CSE 312 Section 9

Maximum Likelihood Estimation

Administrivia

Announcements & Reminders

- Final Exam
  - **Friday 8/18 in class**
  - We will provide a reference sheet with some of the most-used formulas.
  - No personal note sheets are allowed.

- HW6
  - Was due Tuesday, 8/8

- HW7
  - Due next Tuesday, 8/15

Review & Questions

Review of Main Concepts

- **Realization/Sample**: A realization/sample x of a random variable X is the value that is actually observed.
- **Likelihood**: Let $x_1, \dots x_n$ be iid realizations from probability mass function $p_X(\mathrm{x}\,;\theta)$ (if X discrete) or density $f_X(\mathrm{x}\,;\theta)$ (if X continuous), where θ is a parameter (or a vector of parameters). We define the likelihood function to be the probability of seeing the data.

  If X is discrete:

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

  If X is continuous:

$$L(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

Review of Main Concepts

- **Maximum Likelihood Estimator (MLE)**: We denote the MLE of θ as $\hat{\theta}_{\text{MLE}}$ or simply $\hat{\theta}$, the parameter (or vector of parameters) that maximizes the likelihood function (probability of seeing the data).

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L\left(x_1, \ldots, x_n \mid \theta\right) = \arg\max_{\theta} \ln L\left(x_1, \ldots, x_n \mid \theta\right)$$

- **Log-Likelihood**: We define the log-likelihood as the natural logarithm of the likelihood function. 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.

  If X is discrete:

$$\ln L\left(x_1, \ldots, x_n \mid \theta\right) = \sum_{i=1}^{n} \ln p_X\left(x_i \mid \theta\right)$$

  If X is continuous:

$$\ln L\left(x_1, \ldots, x_n \mid \theta\right) = \sum_{i=1}^{n} \ln f_X\left(x_i \mid \theta\right)$$

Review of Main Concepts

- **Bias**: The bias of an estimator $\hat{\theta}$ for a true parameter θ is defined as $\text{Bias}\left(\hat{\theta}, \theta\right) = \mathbb{E}[\hat{\theta}] - \theta$. An estimator $\hat{\theta}$ of θ is unbiased iff $\text{Bias}\left(\hat{\theta}, \theta\right) = 0$, or equivalently $\mathbb{E}[\hat{\theta}] = \theta$.
- **Steps to find the maximum likelihood estimator, $\hat{\theta}$:**
  (a) Find the likelihood and log-likelihood of the data.
  (b) Take the derivative of the log-likelihood and set it to 0 to find a candidate for the MLE, $\hat{\theta}$.
  (c) Take the second derivative and show that $\hat{\theta}$ indeed is a maximizer, that $\frac{\partial^2 L}{\partial \theta^2} < 0$ at $\hat{\theta}$. Also ensure that it is the global maximizer: check points of non-differentiability and boundary values.

Problem 4 – Y me?

4

Let $y_1, y_2, ...y_n$ be i.i.d. samples of a random variable with density function

$$f_Y(y|\theta) = \frac{1}{2\theta} \exp\left(-\frac{|y|}{\theta}\right)$$

.

Find the MLE for θ in terms of $|y_i|$ and n.

Work on this problem with the people around you, and then we'll go over it together!

4

4

Since the samples are i.i.d., the likelihood of seeing n samples of them is just their PDFs multiplied together.

From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$
\begin{aligned}
L\left(y_1, \ldots, y_n \mid \theta\right) &= \prod_{i=1}^{n} \frac{1}{2\theta} \exp(-\frac{|y_i|}{\theta}) \\
\ln L\left(y_1, \ldots, y_n \mid \theta\right) &= \sum_{i=1}^{n} \left[-\ln 2 - \ln\theta - \frac{|y_i|}{\theta}\right] \\
\frac{\partial}{\partial\theta} \ln L\left(y_1, \ldots, y_n \mid \theta\right) &= \sum_{i=1}^{n} \left[-\frac{1}{\theta} + \frac{|y_i|}{\theta^2}\right] \\
\sum_{i=1}^{n} \left[-\frac{1}{\hat{\theta}} + \frac{|y_i|}{\hat{\theta}^2}\right] &= 0 \\
-\frac{n}{\hat{\theta}} + \frac{\Sigma_{i=1}^{n}|y_i|}{\hat{\theta}^2} &= 0 \\
\hat{\theta} = \frac{\Sigma_{i=1}^{n}|y_i|}{n}
\end{aligned}
$$

Problem 5 – A Biased Estimator

5

In class, we showed that the maximum likelihood estimate of the variance θ_2 of a normal distribution (when both the true mean μ and true variance σ^2 are unknown) is what's called the *population variance*. That is

$$\hat{\theta}_2 = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2) \right)$$

where $\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the MLE of the mean. Is $\hat{\theta}_2$ unbiased?

Work on this problem with the people around you, and then we'll go over it together!

5

5

Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$E(\hat{\theta}_2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)\right)$$

which by linearity of expectation (and distributing the sum) is

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E\left(\frac{2}{n}\overline{X}\sum_{i=1}^{n} X_i\right) + E(\overline{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - 2E(\overline{X}^2) + E(\overline{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2). \qquad (**)
\end{aligned}$$

We know that for any random variable Y, since $Var(Y) = E(Y^2) - (E(Y))^2$ it holds that

$$E(Y^2) = Var(Y) + (E(Y))^2.$$

5

Also, we have $E(X_i) = \mu,\ Var(X_i) = \sigma^2\ \forall i$ and $E(\overline{X}) = \mu,\ Var(\overline{X}) = \frac{\sigma^2}{n}$. Combining these facts, we get

$$E(X_i^2) = \sigma^2 + \mu^2 \ \ \forall i \quad \text{and} \quad E(\overline{X}^2) = \frac{\sigma^2}{n} + \mu^2.$$

Substituting these equations into (**) we get

$$\begin{aligned} E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2)\right) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \left(1 - \frac{1}{n}\right)\sigma^2. \end{aligned}$$

Thus $\hat{\theta}_2$ is not unbiased.

Problem 1 – Mystery Dish

1

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 served a dish independently. Let x_A be the number of people who received dish A, x_B the number of people who received dish B, etc, where $x_A + x_B + x_C + x_D = n$. Find the MLE for θ, $\hat{\theta}$.

1

The data tells us, for each diner in the restaurant, what their dish was. We begin by computing the likelihood of seeing the given data given our parameter θ. Because each diner is assigned a dish independently, the likelihood is equal to the product over diners of the chance they got the particular dish they got, which gives us:

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

From there, we just use the MLE process to get the log-likelihood, take the first derivative, set it equal to 0, and solve for $\hat{\theta}$.

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

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

$$\frac{x_B}{\hat{\theta}} + \frac{x_C}{\hat{\theta}} - \frac{3x_D}{0.5-3\hat{\theta}} = 0$$

Solving yields $\hat{\theta} = \frac{x_B+x_C}{6(x_B+x_C+x_D)}$.

Problem 3 – Independent Shreds

3

You are given 100 independent samples $x_1, x_2, \ldots, x_{100}$ 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. 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?

Work on this problem with the people around you, and then we'll go over it together!

3

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

3

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

Note that $\Sigma_{i\in[n]} x_i = 30$, as given in the problem spec. Therefore, there are 30 **1**s and 70 **0**s. (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 for $\hat{\theta}$.

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

3

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

3

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

An estimator is unbiased if the expectation of the estimator is equal to the original parameter, i.e.: $E[\hat{\theta}] = \theta$. Setting up the expectation of our estimator and plugging it in for the generic case, we get the following, which we can then reduce with linearity of expectation:

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

so it is unbiased.

Problem 6 – LTP Review

6

(a) (Discrete version) Suppose we flip a coin with probability U of heads, where U is equally likely to be one of $\Omega_U = \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$ (notice this set has size $n+1$). Let H be the event that the coin comes up heads. What is $\mathbb{P}(H)$?

(b) (Continuous version) Now suppose $U \sim$ Uniform(0,1) has the *continuous* uniform distribution over the interval $[0,1]$. What is $\mathbb{P}(H)$?

(c) Let's generalize the previous result we just used. Suppose E is an event, and X is a continuous random variable with density function $f_X(x)$. Write an expression for $\mathbb{P}(E)$, conditioning on X.

Work on this problem with the people around you, and then we'll go over it together!

6

(a) (Discrete version) Suppose we flip a coin with probability U of heads, where U is equally likely to be one of $\Omega_U = \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$ (notice this set has size $n + 1$). Let H be the event that the coin comes up heads. What is $\mathbb{P}(H)$?

6

(a) (Discrete version) Suppose we flip a coin with probability U of heads, where U is equally likely to be one of $\Omega_U = \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$ (notice this set has size $n+1$). Let H be the event that the coin comes up heads. What is $\mathbb{P}(H)$?

We can use the law of total probability, conditioning on $U = \frac{k}{n}$ for $k = 0, ..., n$. Note that the probability of getting heads conditioning on a fixed U value is U, and that the probability of U taking on any value in its range is $\frac{1}{n+1}$ since it is discretely uniform.

$$\mathbb{P}(H) = \sum_{k=0}^{n} \mathbb{P}(H|U = \frac{k}{n})\mathbb{P}(U = \frac{k}{n}) = \sum_{k=0}^{n} \frac{k}{n} \cdot \frac{1}{n+1} = \frac{1}{n(n+1)} \sum_{k=0}^{n} k = \frac{1}{n(n+1)} \frac{n(n+1)}{2} = \frac{1}{2}$$

6

(b) (Continuous version) Now suppose $U \sim$ Uniform(0,1) has the *continuous* uniform distribution over the interval $[0, 1]$. What is $\mathbb{P}(H)$?

6

(b) (Continuous version) Now suppose $U \sim$ Uniform(0,1) has the *continuous* uniform distribution over the interval $[0, 1]$. What is $\mathbb{P}(H)$?

We do the same thing, this time using the continuous law of total probability. Note, this time, that we're conditioning on $U = u$ and taking the integral with respect to u, and that the density of U for any value in its range is 1 because it is uniformly random.

$$\mathbb{P}(H) = \int_{-\infty}^{\infty} \mathbb{P}(H|U = u) f_U(u) du$$

We can take the integral from 0 to 1 instead because outside of that range the density of U is 0.

$$= \int_0^1 \mathbb{P}(H|U = u) f_U(u) du = \int_0^1 u \cdot 1 du = \frac{1}{2}[u^2]_0^1 = \frac{1}{2}$$

(c) Let's generalize the previous result we just used. Suppose E is an event, and X is a continuous random variable with density function $f_X(x)$. Write an expression for $\mathbb{P}(E)$, conditioning on X.

6

(c) Let's generalize the previous result we just used. Suppose E is an event, and X is a continuous random variable with density function $f_X(x)$. Write an expression for $\mathbb{P}(E)$, conditioning on X.

We use the continuous law of total probability again, this time not deriving it any further and sticking with negative infinity to infinity because we don't know the range of the RV X.

$$\mathbb{P}(E) = \int_{-\infty}^{\infty} \mathbb{P}(E|X = x) f_X(x) dx$$

That's All, Folks!

Thanks for coming to section this week!
Any questions?