Indian Institute of Information Technology, Allahabad

Introduction to Machine Learning

Maximum Likelihood Estimation

By

Dr. Shiv Ram Dubey
Associate Professor
Computer Vision and Biometrics Lab
Department of Information Technology
Indian Institute of Information Technology, Allahabad
Web: https://profile.iiita.ac.in/srdubey/

DISCLAIMER

The content (text, image, and graphics) used in this slide are adopted from many sources for academic purposes. Broadly, the sources have been given due credit appropriately. However, there is a chance of missing out some original primary sources. The authors of this material do not claim any copyright of such material.

Your first consulting job

- Client: I have special coin, if I flip it, what's the probability it will be ***heads***?

- You (a machine learner): I need to collect data.

H H

- You (a frequentist): The probability is:

100% heads!

Your first consulting job

- Client: Uhhhh…. You sure about that? I just got a tails.

- You (a machine learner): I need to collect **more data**.
  - *flips coin 5 times, get HHTHT

- You: The probability is: 60% Heads, 40% Tails!

Your first consulting job

- Client: Uhhhh…. You sure about that? I just got a tails.

- You (a machine learner): I need to collect **more data**.
  - *flips coin 10000 times, it comes up Heads 60% of the time

- You: The probability is: 60% Heads, 40% Tails!

- Client: ***Why should I believe you?***

- You (a machine learner): Let's do some math!

Coin – Bernoulli Distribution

- **Data:** sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
    - Flips are i.i.d.:
        - Independent events
        - Identically distributed according to Bernoulli distribution

$$P(\mathcal{D}|\theta) =$$

$= P(HHTHT \mid \theta)$

$= P(H)P(H)P(T)P(H)P(T)$ # by independence

$= \theta^k (1-\theta)^{n-k}$

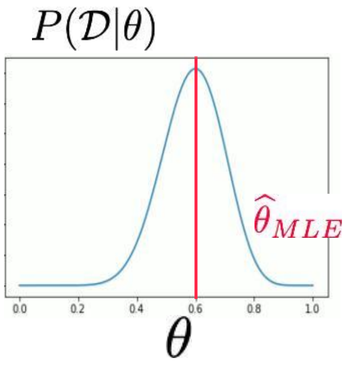
Coin – Bernoulli Distribution

- **Data:** sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
- **Likelihood:**

$$P(\mathcal{D}|\theta) = \theta^k (1-\theta)^{n-k}$$ # likelihood

- **Maximum likelihood estimation (MLE):** Choose θ that maximizes the probability of observed data:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(\mathcal{D}|\theta)$$



Coin – Bernoulli Distribution

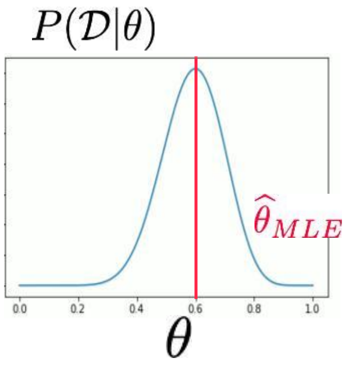
- **Data:** sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
- **Likelihood:**

$$P(\mathcal{D}|\theta) = \theta^k (1-\theta)^{n-k}$$

\# likelihood

- **Maximum likelihood estimation (MLE):** Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta}_{MLE} &= \arg\max_{\theta} \; P(\mathcal{D}|\theta) \\ &= \arg\max_{\theta} \; \log P(\mathcal{D}|\theta)\end{aligned}$$



Why take the log?

- Easier to work with

Coin – Bernoulli Distribution

- **Data:** sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
- **Likelihood:**

$$P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}$$ # likelihood

- **Maximum likelihood estimation (MLE):** Choose θ that maximizes the probability of observed data:

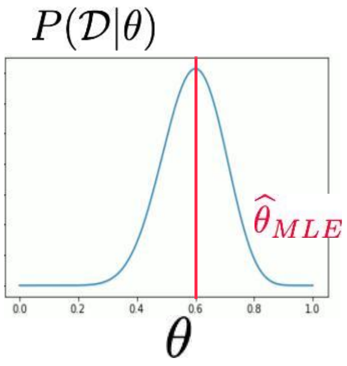
$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \; \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \; \log\left[\theta^k(1-\theta)^{1-k}\right]$$

Why take the log?

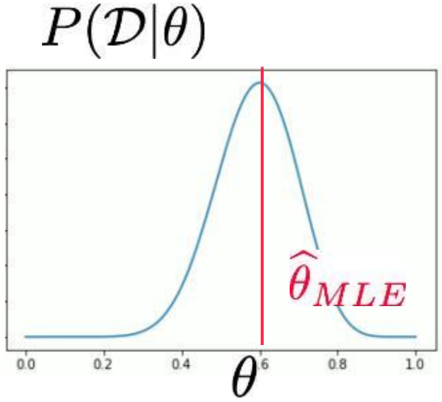
- Easier to work with



MLE: Your first learning algorithm

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; \log P(\mathcal{D}|\theta)$$

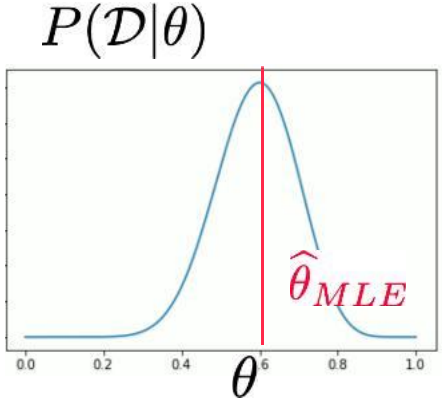
$$= \arg\max_{\theta} \; \log \theta^k (1-\theta)^{n-k}$$



- How do we find θ that maximizes likelihood?

MLE: Your first learning algorithm

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; \log P(\mathcal{D}|\theta)$$
$$= \arg\max_{\theta} \; \log \theta^k (1-\theta)^{n-k}$$



- How do we find θ that maximizes likelihood?
- Use the fact that derivative is zero at maxima (also at minima)
- Set derivative to zero, and find θ satisfying:

$$\boxed{\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0}$$ # MLE

MLE: Your first learning algorithm

- First manipulate the log likelihood to make it easy to work with:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \ \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \ \log \theta^k (1-\theta)^{n-k}$$

$$= \arg\max_{\theta} \ k\log\theta + (n-k)\log(1-\theta)$$

MLE: Your first learning algorithm

- First manipulate the log likelihood to make it easy to work with:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \ \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \ \log \theta^k (1-\theta)^{n-k}$$

$$= \arg\max_{\theta} \ k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find ϑ satisfying:**

$$\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0$$

$$\frac{d}{d\theta}\log P(\mathcal{D}\,|\,\theta) = 0 \longrightarrow \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} = 0$$

for your formula sheet

$$\frac{d}{dx}\log x = \frac{1}{x}$$

MLE: Your first learning algorithm

- First manipulate the log likelihood to make it easy to work with:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \; \log \theta^{k}(1-\theta)^{n-k}$$

$$= \arg\max_{\theta} \; k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find ϑ satisfying:**

$$\frac{d}{d\theta}\log P(\mathcal{D}|\theta) = 0$$

$$\frac{d}{d\theta}\log P(\mathscr{D}\,|\,\theta) = 0 \longrightarrow \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} = 0$$

for your formula sheet

$$\frac{d}{dx}\log x = \frac{1}{x}$$

$$k - k\theta = n\theta - k\theta$$

$$\boxed{\hat{\theta}_{\mathrm{MLE}} = \frac{k}{n}}$$

our example:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{3}{5} = 60\%$$

Your first consulting job

- Client: Uhhhh…. You sure about that? I just got a tails.

- You (a machine learner): I need to collect **more data**.
    - *flips coin 10000 times, it comes up Heads 60% of the time

- You: The probability is: 60% Heads, 40% Tails!

- Client: ***Why should I believe you?***

- You (a machine learner): Let's do some math!

$$\hat{\theta}_{\mathrm{MLE}} = \frac{3}{5} = 60\%$$

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\mathrm{MLE}}$ is a random variable, where there is a ground truth parameter ϑ^* that generates the data

$\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\mathrm{MLE}} = \frac{k}{n}$$ # random variable

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\text{MLE}}$ is a random variable, where there is a ground truth parameter ϑ^* that generates the data

$$\mathscr{D} = (HHTTH\ldots) \text{ of a fixed size } n$$

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n} \quad \text{\# random variable}$$

- What can we say about this random variable $\widehat{\theta}_{\text{MLE}}$?

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\mathrm{MLE}}$ as a random variable, where there is a ground truth parameter ϑ^* that generates the data
  $\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

  $$\hat{\theta}_{\mathrm{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\widehat{\theta}_{\mathrm{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

"true predictor"

$$\mathrm{Bias}(\widehat{\theta}_{\mathrm{MLE}}) := \mathbb{E}_{\mathscr{D}\sim P_{\theta^*}}[\hat{\theta}_{\mathrm{MLE}}] - \theta^* = E\left[\frac{k}{n}\right] - \theta^*$$

Unbiased means bias = 0

$$= \frac{\theta^* \cancel{n}}{\cancel{n}} - \theta^* = 0$$

- Expectation describes how the estimator behaves on average

How many flips do I need?

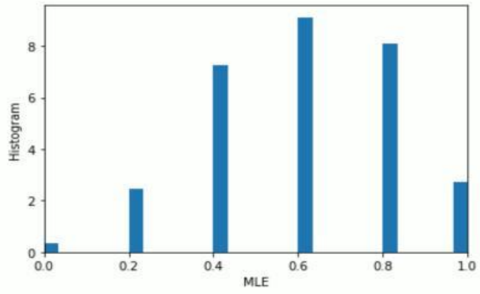
- Consider running many experiments with $\theta^* = \frac{3}{5}$, and observe many instances of the random variable

$$\widehat{\theta}_{MLE} = \frac{k}{n}$$

How many flips do I need?

- Consider running many experiments with $\theta^* = \frac{3}{5}$, and observe many instances of the random variable

$$\widehat{\theta}_{MLE} = \frac{k}{n}$$



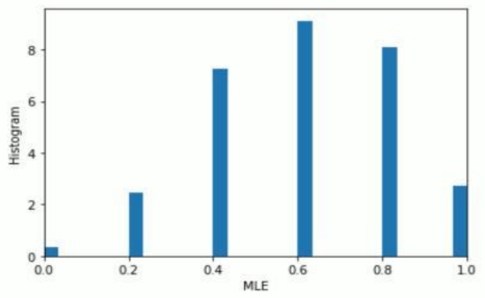
- Client: I flipped the coin 5 times and got 2 heads.

$$\widehat{\theta}_{MLE} =$$

How many flips do I need?

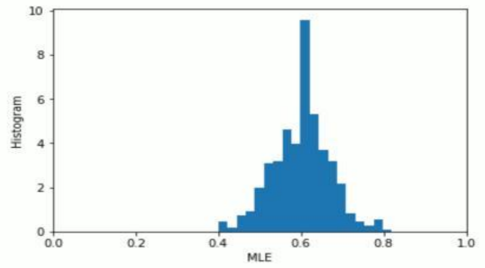
- Consider running many experiments with $\theta^* = \frac{3}{5}$, and observe many instances of the random variable

$$\widehat{\theta}_{MLE} = \frac{k}{n}$$



- Client: I flipped the coin 5 times and got 2 heads.

$$\widehat{\theta}_{MLE} =$$

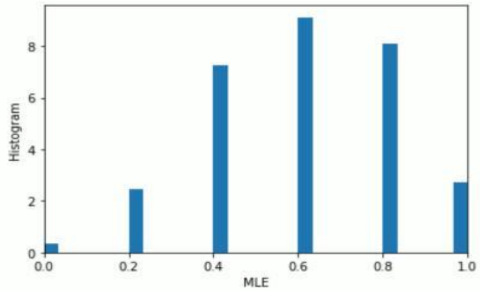


- Client: I flipped the coin 50 times and got 30 heads

$$\widehat{\theta}_{MLE} =$$

How many flips do I need?

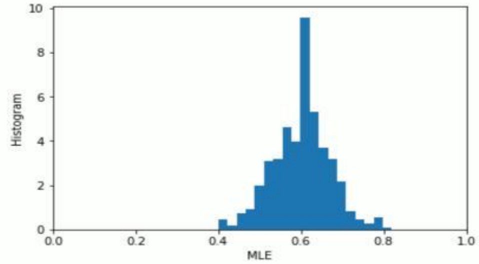
- Consider running many experiments with $\theta^* = \frac{3}{5}$, and observe many instances of the random variable



$$\widehat{\theta}_{MLE} = \frac{k}{n}$$

- Client: I flipped the coin 5 times and got 2 heads.

$$\widehat{\theta}_{MLE} =$$



- Client: I flipped the coin 50 times and got 30 heads

$$\widehat{\theta}_{MLE} =$$

- Client: they are both unbiased, which one is right? Why?
- Variance goes down with larger *n*

$$\sqrt{\mathrm{Var}(\widehat{\theta}_{\mathrm{MLE}})} = \sqrt{\frac{\theta^*(1-\theta^*)}{n}}$$

Fundamental machine learning truth

- More data -> better performance
  - “The Bitter Lesson”
    - https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf

Maximum Likelihood Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

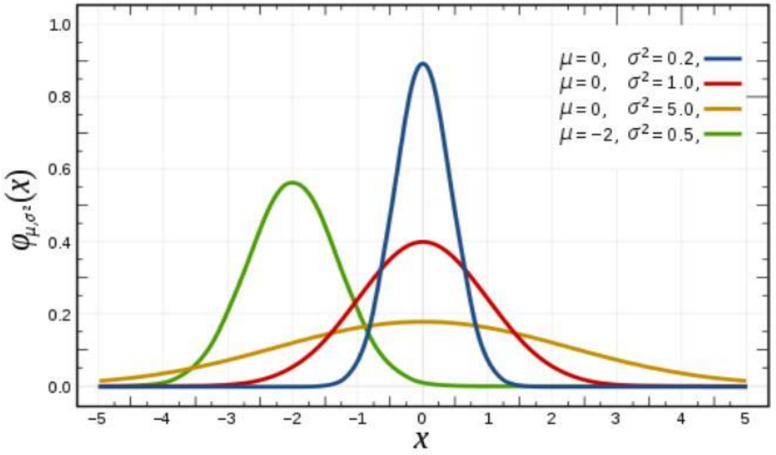
What about continuous variables?

- Client: What if I am measuring a continuous variable?
- You: Let me tell you about Gaussians…

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

PDF of Gaussian

(good candidate for formula sheet)



Given a set of i.i.d. samples from a Gaussian, fit what parameters?

$\theta = [\mu, \sigma]$

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \quad \rightarrow \quad Y \sim N(a\mu+b, a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \quad \rightarrow \quad Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

MLE for Gaussian

- Prob. of i.i.d. samples D=$\{x_1,\ldots,x_n\}$:

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma)$$

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_n\}$:

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma)$$

Likelihood: $$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Wait, why can I just multiply all samples together?

→ By i.i.d. assumption

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_n\}$:

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma)$$

Likelihood: $$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Wait, why can I just multiply all samples together?

→ By i.i.d. assumption

- Log-likelihood of data:

LL: $$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

- What is $\widehat{\theta}_{MLE}$ for $\theta = (\mu,\sigma^2)$? Draw a picture!

MLE for Gaussian

Generate $\mathcal{D} = \{x_1, \ldots, x_n\}$, where

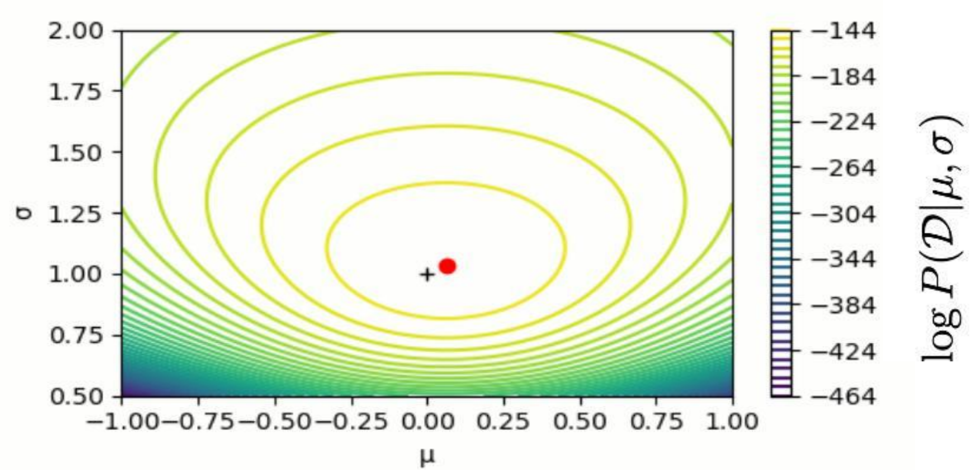
$$n = 100$$

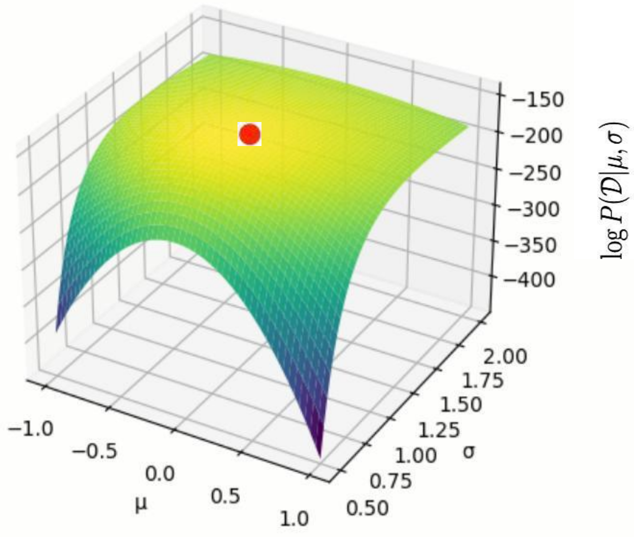
$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu = 0$$

$$\sigma^2 = 1$$

$$\log P(\mathcal{D}|\mu, \sigma) = -n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$





+ $(\mu_{True}, \sigma_{True})$

● $(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE})$

MLE for Mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu}\left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

MLE for Mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu}\left[\cancel{-n \log(\sigma\sqrt{2\pi})} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

$$= \cancel{-}\sum_{i=1}^{n} \frac{\cancel{-}\cancel{2}(x_i - \mu)}{\cancel{2}\sigma^2}$$

$$= \frac{-n\mu + \sum_{i=1}^{n} x_i}{\sigma^2} = 0$$

MLE for Mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu}\left[-n \cancel{\log(\sigma\sqrt{2\pi})} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \cancel{-} \sum_{i=1}^{n} \frac{\cancel{-2}(x_i - \mu)}{\cancel{2}\sigma^2}$$

$$= \frac{-n\mu + \sum_{i=1}^{n} x_i}{\sigma^2} = 0$$

$$\boxed{\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

empirical mean!

MLE for Variance

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma} \left[-n \log(\sigma \sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

MLE for Variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma}\left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

MLE for Variance

\# reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{\cancel{2}(x_i-\mu)^2}{\cancel{2}\sigma^3}$$

$$= \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\cancel{\sigma^3}} = 0$$

MLE for Variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial\sigma}\log P(\mathcal{D}\mid\mu,\sigma) = \frac{\partial}{\partial\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{\cancel{2}(x_i-\mu)^2}{\cancel{2}\sigma^3}$$

$$= \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\cancel{\sigma^3}} = 0$$

$$\boxed{= \sigma^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\mathrm{MLE}})^2}$$

Learning Gaussian parameters

- MLE:

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

- MLE for the variance of a Gaussian is biased

$$\mathbb{E}[\widehat{\sigma^2}_{MLE}] \neq \sigma^2$$

- Unbiased variance estimator:

$$\widehat{\sigma^2}_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

Maximum Likelihood Estimation

Observe $X_1, X_2, \ldots, X_n$ **drawn IID from** $f(x;\theta)$ **for some "true"** $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

Under benign assumptions, as the number of observations $n \to \infty$ we have $\hat{\theta}_{MLE} \to \theta_*$

The MLE is a "recipe" that begins with a *model* for data $f(x;\theta)$

Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., Bernoulli
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE

Applications Preview

Maximum Likelihood Estimation

Why is it useful to recover the “true” parameters θ_* of a probabilistic model?

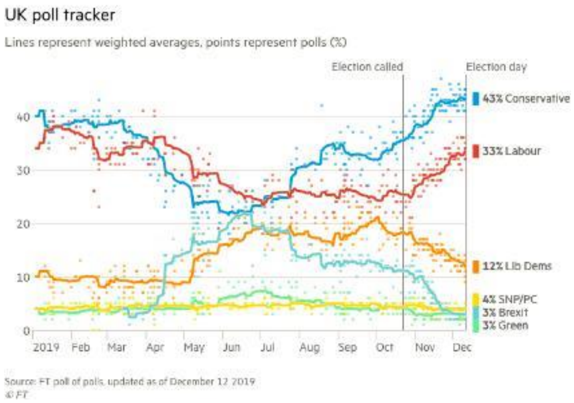
- **Estimation** of the parameters θ_* is the goal
- Help **interpret** or summarize large datasets
- Make **predictions** about future data
- **Generate** new data $X \sim f(\,\cdot\,; \hat{\theta}_{\text{MLE}})$

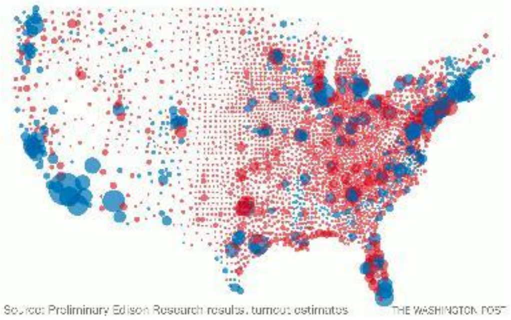
Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Opinion polls
How does the greater population feel about an issue?
Correct for over-sampling?

- θ_* is "true" average opinion
- $X_1, X_2, \ldots$ are sample calls



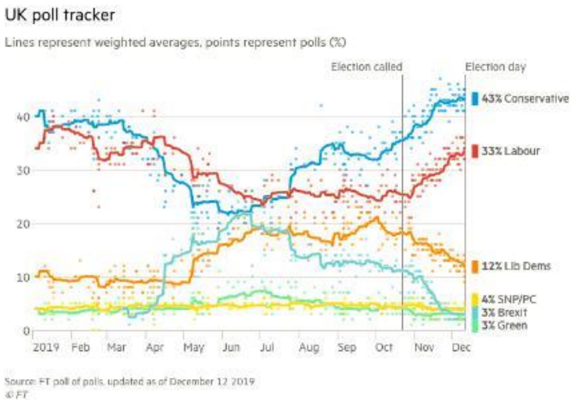


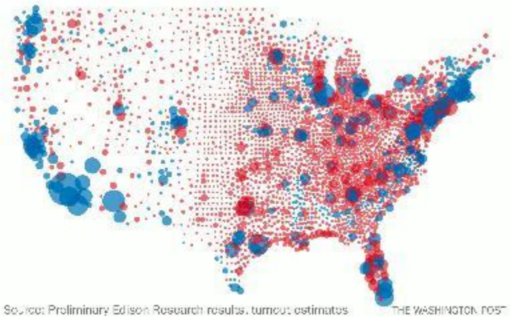
Estimation

Observe $X_1, X_2, \ldots, X_n$ **drawn IID from** $f(x;\theta)$ **for some "true"** $\theta = \theta_*$

Opinion polls
How does the greater population feel about an issue?
Correct for over-sampling?

- θ_* is "true" average opinion
- $X_1, X_2, \ldots$ are sample calls





A/B testing
How do we figure out which ad results in more click-through?

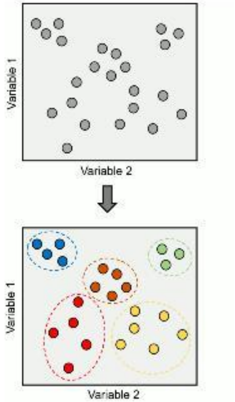
- θ_* are the "true" average rates
- $X_1, X_2, \ldots$ are binary "clicks"

Interpret

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Customer segmentation / clustering
Can we identify distinct groups of customers by their behavior?

- θ_* describes "center" of distinct groups
- $X_1, X_2, \ldots$ are individual customers

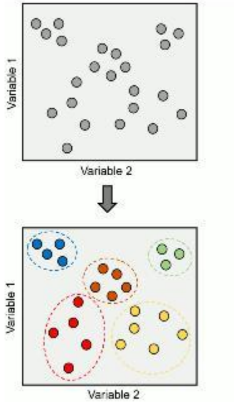


Interpret

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Customer segmentation / clustering
Can we identify distinct groups of customers by their behavior?

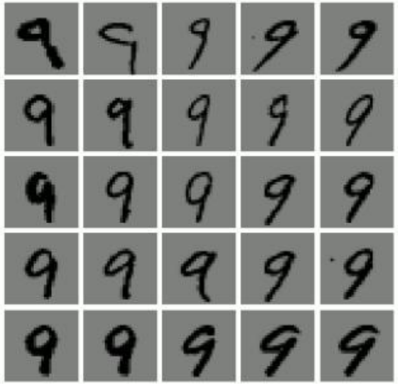
- θ_* describes "center" of distinct groups
- $X_1, X_2, \ldots$ are individual customers



Data exploration

What are the degrees of freedom of the dataset?

- θ_* describes the principle directions of variation
- $X_1, X_2, \ldots$ are the individual images



Predict

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Content recommendation
Can we predict how much someone will like a movie based on past ratings?

- θ_* describes user's preferences
- $X_1, X_2, \ldots$ are (movie, rating) pairs

Predict

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Content recommendation
Can we predict how much someone will like a movie based on past ratings?

- θ_* describes user's preferences
- $X_1, X_2, \ldots$ are (movie, rating) pairs

Object recognition / classification
Identify a flower given just its picture?

- θ_* describes the characteristics of each kind of flower
- $X_1, X_2, \ldots$ are the (image, label) pairs

(a) (b) (c)

Figure 1.1: Three types of Iris flowers: Setosa, Versicolor and Virginica. Used with kind permission of Dennis Kramb and SIGNA.

index	sl	sw	pl	pw	label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
	...				
50	7.0	3.2	4.7	1.4	Versicolor
	...				
149	5.9	3.0	5.1	1.8	Virginica

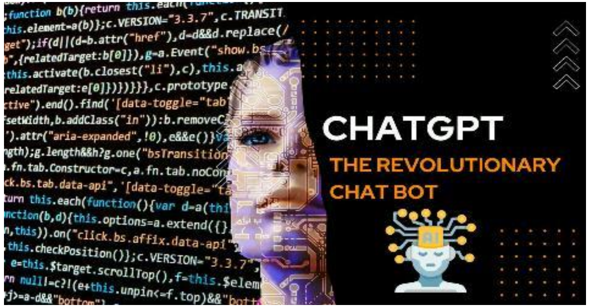
Generate

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Text generation
Can AI generate text that could have been written like a human?

- θ_* describes language structure
- $X_1, X_2, \ldots$ are text snippets found online

"Kaia the dog wasn't a natural pick to go to mars. No one could have predicted she would…"



MLE!

https://chat.openai.com/chat

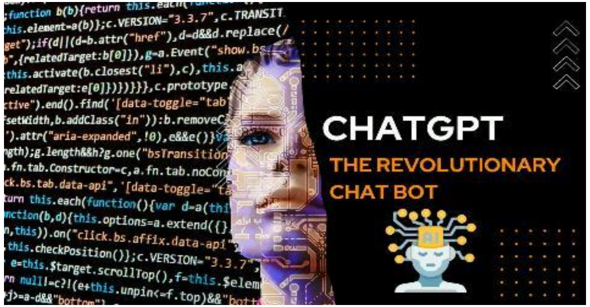
Generate

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Text generation
Can AI generate text that could have been written like a human?

- θ_* describes language structure
- $X_1, X_2, \ldots$ are text snippets found online

"Kaia the dog wasn't a natural pick to go to mars. No one could have predicted she would…"



MLE!

https://chat.openai.com/chat

Image to text generation
Can AI generate an image from a prompt?

- θ_* describes the coupled structure of images and text
- $X_1, X_2, \ldots$ are the (image, caption) pairs found online

"dog talking on cell phone under water, oil painting"

https://labs.openai.com/

Acknowledgement

Most of the slides are based on the ML courses of:

- Natasha Jaques, University of Washington

THANK YOU