

CSE 312 Section 9

MLE

Announcements & Reminders

- HW7
 - Due on Wednesday 11/19
 - Late deadline Saturday 11/22 @ 11:59 pm
 - Check late days remaining!
- HW8
 - Due Wednesday 12/3 @ 11:59 pm

Content review + Problem 1



Review

- + **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 point 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 1a

The Log-Likelihood gives a slightly different estimate, but because it is close enough and easier to compute we use it for our estimate of θ .

- True
- False

Problem 1a

The Log-Likelihood gives a slightly different estimate, but because it is close enough and easier to compute we use it for our estimate of θ .

- True
- **False**

Since the logarithm is a strictly increasing function, the value of θ that maximizes the likelihood will be exactly the same as the value that maximizes the log-likelihood.

Problem 1b

When doing MLE, $\hat{\theta}$ is the true parameter and θ is our estimate.

- True
- False

Problem 1b

When doing MLE, $\hat{\theta}$ is the true parameter and θ is our estimate.

- True
- **False**

It is the other way around. Remember to switch to $\hat{\theta}$ when you set your equation to zero!

Problem 1c

An estimator is unbiased if $\text{Bias}(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta = 0$ or equivalently $E[\hat{\theta}] = \theta$

- True
- False

Problem 1c

An estimator is unbiased if $\text{Bias}(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta = 0$ or equivalently $E[\hat{\theta}] = \theta$

- **True**, by definition of bias
- False

Problem 1d

(d) You flip a coin 10 times and observe HHHTHHTHHH (8 heads, 2 tails). What is the MLE of θ , where θ is the true probability of seeing tails?

☐ $\hat{\theta} = .2$

☐ $\hat{\theta} = .25$

☐ $\hat{\theta} = .8$

☐ $\hat{\theta} = .3$

Problem 1d

(d) You flip a coin 10 times and observe HHHTHHTHHH (8 heads, 2 tails). What is the MLE of θ , where θ is the true probability of seeing tails?

☒ $\hat{\theta} = .2$

☐ $\hat{\theta} = .25$

☐ $\hat{\theta} = .8$

☐ $\hat{\theta} = .3$

Problem 2 – Mystery Dish



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 diner is assigned a dish 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

Law of Total Expectation



LTE Definition

Law of Total Expectation (LTE): Let X, Y be jointly distributed random variables.

If Y is discrete (and X is either discrete or continuous), then:

$$\mathbb{E}[g(X)] = \sum_{y \in \Omega_Y} \mathbb{E}[g(X) \mid Y = y] p_Y(y)$$

If Y is continuous (and X is either discrete or continuous), then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} \mathbb{E}[g(X) \mid Y = y] f_Y(y) dy$$

LTE Definition

Law of Total Expectation (LTE): Let X, Y be jointly distributed random variables.

If Y is discrete (and X is either discrete or continuous), then:

$$\mathbb{E}[g(X)] = \sum_{y \in \Omega_Y} \mathbb{E}[g(X) \mid Y = y] p_Y(y)$$

If Y is continuous (and X is either discrete or continuous), then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} \mathbb{E}[g(X) \mid Y = y] f_Y(y) dy$$

Looks similar to the Law of Total Probability.

LTE Definition

Law of Total Expectation (LTE): Let X, Y be jointly distributed random variables.

If Y is discrete (and X is either discrete or continuous), then:

$$\mathbb{E}[g(X)] = \sum_{y \in \Omega_Y} \mathbb{E}[g(X) \mid Y = y] p_Y(y)$$

If Y is continuous (and X is either discrete or continuous), then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} \mathbb{E}[g(X) \mid Y = y] f_Y(y) dy$$

Looks similar to the Law of Total Probability.

(Just replacing probabilities with expectations).

LTE Definition

Law of Total Expectation (LTE): Let X, Y be jointly distributed random variables.

If Y is discrete (and X is either discrete or continuous), then:

$$\mathbb{E}[g(X)] = \sum_{y \in \Omega_Y} \mathbb{E}[g(X) \mid Y = y] p_Y(y)$$

If Y is continuous (and X is either discrete or continuous), then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} \mathbb{E}[g(X) \mid Y = y] f_Y(y) dy$$

Looks similar to the Law of Total Probability.

(Just replacing probabilities with expectations).

The proof is similar, too.

Problem 0 - Lemonade Stand



Problem 0

- Suppose I run a lemonade stand that costs me \$100 a day to operate. I sell a single lemonade for \$20. Every person who walks by my stand either buys a drink or doesn't (no one buys more than one). If it is raining, n_1 people walk by my stand, and each buys a drink independently with probability p_1 . If it isn't raining, n_2 people walk by my stand, and each buys a drink independent with probability p_2 . It rains each day with probability p_{rain} independently of every other day. Let X be my profit over the next week. In terms of $n_1, n_2, p_1, p_2, p_{rain}$, what is $E[X]$?

Problem 0 Solution

- Let R be the event it rains. Let X_i be how many drinks I sell on day i for $i = 1, \dots, 7$. We are interested in $X = \sum_{i=1}^7 (20X_i - 100)$.

Problem 0 Solution

- Let R be the event it rains. Let X_i be how many drinks I sell on day i for $i = 1, \dots, 7$. We are interested in $X = \sum_{i=1}^7 (20X_i - 100)$.
 - We also know that $X_i|R \sim \text{Bin}(n_1, p_1) \Rightarrow E[X_i|R] = n_1p_1$ from the problem statement.
 - Similarly, $X_i|R^C \sim \text{Bin}(n_2, p_2) \Rightarrow E[X_i|R^C] = n_2p_2$

Problem 0 Solution

■ Let R be the event it rains. Let X_i be how many drinks I sell on day i for $i = 1, \dots, 7$. We are interested in $X = \sum_{i=1}^7 (20X_i - 100)$.

- We also know that $X_i|R \sim \text{Bin}(n_1, p_1) \Rightarrow E[X_i|R] = n_1p_1$ from the problem statement.
- Similarly, $X_i|R^C \sim \text{Bin}(n_2, p_2) \Rightarrow E[X_i|R^C] = n_2p_2$

We then apply the Law of Total Expectation:

$$\mu_{X_i} = E[X_i] = E[X_i|R]P(R) + E[X_i|R^C]P(R^C) = n_1p_1 \cdot p_{\text{rain}} + n_2p_2 \cdot (1 - p_{\text{rain}})$$

Problem 0 Solution

- Let R be the event it rains. Let X_i be how many drinks I sell on day i for $i = 1, \dots, 7$. We are interested in $X = \sum_{i=1}^7 (20X_i - 100)$.

- We also know that $X_i|R \sim \text{Bin}(n_1, p_1) \Rightarrow E[X_i|R] = n_1p_1$ from the problem statement.
- Similarly, $X_i|R^C \sim \text{Bin}(n_2, p_2) \Rightarrow E[X_i|R^C] = n_2p_2$

We then apply the Law of Total Expectation:

$$\mu_{X_i} = E[X_i] = E[X_i|R]P(R) + E[X_i|R^C]P(R^C) = n_1p_1 \cdot p_{rain} + n_2p_2 \cdot (1 - p_{rain})$$

Then Linearity of Expectation:

$$\begin{aligned} E[X] &= E\left[\sum_{i=1}^7 (20X_i - 100)\right] = 20\left(\sum_{i=1}^7 E[X_i]\right) - 700 \\ &= 140(n_1p_1 \cdot p_{rain} + n_2p_2 \cdot (1 - p_{rain})) - 700 \end{aligned}$$

Problem 4 – Independent Shreds



Problem 4

You are given 100 independent samples x_1, x_2, \dots, x_{100} from $\text{Bernoulli}(\theta)$, 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 . Give all answers to 3 significant digits.

(a) What is the maximum likelihood estimator $\hat{\theta}$ of θ ?

(b) Is $\hat{\theta}$ an unbiased estimator of θ ?

Problem 4a

You are given 100 independent samples x_1, x_2, \dots, x_{100} from $\text{Bernoulli}(\theta)$, 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 . Give all answers to 3 significant digits.

(a) What is the maximum likelihood estimator $\hat{\theta}$ of θ ?

Note that $\sum_{i \in [n]} x_i = 30$, as given in the problem spec. Therefore, there are 30 1s and 70 0s. (Note that they come in some specific order.) Therefore, we can setup L as follows, because there is a θ chance of getting a 1, and a $(1 - \theta)$ chance of getting a 0 and they are each i.i.d. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for $\hat{\theta}$.

$$\begin{aligned} L(x_1, \dots, x_n; \theta) &= (1 - \theta)^{70} \theta^{30} \\ \ln L(x_1, \dots, x_n; \theta) &= 70 \ln(1 - \theta) + 30 \ln \theta \\ \frac{\partial}{\partial \theta} \ln L(x_1, \dots, x_n; \theta) &= -\frac{70}{1 - \theta} + \frac{30}{\theta} \\ -\frac{70}{1 - \hat{\theta}} + \frac{30}{\hat{\theta}} &= 0 \\ \frac{30}{\hat{\theta}} &= \frac{70}{1 - \hat{\theta}} \\ 30 - 30\hat{\theta} &= 70\hat{\theta} \\ \hat{\theta} &= \frac{30}{100} \end{aligned}$$

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 6 – Bird Watching



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:

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

- a) What is the likelihood function?
- b) What is the log-likelihood function?
- c) What is the partial derivative of the log-likelihood function with respect to θ_1 ?
- d) What is the partial derivative of the log-likelihood function with respect to θ_2 ?
- e) 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 9 – Covariance Connection



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 $\text{Bin}(2, 0.8)$, and if $X = 1$, Y follows a Binomial distribution $\text{Bin}(2, 0.3)$. The probabilities for X are given by $P(X = 0) = 0.8$ and $P(X = 1) = 0.2$. Find $\text{Cov}(X, Y)$.

Problem 9 Solution

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

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

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!

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

That's All, Folks!

**Thanks for coming to section this week!
Any questions?**