be the true probability of cheating in the population.
What should Y be? Can we draw a margin of error around Y?

$\mathbb{E}[X] = \frac{n}{2} + \frac{1}{2}\mathbb{E}[Y]$

We'll define Y to be: $Y = 2\left(X - \frac{n}{2}\right)$.
This is a definition, based on how the $\mathbb{E}[Y]$ should relate to the $\mathbb{E}[X]$.

But will it be accurate?

$$\mathbb{E}[X] = \frac{n}{2} + \frac{1}{2}\mathbb{E}[Y]$$

$$Y = 2\left(X - \frac{n}{2}\right)$$

$$\mathrm{Var}(Y) = \mathrm{Var}\Big(2\overbrace{\left(X - \frac{n}{2}\right)}^{\text{RV}}\Big)$$

$$= 4\,\mathrm{Var}\left(X - \frac{n}{2}\right)$$

$$= 4\,\mathrm{Var}(X)$$

But will it be accurate?

$\mathbb{E}[X] = \frac{n}{2} + \frac{1}{2}\mathbb{E}[Y]$

$\boxed{Y = 2\left(X - \frac{n}{2}\right)}$

$\mathrm{Var}(X) = \mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i)$

$\mathrm{Var}(X_i)$? It's an indicator with parameter $p + (1-p) \cdot \frac{1}{2} = \frac{1}{2} + \frac{p}{2}$

So $\mathrm{Var}(X_i) = \left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right)$

$\mathrm{Var}(Y) = 4\mathrm{Var}(X) = 4n\mathrm{Var}(X_i) = 4n\left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right) \leq \frac{4n}{4} = n$

The variance is 4 times as much as it would have been for a non-anonymous poll.

Can we use Chernoff?

(Multiplicative) Chernoff Bound

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables.
Let $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \leq \delta \leq 1$

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{3}\right) \text{ and } \mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right)$$

What happens with $\mathrm{n} = 1000$ people?

What range will we be within at least 95% of the time?

☹ Can't bound δ without bounding p

The right tail is the looser bound, so ensuring the right tail is less than 2.5% gives us the needed guarantee.

np

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) = \exp\left(-\frac{\delta^2 1000p}{3}\right) \leq .025$$

★

4

$$-\frac{\delta^2 1000p}{3} \leq \ln(.025)$$

$$-\delta^2 \leq \frac{3 \cdot \ln(.025)}{1000p}$$

$$\boxed{\delta \geq \sqrt{\frac{-3\ln(.025)}{1000p}}}$$

As $p \to 0$, $\delta \to \infty$ – we're not actually making a claim anymore.

A different inequality

If we try to use Chernoff, we'll hit a frustrating block.

Since μ depends on p, p appears in the formula for δ. And we wouldn't get an absolute guarantee unless we could plug in a p.

And it'll turn out that as $p \to 0$ that $\delta \to \infty$ so we don't say anything then.

Luckily, there's always another bound…

Hoeffding's Inequality

Hoeffding's Inequality

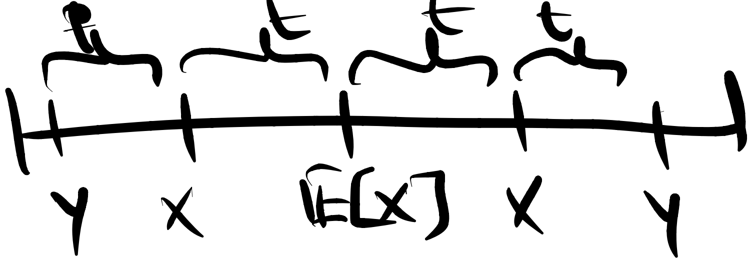
Let $X_1, X_2, \dots, X_n$ be *independent* RVs, each with range $[0,1]$.
Let $\bar{X} = \sum X_i / n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

$2e^{(-2nt^2)}$

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$. Why?

$$Y = 2\left(X - \frac{n}{2}\right) \text{ or } X = \frac{Y+n}{2}$$



$$\begin{aligned}
\boxed{|X - \mathbb{E}[X]|} &= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y+n}{2}\right]\right| \\
&= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y}{2}\right] - \frac{n}{2}\right| \\
&= \left|\frac{Y}{2} - \mathbb{E}\left[\frac{Y}{2}\right]\right| \\
&= \boxed{\frac{1}{2}|Y - \mathbb{E}[Y]|}
\end{aligned}$$

$$|X - \mathbb{E}[X]| \geq t$$

$$\frac{1}{2}|Y - \mathbb{E}[Y]| \geq t$$

$$|Y - \mathbb{E}[Y]| \geq 2t$$

So $|X - \mathbb{E}[X]| \geq t$ if and only if $\frac{1}{2}|Y - \mathbb{E}[Y]| \geq t$ iff $|Y - \mathbb{E}[Y]| \geq 2t$.

Hoeffding's Inequality

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be *independent* RVs, each with range $[0,1]$.
Let $\bar{X} = \sum X_i / n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

How close will we be with *n=1000 with probability at least* .95*?*

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$.

Margin of Error

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) = \mathbb{P}(|X - \mathbb{E}[X]| \geq t/2) \leq 2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) \leq .05$$

Margin of Error

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) = \mathbb{P}(|X - \mathbb{E}[X]| \geq t/2) \leq 2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) \leq .05$$

For $n = 1000$, we get:

$$2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) \leq .05 \Rightarrow -\frac{2000t^2}{4} \leq \ln(.025) \Rightarrow t \leq .086.$$

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq .086) \leq .05$$

So our margin of error is about 8.6%.

To get a margin-of-error of 5% need $2\exp\left(-2n\left(\frac{.05}{2}\right)^2\right) \leq .05$

$$n \geq 2952$$

How much do we lose?

H – cheated or H $\left(\frac{1}{2}\right)$

Roll die 1 – $1/6$

We lose a factor of two in the length of the margin (equivalently, we'd need to talk to 4 times as many people to have the same confidence).

You can also control this tradeoff.

Want more accuracy? Make it roll a die: report 1 if cheated (truth o/w)

Want more security? Make it Bernoulli with probability $p \gg \frac{1}{2}$ or cheated have the same report (e.g. report "die roll 1 [and didn't cheat]" or "die roll 2-6 [or did cheat]"

75%

In The Real World

Injecting randomness to preserve privacy is a real thing.

Instead of having everyone flip a coin, “random noise” can be inserted after all the data has been collected.

Differential privacy is being used to protect Census data.

The overall count of people in each state is exact (well, exactly the data they collected). But the data per block or per city will be randomized to protect against .

This video nicely explains what’s involved. Notice that the accuracy guarantees come in the same “inside-margin-of-error-with-probability” guarantees we’ve been giving for our randomness (just much stronger).

Maximum Likelihood Estimation (MLE)

A tale as old as time…

So far, the probability questions we've asked have followed a particular pattern.

You're given a model with the probabilities you need to make predictions.

A tale as old as time…

Recall… the distributions we have learnt are all parameterized. Where, a distribution is the **model** + **parameter** $\boldsymbol{\theta}$.

1. $Ber(p)$ $\qquad \theta = p$
2. $Poi(\lambda)$ $\qquad \theta = \lambda$
3. $Unif(a, b)$ $\qquad \theta = (a, b)$

Given the model and its parameters, we know how to make predictions!

The Remix

Let’s say you have a coin. You don’t know if it’s fair or not.

It’s reasonable to think that a coin flip follows a Bernoulli distribution.

But what’s the parameter?

The Remix

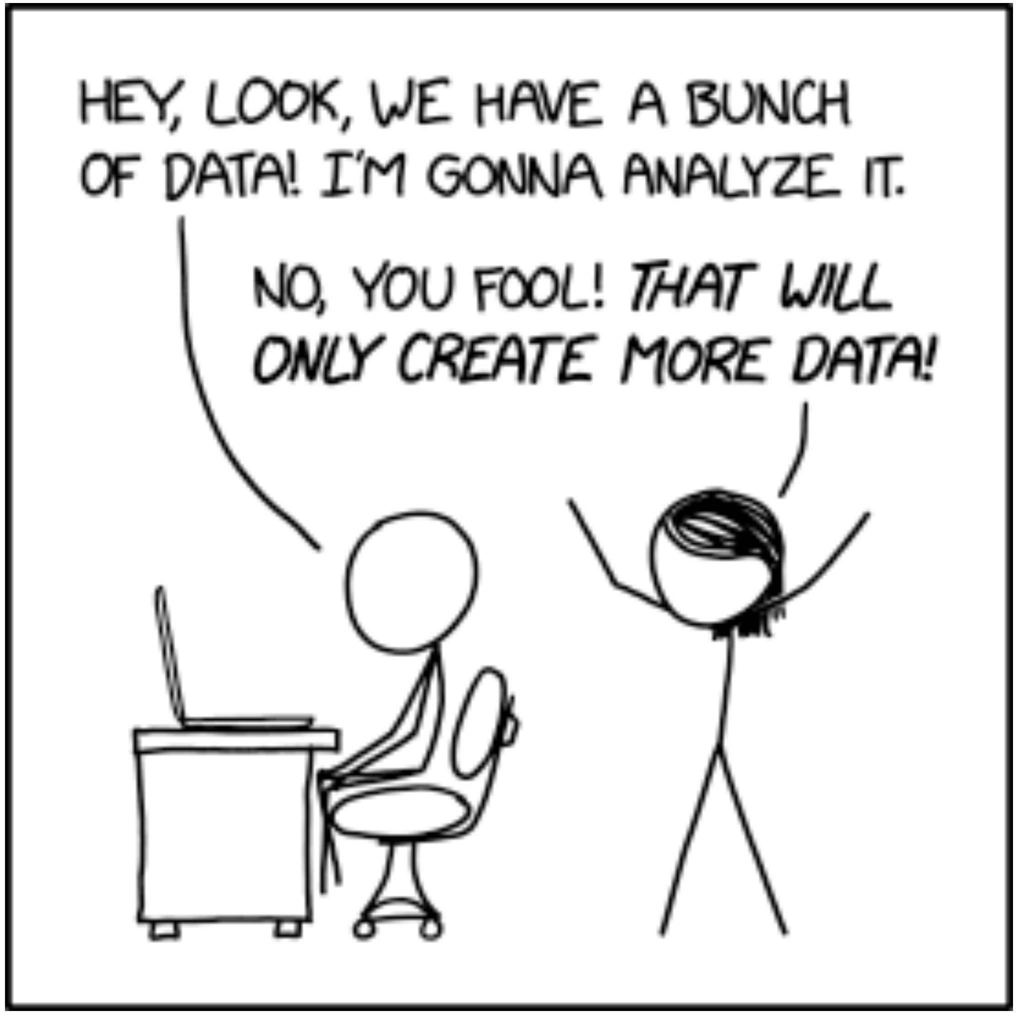
Let's say you have a coin. You don't know if it's fair or not.

It's reasonable to think that a coin flip follows a Bernoulli distribution.

But what's the parameter?

In the real world, you're not given parameters :(

But we do have… data!



https://xkcd.com/2582/

Maximum Likelihood Estimation

We derive an estimator $\hat{\theta}$ for our true parameter θ from the data that we observe.

distribution $\longrightarrow$ predictions

$\uparrow$

data

true param

$X \sim Ber(\theta)$

$\hat{\theta}$ estimator

Back to the coin…

Suppose you flip a coin independently 10 times, and you see

* HTTTHHTHHH

What is your estimate of the probability the coin comes up heads?

$$\mathbb{P}(H) = \frac{\#\text{ of } H}{\#\text{ flips}} = \frac{6}{10}$$

A Coin

Suppose you flip a coin independently 10 times, and you see

HTTTHHTHHH

What is your estimate of the probability the coin comes up heads?

Probably around… 6/10 = 3/5.

But how do we argue "objectively" that this is the best estimate for the Bernoulli parameter?

Maximum Likelihood Estimation

Idea: we got the results we got.

High probability events happen more often than low probability events.

So, guess the rules that maximize the probability of the events we saw (relative to other choices of the rules).

Since that event happened, might as well guess the set of rules for which that event was a ‘high probability event’.

Likelihood

Formally, we are trying to estimate a parameter of the experiment (here: the probability of a coin flip being heads).

The likelihood of an event E given a parameter θ is

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

$$\mathcal{L}(\text{HTH}\,;\theta)$$

Likelihood

Formally, we are trying to estimate a parameter of the experiment (here: the probability of a coin flip being heads).

The likelihood of an event E given a parameter θ is

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

$\theta = 1/3$

We see the outcome HTTTHHTHH.

event param

$\mathcal{L}(\underbrace{\text{HTTTHHTHHH}}_{\text{event}};\underbrace{\theta}_{\text{param}}) = ?$ $\theta^6(1-\theta)^4$

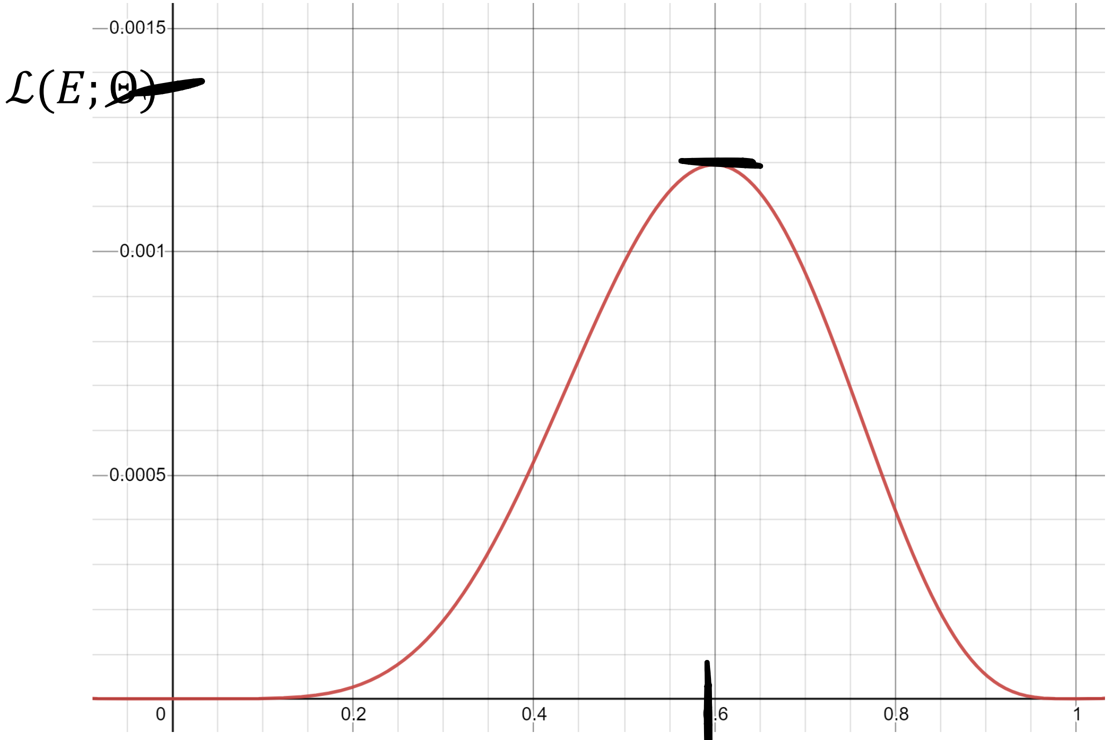
Likelihood

param.

$$\mathcal{L}(\text{HTTTHHTHHH}\,;\,\theta) = \theta^6(1-\theta)^4$$

Probability of observing the outcome HTTTHHTHHH if θ is probability we see heads.

For a fixed outcome, a function of θ.



Θ

$\hat{\theta}$

$\mathcal{L}$ fn. maxim

θ param

Likelihood

(Discrete case)

data sample

Likelihood

The likelihood of i.i.d observations $x_1, \dots, x_n$ from a discrete distribution is

$$\mathcal{L}(x_1, \dots, x_n\,; \theta) = \prod_{i=1}^{n} \mathbb{P}(x_i; \theta)$$

Notation comparison

$\mathbb{P}(X|Y)$ probability of X, conditioned on the **event** Y having happened (Y is a subset of the sample space).

$\neq$ event

★ $\mathbb{P}(X;\theta)$ probability of X, where to properly define our probability space we need to know the extra piece of information θ. Since θ isn't an event, this is not conditioning.

★ $\mathcal{L}(X;\theta)$ the likelihood of event X, given that an experiment was run with parameter θ. Likelihoods don't have all the properties we associate with probabilities (e.g. they don't all sum up to 1) and this isn't conditioning on an event (θ is a parameter/rule of how the event could be generated).

Notation comparison

$\mathbb{P}(X;\theta)$ probability of X, where to properly define our probability space we need to know the extra piece of information θ.

Fixed θ, a function of x_i

$\mathcal{L}(X;\theta)$ the likelihood of event X, given that an experiment was run with parameter θ.

Fixed $x_1, \ldots, x_n$, a function of θ

probability

statistics

model → predict data

data → predict model

Maximum Likelihood Estimation

We will choose $\hat{\theta} = \mathbf{argmax}_{\theta}\ \underline{\mathcal{L}(E;\theta)}$

$\mathbf{argmax}$ is the argument that produces the maximum so the θ that causes $\mathcal{L}(E;\theta)$ to be maximized.

$$\mathcal{L}(E;\theta) = \theta^{6}(1-\theta)^{4}$$

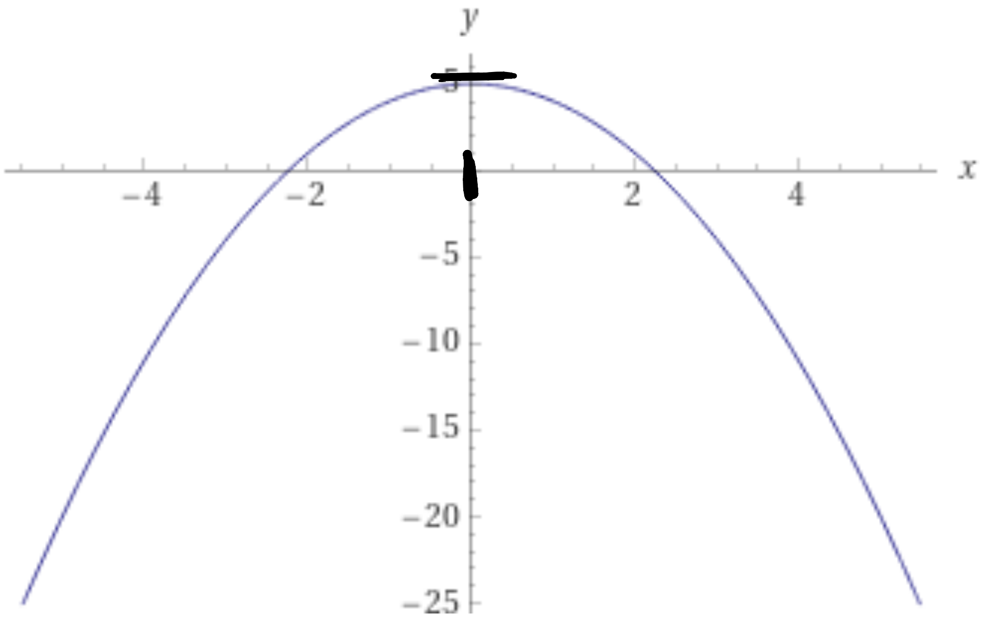
Argmax?

input

For example

$\text{argmax}_x\ (5 - \text{x}^2) = 0$

$\max_x(5 - x^2) = 5$, the input (argument) which produces $\mathbf{5}$ is $\mathbf{0}$, so the argmax is 0



max

MLE

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is:

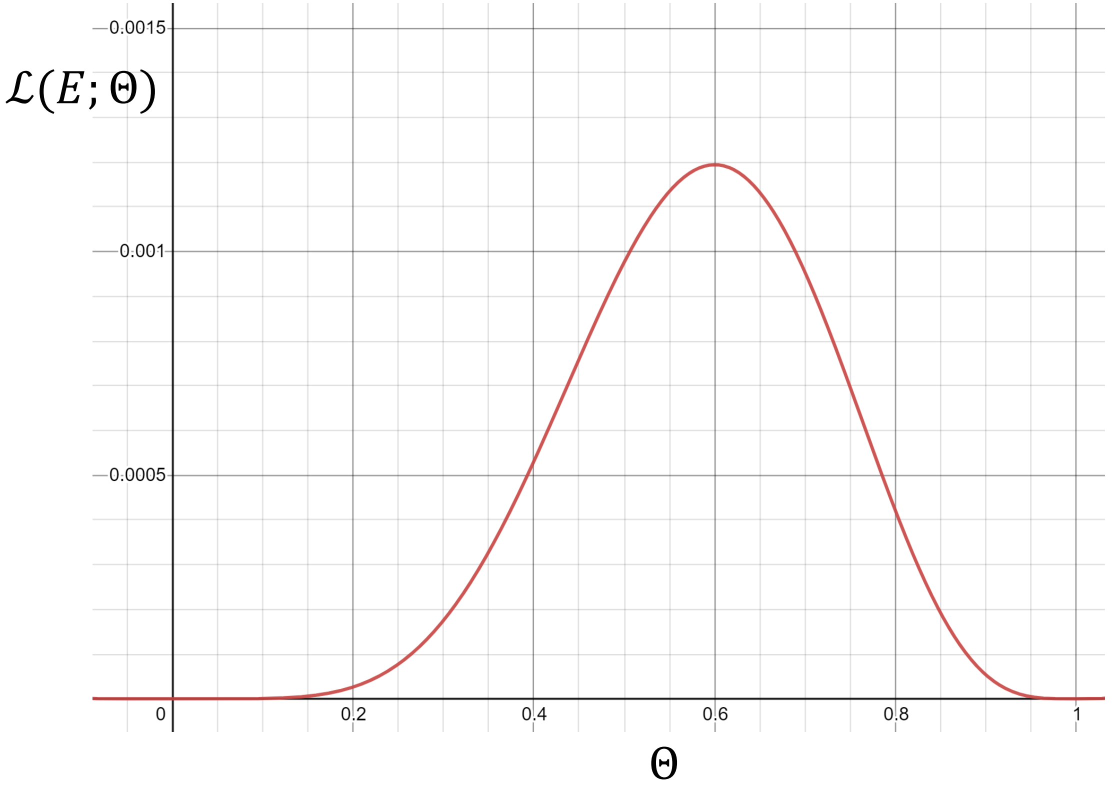
$$\hat{\theta} = \mathrm{argmax}_{\theta} \ \mathcal{L}(E;\theta)$$

θ is a variable, $\hat{\theta}$ is a number (or formula given the event).

We'll also use the notation $\hat{\theta}_{\mathbf{MLE}}$ if we want to emphasize how we found this estimator.

The Coin Example

$\mathcal{L}(\text{HTTTHHTHHH}\,;\,\theta) = \theta^6(1-\theta)^4$



What θ maximizes the likelihood?

How do we usually find a maximum?

Calculus!! $uv' + u'v$

$$\frac{\partial}{\partial\theta}\theta^6(1-\theta)^4 = 6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3$$

Set equal to 0 and solve

$$\boxed{6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3 = 0} \Rightarrow 6(1-\theta) - 4\theta = 0 \Rightarrow -10\theta = -6 \Rightarrow \boxed{\theta = \frac{3}{5}}$$

Is that really the maximum?

Ideally, should use the second derivative test.

Set derivative equal to 0 and solve.

Take the second derivative. If negative everywhere, then the critical point is the maximizer.

Half a step backwards…

We’re going to be taking the derivative of products a lot.

The product rule is not fun. There has to be a better way!

Half a step backwards…

We're going to be taking the derivative of products a lot.

The product rule is not fun. There has to be a better way!

Take the log!

$$\ln(a \cdot b) = \ln(a) + \ln(b)$$

$$\ln\left(a^b\right) = \cancel{a}\, b \ln(\cancel{b}\, a)$$

We don't need the product rule if our expression is a sum!

monotone

Can we still take the max? $\ln()$ is an increasing function, so

$$\text{argmax}_\theta \ln(\mathcal{L}(E;\theta)) = \text{argmax}_\theta \mathcal{L}(E;\theta)$$

$x < y$

$\ln(x) < \ln(y)$

Coin flips is easier

$\ln(\theta^6 (1-\theta)^4)$.

$\ln(\theta^6) + \ln((1-\theta)^4)$

$6 \ln \theta + 4 \ln(1-\theta)$.

- $\mathcal{L}(\text{HTTTHHTHHH}; \theta) = \theta^6 (1-\theta)^4$
- $\ln(\mathcal{L}(\text{HTTTHHTHHH}; \theta) = 6\ln(\theta) + 4\ln(1-\theta)$
- $\frac{d}{d\theta} \ln\big(\mathcal{L}(\cdot)\big) = \frac{6}{\theta} - \frac{4}{1-\theta}$

Set to 0 and solve:

$$\boxed{\frac{6}{\theta} - \frac{4}{1-\theta} = 0} \Rightarrow \frac{6}{\theta} = \frac{4}{1-\theta} \Rightarrow 6 - 6\theta = 4\theta \Rightarrow \boxed{\theta = \frac{3}{5}}$$

$\frac{d^2}{d\theta^2} = \frac{-6}{\theta^2} - \frac{4}{(1-\theta)^2} < 0$ everywhere, so any critical point must be a maximum.

Summary

Given: an event E (usually n i.i.d. samples from a distribution with unknown parameter θ).

1. Find likelihood $\mathcal{L}(E;\theta)$

Usually $\prod \mathbb{P}\,(x_i;\theta)$ for discrete

2. Maximize the likelihood. Usually:

A. Take the log (if it will make the math easier)

B. Set the derivative to 0 and solve

C. Use the second derivative test to confirm you have a maximizer

Summary

Given: an event E (usually n i.i.d. samples from a distribution with unknown parameter θ).

1. Find likelihood $\mathcal{L}(E; \theta)$

Usually $\prod \mathbb{P}\,(x_i; \theta)$ for discrete

2. Maximize the likelihood. Usually:

A. Take the log (if it will make the math easier)

B. Set the derivative to 0 and solve

C. Use the second derivative test to confirm you have a maximizer