Confidence Interval Review

Akhil Vasvani

March 2019

1 Questions

Exercise 1. What are point estimation and function estimation in the context of Machine Learning? What is the relation between them?

Proof. The field of statistics gives us many tools that can be used to achieve the machine learning goal of solving a task not only on the training set but also to generalize. Foundational concepts such as parameter estimation, bias and variance are useful to formally characterize notions of generalization, underfitting and overfitting.

Point estimation is the attempt to provide the single "best" prediction of some quantity of interest. In general the quantity of interest can be a single parameter or a vector of parameters in some parametric model, but it can be a whole function. In order to distinguish estimates of parameters from their true value, our convention will be to denote a point estimate of a parameter $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$. Let $\{\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(m)}\}$ be a set of m independent and identically distributed (i.i.d) data points (or examples). A **point estimator** or **statistic** is any function of the data:

$$\hat{\boldsymbol{\theta}}_m = g(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(m)}). \tag{1}$$

Note: The definition does not require that g return a value that is close to the true $\boldsymbol{\theta}$ or even that the range of g is the same as the set of allowable values of $\boldsymbol{\theta}$. While almost any function thus qualifies as an estimator, a good estimator is a function whose output is close to the true underlying $\boldsymbol{\theta}$ that generated the training data. We assume that the true parameter value $\boldsymbol{\theta}$ is fixed but unknown, while the point estimate $\hat{\boldsymbol{\theta}}$ is a function of the data. Since the data is drawn from a random process, any function of the data is random. Therefore $\hat{\boldsymbol{\theta}}$ is a random variable.

Point estimation can also refer to the estimation of the relationship between input and target variables (**function estimation** or function approximation). We refer to these types of point estimates as function estimators. Here we are trying to predict a variable $\boldsymbol{y}$ given an input vector $\boldsymbol{x}$. We assume that there is a function $f(\boldsymbol{x})$ that describes the approximate relationship between $\boldsymbol{y}$ and $f(\boldsymbol{x})$. For example, we may assume that $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ stands for the part of $\boldsymbol{y}$ that is not predictable from $\boldsymbol{x}$. In function estimation, we are interested in approximating f with a model or estimate $\hat{f}$. Function estimation is really just the same as estimating a parameter θ; the function estimator $\hat{f}$ is simply a point estimator in function space.

Example. The linear regression example and the polynomial regression example are both scenarios that may be interpreted either as estimating a parameter $\boldsymbol{w}$ or estimating a function $\hat{f}$ mapping from $\boldsymbol{x}$ to $\boldsymbol{y}$.

□

Exercise 2. What is the bias of an estimator?

Proof. The bias of an estimator is defined as:

$$\text{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\theta} \tag{2}$$

where the expectation is over the data (seen as samples from a random variable) and $\boldsymbol{\theta}$ is the true underlying value of $\boldsymbol{\theta}$ used to define the data generating distribution. An estimator $\hat{\boldsymbol{\theta}}_m$ is said to be **unbiased** if $\text{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbf{0}$, which implies that $\mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \mathbf{0}$. An estimator $\hat{\boldsymbol{\theta}}_m$ is said to be **asymptotically unbiased** if $\lim_{m\to\infty} \text{bias}(\hat{\boldsymbol{\theta}}_m) = \boldsymbol{\theta}$, which implies that $\lim_{m\to\infty} \mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \boldsymbol{\theta}$.

Example. Let $Y_1, \ldots, Y_n$ be a random sample from a population whose density is

$$f(y|\theta) = \begin{cases} 3\theta^3 y^{-4}, & \theta \leq y \\ 0, & \text{otherwise} \end{cases}$$

where $\theta > 0$ is a parameter. Suppose that we wish to estimate θ using the estimator $\hat{\theta} = \min\{Y_1, \ldots, Y_n\}$. Now, we wish to compute $B(\hat{\theta})$—the bias of $\hat{\theta}$.

Since $B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$, we must first compute $\mathbb{E}(\hat{\theta})$. To determine $\mathbb{E}(\hat{\theta})$, we need to find the density function of $\hat{\theta}$, which requires us first to find the distribution function of $\hat{\theta}$. There is a "trick" for computing the distribution function of a minimum of random variables. That is, since $Y_1, \ldots, Y_n$ we find

$$\begin{aligned} P(\hat{\theta} > x) = P\left(\min\left\{Y_1, \ldots, Y_n\right\} > x\right) &= P\left(Y_1 > x, \ldots, Y_n > x\right) \\ &= \left[P\left(Y_1 > x\right)\right]^n \end{aligned}$$

We know the density of Y_1, and so if $x \geq \theta$, we compute

$$P\left(Y_1 > x\right) = \int_x^\infty f(y|\theta)\mathrm{d}y = \int_x^\infty 3\theta^3 y^{-4}\mathrm{d}y = \theta^3 x^{-3}.$$

Therefore, we find

$$P(\hat{\theta} > x) = \left[P\left(Y_1 > x\right)\right]^n = \theta^{3n} x^{-3n} \qquad \text{for } x \geq \theta$$

and so the distribution function for $\hat{\theta}$ is

$$F(x) = P(\hat{\theta} \leq x) = 1 - P(\hat{\theta} > x) = 1 - \theta^{3n} x^{-3n}$$

for $x \geq \theta$, and $F(x) = 0$ for $x < \theta$. Finally, we differentiate to conclude that the density function for $\hat{\theta}$ is

$$f(x) = F'(x) = \begin{cases} 3n\theta^{3n} x^{-3n-1}, & x \geq \theta \\ 0, & x < \theta \end{cases}$$

Now, we can determine $\mathbb{E}(\hat{\theta})$ via

$$\begin{aligned}\mathbb{E}(\hat{\theta}) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x = \int_{\theta}^{\infty} x \cdot 3n\theta^{3n}x^{-3n-1}\mathrm{d}x &= 3n\theta^{3n}\int_{\theta}^{\infty} x^{-3n}\mathrm{d}x \\ &= 3n\theta^{3n} \cdot \frac{\theta^{-3n+1}}{3n-1} \\ &= \frac{3n}{3n-1}\theta\end{aligned}$$

Hence, the bias of $\hat{\theta}$ is given by

$$B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = \frac{3n}{3n-1}\theta - \theta = \frac{\theta}{3n-1}.$$

Observe that $\hat{\theta}$ is not unbiased; that is $B(\hat{\theta}) \neq 0$. This particular $\hat{\theta}$ is not preferred. However, it might not be possible to find any unbiased estimators of θ. Thus, we will be forced to settle on one which is biased. Since

$$\lim_{n\to\infty} B(\hat{\theta}) = \lim_{n\to\infty} \frac{\theta}{3n-1} = 0$$

$\hat{\theta}$ is asymptotically unbiased. If no unbiased estimators can be found, the next best thing is to find asymptotically unbiased estimators.

□

Exercise 3. What is population mean and sample mean?

Proof. A **sample mean** is the mean of the statistical samples while a **population mean** is the mean of the total population. In other words, the sample mean provides an estimate of the population mean.

Usually the population mean is denoted as μ, while the sample mean is denoted as:

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}.$$

Thus, the sample mean increases its accuracy to the population mean with the increased number of observations. □

Exercise 4. What is population standard deviation and sample standard deviation?

Proof. Similarly, the **unbiased sample variance** is the variance of the statistical samples, while the **population variance** is the variance of the total population.

Most often the population variance is denoted as σ^2. The sample variance is denoted as:

$$\tilde{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2.$$

□

Exercise 5. Why population standard deviation has N degrees of freedom while sample standard deviation has $N-1$ degrees of freedom? In other words, why $\frac{1}{N}$ inside root for population standard deviation and $\frac{1}{N-1}$ inside root for sample standard deviation?

Proof. To explain why we use $\frac{1}{N-1}$ inside the root of the sample standard deviation, let's compare two different estimators of the variance parameter σ^2 of a Gaussian distribution. We are interested in knowing if either estimator is biased.

The first estimator of σ^2:

$$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2,$$

where $\hat{\mu}_m$ is the sample mean. More formally, we are interested in computing:

$$\begin{aligned}\text{bias}(\hat{\sigma}_m^2) &= \mathbb{E}[\hat{\sigma}_m^2] - \sigma^2\\ &= \frac{m-1}{m}\sigma^2 - \sigma^2\\ &= -\frac{\sigma^2}{m} \neq 0\end{aligned}$$

Therefore, this is biased. So, to get the unbiased sample variance:

$$\tilde{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2.$$

Plugging this back into the main bias equation, we get:

$$\begin{aligned}\text{bias}(\tilde{\sigma}_m^2) &= \mathbb{E}[\tilde{\sigma}_m^2] - \sigma^2\\ &= \left(\frac{m}{m-1}\mathbb{E}[\hat{\sigma}_m^2]\right) - \sigma^2\\ &= \frac{m}{m-1}\left(\frac{m-1}{m}\right)\sigma^2 - \sigma^2\\ &= \sigma^2 - \sigma^2\\ &= 0.\end{aligned}$$

Note. Full Proof can be found here.

Thus, $\frac{1}{N-1}$ is inside the root of the sample standard deviation (which is the square root of the sample variance), so that it corrects the bias in the estimation of the population variance. Note, the important change from N to $N-1$ is called **Bessel's Correction**.

Note. It is not possible to find an estimate of the standard deviation which is unbiased for all population distributions, as the bias depends on the particular distribution. Much of the following relates to estimation assuming a normal distribution.

□

Exercise 6. What is the formula for calculating the standard deviation of the sample mean?

Proof. The formula for calculating the standard deviation of the sample mean:

$$\hat{\sigma} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2} \tag{3}$$

□

Exercise 7. What is the variance of an estimator? What is standard error?

Proof. We should consider how much we expect the estimator to vary as a function of the data sample. The **variance** of the estimator is just simply $\mathrm{Var}(\hat{\theta})$, where the random variable is the training set. Alternately, the square root of the variance is called the **standard error**, denoted $\mathrm{SE}(\theta)$.

The variance or the standard error of an estimator provides a measure of how we would expect the estimate we compute from data to vary as we independently resample the dataset from the underlying data generating process. When we compute any statistic using a finite number of samples, our estimate of the true underlying parameter is uncertain, in the sense that we could have obtained other samples from the same distribution and their statistics would have been different. The expected degree of variation in any estimator is a source of error that we want to quantify.

The standard error of the sample mean is given by

$$\mathrm{SE}(\hat{\mu}_m) = \sqrt{\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}} \tag{4}$$

where σ^2 is the true variance of the samples x^i . The standard error is often estimated by using an estimate of σ.

Note. Unfortunately, neither the square root of the sample variance nor the square root of the unbiased estimator of the variance provide an unbiased estimate of the standard deviation. Both approaches tend to underestimate the true standard deviation, but are still used in practice. The square root of the unbiased estimator of the variance is less of an underestimate. For large m, the approximation is quite reasonable.

Example. Let $Y_1, \ldots, Y_n$ be a random sample from a population whose density is

$$f(y|\theta) = \begin{cases} 3\theta^3 y^{-4}, & \theta \leq y \\ 0, & \text{otherwise} \end{cases}$$

where $\theta > 0$ is a parameter. Suppose that we wish to estimate θ using the estimator $\hat{\theta} = \min\{Y_1, \ldots, Y_n\}$. We wish to compute $\sigma_{\hat{\theta}}$— the standard error of $\hat{\theta}$.

Now, the standard error $\sigma_{\hat{\theta}} = \sqrt{\mathrm{Var}(\hat{\theta})} = \sqrt{\mathrm{MSE}(\hat{\theta}) - [B(\hat{\theta})]^2}$.

Let's first calculate the $\mathrm{MSE}(\hat{\theta})$. For the mean-square error, we have by definition $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2$ and so

$$\mathrm{MSE}(\hat{\theta}) = \int_{-\infty}^{\infty}(x-\theta)^2 f(x)\mathrm{d}x = \int_{\theta}^{\infty}(x-\theta)^2 \cdot 3n\theta^{3n}x^{-3n-1}\mathrm{d}x$$

$$
\begin{aligned}
&= 3n\theta^{3n}\left[\int_\theta^\infty x^{1-3n}\mathrm{d}x - 2\theta\int_\theta^\infty x^{-3n}\mathrm{d}x + \theta^2\int_\theta^\infty x^{-3n-1}\mathrm{d}x\right] \\
&= 3n\theta^{3n}\left[\frac{\theta^{2-3n}}{3n-2} - 2\theta\cdot\frac{\theta^{1-3n}}{3n-1} + \theta^2\cdot\frac{\theta^{-3n}}{3n}\right] \\
&= 3n\theta^2\left[\frac{1}{3n-2} - \frac{2}{3n-1} + \frac{1}{3n}\right] \\
&= \theta^2\left[\frac{(3n-1)(3n-2)-9n(n-1)}{(3n-1)(3n-2)}\right] \\
&= \frac{2\theta^2}{(3n-1)(3n-2)}.
\end{aligned}
$$

Now, let's find $\mathrm{Var}(\hat{\theta}) = \mathrm{MSE}(\hat{\theta}) - [B(\hat{\theta})]^2$

$$
\begin{aligned}
&= \frac{2\theta^2}{(3n-1)(3n-2)} - \left[\frac{\theta}{3n-1}\right]^2 \\
&= \frac{\theta^2}{3n-1}\left[\frac{2}{3n-2} - \frac{1}{3n-1}\right] \\
&= \frac{3n\theta^2}{(3n-1)^2(3n-2)}.
\end{aligned}
$$

Therefore, the standard error of $\hat{\theta}$ is

$$
\sigma_{\hat{\theta}} = \sqrt{\mathrm{Var}(\hat{\theta})} = \frac{\theta}{(3n-1)}\sqrt{\frac{3n}{3n-2}}.
$$

□

Exercise 8. What is a confidence interval?

Proof. A **confidence interval** (CI) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter. The interval has an associated confidence level that, loosely speaking, quantifies the level of confidence that the parameter lies in the interval. More strictly speaking, the confidence level represents the frequency (i.e. the proportion) of possible confidence intervals that contain the true value of the unknown population parameter. In other words, if confidence intervals are constructed using a given confidence level from an infinite number of independent sample statistics, the proportion of those intervals that contain the true value of the parameter will be equal to the confidence level.

Note. This is important in machine learning because we often estimate the generalization error by computing the sample mean of the error on the test set. The number of examples in the test set determines the accuracy of this estimate. Taking advantage of the central limit theorem, which tells us that the mean will be approximately distributed with a normal distribution, we can use the standard error to compute the probability that the true expectation falls in any chosen interval.

Example. For example, the 95% confidence interval centered on the mean $\hat{\mu}_m$ is

$$(\hat{\mu}_m - 1.96\mathrm{SE}(\hat{\mu}_m), \hat{\mu}_m + 1.96\mathrm{SE}(\hat{\mu}_m))$$

under the normal distribution with mean $\hat{\mu}_m$ and variance $\mathrm{SE}(\hat{\mu}_m)^2$. In machine learning experiments, it is common to say that algorithm A is better than algorithm B if the upper bound of the 95% confidence interval for the error of algorithm A is less than the lower bound of the 95% confidence interval for the error of algorithm B.

□