



CSE 312 SP24

Section 9: Maximum Likelihood Estimation



**Any lingering questions
from this last week?**

Kahoot!

+ Questions are under Task 1 on the handout

Review

+ **Weak Law of Large Numbers** (not covered): Let X_1, \dots, X_n be iid random variables with common mean μ and variance σ^2 . Let

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ be the sample mean for a sample of size n . Then

for any $\epsilon > 0$, $\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$.

+ Basically, as we sample more, the sample mean should converge to the population mean.

+ **Realization/Sample**: A realization/sample x of a random variable X is the value that is actually observed.

Review (cont.)

- + **Likelihood:** Let x_1, \dots, x_n be iid samples from pmf $p_X(x; \theta)$ where θ are the distribution's parameters. We define the likelihood function to be the probability of seeing the data given the parameters as

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

- + The continuous case just replaces a pmf with a pdf.
- + Used in statistical learning
- + Bar notation and semicolon notation mean the same thing (semicolon attempts to clear up confusion about conditional probability vs parameters)
 - + Re-typesetting takes a long time in Office equations—sorry
- + Sometimes (most times), you'll also see pmfs represented with p and pdfs represented with f —this class used f for both to emphasize they are analogous.

Review (cont.)

+ **Maximum Likelihood Estimator:** We denote the MLE of θ as $\hat{\theta}$, the parameters that maximize the likelihood function. Expressed mathematically, we have

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}(x_1, \dots, x_n; \theta)$$

+ Note that we often take the argmax of the *(natural) log likelihood*. This is equivalent to the above since the log function is monotone increasing. This is better for numerical stability: we often get floating pointer underflow when multiplying small numbers together, so taking the log lets us add them instead (think back to Naïve Bayes)

Review (cont.)

- + **Bias:** The bias of an estimator $\hat{\theta}$ for a **true** parameter θ is defined as $E[\hat{\theta}] - \theta$. An estimator is unbiased iff $E[\hat{\theta}] = \theta$.
- + **Steps to find the MLE:**
 - + (a) Find the likelihood and log-likelihood of the data
 - + (b) Take the derivative of the log-likelihood wrt θ and set it to 0 to find $\hat{\theta}$.
 - + (c) Take the second derivative and show that $\hat{\theta}$ is a global maximizer (calc 2: second-derivative test)

Problem 2

+A fancy new restaurant has opened up which features only 4 dishes. The unique feature of dining here is that they will serve you any of the four dishes randomly according to the following probability distribution: give dish A with probability 0.5, dish B with probability θ , dish C with probability 2θ , and dish D with probability $0.5 - 3\theta$. Each student is assigned a grad independently. Let x_A be the number of people who received an A, etc., so $x_A + x_B + x_C + x_D = n$. Find the MLE for θ .

Problem 2 Solution

+1. Compute likelihood:

$$\mathcal{L}(x|\theta) = 0.5^{x_A} \theta^{x_B} (2\theta)^{x_C} (0.5 - 3\theta)^{x_D}$$

Problem 2 Solution

+1. Compute likelihood:

$$\mathcal{L}(x|\theta) = 0.5^{x_A} \theta^{x_B} (2\theta)^{x_C} (0.5 - 3\theta)^{x_D}$$

$$\Rightarrow \ln \mathcal{L}(x|\theta) = x_A \ln(0.5) + x_B \ln(\theta) + x_C \ln(2\theta) + x_D \ln(0.5 - 3\theta)$$

Problem 2 Solution

+1. Compute likelihood:

$$\mathcal{L}(x|\theta) = 0.5^{x_A} \theta^{x_B} (2\theta)^{x_C} (0.5 - 3\theta)^{x_D}$$

$$\Rightarrow \ln \mathcal{L}(x|\theta) = x_A \ln(0.5) + x_B \ln(\theta) + x_C \ln(2\theta) + x_D \ln(0.5 - 3\theta)$$

+2. Take derivative & set equal to 0:

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x|\theta) = \frac{x_B}{\hat{\theta}} + \frac{x_C}{\hat{\theta}} - \frac{3x_D}{(0.5 - 3\hat{\theta})} = 0$$

Problem 2 Solution

+1. Compute likelihood:

$$\mathcal{L}(x|\theta) = 0.5^{x_A} \theta^{x_B} (2\theta)^{x_C} (0.5 - 3\theta)^{x_D}$$

$$\Rightarrow \ln \mathcal{L}(x|\theta) = x_A \ln(0.5) + x_B \ln(\theta) + x_C \ln(2\theta) + x_D \ln(0.5 - 3\theta)$$

+2. Take derivative & set equal to 0:

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x|\theta) = \frac{x_B}{\hat{\theta}} + \frac{x_C}{\hat{\theta}} - \frac{3x_D}{(0.5 - 3\hat{\theta})} = 0$$

$$\Rightarrow \hat{\theta} = \frac{x_B + x_C}{6(x_B + x_C + x_D)}$$

+ Normally, you'd perform the second-derivative test for completeness, but we skip it here for brevity's sake

Problem 6

- + You are an ornithologist studying a rare species of birds in a nature reserve. Over a period of 50 days, you record the number of sightings of this bird. Your research has shown that the number of sightings on this species depends on the number of monkeys living in the reserve, θ_1 , and the θ_2 . After years of studying this rare species in other environments, you've found the number of birds observed on a particular day follows the following distribution:

$$p_X(k) = \frac{1}{k!} (\theta_1^k \cdot e^{-\theta_1} \cdot \theta_2^k \cdot e^{-3\theta_2})$$

- What is the likelihood function?
- What is the log-likelihood function?
- What is the partial derivative of the log-likelihood function with respect to θ_1 ?
- What is the partial derivative of the log-likelihood function with respect to θ_2 ?
- Set both partial derivatives to 0 and solve for $\hat{\theta}_1$ and $\hat{\theta}_2$.

Problem 6 Solution

+ What is the likelihood function?

+ Multiply the PDFs together:

$$L(x; \theta_1, \theta_2) = \prod_{i=1}^{50} \left(\frac{1}{x_i!} (\theta_1^{x_i} \cdot e^{-\theta_1} \cdot \theta_2^{x_i} \cdot e^{-3\theta_2}) \right)$$

Problem 6 Solution

+ What is the log-likelihood function?

+ Take the log of the previous part and simplify

$$\begin{aligned}\ln(L(x; \theta_1, \theta_2)) &= \ln \left(\prod_{i=1}^{50} \left(\frac{1}{x_i!} (\theta_1^{x_i} \cdot e^{-\theta_1} \cdot \theta_2^{x_i} \cdot e^{-3\theta_2}) \right) \right) \\ &= \sum_{i=1}^{50} \left(\ln \frac{1}{x_i!} + \ln \theta_1^{x_i} + \ln \theta_2^{x_i} + \ln e^{-3\theta_2} \right) \\ &= \sum_{i=1}^{50} \left(\ln \frac{1}{x_i!} + x_i \ln \theta_1 - \theta_1 + x_i \ln \theta_2 - 3\theta_2 \right)\end{aligned}$$

Problem 6 Solution

- + What is the partial derivative of the log-likelihood function with respect to θ_1 ?
- + Take the derivative as usual but treat θ_2 as a constant and derivate with respect to θ_1 .

$$\begin{aligned}\frac{\partial}{\partial \theta_1} (L(x; \theta_1, \theta_2)) &= \sum_{i=1}^{50} \left(\frac{x_i}{\theta_1} - 1 \right) \\ &= \sum_{i=1}^{50} \left(\frac{x_i}{\theta_1} \right) - 50\end{aligned}$$

Problem 6 Solution

- + What is the partial derivative of the log-likelihood function with respect to θ_2 ?
- + Take the derivative as usual but treat θ_1 as a constant and derivate with respect to θ_2 .

$$\begin{aligned}\frac{\partial}{\partial \theta_2} (L(x; \theta_1, \theta_2)) &= \sum_{i=1}^{50} \left(\frac{x_i}{\theta_2} - 3 \right) \\ &= \sum_{i=1}^{50} \left(\frac{x_i}{\theta_2} \right) - 150\end{aligned}$$

Problem 6 Solution

- + Set both partial derivatives to 0 and solve for $\hat{\theta}_1$ and $\hat{\theta}_2$.
- + Add hats to thetas here, since we are finally solving for the *maximum* likelihood estimator:

$$\sum_{i=1}^{50} \left(\frac{x_i}{\hat{\theta}_1} \right) - 50 = 0 \Rightarrow \hat{\theta}_1 = \frac{\sum_{i=1}^{50} x_i}{50}$$

$$\sum_{i=1}^{50} \left(\frac{x_i}{\hat{\theta}_2} \right) - 150 = 0 \Rightarrow \hat{\theta}_2 = \frac{\sum_{i=1}^{50} x_i}{150}$$

Problem 4b

- + You are given 100 independent samples x_1, x_2, \dots, x_{100} from Bernoulli(θ), where θ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. You would like to estimate the distribution's parameter θ .
- + Is $\hat{\theta}$ an unbiased estimator of θ ?

Problem 4b solution

+ Check if $E[\hat{\theta}] = \theta$:

$$\begin{aligned} E[\hat{\theta}] &= E\left[\frac{1}{100} \sum_{i=1}^{100} X_i\right] \\ &= \frac{1}{100} \sum_{i=1}^{100} E[X_i] \\ &= \frac{1}{100} \cdot 100\theta = \theta \end{aligned}$$

$E[\hat{\theta}] = \theta$, so $\hat{\theta}$ is unbiased.

Problem 9

+ Let X be the network connection status, where $X = 0$ represents a stable connection and $X = 1$ represents an unstable connection. Let Y be the number of successes in data transmission, taking values in the set $\{0, 1, 2\}$. If $X = 0$, Y follows a Binomial distribution $Bin(2, 0.8)$, and if $X = 1$, Y follows a Binomial distribution $Bin(2, 0.3)$. The probabilities for X are given by $P(X = 0) = 0.8$ and $P(X = 1) = 0.2$. Find $Cov(X, Y)$.

Problem 9 Solution

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

$$E[X] = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0 \cdot 0.8 + 1 \cdot 0.2 = 0.2$$

Use law of total expectation and expectation of binomials for $E[Y]$:

$$\begin{aligned} E[Y] &= E[Y|X = 0]P(X = 0) + E[Y|X = 1]P(X = 1) \\ &= (2 \cdot 0.8) \cdot 0.8 + (2 \cdot 0.3) \cdot 0.2 = 1.4 \end{aligned}$$

Problem 9 Solution

+ Compute $E[XY]$ by computing probabilities for each possible XY

$$\begin{aligned}P(XY = 2) &= P(X = 1 \cap Y = 2) = P(X = 1)P(Y = 2|X = 1) \\ &= 0.2(0.3^2) = 0.018\end{aligned}$$

$$\begin{aligned}P(XY = 1) &= P(X = 1 \cap Y = 1) = P(X = 1)P(Y = 1|X = 1) \\ &= 0.2 \cdot (2 \cdot 0.3 \cdot 0.7) = 0.084\end{aligned}$$

$$\begin{aligned}P(XY = 0) &= 1 - P(XY = 1) - P(XY = 2) \\ &= 1 - 0.018 - 0.084 = 0.68\end{aligned}$$

+ Apply definition of expectation:

$$E[XY] = 0 \cdot 0.68 + 1 \cdot 0.084 + 2 \cdot 0.018 = 0.12$$

Problem 9 Solution

+ Plug in the numbers!

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0.12 - 0.2 \cdot 1.4 = -0.16$$