Section 9 Slides

Solution by Anna Karlin, William Howard-Snyder and Aleks Jovcic

Slides by Leiyi Zhang and Scott Ni

Administrivia & Agenda

- Pset 8 is out
- The quarter is almost over, you've got this!!

Law of total probability and law of total expectation

1) **Law of Total Probability (partition based on value of a r.v.):** If X is a discrete random variable, then

$$\mathbb{P}(A) = \sum_{x \in \Omega_X} \mathbb{P}(A|X = x) p_X(x)$$

If X is a continuous random variable, then

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

2) **Conditional Expectation**: Let X and Y be random variables. Then, the conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X|Y = y] = \sum_{x \in \Omega_X} x \cdot \mathbb{P}\left(X = x|Y = y\right) \qquad X \text{ discrete}$$

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x \cdot \mathbb{P}\left(X = x|Y = y\right)\ dx \qquad X \text{ continuous}$$

and for any event A,

$$\mathbb{E}[X|A] = \sum_{x \in \Omega_X} x \cdot \mathbb{P}\left(X = x|A\right) \qquad X \text{ discrete}$$

$$\mathbb{E}[X|A] = \int_{-\infty}^{\infty} x \cdot \mathbb{P}\left(X = x|A\right)\ dx \qquad X \text{ continuous}$$

Note that linearity of expectation still applies to conditional expectation: $\mathbb{E}[X + Y|A] = \mathbb{E}[X|A] + \mathbb{E}[Y|A]$

3) **Law of Total Expectation (Event Version)**: Let X be a random variable, and let events $A_1, ..., A_n$ partition the sample space. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i]\mathbb{P}(A_i)$$

4) **Law of Total Expectation (RV Version)**: Suppose X and Y are random variables. Then,

$$\mathbb{E}[X] = \sum_{y} \mathbb{E}[X|Y=y]p_Y(y) \qquad Y \text{ discrete r.v..}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y]f_Y(y)dy \qquad Y \text{ continuous r.v.}$$

Maximum Likelihood Estimation

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

2) **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:

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

   If X is continuous:

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

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

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

4) **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:

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

5) **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.

(d) If we are finding the MLE for a set of parameters, then we set up the system of equations obtained by taking the partial derivative of the log-likelihood function with respect to each of the parameters and setting it equal to 0. We then solve this system to get the MLEs. (And again, second order conditions need to be checked.)

6) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

Markov Chains: (to be covered)

1) A **discrete-time stochastic process (DTSP)** is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \ldots$, where $X^{(t)}$ is the value at time t. For example, the temperature in Seattle or stock price of TESLA each day, or which node you are at after each time step on a random walk on a graph.

2) A **Markov Chain** is a DTSP, with the additional following three properties:

(a) ...has a finite (or countably infinite) **state space** $\mathcal{S} = \{s_1, \ldots, s_n\}$ which it bounces between, so each $X^{(t)} \in \mathcal{S}$.

(b) ...satisfies the **Markov property**. A DTSP satisfies the Markov property if the future is (conditionally) independent of the past given the present. Mathematically, it means,

$$\mathbb{P}\left(X^{(t+1)} = x_{t+1} \mid X^{(0)} = x_0, X^{(1)} = x_1, \ldots, X^{(t-1)} = x_{t-1}, X^{(t)} = x_t\right) = \mathbb{P}\left(X^{(t+1)} = x_{t+1} \mid X^{(t)} = x_t\right) .$$

(c) ...has **fixed transition probabilities**. Meaning, if we are at some state s_i, we transition to another state s_j with probability *independent* of the current time. Due to this property and the previous, the transitions are governed by n^2 probabilities: the probability of transitioning from one of n current states to one of n next states. These are stored in a square $n \times n$ **transition probability matrix (TPM)** $\mathbf{M}$, where $M_{ij} = \mathbb{P}\left(X^{(t+1)} = s_j \mid X^{(t)} = s_i\right)$ is the probability of transitioning from state s_i to state s_j for any/every value of t.

3) A **stationary distribution** of a Markov chain is a probability distribution on states that is unchanged by taking one step of the Markov chain. It is obtained by solving the matrix equation $\pi M = \pi$, where $\pi = (\pi_1, \ldots, \pi_n)$ is a row vector, with π_i the stationary probability of being in state i. Note that we must have $\sum_i \pi_i = 1$.

Question 2: Lemonade Stand

Suppose I run a lemonade stand, which costs me \$100 a day to operate. I sell a drink of 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 independently with probability p_2. It rains each day with probability p_3, 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 and p_3, what is $\mathbb{E}[X]$? Use the law of total expectation.

Solution 2: Lemonade Stand

Let R be the event it rains. Let X_i be how many drinks I sell on day i for $i = 1, ..., 7$. We are interested in $X = \sum_{i=1}^{7} (20X_i - 100)$. We have $X_i|R \sim$ Binomial(n_1, p_1), so $\mathbb{E}[X_i|R] = n_1 p_1$. Similarly, $X_i|R^C \sim$ Binomial(n_2, p_2), so $\mathbb{E}[X_i|R^C] = n_2 p_2$. By the law of total expectation,

$$\mu = \mathbb{E}[X_i] = \mathbb{E}[X_i|R]\mathbb{P}(R) + \mathbb{E}[X_i|R^C]\mathbb{P}\left(R^C\right) = n_1 p_1 p_3 + n_2 p_2 (1 - p_3)$$

Hence, by linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{7} (20X_i - 100)] = 20 \sum_{i=1}^{7} \mathbb{E}[X_i] - 700 = 140\mu - 700$$

$$= 140 \cdot (n_1 p_1 p_3 + n_2 p_2 (1 - p_3)) - 700.$$

Question 3: Mystery Dish!

A fancy new restaurant has opened up that 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 $\hat{\theta}$ for θ.

Solution 3: Mystery Dish!

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:

Solution 3: Mystery Dish!

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:

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

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 $\hat{\theta}$.

Solution 3: Mystery Dish!

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:

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

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 $\hat{\theta}$.

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

Solution 3: Mystery Dish!

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:

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

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 $\hat{\theta}$.

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

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(x \mid \theta) = \frac{x_B}{\theta} + \frac{x_C}{\theta} - \frac{3x_D}{0.5-3\theta}$$

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

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

Question 5: A biased estimator

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?

Question 5: A biased estimator

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?

6) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

6) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

Solution 5: A biased estimator

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

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

6) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

Solution 5: A biased estimator

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

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

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

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\frac{2}{n}\overline{X}\sum_{i=1}^{n}X_i\right] + \mathbb{E}\left[\overline{X}^2\right]$$

6) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

Solution 5: A biased estimator

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

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

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

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

6) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

Solution 5: A biased estimator

We know that for any random variable Y, since $\mathrm{Var}(Y) = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2$ it holds that

$$\mathbb{E}\left[Y^2\right] = \mathrm{Var}(Y) + (\mathbb{E}[Y])^2.$$

6) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

Solution 5: A biased estimator

We know that for any random variable Y, since $\text{Var}(Y) = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2$ it holds that

$$\mathbb{E}\left[Y^2\right] = \text{Var}(Y) + (\mathbb{E}[Y])^2.$$

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

$$\mathbb{E}\left[X_i^2\right] = \sigma^2 + \mu^2 \quad \forall i \qquad \text{and} \qquad \mathbb{E}\left[\overline{X}^2\right] = \frac{\sigma^2}{n} + \mu^2.$$

6) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

Solution 5: A biased estimator

Substituting these equations into (**) we get

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}^2\right] = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= \left(1 - \frac{1}{n}\right)\sigma^2.$$

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