Bias, Variance, and MSE of Estimators

Guy Lebanon

September 4, 2010

We assume that we have iid (independent identically distributed) samples $X^{(1)}, \ldots, X^{(n)}$ that follow some unknown distribution. The task of statistics is to estimate properties of the unknown distribution. In this note we focus one estimating a parameter of the distribution such as the mean or variance. In some cases the parameter completely characterizes the distribution and estimating it provides a probability estimate.

In this note, we assume that the parameter is a real vector $\theta \in \mathbb{R}^d$. To estimate it, we use an estimator which is a function of our observations $\hat{\theta}(x^{(1)}, \ldots, x^{(n)})$. We follow standard practice and omit (in notation only) the dependency of the estimator on the samples, i.e. we write $\hat{\theta}$. However, note that $\hat{\theta} = \hat{\theta}(X^{(1)}, \ldots, X^{(n)})$ is a random variable since it is a function of n random variables.

A desirable property of an estimator is that it is correct on average. That is, if there are repeated samplings of n samples $X^{(1)}, \ldots, X^{(n)}$, the estimator $\hat{\theta}(X^{(1)}, \ldots, X^{(n)})$ will have, on average, the correct value. Such estimators are called unbiased.

Definition 1. *The bias of $\hat{\theta}$ is*[1] $\mathit{Bias}(\hat{\theta}) = \mathit{E}(\hat{\theta}) - \theta$. *If it is 0, the estimator $\hat{\theta}$ is said to be unbiased.*

There is, however, more important performance characterizations for an estimator than just being unbiased. The mean squared error is perhaps the most important of them. It captures the error that the estimator makes. However, since the estimator is a RV, we need to average over its distribution thus capturing the average performance if there are many repeated samplings of $X^{(1)}, \ldots, X^{(n)}$.

Definition 2. *The mean squared error (MSE) of an estimator is* $\mathit{E}(\|\hat{\theta} - \theta\|^2) = \mathit{E}(\sum_{j=1}^{d} (\hat{\theta}_j - \theta_j)^2)$.

Theorem 1.

$$\mathit{E}(\|\hat{\theta} - \theta\|^2) = \mathit{trace}(\mathit{Var}(\hat{\theta})) + \|\mathit{Bias}(\hat{\theta})\|^2.$$

Note that $\mathsf{Var}(\hat{\theta})$ is the covariance matrix of $\hat{\theta}$ and so its trace is $\sum_{j=1}^{d} \mathsf{Var}(\hat{\theta}_j)$.

Proof. Since the MSE equals $\sum_{j=1}^{d} \mathsf{E}((\hat{\theta}_j - \theta_j)^2)$ it is sufficient to prove for a scalar θ, $\mathsf{E}((\hat{\theta} - \theta)^2) = \mathsf{Var}(\hat{\theta}) + \mathsf{Bias}^2(\hat{\theta})$:

$$\begin{aligned}
\mathsf{E}((\hat{\theta} - \theta)^2) &= \mathsf{E}(((\hat{\theta} - \mathsf{E}(\hat{\theta})) + (\mathsf{E}(\hat{\theta}) - \theta)^2) = \mathsf{E}\{(\hat{\theta} - \mathsf{E}(\hat{\theta}))^2 + (\mathsf{E}(\hat{\theta}) - \theta)^2 + (\hat{\theta} - \mathsf{E}(\hat{\theta}))(\mathsf{E}(\hat{\theta}) - \theta)\} \\
&= \mathsf{Var}(\hat{\theta}) + \mathsf{Bias}^2(\hat{\theta}) + \mathsf{E}((\hat{\theta} - \mathsf{E}(\hat{\theta}))(\mathsf{E}(\hat{\theta}) - \theta)) = \mathsf{Var}(\hat{\theta}) + \mathsf{Bias}^2(\hat{\theta}) + \mathsf{E}(\hat{\theta}\mathsf{E}(\hat{\theta}) - (\mathsf{E}(\hat{\theta}))^2 - \theta\hat{\theta} + \mathsf{E}(\hat{\theta})\theta) \\
&= \mathsf{Var}(\hat{\theta}) + \mathsf{Bias}^2(\hat{\theta}) + (\mathsf{E}(\hat{\theta}))^2 - (\mathsf{E}(\hat{\theta}))^2 - \theta\mathsf{E}(\hat{\theta}) + \theta\mathsf{E}(\hat{\theta}) = \mathsf{Var}(\hat{\theta}) + \mathsf{Bias}^2(\hat{\theta}).
\end{aligned}$$

□

Since the MSE decomposes into a sum of the bias and variance of the estimator, both quantities are important and need to be as small as possible to achieve good estimation performance. It is common to trade-off some increase in bias for a larger decrease in the variance and vice-verse.

[1] Note here and in the sequel all expectations are with respect to $X^{(1)}, \ldots, X^{(n)}$.

Two important special cases are the mean $\hat{\theta} = \bar{X} = \frac{1}{n}\sum X^{(i)}$ which estimates the vector $\mathsf{E}(X)$ and

$$\hat{\theta} = S^2, \quad \text{where} \quad S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_j^{(i)} - \bar{X}_j)^2 \quad j = 1, \ldots, d$$

which estimates the diagonal of the covariance matrix $\mathsf{Var}(X)$. We show below that both are unbiased and therefore their MSE is simply their variance.

Theorem 2. *$\bar{X}$ is an unbiased estimator of $\mathsf{E}(X)$ and S^2 is an unbiased estimator of the diagonal of the covariance matrix $\mathsf{Var}(X)$.*

Proof.

$$\mathsf{E}(\bar{X}) = \mathsf{E}\left(n^{-1}\sum_{i=1}^{n} X^{(i)}\right) = \sum_{i=1}^{n}\mathsf{E}(X^{(i)})/n = n\mathsf{E}(X^{(i)})/n.$$

To prove that S^2 is unbiased we show that it is unbiased in the one dimensional case i.e., X, S^2 are scalars (if this holds, we can apply this result to each component separately to get unbiasedness of the vector S^2). We first need the following result (recall that below X is a scalar)

$$\sum_{i=1}^{n}(X^{(i)} - \bar{X})^2 = \sum(X^{(i)})^2 - 2\bar{X}\sum X^{(i)} + n\bar{X}^2 = \sum(X^{(i)})^2 - 2n\bar{X}\bar{X} + n\bar{X}^2 = \sum(X^{(i)})^2 - n\bar{X}^2$$

and therefore

$$\mathsf{E}\left(\sum_{i=1}^{n}(X^{(i)} - \bar{X})^2\right) = \mathsf{E}\left(\sum_{i=1}^{n}(X^{(i)})^2 - n\bar{X}^2\right) = \sum_{i=1}^{n}\mathsf{E}((X^{(i)})^2) - n\mathsf{E}(\bar{X}^2) = n\mathsf{E}((X^{(1)})^2) - n\mathsf{E}(\bar{X}^2).$$

Substituting the expectations $\mathsf{E}(X^2) = \mathsf{Var}(X) + (\mathsf{E}(X))^2 = \mathsf{Var}(X) + (\mathsf{E}X)^2$ and $\mathsf{E}(\bar{X}^2) = \mathsf{Var}(\bar{X}) + (\mathsf{E}(\bar{X}))^2 = \frac{\mathsf{Var}(X)}{n} + (\mathsf{E}X)^2$ in the above equation we have

$$\mathsf{E}\left(\sum_{i=1}^{n}(X^{(i)} - \bar{X})^2\right) = n(\mathsf{Var}(X) + (\mathsf{E}X)^2) - n\left(\frac{\mathsf{Var}(X)}{n} + (\mathsf{E}X)^2\right) = (n-1)\mathsf{Var}(X).$$

Returning now to the multivariate case, this implies $\mathsf{E}(S_j^2) = E(\sum_{i=1}^{n}(X_j^{(i)} - \bar{X}_j)^2/(n-1)) = \mathsf{Var}(X_j)$ for all j and therefore $\mathsf{E}S^2 = \mathsf{diag}(\mathsf{Var}(X))$. □

Since the bias is zero, the MSE of $\bar{X}$ as an estimator of $\mathsf{E}X$ is $\sum_{j=1}^{d}\mathsf{Var}(\bar{X}_j) = \sum_{j=1}^{d}\frac{n}{n^2}\mathsf{Var}(X_j)$. Thus, if d is fixed and $n \to \infty$ the $\mathsf{MSE}(\bar{X}) \to 0$. For S^2, the MSE is $\mathsf{trace}(\mathsf{Var}(S^2))$ which may be computed with some tedious algebra (it also decreases to 0 as $n \to \infty$).

Another performance measure for estimators $\hat{\theta}$ is the probability that the estimator's error is outside some acceptable error range $P(\|\hat{\theta} - \theta\| \geq \epsilon)$. However, to evaluate the above quantity, we need (i) the pdf $f_{\hat{\theta}}$ which depends on the pdf of X (which is typically unknown) and (ii) the true value θ (also typically unknown). If $\hat{\theta}$ is unbiased we may obtain the following bound

$$\begin{aligned} P(\|\hat{\theta} - \theta\|^2 \geq \epsilon) &= P\left(\sum_{j=1}^{d}|\hat{\theta}_j - \theta_j|^2 \geq \epsilon\right) \leq \sum_{j=1}^{d}P(|\hat{\theta}_j - \theta_j|^2 \geq \epsilon/d) = \sum_{j=1}^{d}P(|\hat{\theta}_j - \theta_j| \geq \sqrt{\epsilon/d}) \\ &\leq \frac{d}{\epsilon}\sum_{j=1}^{d}\mathsf{Var}(\hat{\theta}_j) = \frac{d}{\epsilon}\,\mathsf{trace}(\mathsf{Var}(\hat{\theta})) \end{aligned}$$

where we used Boole's and Chebyshev's inequalities. This again shows (but in a different way than the bias variance decomposition of the MSE) that the quality of unbiased estimators is determined by $\mathsf{trace}(\mathsf{Var}(\hat{\theta}))$.