Maximum Likelihood and Maximum-a-Posteriori Estimation

CS21206: Foundations of AI and ML

Abir Das & Ayan Chaudhury

IIT Kharagpur

Jan 15, 2026

Agenda

- Inferring probability models from data
- Estimating model parameters with Maximum Likelihood
- Estimating model parameters with Maximum *a-posteriori*

Resources

- "Probability and Statistics for Computer Science", David Forsyth - [PSCE]
- "Applied Machine Learning", David Forsyth - [AML]
- "Machine Learning: A Probabilistic Perspective", Kevin P. Murphy - [MLAPP]:

Bayes Rule (RECAP)

§ Let the events $B_1, B_2, \cdots, B_n$ partitions a sample space such that each of the $P(B_i)$'s are non-negative. The Bayes' rule states,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \tag{1}$$

§ A can be thought of as the "effect" and B_i's are several "causes".

§ From the probability of the effect due to the causes $\big(P(A|B_i)\big)$ and the probability of the cause $\big(P(B_i)\big)$ to occur frequently, the probability of a cause is the reason behind the effect $\big(P(B_i|A)\big)$ is computed.

- $P(A|B_i) \rightarrow$ "Likelihood"
- $P(B_i) \rightarrow$ "Prior"
- $P(B_i|A) \rightarrow$ "Posterior"

Bayes Rule (RECAP)

§ Let the events $B_1, B_2, \cdots, B_n$ partitions a sample space such that each of the $P(B_i)$'s are non-negative. The Bayes' rule states,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \tag{1}$$

§ A can be thought of as the "effect" and B_i's are several "causes".

§ From the probability of the effect due to the causes $\big(P(A|B_i)\big)$ and the probability of the cause $\big(P(B_i)\big)$ to occur frequently, the probability of a cause is the reason behind the effect $\big(P(B_i|A)\big)$ is computed.

- $P(A|B_i) \rightarrow$ "Likelihood"
- $P(B_i) \rightarrow$ "Prior"
- $P(B_i|A) \rightarrow$ "Posterior"

Applying Bayes Rule

- The probabilities that X, Y and Z becoming the director of a technological institute are $0.3, 0.5$ and 0.2 respectively. The probability that there will not be classes on the tech-fest day if X, Y and Z is the director are $0.4, 0.6$ and 0.1 respectively. Given that the classes are called off on the tech-fest day, find the probability that Y has been appointed as the director.
- $P(\text{No Class}|X) = 0.4, P(\text{No Class}|Y) = 0.6, P(\text{No Class}|Z) = 0.1$
- $P(X) = 0.3, P(Y) = 0.5, P(Z) = 0.2$

$$P(Y|\text{No Class}) = \frac{P(\text{No Class}|Y)P(Y)}{P(\text{No Class}|X)P(X) + P(\text{No Class}|Y)P(Y) + P(\text{No Class}|Z)P(Z)}$$

$$= \frac{0.6 \times 0.5}{0.4 \times 0.3 + 0.6 \times 0.5 + 0.1 \times 0.2} = 0.68 \quad (2)$$

Applying Bayes Rule

- The probabilities that X, Y and Z becoming the director of a technological institute are $0.3, 0.5$ and 0.2 respectively. The probability that there will not be classes on the tech-fest day if X, Y and Z is the director are $0.4, 0.6$ and 0.1 respectively. Given that the classes are called off on the tech-fest day, find the probability that Y has been appointed as the director.
- $P(\text{No Class}|X) = 0.4, P(\text{No Class}|Y) = 0.6, P(\text{No Class}|Z) = 0.1$
- $P(X) = 0.3, P(Y) = 0.5, P(Z) = 0.2$

$$P(Y|\text{No Class}) = \frac{P(\text{No Class}|Y)P(Y)}{P(\text{No Class}|X)P(X) + P(\text{No Class}|Y)P(Y) + P(\text{No Class}|Z)P(Z)}$$
$$= \frac{0.6 \times 0.5}{0.4 \times 0.3 + 0.6 \times 0.5 + 0.1 \times 0.2} = 0.68 \quad (2)$$

- what does all this have to do with machine learning?

Applying Bayes Rule

- The probabilities that X, Y and Z becoming the director of a technological institute are $0.3, 0.5$ and 0.2 respectively. The probability that there will not be classes on the tech-fest day if X, Y and Z is the director are $0.4, 0.6$ and 0.1 respectively. Given that the classes are called off on the tech-fest day, find the probability that Y has been appointed as the director.
- $P(\text{No Class}|X) = 0.4, P(\text{No Class}|Y) = 0.6, P(\text{No Class}|Z) = 0.1$
- $P(X) = 0.3, P(Y) = 0.5, P(Z) = 0.2$

$$P(Y|\text{No Class}) = \frac{P(\text{No Class}|Y)P(Y)}{P(\text{No Class}|X)P(X) + P(\text{No Class}|Y)P(Y) + P(\text{No Class}|Z)P(Z)}$$
$$= \frac{0.6 \times 0.5}{0.4 \times 0.3 + 0.6 \times 0.5 + 0.1 \times 0.2} = 0.68 \quad (2)$$

- what does all this have to do with machine learning?
- instead of $F : X \rightarrow Y$, learn $P(Y|X)$.

Applying Bayes Rule

- The probabilities that X, Y and Z becoming the director of a technological institute are $0.3, 0.5$ and 0.2 respectively. The probability that there will not be classes on the tech-fest day if X, Y and Z is the director are $0.4, 0.6$ and 0.1 respectively. Given that the classes are called off on the tech-fest day, find the probability that Y has been appointed as the director.
- $P(\text{No Class}|X) = 0.4, P(\text{No Class}|Y) = 0.6, P(\text{No Class}|Z) = 0.1$
- $P(X) = 0.3, P(Y) = 0.5, P(Z) = 0.2$

$$P(Y|\text{No Class}) = \frac{P(\text{No Class}|Y)P(Y)}{P(\text{No Class}|X)P(X) + P(\text{No Class}|Y)P(Y) + P(\text{No Class}|Z)P(Z)}$$
$$= \frac{0.6 \times 0.5}{0.4 \times 0.3 + 0.6 \times 0.5 + 0.1 \times 0.2} = 0.68 \quad (2)$$

- what does all this have to do with machine learning?
- instead of $F : X \rightarrow Y$, learn $P(Y|X)$.

Applying Bayes Rule

- The probabilities that X, Y and Z becoming the director of a technological institute are $0.3, 0.5$ and 0.2 respectively. The probability that there will not be classes on the tech-fest day if X, Y and Z is the director are $0.4, 0.6$ and 0.1 respectively. Given that the classes are called off on the tech-fest day, find the probability that Y has been appointed as the director.
- $P(\text{No Class}|X) = 0.4, P(\text{No Class}|Y) = 0.6, P(\text{No Class}|Z) = 0.1$
- $P(X) = 0.3, P(Y) = 0.5, P(Z) = 0.2$

$$P(Y|\text{No Class}) = \frac{P(\text{No Class}|Y)P(Y)}{P(\text{No Class}|X)P(X) + P(\text{No Class}|Y)P(Y) + P(\text{No Class}|Z)P(Z)}$$
$$= \frac{0.6 \times 0.5}{0.4 \times 0.3 + 0.6 \times 0.5 + 0.1 \times 0.2} = 0.68 \quad (2)$$

- what does all this have to do with machine learning?
- instead of $F : X \rightarrow Y$, learn $P(Y|X)$.

Estimating Probabilities

- With your knowledge of probability, if you are given a model/distribution, with all the involved probabilities, you can make predictions.
- But where do these probability values come from?
- You have to learn/estimate them from data/observations.
  - maximum likelihood estimates (MLE)
  - maximum a posteriori estimates (MAP)
- Rooted deep in *parameter estimation theory*.

Estimating Probabilities

- With your knowledge of probability, if you are given a model/distribution, with all the involved probabilities, you can make predictions.
- But where do these probability values come from?
- You have to learn/estimate them from data/observations.
  - maximum likelihood estimates (MLE)
  - maximum a posteriori estimates (MAP)
- Rooted deep in *parameter estimation theory*.
- Wait! What are parameters?

Estimating Probabilities

- With your knowledge of probability, if you are given a model/distribution, with all the involved probabilities, you can make predictions.
- But where do these probability values come from?
- You have to learn/estimate them from data/observations.
  - maximum likelihood estimates (MLE)
  - maximum a posteriori estimates (MAP)
- Rooted deep in *parameter estimation theory*.
- Wait! What are parameters?

Estimating Probabilities

- With your knowledge of probability, if you are given a model/distribution, with all the involved probabilities, you can make predictions.
- But where do these probability values come from?
- You have to learn/estimate them from data/observations.
  - maximum likelihood estimates (MLE)
  - maximum a posteriori estimates (MAP)
- Rooted deep in *parameter estimation theory*.
- Wait! What are parameters?

What are Parameters?

- Consider some probability distributions.
  - Ber(p) $\theta = p$
  - Poisson(λ) $\theta = \lambda$
  - Exp(λ) $\theta = \lambda$
  - Bin(n, p) $\boldsymbol{\theta} = (n, p)$
  - Uni(a, b) $\boldsymbol{\theta} = (a, b)$
  - $\mathcal{N}(\mu, \sigma^2)$ $\boldsymbol{\theta} = (\mu, \sigma^2)$
- These are called 'parametric' models.
- Usually parameters are denoted by the symbol θ.
- Note, θ can be a vector of parameters also.

What are Parameters?

- Consider some probability distributions.
  - Ber(p) $\quad \theta = p$
  - Poisson(λ) $\quad \theta = \lambda$
  - Exp(λ) $\quad \theta = \lambda$
  - Bin(n, p) $\quad \boldsymbol{\theta} = (n, p)$
  - Uni(a, b) $\quad \boldsymbol{\theta} = (a, b)$
  - $\mathcal{N}(\mu, \sigma^2)$ $\quad \boldsymbol{\theta} = (\mu, \sigma^2)$
- These are called 'parametric' models.
- Usually parameters are denoted by the symbol θ.
- Note, $\boldsymbol{\theta}$ can be a vector of parameters also.

Why do we care?

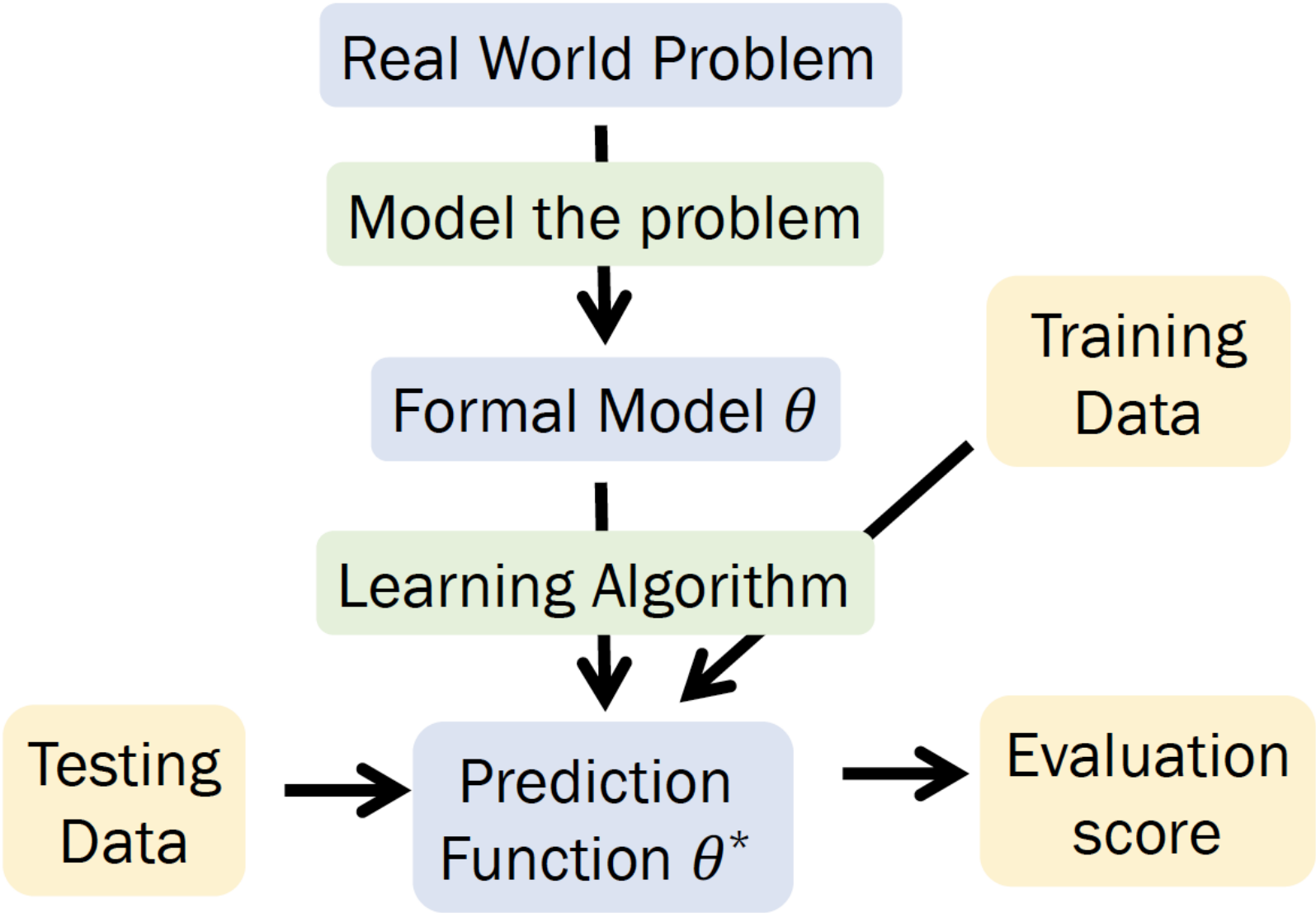


Fig credit: Stanford Course CS109: Lecture 21

Why do we care?

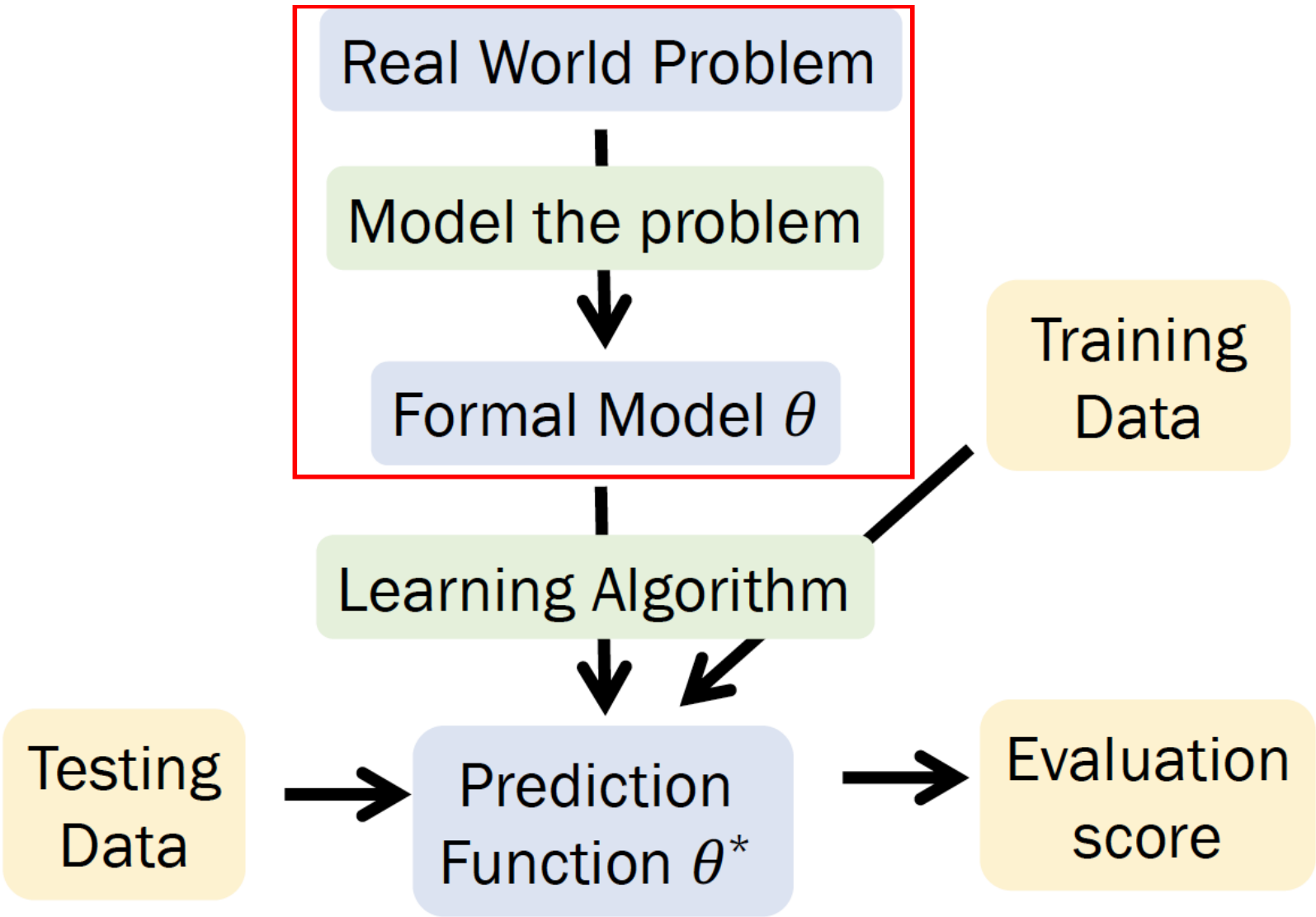


Fig credit: Stanford Course CS109: Lecture 21

Why do we care?

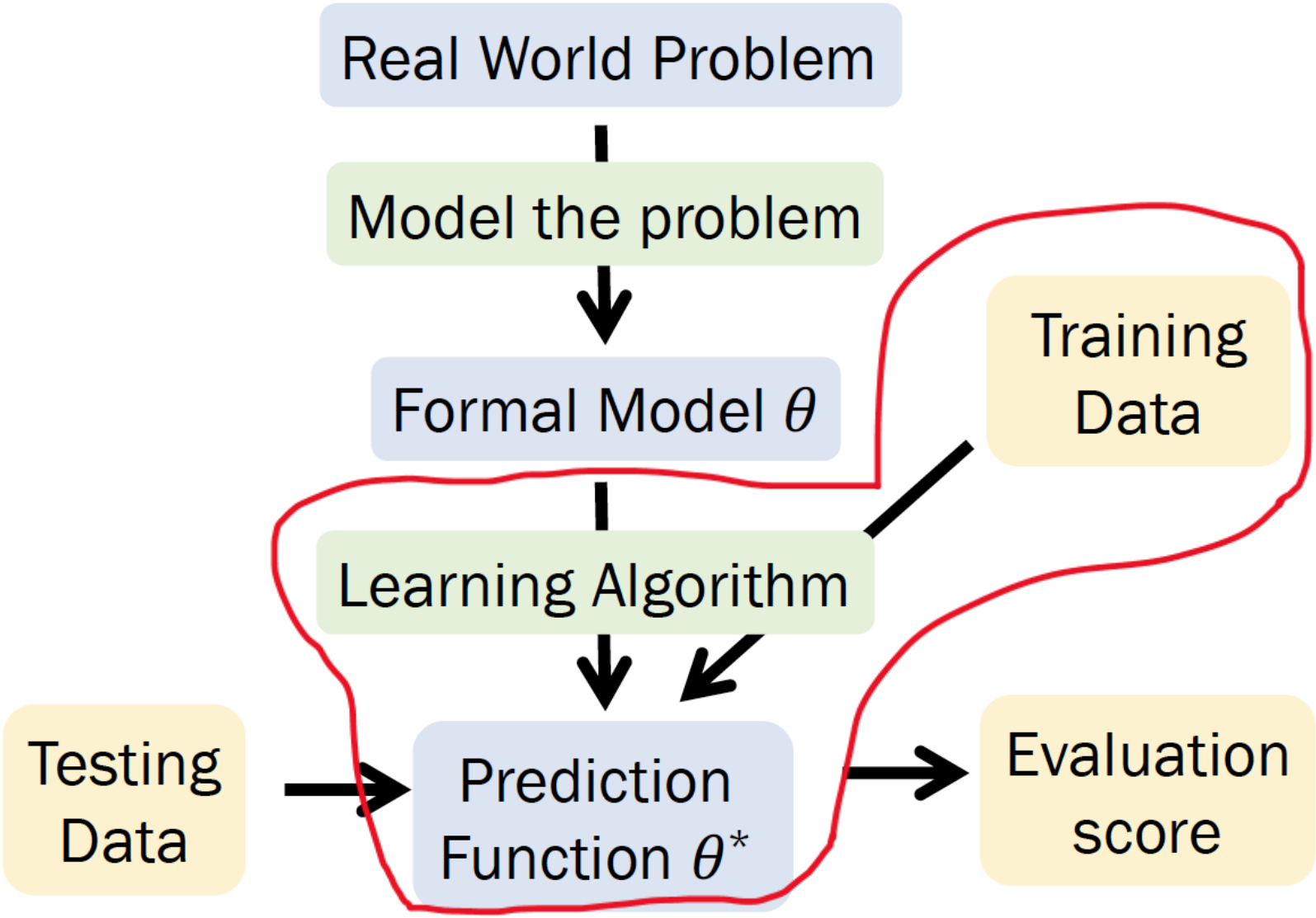


Fig credit: Stanford Course CS109: Lecture 21

Maximum Likelihood Estimation

We have $n = 3$ data points $y_1 = \mathbf{1}, y_2 = \mathbf{0.5}, y_3 = \mathbf{1.5}$, which are independent and Gaussian with unknown $mean = \theta$ and $variance = 1$:

$$y_i \sim \mathcal{N}(\theta, 1)$$

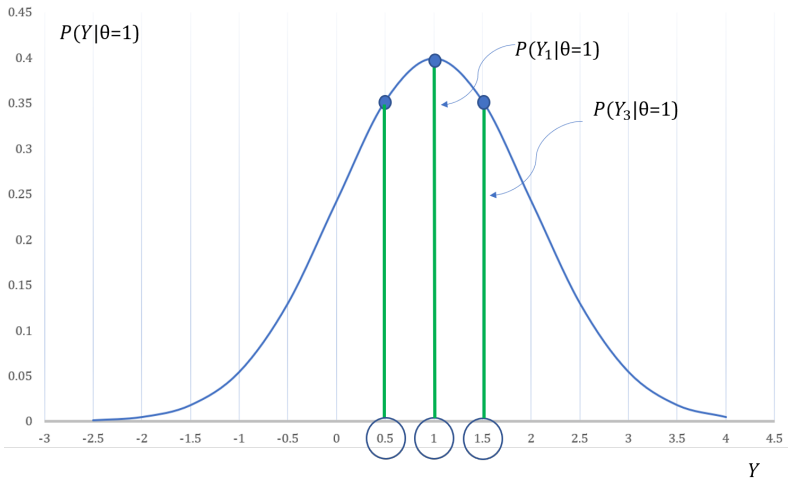
with **likelihood** $\boldsymbol{P}(y_1, y_2, y_3; \theta) = \boldsymbol{P}(y_1; \theta)\boldsymbol{P}(y_2; \theta)\boldsymbol{P}(y_3; \theta)$. Consider two guesses of θ, 1 and 2.5. Which has higher likelihood (probability of generating the three observations)?

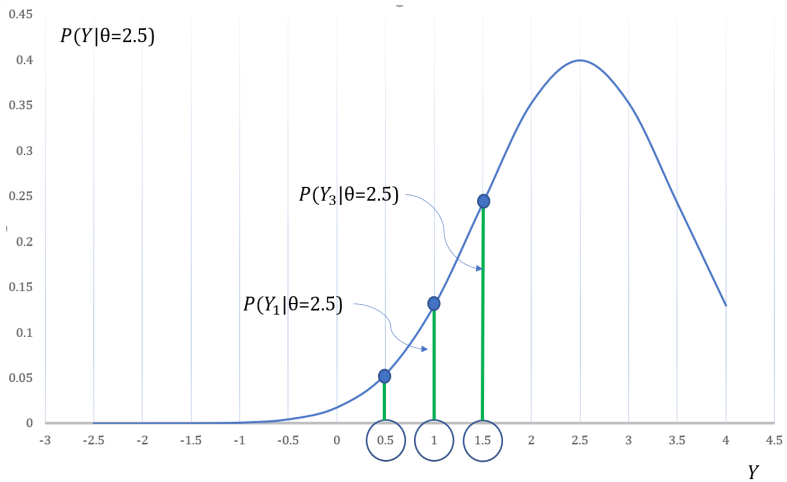
Maximum Likelihood Estimation

We have $n = 3$ data points $y_1 = \mathbf{1}, y_2 = \mathbf{0.5}, y_3 = \mathbf{1.5}$, which are independent and Gaussian with unknown $mean = \theta$ and $variance = 1$:

$$y_i \sim \mathcal{N}(\theta, 1)$$

with **likelihood** $\boldsymbol{P}(y_1, y_2, y_3; \theta) = \boldsymbol{P}(y_1; \theta)\boldsymbol{P}(y_2; \theta)\boldsymbol{P}(y_3; \theta)$. Consider two guesses of θ, 1 and 2.5. Which has higher likelihood (probability of generating the three observations)?





Finding the θ that maximizes the likelihood is equivalent to moving the Gaussian until the product of 3 green bars (likelihood)is maximized.

Slide Motivation: *Nando de Freitas [Link]*

Maximum Likelihood Estimation

- **Intuitive example: Estimate a coin toss**
  I have seen 3 flips of heads, 2 flips of tails, what is the chance of head (or tail) of my next flip?
- **Model:**
  Each flip is a Bernoulli random variable x.
  x can take only *two* values: 1(head), 0(tail)

$$p(x;\theta) = \begin{cases} \theta, & if\ x = 1 \\ 1-\theta, & if\ x = 0 \end{cases} \tag{3}$$

  where, $\theta \in [0, 1]$, is a parameter to be defined from data
- We can write this probability more succinctly as

$$p(x;\theta) = \theta^x (1-\theta)^{1-x} \tag{4}$$

Maximum Likelihood: Example

- Let us now assume, that we have flipped the coin a few times and got the results $x_1, ..., x_n$, which are either 0 or 1. The question is what is the value of the probability θ?

Maximum Likelihood: Example

- Let us now assume, that we have flipped the coin a few times and got the results $x_1, ..., x_n$, which are either 0 or 1. The question is what is the value of the probability θ?
- Intuitively, one could assume that it is the number of heads we got divided by the total number of coin throws.

Maximum Likelihood: Example

- Let us now assume, that we have flipped the coin a few times and got the results $x_1, ..., x_n$, which are either 0 or 1. The question is what is the value of the probability θ?
- Intuitively, one could assume that it is the number of heads we got divided by the total number of coin throws.
- We will prove in the following that the intuition in this case is correct, by proving that the guess $\theta = \sum_i x_i/n$ is the "most likely" value for the real θ.

Maximum Likelihood: Example

- Let us now assume, that we have flipped the coin a few times and got the results $x_1, ..., x_n$, which are either 0 or 1. The question is what is the value of the probability θ?
- Intuitively, one could assume that it is the number of heads we got divided by the total number of coin throws.
- We will prove in the following that the intuition in this case is correct, by proving that the guess $\theta = \sum_i x_i/n$ is the "most likely" value for the real θ.
- Then the joint probability is

$$f(x_1, ..., x_n; \theta) = \prod_i f(x_i; \theta) = \theta^{\sum_i x_i} (1-\theta)^{n - \sum_i x_i} \tag{5}$$

Maximum Likelihood: Example

- We now want to find the θ which makes this probability the highest.
- It is easier to maximize the *log* of the joint probabilities $log\,\mathcal{L}(\theta) = \sum_i x_i log\,\theta + (n - \sum_i x_i) log\,(1-\theta)$, which yields the same result, since the *log* is monotonously increasing.
- As we may remember, maximizing a function means setting its first derivative to 0.

$$\begin{aligned}
\frac{\partial log\,\mathcal{L}(\theta)}{\partial\theta} &= \frac{\sum_i x_i}{\theta} - \frac{(n - \sum_i x_i)}{1-\theta} \\
&= \frac{(1-\theta)\sum_i x_i - \theta n + \theta \sum_i x_i}{\theta(1-\theta)} \\
&= \frac{\sum_i x_i - \theta n}{\theta(1-\theta)} = 0 \\
\implies \theta &= \frac{\sum_i x_i}{n}
\end{aligned} \tag{6}$$

Maximum Likelihood Estimation: Recipe

- In general, we have observations, $\mathcal{D} = \{u^{(1)}, u^{(2)}, \cdots, u^{(N)}\}$
- We assume data is generated by some distribution $U \sim p(U; \theta)$
- Compute the likelihood function

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p(u^{(i)}; \theta) \leftarrow \texttt{Likelihood Function} \tag{7}$$

$$\begin{aligned} \theta_{ML} &= \arg\max_{\theta} \mathcal{L}(\theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log p(u^{(i)}; \theta) \leftarrow \texttt{Log Likelihood} \end{aligned} \tag{8}$$

- $\log(f(x))$ is monotonic/ increasing, same $\arg\max$ as $f(x)$

Maximum Likelihood with Poisson

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \text{Poisson}(\lambda)$.
- Goal: Use Maximum Likelihood Estimation to find the optimal λ.
- PMF can be written as $p(x_i; \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $\mathcal{L}(\lambda) = p(x_1, x_2, \cdots, x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

Maximum Likelihood with Poisson

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \text{Poisson}(\lambda)$.
- Goal: Use Maximum Likelihood Estimation to find the optimal λ.
- PMF can be written as $p(x_i; \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $\mathcal{L}(\lambda) = p(x_1, x_2, \cdots, x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Log likelihood

$$\log \mathcal{L}(\lambda) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} \left(-\lambda + x_i \log \lambda - \log x_i!\right)$$

Maximum Likelihood with Poisson

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \text{Poisson}(\lambda)$.
- Goal: Use Maximum Likelihood Estimation to find the optimal λ.
- PMF can be written as $p(x_i; \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $\mathcal{L}(\lambda) = p(x_1, x_2, \cdots, x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Log likelihood

$$\log \mathcal{L}(\lambda) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} (-\lambda + x_i \log \lambda - \log x_i!)$$

- Differentiate w.r.t. λ, and set to 0

$$\frac{\partial log\, \mathcal{L}(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \left(-1 + \frac{x_i}{\lambda}\right) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{9}$$

Maximum Likelihood with Poisson

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \text{Poisson}(\lambda)$.
- Goal: Use Maximum Likelihood Estimation to find the optimal λ.
- PMF can be written as $p(x_i; \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $\mathcal{L}(\lambda) = p(x_1, x_2, \cdots, x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Log likelihood

$$\log \mathcal{L}(\lambda) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} \left(-\lambda + x_i \log \lambda - \log x_i!\right)$$

- Differentiate w.r.t. λ, and set to 0

$$\frac{\partial log\, \mathcal{L}(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \left(-1 + \frac{x_i}{\lambda}\right) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{9}$$

Maximum Likelihood with Poisson

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \text{Poisson}(\lambda)$.
- Goal: Use Maximum Likelihood Estimation to find the optimal λ.
- PMF can be written as $p(x_i; \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $\mathcal{L}(\lambda) = p(x_1, x_2, \cdots, x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Log likelihood

$$\log \mathcal{L}(\lambda) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} \left(-\lambda + x_i \log \lambda - \log x_i!\right)$$

- Differentiate w.r.t. λ, and set to 0

$$\frac{\partial log\, \mathcal{L}(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \left(-1 + \frac{x_i}{\lambda}\right) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{9}$$

Maximum Likelihood with Normal

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
- Goal: Use Maximum Likelihood Estimation to find optimal μ and σ^2.
- PDF can be written as $p(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Likelihood: $\mathcal{L}(\mu, \sigma^2) = p(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

Maximum Likelihood with Normal

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
- Goal: Use Maximum Likelihood Estimation to find optimal μ and σ^2.
- PDF can be written as $p(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Likelihood: $\mathcal{L}(\mu, \sigma^2) = p(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Log likelihood:

$$\log \mathcal{L}(\mu, \sigma^2) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = -\sum_{i=1}^{n} \frac{1}{2} \log 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= \text{const.} - \sum_{i=1}^{n} \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \tag{10}$$

Maximum Likelihood with Normal

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
- Goal: Use Maximum Likelihood Estimation to find optimal μ and σ^2.
- PDF can be written as $p(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Likelihood: $\mathcal{L}(\mu, \sigma^2) = p(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Log likelihood:

$$\log \mathcal{L}(\mu, \sigma^2) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = -\sum_{i=1}^{n} \frac{1}{2} \log 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= \text{const.} - \sum_{i=1}^{n} \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \tag{10}$$

- Differentiate $\log \mathcal{L}(\mu, \sigma^2)$ w.r.t. μ and σ separately and set to 0.

Maximum Likelihood with Normal

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
- Goal: Use Maximum Likelihood Estimation to find optimal μ and σ^2.
- PDF can be written as $p(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Likelihood: $\mathcal{L}(\mu, \sigma^2) = p(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Log likelihood:

$$\log \mathcal{L}(\mu, \sigma^2) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = -\sum_{i=1}^{n} \frac{1}{2} \log 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= \text{ const.} - \sum_{i=1}^{n} \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \tag{10}$$

- Differentiate $\log \mathcal{L}(\mu, \sigma^2)$ w.r.t. μ and σ separately and set to 0.

Maximum Likelihood with Normal

- Consider we observed I.I.D. random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
- Goal: Use Maximum Likelihood Estimation to find optimal μ and σ^2.
- PDF can be written as $p(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Likelihood: $\mathcal{L}(\mu, \sigma^2) = p(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Log likelihood:

$$\log \mathcal{L}(\mu, \sigma^2) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = -\sum_{i=1}^{n} \frac{1}{2} \log 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= \text{ const.} - \sum_{i=1}^{n} \log \sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \tag{10}$$

- Differentiate $\log \mathcal{L}(\mu, \sigma^2)$ w.r.t. μ and σ separately and set to 0.

Maximum Likelihood with Normal

- $\log \mathcal{L}(\mu, \sigma^2) = \text{ const. } - \sum_{i=1}^{n} \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$

$$\frac{\partial log\, \mathcal{L}(\mu, \sigma^2)}{\partial \mu} = 0 - 0 + \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} (x_i - \mu) = 0 \tag{11}$$

- So, $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$, again the *sample mean*!
- Sometimes this is denoted as $\hat{\mu}_{MLE}$

Maximum Likelihood with Normal

- $\log \mathcal{L}(\mu, \sigma^2) = \text{ const.} - \sum_{i=1}^{n} \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$

$$\frac{\partial log\, \mathcal{L}(\mu, \sigma^2)}{\partial \mu} = 0 - 0 + \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} (x_i - \mu) = 0 \tag{11}$$

- So, $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$, again the *sample mean*!
- Sometimes this is denoted as $\hat{\mu}_{MLE}$

Maximum Likelihood with Normal

- $\log \mathcal{L}(\mu, \sigma^2) = \text{ const.} - \sum_{i=1}^{n} \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$

$$\frac{\partial log\, \mathcal{L}(\mu, \sigma^2)}{\partial \sigma} = 0 - \frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

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

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{12}$$

- So, $\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})^2$ – sample variance!

Why MLE is Loved So Much!

- Some desirable properties of an estimator
  - Giving the true value of the parameters on average.
  - The estimated value does not vary too much about the true value.
  - Even if the data doesn't follow the particular distribution we thought it to follow, still the estimated value corresponds to the model that is the closest to the source of the data.
- These correspond to the following properties of estimators.
  - Unbiased estimator.
  - Low (or minimum possible) variance estimator.
  - Consistent estimator (given the number of data samples $\rightarrow \infty$).
- Detailed discussion is beyond the scope.
- MLE is "asymptotically optimal". Its great properties (consistency and efficiency) are guaranteed as data $\rightarrow \infty$
- but it can be problematic with small datasets

Why MLE is Loved So Much!

- Some desirable properties of an estimator
  - Giving the true value of the parameters on average.
  - The estimated value does not vary too much about the true value.
  - Even if the data doesn't follow the particular distribution we thought it to follow, still the estimated value corresponds to the model that is the closest to the source of the data.
- These correspond to the following properties of estimators.
  - Unbiased estimator.
  - Low (or minimum possible) variance estimator.
  - Consistent estimator (given the number of data samples $\rightarrow \infty$).
- Detailed discussion is beyond the scope.
- MLE is "asymptotically optimal". Its great properties (consistency and efficiency) are guaranteed as data $\rightarrow \infty$
- but it can be problematic with small datasets

Why MLE is Loved So Much!

- Some desirable properties of an estimator
  - Giving the true value of the parameters on average.
  - The estimated value does not vary too much about the true value.
  - Even if the data doesn't follow the particular distribution we thought it to follow, still the estimated value corresponds to the model that is the closest to the source of the data.
- These correspond to the following properties of estimators.
  - Unbiased estimator.
  - Low (or minimum possible) variance estimator.
  - Consistent estimator (given the number of data samples $\rightarrow \infty$).
- Detailed discussion is beyond the scope.
- MLE is "asymptotically optimal". Its great properties (consistency and efficiency) are guaranteed as data $\rightarrow \infty$
- but it can be problematic with small datasets

Weighing Pros and Cons of MLE

- MLE can be biased, especially for finite samples.
- The bias often decreases as sample size increases.
- MLE is not always minimum variance.
- For large datasets, MLE variance approaches the best possible.
- If the data truly comes from the assumed model
- And you collect more and more data
- Then MLE converges to the true parameter

Weighing Pros and Cons of MLE

- MLE can be biased, especially for finite samples.
- The bias often decreases as sample size increases.
- MLE is not always minimum variance.
- For large datasets, MLE variance approaches the best possible.
- If the data truly comes from the assumed model
- And you collect more and more data
- Then MLE converges to the true parameter

Weighing Pros and Cons of MLE

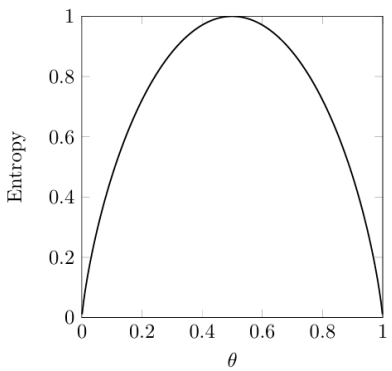
- MLE can be biased, especially for finite samples.
- The bias often decreases as sample size increases.
- MLE is not always minimum variance.
- For large datasets, MLE variance approaches the best possible.
- If the data truly comes from the assumed model
- And you collect more and more data
- Then MLE converges to the true parameter

Incorporating Prior Beliefs in Estimation

- (Till now): MLE assumes all values of the parameter θ are equally likely.
- However, we can, very well, have a belief about the value of θ before any data is observed
- However, before delving into how prior beliefs can be incorporated in estimation, lets wrap up MLE by exploring a relationship.

Entropy

- The entropy of a discrete random variable with PMF P[1] over K states is defined by, $H(p) = -\mathbb{E}_P[\log P(x)] = -\sum_{i=1}^{K} P(x_i) \log P(x_i)$
- The entropy of a probability distribution can be interpreted as a measure of uncertainty or lack of predictability
- For a Bernoulli random variable $p(x;\theta) = \theta^x (1-\theta)^{1-x}$, the entropy is: $H(p) = -\big(\theta \log \theta + (1-\theta)\log(1-\theta)\big)$

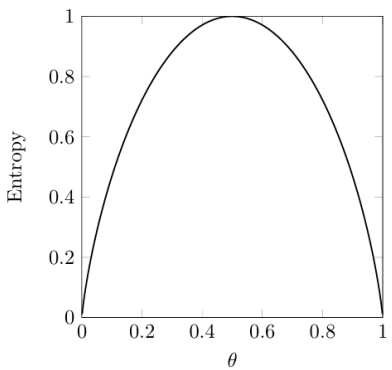


  - Entropy of a Bernoulli random variable as a function of θ and with log base 2.
  - Maximum entropy is when the distribution is uniform, $\theta = 0.5$.
  - Minimum entropy is when one of the probabilities is 1, *i.e.*, there is no uncertainty.
- An uniform distribution has the highest entropy while a distribution with all probability mass concentrated at a single value has the minimum entropy.

[1] Can be extended to continuous random variables by using the integral instead of the sum and probability density function by mass function.

Entropy

- The entropy of a discrete random variable with PMF P[1] over K states is defined by, $H(p) = -\mathbb{E}_P[\log P(x)] = -\sum_{i=1}^{K} P(x_i) \log P(x_i)$
- The entropy of a probability distribution can be interpreted as a measure of uncertainty or lack of predictability
- For a Bernoulli random variable $p(x;\theta) = \theta^x (1-\theta)^{1-x}$, the entropy is: $H(p) = -\big(\theta \log \theta + (1-\theta)\log(1-\theta)\big)$



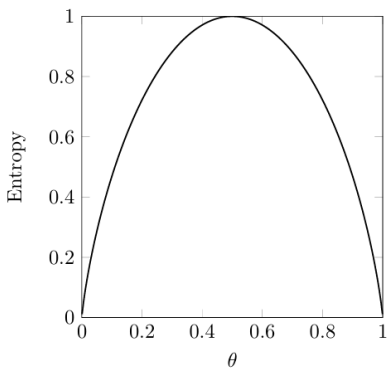
- Entropy of a Bernoulli random variable as a function of θ and with log base 2.
- Maximum entropy is when the distribution is uniform, $\theta = 0.5$.
- Minimum entropy is when one of the probabilities is 1, *i.e.*, there is no uncertainty.

- An uniform distribution has the highest entropy while a distribution with all probability mass concentrated at a single value has the minimum entropy.

[1] Can be extended to continuous random variables by using the integral instead of the sum and probability density function by mass function.

Entropy

- The entropy of a discrete random variable with PMF P[1] over K states is defined by, $H(p) = -\mathbb{E}_P[\log P(x)] = -\sum_{i=1}^{K} P(x_i) \log P(x_i)$
- The entropy of a probability distribution can be interpreted as a measure of uncertainty or lack of predictability
- For a Bernoulli random variable $p(x;\theta) = \theta^x (1-\theta)^{1-x}$, the entropy is: $H(p) = -\big(\theta \log \theta + (1-\theta)\log(1-\theta)\big)$



- Entropy of a Bernoulli random variable as a function of θ and with log base 2.
- Maximum entropy is when the distribution is uniform, $\theta = 0.5$.
- Minimum entropy is when one of the probabilities is 1, *i.e.*, there is no uncertainty.

- An uniform distribution has the highest entropy while a distribution with all probability mass concentrated at a single value has the minimum entropy.

[1] Can be extended to continuous random variables by using the integral instead of the sum and probability density function by mass function.

Relative Entropy or Kullback-Leibler Divergence

§ Given two distributions p and q, it is often useful to measure how "close" or "similar" they are.

$$KL(p||q) = \mathbb{E}_p\Big[\log \frac{p(x)}{q(x)}\Big] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) = -H(p) + H(p,q) \quad (13)$$

where $H(p) = -\sum_x p(x) \log p(x)$ is the entropy of the distribution p and $H(p,q) = -\sum_x p(x) \log q(x)$ is called the cross-entropy.

§ KL divergence satisfies the following two properties.

- $KL(p||q) \geq 0$
- $KL(p||q) = 0$ if and only if $p = q$

§ It is worth mentioning that the KL-divergence is not a distance.

§ It is not symmetric and it does not satisfy the triangle inequality in general.

Relative Entropy or Kullback-Leibler Divergence

§ Given two distributions p and q, it is often useful to measure how "close" or "similar" they are.

$$KL(p||q) = \mathbb{E}_p\Big[\log \frac{p(x)}{q(x)}\Big] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) = -H(p) + H(p,q) \quad (13)$$

where $H(p) = -\sum_x p(x) \log p(x)$ is the entropy of the distribution p and $H(p,q) = -\sum_x p(x) \log q(x)$ is called the cross-entropy.

§ KL divergence satisfies the following two properties.

- $KL(p||q) \geq 0$
- $KL(p||q) = 0$ if and only if $p = q$

§ It is worth mentioning that the KL-divergence is not a distance.

§ It is not symmetric and it does not satisfy the triangle inequality in general.

Relative Entropy or Kullback-Leibler Divergence

§ Given two distributions p and q, it is often useful to measure how "close" or "similar" they are.

$$KL(p||q) = \mathbb{E}_p\Big[\log \frac{p(x)}{q(x)}\Big] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) = -H(p) + H(p,q) \quad (13)$$

where $H(p) = -\sum_x p(x) \log p(x)$ is the entropy of the distribution p and $H(p,q) = -\sum_x p(x) \log q(x)$ is called the cross-entropy.

§ KL divergence satisfies the following two properties.

- $KL(p||q) \geq 0$
- $KL(p||q) = 0$ if and only if $p = q$

§ It is worth mentioning that the KL-divergence is not a distance.

§ It is not symmetric and it does not satisfy the triangle inequality in general.

Relation to Maximum Likelihood

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i; \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log q(x_i; \theta)$$

Relation to Maximum Likelihood

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i;\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log q(x_i;\theta)$$

$$= \arg\max_{\theta} \Big[\frac{1}{N}\sum_{i=1}^{N} \log q(x_i;\theta) - \frac{1}{N}\sum_{i=1}^{N} \log p(x_i)\Big]$$

§ Replace sum by average and add a term that does not depend on θ

Relation to Maximum Likelihood

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i;\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log q(x_i;\theta)$$

$$= \arg\max_{\theta} \Big[\frac{1}{N}\sum_{i=1}^{N} \log q(x_i;\theta) - \frac{1}{N}\sum_{i=1}^{N} \log p(x_i)\Big]$$

$$= \arg\min_{\theta} \Big[\frac{1}{N}\sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i;\theta)}\Big]$$

- Replace sum by average and add a term that does not depend on θ
- Switched the signs and changed argmax to argmin

Relation to Maximum Likelihood

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i;\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log q(x_i;\theta)$$

$$= \arg\max_{\theta} \left[\frac{1}{N}\sum_{i=1}^{N} \log q(x_i;\theta) - \frac{1}{N}\sum_{i=1}^{N} \log p(x_i)\right]$$

$$= \arg\min_{\theta} \left[\frac{1}{N}\sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i;\theta)}\right]$$

$$\xrightarrow[N\to\infty]{} \arg\min_{\theta} \mathbb{E}_p\left[\log \frac{p(x_i)}{q(x_i;\theta)}\right] = KL(p||q_\theta)$$

- Replace sum by average and add a term that does not depend on θ
- Switched the signs and changed argmax to argmin
- Average replaced by the expectation under the distribution p with near infinite i.i.d. data from p

Maximum-a-Posteriori

- We have seen that MLE only looks at the data.
- MAP also asks: what did we believe before seeing the data?
- This is incorporated by modelling θ as a random variable and imposing a distribution on θ as our prior belief before any observation is obtained.
- This is denoted as $p(\theta)$, called the **prior distribution** or simply **prior** of θ.
- The goal is to get a distribution of θ after observing the data x. This distribution is called the **posterior** distribution (or simply **posterior**) of θ and is denoted by $p(\theta|x)$.
- The idea is that we have some prior belief $p(\theta)$ about what θ can be. We observe the data and the prior belief (most likely) changes to a posterior $p(\theta|x)$.

Maximum-a-Posteriori

- We have seen that MLE only looks at the data.
- MAP also asks: what did we believe before seeing the data?
- This is incorporated by modelling θ as a random variable and imposing a distribution on θ as our prior belief before any observation is obtained.
- This is denoted as $p(\theta)$, called the **prior distribution** or simply **prior** of θ.
- The goal is to get a distribution of θ after observing the data x. This distribution is called the **posterior** distribution (or simply **posterior**) of θ and is denoted by $p(\theta|x)$.
- The idea is that we have some prior belief $p(\theta)$ about what θ can be. We observe the data and the prior belief (most likely) changes to a posterior $p(\theta|x)$.
- The posterior can be obtained by using the Bayes' rule as,

$$p(\theta|x) = \frac{\overbrace{p(x|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\int p(x|\theta)p(\theta)d\theta} \tag{14}$$

Maximum-a-Posteriori

- We have seen that MLE only looks at the data.
- MAP also asks: what did we believe before seeing the data?
- This is incorporated by modelling θ as a random variable and imposing a distribution on θ as our prior belief before any observation is obtained.
- This is denoted as $p(\theta)$, called the **prior distribution** or simply **prior** of θ.
- The goal is to get a distribution of θ after observing the data x. This distribution is called the **posterior** distribution (or simply **posterior**) of θ and is denoted by $p(\theta|x)$.
- The idea is that we have some prior belief $p(\theta)$ about what θ can be. We observe the data and the prior belief (most likely) changes to a posterior $p(\theta|x)$.
- The posterior can be obtained by using the Bayes' rule as,

$$p(\theta|x) = \frac{\overbrace{p(x|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\int p(x|\theta)p(\theta)d\theta} \tag{14}$$

Maximum-a-Posteriori

- We have seen that MLE only looks at the data.
- MAP also asks: what did we believe before seeing the data?
- This is incorporated by modelling θ as a random variable and imposing a distribution on θ as our prior belief before any observation is obtained.
- This is denoted as $p(\theta)$, called the **prior distribution** or simply **prior** of θ.
- The goal is to get a distribution of θ after observing the data x. This distribution is called the **posterior** distribution (or simply **posterior**) of θ and is denoted by $p(\theta|x)$.
- The idea is that we have some prior belief $p(\theta)$ about what θ can be. We observe the data and the prior belief (most likely) changes to a posterior $p(\theta|x)$.
- The posterior can be obtained by using the Bayes' rule as,

$$p(\theta|x) = \frac{\overbrace{p(x|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\int p(x|\theta)p(\theta)d\theta} \tag{14}$$

Maximum-a-Posteriori

- This seemingly simple formulation is often very hard to compute. The main culprit is the integration in the denominator.
- Thus an approximation is done. Instead of finding the complete posterior distribution, only its mode is computed. The θ for which $p(\theta|x)$ attains its maximum value is called the Maximum-a-Posteriori (MAP) estimate of θ.

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|x) = \underset{\theta}{\operatorname{argmax}}\ \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} = \underset{\theta}{\operatorname{argmax}}\ p(x|\theta)p(\theta) \quad (15)$$

The last step is a result of the fact that the denominator is not a function of θ.

- Though we started with an assumption of a distribution of θ, MAP gives a single (most probable) value of θ.

Maximum-a-Posteriori

- This seemingly simple formulation is often very hard to compute. The main culprit is the integration in the denominator.
- Thus an approximation is done. Instead of finding the complete posterior distribution, only its mode is computed. The θ for which $p(\theta|x)$ attains its maximum value is called the Maximum-a-Posteriori (MAP) estimate of θ.

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ p(\theta|x) = \underset{\theta}{\text{argmax}}\ \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} = \underset{\theta}{\text{argmax}}\ p(x|\theta)p(\theta) \quad (15)$$

The last step is a result of the fact that the denominator is not a function of θ.

- Though we started with an assumption of a distribution of θ, MAP gives a single (most probable) value of θ.

MAP Estimation & Conjugacy

- In some cases, the likelihood and prior can be such that when multiplied together, the result *i.e.*, the posterior turns out to be same distribution as the prior, just with different values of the parameters compared to the prior.
- Such type of a prior is called a **conjugate prior**.

Distribution parameter	Conjugate distribution
Bernoulli p	Beta
Binomial p	Beta
Multinomial p_i	Dirichlet
Poisson λ	Gamma
Exponential λ	Gamma
Normal μ	Normal
Normal σ^2	Inverse Gamma

Fig credit: Stanford Course CS109: Lecture 22

Bernoulli-Beta Conjugacy

- In the Bernoulli coin toss, θ is the probability of Heads.
- In MAP, we treat θ as a random variable.
- **Prior:** We need a distribution where $\theta \in [0, 1]$.

The Beta Distribution: $Beta(\theta|\alpha, \beta)$

It describes a probability distribution *over probabilities*.

$$P(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where, $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$

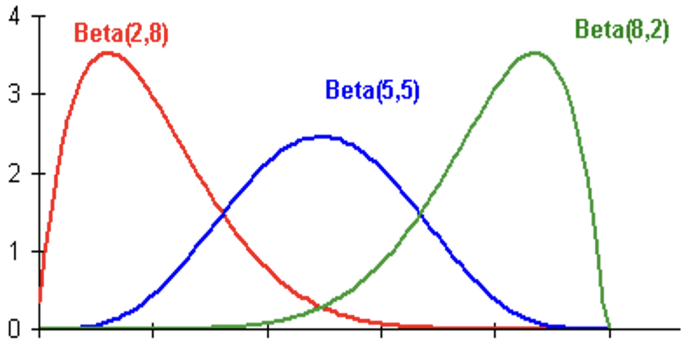


Fig credit: Stanford Course CS109: Lecture 16

- Think of α as "imaginary heads" and β as "imaginary tails" observed before the experiment.

Bernoulli-Beta Conjugacy

- In the Bernoulli coin toss, θ is the probability of Heads.
- In MAP, we treat θ as a random variable.
- **Prior:** We need a distribution where $\theta \in [0, 1]$.

The Beta Distribution: $Beta(\theta|\alpha, \beta)$

It describes a probability distribution *over probabilities*.

$$P(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where, $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$

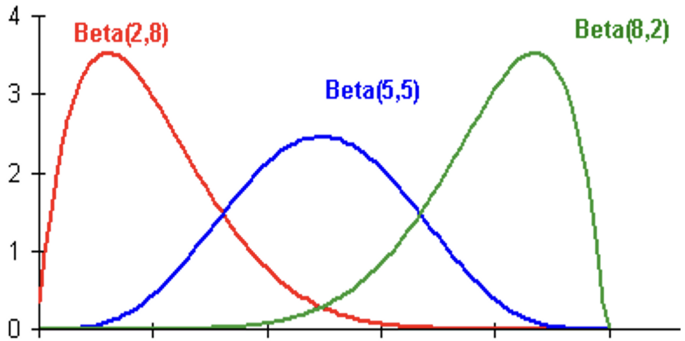


Fig credit: Stanford Course CS109: Lecture 16

- Think of α as "imaginary heads" and β as "imaginary tails" observed before the experiment.

The Derivation

We flip a coin N times and get H heads and T tails ($N = H + T$).

The MAP Objective

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- **Likelihood (Bernoulli):** $P(X|\theta) \propto \theta^H (1-\theta)^T$
- **Prior (Beta):** $P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

Let's multiply them:

$$\begin{aligned} P(\theta|X) &\propto \left[\theta^H (1-\theta)^T\right] \times \left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] \\ &= \theta^{(H+\alpha-1)}(1-\theta)^{(T+\beta-1)} \end{aligned}$$

Look closely at the result. Does the form look familiar?

Conjugacy

The Result

The Posterior is **also** a Beta distribution!

$$P(\theta|X) = \text{Beta}(\theta \mid \alpha', \beta')$$

Where $\alpha' = \alpha + H$ and $\beta' = \beta + T$.

- The Beta prior is a *conjugate prior* for the Bernoulli likelihood.
- **Intuition:** The posterior simply updates our "counts."
  - New α = Old α + Observed Heads
  - New β = Old β + Observed Tails

Why does this matter? We avoided complex integration! We just did simple addition to update our beliefs.

Finding the MAP Estimate

We have the Posterior: $P(\theta|X) = \text{Beta}(\theta \mid \alpha', \beta')$ where $\alpha' = H + \alpha, \beta' = T + \beta$.
Goal: Find the θ that maximizes this posterior (the **Mode**).

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \left[\theta^{\alpha'-1}(1-\theta)^{\beta'-1}\right] \tag{16}$$

Taking the derivative of the log-posterior w.r.t θ and setting to 0:

$$\begin{aligned} \frac{\partial}{\partial\theta}\left[(\alpha'-1)\ln\theta + (\beta'-1)\ln(1-\theta)\right] &= 0 \\ \frac{\alpha'-1}{\theta} - \frac{\beta'-1}{1-\theta} &= 0 \\ \alpha' - 1 - \theta\alpha' + \theta - \theta\beta' + \theta &= 0 \\ (\alpha' + \beta' - 2)\theta &= \alpha' - 1 \end{aligned} \tag{17}$$

The MAP Estimator for Bernoulli

$$\hat{\theta}_{MAP} = \frac{\alpha'-1}{(\alpha'-1)+(\beta'-1)} = \frac{H+\alpha-1}{N+\alpha+\beta-2}$$

Comparison: MAP vs. MLE

Let's compare our two estimators side-by-side:

MLE (Data Only)

$$\hat{\theta}_{MLE} = \frac{H}{N}$$

Relying purely on the observed coin flips.

MAP (Data + Prior)

$$\hat{\theta}_{MAP} = \frac{H + (\alpha - 1)}{N + (\alpha + \beta - 2)}$$

Data counts plus "pseudo-counts" from prior.

Key Insights:

- **Smoothing:** If $\alpha, \beta > 1$, MAP "smooths" the estimate away from 0 or 1 (helps when data is scarce!). "smoothing" means "preventing the model from assigning zero probability to an event just because it hasn't seen it yet."
- **Asymptotics:** As $N \to \infty$ (lots of data), the H and N terms dominate the constants α and β.

$$\lim_{N \to \infty} \hat{\theta}_{MAP} \approx \frac{H}{N} = \hat{\theta}_{MLE}$$

With enough data, the prior doesn't matter!

Gaussian–Gaussian Conjugacy: Setup

Goal: Estimate the mean of a Gaussian distribution using MAP.
Assumptions:

- Data: $\mathcal{D} : x_1, x_2, \ldots, x_N \sim \mathcal{N}(\mu, \sigma^2)$
- Mean μ is *unknown*, variance σ^2 is *known*
- We place a Gaussian prior on μ, *i.e.*, $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$
- This implies that the data comes from a Gaussian and the prior on mean of that Gaussian is also another Gaussian

The Posterior is also Gaussian!

$$P(\mu|D) = \mathcal{N}(\mu_{post}, \sigma_{post}^2)$$

- How do we get the updated parameters $(\mu_0 \rightarrow \mu_{post})$ and $(\sigma_0 \rightarrow \sigma_{post})$?

Gaussian–Gaussian Conjugacy: Setup

Goal: Estimate the mean of a Gaussian distribution using MAP.
Assumptions:

- Data: $\mathcal{D} : x_1, x_2, \ldots, x_N \sim \mathcal{N}(\mu, \sigma^2)$
- Mean μ is *unknown*, variance σ^2 is *known*
- We place a Gaussian prior on μ, *i.e.*, $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$
- This implies that the data comes from a Gaussian and the prior on mean of that Gaussian is also another Gaussian

The Posterior is also Gaussian!

$$P(\mu|D) = \mathcal{N}(\mu_{post}, \sigma_{post}^2)$$

- How do we get the updated parameters $(\mu_0 \rightarrow \mu_{post})$ and $(\sigma_0 \rightarrow \sigma_{post})$?

Updated Mean and Standard Deviation

Update Rules

- $\frac{1}{\sigma^2_{post}} = \frac{1}{\sigma^2_0} + \frac{N}{\sigma^2}$
- $\mu_{post} = \frac{\frac{\mu_0}{\sigma^2_0} + \frac{\sum_{i=1}^{N} x_i}{\sigma^2}}{\frac{1}{\sigma^2_0} + \frac{N}{\sigma^2}}$

- We will defer the derivation for a later class/tutorial.
- However, let us interpret the results in terms of **Precision** (certainty).
- Define **Precision** as the inverse variance: $\lambda = 1/\sigma^2$.
  - Prior Precision: $\lambda_0 = 1/\sigma^2_0$
  - Data Precision (for N points): $\lambda_{data} = N/\sigma^2$

Updated Mean and Standard Deviation

Update Rules

- $\frac{1}{\sigma^2_{post}} = \frac{1}{\sigma^2_0} + \frac{N}{\sigma^2}$
- $\mu_{post} = \frac{\frac{\mu_0}{\sigma^2_0} + \frac{\sum_{i=1}^{N} x_i}{\sigma^2}}{\frac{1}{\sigma^2_0} + \frac{N}{\sigma^2}}$

- We will defer the derivation for a later class/tutorial.
- However, let us interpret the results in terms of **Precision** (certainty).
- Define **Precision** as the inverse variance: $\lambda = 1/\sigma^2$.
  - Prior Precision: $\lambda_0 = 1/\sigma^2_0$
  - Data Precision (for N points): $\lambda_{data} = N/\sigma^2$

Update Rules (in terms of Precision)

- $\lambda_{post} = \lambda_0 + \lambda_{data}$
- $\mu_{post} = \frac{\lambda_0 \mu_0 + \lambda_{data} \bar{x}}{\lambda_0 + \lambda_{data}}$

Updated Mean and Standard Deviation

Update Rules

- $\frac{1}{\sigma_{post}^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$
- $\mu_{post} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{N} x_i}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}$

- We will defer the derivation for a later class/tutorial.
- However, let us interpret the results in terms of **Precision** (certainty).
- Define **Precision** as the inverse variance: $\lambda = 1/\sigma^2$.
  - Prior Precision: $\lambda_0 = 1/\sigma_0^2$
  - Data Precision (for N points): $\lambda_{data} = N/\sigma^2$

Update Rules (in terms of Precision)

- $\lambda_{post} = \lambda_0 + \lambda_{data}$
- $\mu_{post} = \frac{\lambda_0 \mu_0 + \lambda_{data} \bar{x}}{\lambda_0 + \lambda_{data}}$

Updated Mean and Standard Deviation

Update Rules

- $\frac{1}{\sigma^2_{post}} = \frac{1}{\sigma^2_0} + \frac{N}{\sigma^2}$
- $\mu_{post} = \frac{\frac{\mu_0}{\sigma^2_0} + \frac{\sum_{i=1}^{N} x_i}{\sigma^2}}{\frac{1}{\sigma^2_0} + \frac{N}{\sigma^2}}$

- We will defer the derivation for a later class/tutorial.
- However, let us interpret the results in terms of **Precision** (certainty).
- Define **Precision** as the inverse variance: $\lambda = 1/\sigma^2$.
  - Prior Precision: $\lambda_0 = 1/\sigma^2_0$
  - Data Precision (for N points): $\lambda_{data} = N/\sigma^2$

Update Rules (in terms of Precision)

- $\lambda_{post} = \lambda_0 + \lambda_{data}$
- $\mu_{post} = \frac{\lambda_0 \mu_0 + \lambda_{data} \bar{x}}{\lambda_0 + \lambda_{data}}$

Intuition

Define **Precision** as the inverse variance: $\lambda = 1/\sigma^2$.

- Prior Precision: $\lambda_0 = 1/\sigma_0^2$
- Data Precision (for N points): $\lambda_{data} = N/\sigma^2$

Update Rules (The "Tug of War")

1. **Precisions Add:** We simply become more certain.

$$\lambda_{post} = \lambda_0 + \lambda_{data}$$

2. **Mean is a Weighted Average:**

$$\mu_{post} = \frac{\lambda_0 \mu_0 + \lambda_{data} \bar{x}}{\lambda_0 + \lambda_{data}}$$

Inferring Variance and Both

Inferring Variance (σ^2) with known Mean (μ)

- We cannot use a Gaussian prior (variance must be positive!).
- **Conjugate Prior:** Inverse-Gamma distribution.

$$\sigma^2 \sim \text{Inv-Gamma}(\alpha, \beta)$$

- The Posterior is also Inverse-Gamma.

Inferring Both (μ and σ^2)

- Parameters are coupled! The uncertainty in μ depends on σ^2.
- **Conjugate Prior:** Normal-Inverse-Gamma (NIG) distribution.

Reference

For full derivation and update formulas, see: *"Machine Learning: A Probabilistic Perspective", Kevin P. Murphy, Section 7.2.3.*

Summary: MLE vs. MAP

MLE (Frequentist)

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

- **Philosophy:** Parameters are fixed constants; data is random.
- **Pros:** Unbiased (asymptotically), simple to compute.
- **Cons:** Prone to overfitting on small data (e.g., observing 3 Heads $\rightarrow P(H) = 1$).

MAP (Bayesian)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\mathcal{D}|\theta)P(\theta)$$

- **Philosophy:** Parameters are random variables with beliefs.
- **Pros:** Incorporates prior knowledge, "smooths" estimates (regularization).
- **Cons:** Need to choose a valid Prior; computation can be hard (without conjugacy).

Looking Ahead: Why does this theory matter?

We are moving to **Linear** and **Logistic Regression**. You will see that these "algorithms" are actually just MLE and MAP in disguise!

- **Linear Regression as MLE:** We will maximize the likelihood of data which is observed from a data generating function with a Gaussian noise
- **Regularization (The "L2 Penalty"):** We often add a penalty term $\lambda||\mathbf{w}||^2$ to stop overfitting.

$$\text{L2 Regularization} \equiv \text{MAP with a Gaussian Prior}$$

Assuming weights $w \sim \mathcal{N}(0, \lambda^{-1})$ forces them to stay small.

Next Class: Linear Regression