Math 170S

Lecture Notes Section 6.4 *†

Maximum likelihood estimation

Instructor: Swee Hong Chan

WARNING: We have now reached possibly the most important concept in 170S. Please please please ask me questions if you are lost at any point.

---

*Version date: Sunday 11^{th} October, 2020, 20:41.

†This notes is based on Hanbaek Lyu's and Liza Rebrova's notes from the previous quarter, and I would like to thank them for their generosity. "*Nanos gigantum humeris insidentes* (I am but a dwarf standing on the shoulders of giants)".

1 Example of MLE: motivation

The friendly instructor owns an unknown random variable X for you to guess. He gives you three hints:

- X is a Bernoulli random variable for some unknown parameter p, with pmf

$$f_X(x) = P[X = x] \;=\; p^x(1-p)^{1-x} \qquad x = 0, 1.$$

- p is one of these four numbers,

$$p \in \{0,\ 0.2,\ 0.7,\ 1\}.$$

- 5 sample values x_1, x_2, x_3, x_4, x_5 are given.

Your job is to **guess the unknown parameter p**.

Here are some possible scenarios:

- $x_1 = x_2 = \ldots = x_5 = 0$. In this case, most people would guess $p = 0$.

- $x_1 = x_2 = \ldots = x_5 = 1$. In this case, most people would guess $p = 1$.

- $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, $x_4 = 1$, $x_5 = 1$. Intuitively, most people would guess $p = \frac{3}{5} = 0.6$. However, 0.6 is not an option for p, so we choose the closest value $p = 0.7$.

These guesses are indeed the best guess one can make, in a manner to be made precise.

Let X be the Bernoulli random variable with unknown parameter p.

Suppose that the samples are $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, $x_4 = 1$, $x_5 = 1$.

The probability for this particular outcome is: (BT)

$$
\begin{aligned}
& P[X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5] \\
&= P[X_1 = x_1]\ P[X_2 = x_2]\ \ldots\ P[X_5 = x_5] \\
&= p^{x_1}(1-p)^{1-x_1}\ p^{x_2}(1-p)^{1-x_2}\ \ldots\ p^{x_5}(1-p)^{1-x_5} \\
&= p^{x_1+x_2+x_3+x_4+x_5}\ (1-p)^{5-(x_1+x_2+x_3+x_4+x_5)} \\
&= p^3(1-p)^2.
\end{aligned}
$$

We call this function $L(p)$.

Let us test the value of $L(p)$ for four choices of p:

- When $p = 0$,

  $$L(p) = 0^3(1-0)^2 = 0.$$

  So if $p = 0$, the probability to see this particular outcome is 0. Clearly not a good choice.

- When $p = 1$, we have

  $$L(p) = 1^3(1-1)^2 = 0.$$

  Again the probability to see this particular outcome is 0. Also not a good choice.

- When $p = 0.2$, we have

$$L(p) = (0.2)^3(1 - 0.2)^2 = 0.00512.$$

- When $p = 0.7$, we have

$$L(p) = (0.7)^3(1 - 0.7)^2 = 0.03087.$$

Our particular outcomes can occur for both $p = 0.2$ and $p = 0.7$, but it is six times as likely if $p = 0.7$! That is why we choose $p = 0.7$.

The best guess for p would be the value that maximizes the function $L(p)$.

2 The problem

- **Assumption:** X is a random variable with distribution $f_{X;\theta}$ with unknown θ.
- **Problem:** Predict the unknown θ, and thus the unknown random variable X;
- **Input:** Samples $x_1, \ldots, x_n$.

Our strategy is to find θ that maximizes the likelihood function.

3 Maximum likelihood estimate

The **likelihood function** $L(\theta) := L(x_1, \ldots, x_n; \theta)$ is

$$L(\theta) \quad := \quad f_{X;\theta}(x_1) f_{X;\theta}(x_2) \ldots f_{X;\theta}(x_n).$$

Here $x_1, x_2, \ldots, x_n$ are fixed values, only θ is variable.

The **log likelihood function** $\ell(\theta)$ is

$$\ell(\theta) := \log L(\theta).$$

The **maximum likelihood estimate (MLE)** is the value of θ that maximizes $L(\theta)$, or equivalently $\ell(\theta)$. The MLE is usually denoted $\widehat{\theta}$.

4 Example: Bernoulli

Let X be a Bernoulli RV with unknown $p \in [0, 1]$.

Let $x_1, x_2, \ldots, x_n$ be unknown samples of X.

Let $\overline{\mathrm{x}}$ be the sample mean $\frac{x_1+\ldots+x_n}{n}$.

The likelihood function $L(p)$[1] for Bernoulli is (BT)

$$\begin{aligned} L(p) &= f_{X;p}(x_1)\ldots f_{X;p}(x_n) \\ &= p^{x_1}(1-p)^{1-x_1}\, p^{x_2}(1-p)^{1-x_2}\ \ldots\ p^{x_n}(1-p)^{1-x_n} \\ &= p^{x_1+\ldots+x_n}(1-p)^{n-(x_1+\ldots+x_n)} \\ &= p^{n\,\overline{\mathrm{x}}}(1-p)^{n-n\,\overline{\mathrm{x}}}. \end{aligned}$$

The log likelihood function is

$$\ell(p) = \log\left(p^{n\,\overline{\mathrm{x}}}(1-p)^{n-n\,\overline{\mathrm{x}}}\right) = n\,\overline{\mathrm{x}}\log p + (n - n\,\overline{\mathrm{x}})\log(1-p).$$

[1]We substitute θ for p here

To maximize the log likelihood function, we take the partial derivatives in p:

$$\frac{\partial \ell(p)}{dp} = \frac{n\,\overline{\mathrm{x}}}{p} - \frac{n - n\,\overline{\mathrm{x}}}{1-p}.$$

The MLE $\widehat{\mathrm{p}}$ is a maximizer of $\ell(p)$, so $\frac{\partial \ell(\widehat{\mathrm{p}})}{dp} = 0$. This gives us (BT)

$$\begin{aligned}
0 &= \frac{n\,\overline{\mathrm{x}}}{\widehat{\mathrm{p}}} - \frac{n - n\,\overline{\mathrm{x}}}{1-\widehat{\mathrm{p}}} \\
\frac{n\,\overline{\mathrm{x}}}{\widehat{\mathrm{p}}} &= \frac{n - n\,\overline{\mathrm{x}}}{1-\widehat{\mathrm{p}}} \\
\frac{1-\widehat{\mathrm{p}}}{\widehat{\mathrm{p}}} &= \frac{n - n\,\overline{\mathrm{x}}}{n\,\overline{\mathrm{x}}} \\
\frac{1}{\widehat{\mathrm{p}}} - 1 &= \frac{1}{\overline{\mathrm{x}}} - 1 \\
\widehat{\mathrm{p}} &= \overline{\mathrm{x}}\,.
\end{aligned}$$

So $\widehat{\mathrm{p}}$ is either 0, $\overline{\mathrm{x}}$, or 1.

These three options give us

$$
\begin{aligned}
\ell(0) &= -\infty; \\
\ell(p) &= n\,\overline{\mathrm{x}}\log\overline{\mathrm{x}} - (n - n\,\overline{\mathrm{x}})\log(1 - \overline{\mathrm{x}}); \\
\ell(1) &= -\infty;
\end{aligned}
$$

So the MLE for Bernoulli is $\widehat{\mathrm{p}} = \overline{\mathrm{x}}$, which is the sample mean. This confirms our usual intuition.

5 Example: Exponential

Let X be the exponential random variable with unknown parameter $\theta \in [0, \infty]$,

$$f_{X;\theta}(x) = \begin{cases} \theta^{-1} e^{-x/\theta} & \text{if } x \geq 0; \\ \\ 0 & \text{if } x < 0. \end{cases}$$

Let $x_1, \ldots, x_n$ be (nonnegative) samples.

The likelihood function is (BT)

$$\begin{aligned} L(\theta) =& f_{X;\theta}(x_1) f_{X;\theta}(x_2) \ldots f_{X;\theta}(x_n) \\ =& \theta^{-1} e^{-x_1/\theta} \; \theta^{-1} e^{-x_2/\theta} \; \ldots \; \theta^{-1} e^{-x_n/\theta} \\ =& \theta^{-n} e^{-(x_1 + \ldots + x_n)/\theta} = \theta^{-n} e^{-n \, \overline{\mathrm{x}} / \theta}. \end{aligned}$$

The log likelihood function is

$$\ell(\theta) = \log \left(\theta^{-n} e^{-\theta n \, \overline{\mathrm{x}}} \right) \; = \; -n \log \theta - \frac{n \, \overline{\mathrm{x}}}{\theta}.$$

To maximize the $\ell(\theta)$, we take partial derivatives in θ:

$$\frac{\partial \ell(\theta)}{d\theta} = -\frac{n}{\theta} + \frac{n\,\overline{\mathrm{x}}}{\theta^2}.$$

The MLE $\widehat{\theta}$ is a maximizer of $\ell(\widehat{\theta})$, and therefore $\frac{\partial \ell(\widehat{\theta})}{d\theta} = 0$. This gives us (BT)

$$\begin{aligned} 0 &= -\frac{n}{\widehat{\theta}} + \frac{n\,\overline{\mathrm{x}}}{\widehat{\theta}^2} \\ \widehat{\theta} &= \overline{\mathrm{x}} \end{aligned}$$

Therefore the MLE for exponential RVs is sample mean, again confirming our intuition.

6 Example: Normal

Suppose that X is a normal RV with unknown mean μ and variance σ^2,

$$f_{X;\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Let $\overline{\mathrm{x}}$ be the sample mean, and v be the variance of the empirical distribution,

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{\mathrm{x}})^2.$$

The likelihood function $L(\mu, \sigma^2)$ is[2] (BT)

$$\begin{aligned} L(\mu, \sigma^2) &= f_{X;\mu,\sigma^2}(x_1) \dots f_{X;\mu,\sigma^2}(x_n) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_1 - \mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right). \end{aligned}$$

The log likelihood function is

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

[2]the variable θ is replaced by two variables μ and σ^2

To maximize the log likelihood function, we take partial derivatives in μ and σ^2: (BT)

$$\begin{aligned}
\frac{\partial \ell(\mu, \sigma^2)}{d\mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \;=\; \frac{n}{\sigma^2}(\overline{\mathrm{x}} - \mu); \\
\frac{\partial \ell(\mu, \sigma^2)}{d\sigma^2} &= -\frac{n}{2} \frac{2\pi}{2\pi\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \\
&= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2.
\end{aligned}$$

The MLE $\widehat{\mu}$ and $\widehat{\sigma}^2$ satisfies $\frac{\partial \ell(\widehat{\mu}, \widehat{\sigma}^2)}{d\mu} = \frac{\partial \ell(\widehat{\mu}, \widehat{\sigma}^2)}{d\sigma^2} = 0$. Solving for $\frac{\partial \ell(\widehat{\mu}, \widehat{\sigma}^2)}{d\mu} = 0$ gives us (BT)

$$\begin{aligned}
0 &= \frac{n}{\widehat{\sigma}^2}(\overline{\mathrm{x}} - \widehat{\mu}) \\
\overline{\mathrm{x}} &= \widehat{\mu}\,.
\end{aligned}$$

Solving for $\frac{\partial \ell(\widehat{\mu}, \widehat{\sigma}^2)}{d\sigma^2} = 0$ gives us (BT)

$$
\begin{aligned}
0 &= -\frac{n}{2}\frac{1}{\widehat{\sigma}^2} + \frac{1}{2\,\widehat{\sigma}^4}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2 \\
n\,\widehat{\sigma}^2 &= \sum_{i=1}^{n}(x_i - \widehat{\mu})^2 \\
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{\mathrm{x}})^2 \;=\; v.
\end{aligned}
$$

Thus the maximum likelihood estimate $\widehat{\mu}$ is the sample mean, and $\widehat{\sigma}^2$ is the variance of the empirical distribution (NOT sample variance!).

7 Unbiased estimator: Bernoulli

Let X be a Bernoulli random variable with unknown parameter p. We have seen that the MLE is equal to

$$\widehat{\mathrm{p}} = \frac{x_1 + \ldots + x_n}{n}.$$

We make the following observations:

1. $\widehat{\mathrm{p}} := \widehat{\mathrm{p}}(x_1, \ldots, x_n)$ is a function that depends on the value of $x_1, x_2, \ldots, x_n$;

2. For small n, the MLE $\widehat{\mathrm{p}}$ can be very different from the real value p. However, as n gets larger, $\widehat{\mathrm{p}}$ should be very close to p.

The samples $x_1, \ldots, x_n$ is deterministic **AFTER** we complete the experiment, but **BEFORE** the experiment they are unknown and it stands to reason to assume that they are random variables.

Therefore, we replace deterministic numbers $x_1, \ldots, x_n$ with random numbers $X_1, \ldots, X_n$, which are are independent Bernoullis with (unknown) parameter p. Then

$$\widehat{\mathrm{p}}(X_1, \ldots, X_n) = \frac{X_1 + \ldots + X_n}{n}.$$

Taking the expectation of $\widehat{\mathrm{p}}$ as a random variable,

$$\begin{aligned} E[\widehat{\mathrm{p}}(X_1, \ldots, X_n)] &= E\left[\frac{X_1 + \ldots + X_n}{n}\right] \\ &= \frac{E[X_1] + \ldots + E[X_n]}{n} \\ &= \frac{p + p + \ldots + p}{n} = p. \end{aligned}$$

Therefore, if we repeat the experiment often enough, the MLE $\widehat{\mathrm{p}}(X_1, \ldots, X_n)$ is indeed equal to p on average!

Not all MLEs have this property, and the one that does is called unbiased estimator.

8 Unbiased estimator

Let X be a random variable with parameter θ.

Let $\widehat{\theta} := \theta(x_1, \ldots, x_n)$ be an MLE for X.

Let $X_1, \ldots, X_n$ be independent random variables with the same distribution as X.

Then $\widehat{\theta}$ is an **unbiased estimator** if

$$E\left[\,\widehat{\theta}(X_1, \ldots, X_n)\right] = \theta.$$

9 Unbiased estimator: Normal

Let X be a normal random variable with mean μ and variance σ^2. The MLEs are

$$\begin{aligned}
\widehat{\mu}(x_1, \ldots, x_n) &= \frac{x_1 + x_2 + \ldots + x_n}{n}; \\
\widehat{\sigma}^2(x_1, \ldots, x_n) &= \frac{x_1^2 + \ldots + x_n^2}{n} - \left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right)^2.
\end{aligned}$$

We check if $\widehat{\mu}$ is an unbiased estimator: (BT)

$$\begin{aligned}
E[\widehat{\mu}(X_1, \ldots, X_n)] &= E\left[\frac{X_1 + \ldots + X_n}{n}\right] \\
&= \frac{E[X_1] + \ldots + E[X_n]}{n} \\
&= \frac{\mu + \mu + \ldots + \mu}{n} = \mu.
\end{aligned}$$

Indeed $\widehat{\mu}$ is an unbiased estimator for μ.

We check if $\widehat{\sigma}^2$ is an unbiased estimator: (BT)

$$
\begin{aligned}
& E[\widehat{\sigma}^2(X_1,\ldots,X_n)] \\
&= E\left[\frac{X_1^2+\ldots+X_n^2}{n} - \left(\frac{X_1+\ldots+X_n}{n}\right)^2\right] \\
&= E\left[\frac{X_1^2+\ldots+X_n^2}{n} - \left(\frac{X_1^2+\ldots+X_n^2}{n^2} + \sum_{1\le i\neq j\le n}\frac{X_iX_j}{n^2}\right)\right] \\
&= E\left[\frac{n-1}{n^2}\left(X_1^2+\ldots+X_n^2\right) - \sum_{1\le i\neq j\le n}\frac{X_iX_j}{n^2}\right] \\
&= \frac{n-1}{n^2}\left(E[X_1^2]+\ldots+E[X_n^2]\right) - \sum_{1\le i\neq j\le n}\frac{E[X_i]E[X_j]}{n^2} \\
&= \frac{n-1}{n^2}\left(E[X^2]+\ldots+E[X^2]\right) - \sum_{1\le i\neq j\le n}\frac{E[X]E[X]}{n^2} \\
&= \frac{n-1}{n^2}nE[X^2] - n(n-1)\frac{E[X]E[X]}{n^2} \\
&= \frac{n-1}{n}\left(E[X^2]-(E[X])^2\right) \;=\; \frac{n-1}{n}\sigma^2.
\end{aligned}
$$

So $\widehat{\sigma}$ is a **biased** estimator for the variance σ^2!

To correct this bias, we use the sample variance,

$$s^2 \ = \ \frac{n}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n^2}\right)$$

as the MLE $\widehat{\sigma}^2$. Indeed, by using $\widehat{\sigma}^2 = s^2$:

$$\begin{aligned} E[\widehat{\sigma}^2] \ &= E\left[\frac{n}{n-1}\left(\frac{X_1^2 + \ldots + X_n^2}{n} - \left(\frac{X_1 + \ldots + X_n}{n}\right)^2\right)\right] \\ &= \frac{n}{(n-1)}\frac{(n-1)}{n}\sigma^2 = \sigma^2, \end{aligned}$$

which is indeed an unbiased estimator.

10 Method of moments: Intro

Before the method of MLE was popularized by Fischer in the beginning of 20th century, the **method of moments** was widely used.

In the modern era we still use method of moments sometimes, as it is more computationally efficient than MLE.

Recall that the k-th moment of X is the quantity $E[X^k]$.

11 Method of moments

- **Assumption:** X is a random variable with unknown parameter θ and moments $M_1(\theta) := E[X]$, $M_2(\theta) := E[X^2]$, ...

- **Problem:** Predict the unknown parameter θ.

- **Input:** Samples $x_1, \ldots, x_n$ sampled from X.

- **Solution:** Our guess for θ is a value $\widetilde{\theta}$ such that

$$\begin{aligned} M_1(\theta) &= \frac{x_1 + \ldots x_n}{n}; \\ M_2(\theta) &= \frac{x_1^2 + \ldots x_n^2}{n}; \\ &\vdots \end{aligned}$$

It can be impractical to check for all moments of X, and one usually stops at the second moment.

12 Method of moments: Gamma

Let X be a gamma random variable with unknown parameters α, β,

$$f_{X;\alpha,\beta}(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if } x \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

where $\Gamma(x)$ is the function

$$\Gamma(x) := \int_0^\infty t^x e^{-t} \, dt.$$

The MLE for this random variable is

$$L(\alpha, \beta) = \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \right]^n (x_1 x_2 \ldots x_n)^{\alpha-1} \exp\left(-\beta \sum_{i=1}^n x_i \right).$$

Maximizing the function above is not meant for mortals, because the term $\Gamma(x)$ in $L(\alpha, \beta)$ is hard to compute.

Instead, we use the fact that $\Gamma(\alpha, \beta)$ has moments:

$$M_1(\alpha, \beta) = \frac{\alpha}{\beta}; \qquad M_2(\alpha, \beta) = \frac{\alpha + \alpha^2}{\beta^2}.$$

Solving for the first moment, (BT)

$$\begin{aligned} M_1(\widetilde{\alpha}, \widetilde{\beta}) &= \frac{x_1 + \ldots x_n}{n} \\ \frac{\widetilde{\alpha}}{\widetilde{\beta}} &= \frac{x_1 + \ldots x_n}{n} = \overline{\mathrm{x}}\,. \end{aligned}$$

Solving for the second moment,

$$\begin{aligned} M_2(\widetilde{\alpha}, \widetilde{\beta}) &= \frac{x_1^2 + \ldots x_n^2}{n} \\ \frac{\widetilde{\alpha} + \widetilde{\alpha}^2}{\widetilde{\beta}^2} &= v + \overline{\mathrm{x}}^2 \end{aligned}$$

Substituting the first equation into the second equation,

$$\frac{\widetilde{\alpha} + \widetilde{\alpha}^2}{\widetilde{\beta}^2} = v + \frac{\widetilde{\alpha}^2}{\widetilde{\beta}^2}$$
$$\frac{\widetilde{\alpha}}{\widetilde{\beta}^2} = v.$$

Combining $\frac{\widetilde{\alpha}}{\widetilde{\beta}^2} = v$ and $\frac{\widetilde{\alpha}}{\widetilde{\beta}} = \overline{\mathrm{x}}$, we then conclude that

$$\widetilde{\alpha} = \frac{\overline{\mathrm{x}}^2}{v}; \qquad \widetilde{\beta} = \frac{\overline{\mathrm{x}}}{v}.$$