CS 536 : Estimation

16:198:536

In the usual way, we imagine a sequence of i.i.d. random samples $X_1, X_2, \ldots, X_n$ samples from some underlying distribution P. For the moment, it is convenient to take the X_i to be distributed over the real numbers, with density f. It is a problem of general interest to take a sequence of samples $\{X_i\}_{i=1,\ldots,n}$ and try to construct an estimate of the properties of the underlying distribution, if not the distribution itself. That is, if $\mathcal{F}$ represents a class of distributions or hypotheses, we want to construct a $\hat{f} \in \mathcal{F}$ such that $\hat{f} \approx f$. This is generally a hard problem, but we are assisted somewhat if the class of distributions we are interested in is *parameterized*, in which case we can try to estimate the parameters rather than the density (or mass) function itself. Consider the situation where the X_i are normally distributed, with mean μ and variance σ^2:

$$\begin{aligned} f(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x) dx = \mu \\ \mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu)^2\right] = \sigma^2. \end{aligned} \tag{1}$$

Given a sequence of i.i.d. samples with the above distribution, we want to construct estimators for μ and σ^2.

Method of Moments

The method of moments is to try to approximate the parameters of the underlying distribution by first approximating the *moments* with the sample moments. The k-th moment is defined to be the expected value $\mathbb{E}[X^k]$, and we can approximate it with the k-th sample moment $(1/n)\sum X_i^k$. In the case of a normal distribution, the first two moments are given by

$$\begin{aligned} \mathbb{E}[X] &= \mu \\ \mathbb{E}\left[X^2\right] &= \sigma^2 + \mu^2 \textit{(why?)}. \end{aligned} \tag{2}$$

The method of moments then sets up the following relationships for the estimators $\hat{\mu}$ and $\hat{\sigma}^2$:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i &= \hat{\mu} \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \hat{\sigma}^2 + \hat{\mu}^2, \end{aligned} \tag{3}$$

which solves to yield the following estimators:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2, \end{aligned} \tag{4}$$

i.e., yields the estimators of the usual sample mean and sample variance. Note, denoting the sample mean as $\bar{X}_n$, the expression above for $\hat{\sigma}^2$ un-simplifies to

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\frac{1}{n}\sum_{i=1}^{n} X_i\bar{X}_n + \bar{X}_n^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i\bar{X}_n + n\bar{X}_n^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2\bar{X}_n + \bar{X}_n^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2,
\end{aligned} \tag{5}$$

i.e., the standard (though biased) sample variance.

In general, the Method of Moments requires setting up a system of equations relating some number of moments to the parameters of the distribution, solving for the parameters in terms of the moments, and estimating the results based on the sample moments.

Maximum Likelihood Estimation

The *likelihood* of a set of data can be though of as the probability that the data occurred (or in the continuous case, the joint density of that particular set of data). Let f be the p.d.f. or p.m.f. of a distribution, then we can define the likelihood as the probability or joint density of receiving that specific data set:

$$\text{Lik}(f) = \prod_{i=1}^{n} f\left(X_i\right). \tag{6}$$

Over a family of distributions $\mathcal{F}$, the *maximum likelihood estimator* asks which would make the data that occurred as likely as possible? In other words,

$$\hat{f}_{\text{MLE}} = \text{argmax}_{f\in\mathcal{F}}\text{Lik}(f). \tag{7}$$

The mechanics of MLEs are nice in general, because it exchanges solving the system of estimator equations for an optimization problem, namely,

$$\max_{f\in\mathcal{F}} \prod_{i=1}^{n} f(X_i), \tag{8}$$

and the standard techniques and algorithms of optimization may be applied here. In general, instead of maximizing the product, this is generally expressed as minimizing the negative log of the likelihood, to convert the product into a sum:

$$\min_{f\in\mathcal{F}} -\ln \text{Lik}(f) = \min_{f\in\mathcal{F}} \sum_{i=1}^{n} -\ln f(X_i). \tag{9}$$

In the case of normal distributions, we have

$$\begin{aligned}
\ln \mathrm{Lik}\left(\mu, \sigma^2\right) &= \sum_{i=1}^{n}\left[-\frac{1}{2\sigma^2}\left(X_i-\mu\right)^2 - \ln\left(\sigma\sqrt{2\pi}\right)\right] \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2 - \frac{1}{2}n\ln\left(\sigma^2 2\pi\right) \\
&= -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}X_i^2 - 2\mu\sum_{i=1}^{n}X_i + n\mu^2\right] - \frac{1}{2}n\ln\left(\sigma^2 2\pi\right) \\
&= -\frac{1}{2\sigma^2}n\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\mu\bar{X}_n + \mu^2\right] - \frac{1}{2}n\ln\left(\sigma^2 2\pi\right) \\
&= -\frac{1}{2\sigma^2}n\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}_n^2 + \bar{X}_n^2 - 2\mu\bar{X}_n + \mu^2\right] - \frac{1}{2}n\ln\left(\sigma^2 2\pi\right) \\
&= -\frac{1}{2\sigma^2}n\left[\bar{X}_n-\mu\right]^2 - \frac{1}{2\sigma^2}n\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}_n^2\right] - \frac{1}{2}n\ln\left(\sigma^2 2\pi\right)
\end{aligned} \tag{10}$$

Using the algebra from the previous section, this simplifies to

$$-\ln \mathrm{Lik}\left(\mu, \sigma^2\right) = \frac{1}{2}n\left[\left(\frac{\bar{X}_n-\mu}{\sigma}\right)^2 + \frac{1}{\sigma^2}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)^2\right] + \ln\left(\sigma^2 2\pi\right)\right], \tag{11}$$

which is minimized at

$$\begin{aligned}
\hat{\mu} &= \bar{X}_n \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)^2,
\end{aligned} \tag{12}$$

i.e., exactly the same estimators as under the Method of Moments.

In general however, it is not guaranteed that the method of moments estimators and the maximum likelihood estimators will agree.

An Example when MOM and MLE Disagree: Uniform Random Variables

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables distributed uniformly over $[0, L]$, i.e., with a p.d.f. of $f(x) = 1/L$ on this interval, $f(x) = 0$ off this interval. In this case, the family of distributions is parameterized by the value L - thus we want to estimate this parameter based on the values of our samples.

Estimating via moments, it's relatively easy to show that

$$\mathbb{E}\left[X\right] = \frac{1}{2}L, \tag{13}$$

which gives us a fairly immediate and natural estimator

$$\hat{L} = 2\bar{X}_n. \tag{14}$$

The maximum likelihood estimator takes some more thought, however. Note that for a given value of L, the density at a value x is not simply $1/L$ - if $x > L$ the density is 0!

$$f_L(x) = \begin{cases} \frac{1}{L} & \text{if } x \le L \\ 0 & \text{else} \end{cases} \tag{15}$$

. Taking this view then, the likelihood of a given value of L is $1/L^n$ - *as long as all the X_i are at most L!*

$$\text{Lik}(L) = \begin{cases} \frac{1}{L^n} & \text{if } \max_i X_i \leq L \\ 0 & \text{else} \end{cases}. \tag{16}$$

Here, the maximum likelihood is achieved using the smallest possible L for which the likelihood is non-zero, i.e., giving an estimator of

$$\hat{L} = \max_i X_i. \tag{17}$$