Proof of Bayes' Estimator

Dr. Shafayat Abrar

We initiate our discussions of estimation theory by posing and solving a simple (almost trivial) estimation problem. Thus suppose that all we know about a real-valued random variable $\boldsymbol{x}$ is its mean $\bar{\boldsymbol{x}}$ and its variance σ_X^2, and that we wish to estimate the value that X will assume in a given experiment. We shall denote the estimate of X by $\widehat{X}$; it is a deterministic quantity (i.e., a number). But how do we come up with a value for $\hat{\boldsymbol{x}}$? And how do we decide whether this value is optimal or not? And if optimal, in what sense? These inquiries are at the heart of every estimation problem.

To answer these questions, we first need to choose a cost function to penalize the estimation error. The resulting estimate $\widehat{X}$ will be optimal only in the sense that it leads to the smallest cost value. Different choices for the cost (or risk) function will in general lead to different choices for $\widehat{X}$, each of which will be optimal in its own way.

The design criterion we shall adopt is the so-called mean-square-error criterion. It is based on introducing the error signal

$$\widetilde{X} \triangleq X - \widehat{X}$$

and then determining $\widehat{X}$ by minimizing the mean-square-error (m.s.e.), which is defined as the expected value of $\widetilde{X}^2$, i.e.,

$$\widehat{X} = \arg_{\widehat{X}} \min E\left[\widetilde{X}^2\right]$$

The error $\widetilde{X}$ is a random variable since X is random. The resulting estimate, $\widehat{X}$, will be called the least-mean-squares estimate of X.

Lemma (Lack of observations)

The least-mean-squares estimate of X given knowledge of only $\left(\overline{X}, \sigma_X^2\right)$ is $\widehat{X} = \overline{X}$.
The resulting minimum cost is $\mathbf{E}\,\widetilde{X}^2 = \sigma_X^2$. Note $\widetilde{X} = X - \widehat{X}$.

Proof: Expand the mean-square error by subtracting and adding $\overline{X}$ as follows:

$$\mathrm{E}\widetilde{X}^2 = \mathrm{E}(X - \widehat{X})^2 = \mathrm{E}[(X - \overline{X}) + (\overline{X} - \widehat{X})]^2 = \sigma_X^2 + (\overline{X} - \widehat{X})^2$$

The choice of $\widehat{X}$ that minimizes the m.s.e. is now evident.
Only the term $(\overline{X} - \widehat{X})^2$ is dependent on $\widehat{X}$ and this term can be annihilated by choosing $\widehat{X} = \overline{X}$. The resulting minimum mean-square error (m.m.s.e.) is then

$$\text{m.m.s.e.} \triangleq E\widetilde{X}^2 = \sigma_x^2$$

An alternative derivation would be to expand the cost function as

$$\mathrm{E}(X - \widehat{X})^2 = \mathrm{E}X^2 - 2\,\overline{X}\,\widehat{X} + \widehat{X}^2$$

and to differentiate it with respect to $\widehat{X}$. By setting the derivative equal to zero we arrive at the same conclusion, namely, $\widehat{X} = \overline{X}$.

The effectiveness of this estimation procedure can be measured by examining the value of the resulting minimum cost, which is the variance of the resulting estimation error.
The above lemma tells us that the minimum cost is equal to σ_X^2. That is,

$$\sigma_{\widetilde{X}}^2 = \sigma_X^2$$

so that the estimate $\widehat{X} = \overline{X}$ does not reduce our initial uncertainty about X since the error variable still has the same variance as $\boldsymbol{x}$ itself! We thus find that the performance of the mean-square-error design procedure is rather limited in this case. Of course, we are more interested in estimation procedures that result in error variances that are smaller than the original signal variance.

Theorem (Optimal mean-square-error estimator)

The least-mean-squares estimator (l.m.s.e.) of X given Y is the conditional expectation of X given Y, i.e., $\widehat{X} = \mathrm{E}[X \mid Y]$. The resulting estimate is

$$\widehat{X} = \mathrm{E}(X \mid Y) = \int_{\mathcal{S}_X} x\, f_{X|Y}(x \mid y) \mathrm{d}x$$

where $\mathcal{S}_X$ denotes the support (or domain) of the random variable X. Moreover, the estimator is unbiased, i.e., $\mathrm{E}\widehat{X} = \overline{X}$, and the resulting minimum cost is given by

$$\mathrm{E}[\widetilde{X}^2] = \mathrm{E}[X^2] - \mathrm{E}[\widehat{X}^2] = \sigma_X^2 - \sigma_{\widehat{X}}^2$$

$$
\begin{aligned}
E_x[X] &= E_y[E_x(X \mid Y)] \\
Proof &: \\
E_x[X] &= \int x f_X(x) dx \\
&= \int x \int f_{X,Y}(x, y) dy dx \\
&= \int x \int f_{X|Y}(x \mid y) f_Y(y) dy dx \\
&= \int \left[\int x f_{X|Y}(x \mid y) dx \right] f_Y(y) dy \\
&= \int E_x(X \mid Y) f_Y(y) dy \\
&= E_y[E_x(X \mid Y)]
\end{aligned}
$$

Uncorrelatedness property

$$E_{x,y}[Xg(Y)] = \iint xg(y)f_{X,Y}(x,y)dxdy$$

$$= \iint xg(y)f_{X|Y}(x \mid y)f_Y(y)dxdy$$

$$= \iint \left[xf_{X|Y}(x \mid y)dx \right] g(y)f_Y(y)dy$$

$$= \int E_x(X \mid Y)g(y)\left[\int f_{X,Y}(x,y)dx \right] dy$$

$$= \iint E_x(X \mid Y)g(y)f_{X,Y}(x,y)dxdy$$

$$= E_{x,y}[E_x(X \mid Y)g(Y)]$$

or we may say

$$E_{x,y}\left[\Big(X - E_x(X \mid Y) \Big) g(Y) \right] = 0$$

Uncorrelatedness property
Some discussions:

NOTE: two random variables X and Y are **uncorrelated** if, and only if, their **crosscorrelation** is zero, i.e., $\mathrm{E}(X - \overline{X})(Y - \overline{Y}) = 0$.

On the other hand, the random variables are said to be **orthogonal** if and only if, $E[XY] = 0$.

It is easy to verify that the concepts of **orthogonality** and **uncorrelatedness** coincide if at least one of the random variables is zero mean.

From the equation

$$E_{x,y}\left[\left(X - E_x(X \mid Y)\right)g(Y)\right] = 0$$

we conclude that the variables $X - \mathrm{E}(X \mid Y)$ and $g(Y)$ are orthogonal.

However, since $X - \mathrm{E}(X \mid Y)$ is zero mean, then we can also say that they are uncorrelated.

Using the intermediate result, we return to the m.s.e. cost function, add and subtract $E(X \mid Y)$ to its argument, and express it as

$$E[(X - \widehat{X})^2] = E\left[\left(X - E(X \mid Y) + E(X \mid Y) - \widehat{X}\right)^2\right]$$

The term $E(X \mid Y) - \widehat{X}$ is a function of Y. Therefore, if we choose $g(Y) = \mathbf{E}(X \mid Y) - \widehat{X}$, then from the uncorrelatedness property we conclude that

$$E(X - \widehat{X})^2 = E\left[\left(X - E(X \mid Y)\right)^2\right] + E\left[\left(\widehat{X} - E(X \mid Y)\right)^2\right]$$

Only the second term on the right-hand side is dependent on $\widehat{X}$ and the m.s.e. is minimized by choosing $\widehat{X} = \mathbf{E}(X \mid Y)$.

$$
\begin{aligned}
E\widetilde{X}^2 &= E[X-\widehat{X}][X-\widehat{X}] \\
&= E[X-\widehat{X}]X \quad \text{(because of uncorreletedness)} \\
&= EX^2 - E\widehat{X}X \\
&= EX^2 - E\widehat{X}[\widetilde{X}+\widehat{X}] \\
&= EX^2 - E\widehat{X}^2 \quad \text{(because of uncorreletedness)} \\
&= \left(E\left[X^2\right] - \overline{X}^2\right) + \left(\overline{X}^2 - E\left[\widehat{X}^2\right]\right) \\
&= \sigma_X^2 - \sigma_{\widehat{X}}^2 = \sigma_{\widetilde{X}}^2
\end{aligned}
$$