Bias and variance for linear regression

Fraida Fund

Contents

In this set of notes, we derive the bias and variance for linear regression models, including transformed linear models.

Transformed linear model

Consider the linear model in general transformed feature space:

$$\hat{y} = f(x, \beta) = \phi(x)^T \beta = \beta_1 \phi_1(x) + ... + \beta_p \phi_p(x)$$

Assume the true function is

$$y = f_0(x) + \epsilon, \quad \epsilon \sim N(0, \sigma_\epsilon^2)$$

When there is no under-modeling,

$$f_0(x) = f(x, \beta^0) = \phi(x)^T \beta_0$$

where $\beta_0 = (\beta_0^0, \cdots, \beta_k^0)$ is the true parameter.

For data $(x_i, y_i), i = 1, ..., N$, the least squares fit is

$$\hat{\beta} = (A^T A)^{-1} A^T y$$

where

$$A = \begin{bmatrix} \phi_1(\mathbf{x_1}) & \cdots & \phi_p(\mathbf{x_1}) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x_N}) & \cdots & \phi_p(\mathbf{x_N}) \end{bmatrix}$$

Unique solution to LS estimate

There is a unique solution to the LS estimate only if A^TA is invertible. Since $A \in R^{N\times p}$, the solution is unique only if $Rank(A) \geq p$, and since $Rank(A) \leq min(N, p)$, we need $N \geq p$.

In other words, the unique solution exists only if the number of data samples for training (N) is greater than or equal to the number of parameters p.

This limits the model complexity you can use (greater $p \rightarrow$ greater model complexity).

Linear transforms of random vectors

First, some review of terminology of random vectors:

- A **random vector** $\mathbf{x} = (x_1, \dots, x_d)^T$ is a vector where each component x_j is a random variable.
- The **vector of means** of the components is $\boldsymbol{\mu} = (E[x_1], \dots, E[x_d])^T = (u_1, \dots, u_d)^T$.
- The covariance of x_i, x_j is $Cov(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$
- The variance matrix (which is a $d \times d$ matrix) is:

$$Var(\mathbf{x}) := E[(\mathbf{x} - \mathbf{u})(\mathbf{x} - \mathbf{u})^T] = \begin{bmatrix} Cov(x_1, x_1) & \cdots & Cov(x_1, x_d) \\ \vdots & \ddots & \vdots \\ Cov(x_d, x_1) & \cdots & Cov(x_d, x_d) \end{bmatrix}$$

- In a **linear transform** $y = Ax + b$, the input $x \in R^N$ is mapped to $Ax \in R^M$ by $A \in R^{M\times N}$
- The mean and variance matrix under this linear transform are given by $E(y) = AE(x) + b$ and $\text{Var}(y) = A\,\text{Var}(x)A^T$, respectively.

Bias of linear model

Suppose that there is no under-modeling, i.e. $f_0(x) = \phi(x)^T\beta^0$. Then each training sample output is $y_i = \phi(x_i)^T\beta_0 + \epsilon_i$. The "true" data vector is $y = A\beta^0 + \epsilon$.

Under these circumstances, the parameter estimate will be

$$\hat{\beta} = (A^TA)^{-1}A^Ty = (A^TA)^{-1}A^T(A\beta^0 + \epsilon) = \beta^0 + (A^TA)^{-1}A^T\epsilon$$

Since $E[\epsilon] = 0$, $E[\hat{\beta}] = \beta^0$: the average of the parameter estimate matches the true parameter.

Then $E[f(x_{test}, \hat{\beta})] = \phi(x_{test})^T E[\hat{\beta}] = \phi(x_{test})^T\beta^0 = f_0(x_{test})$.

Recall the definition of bias:

$$Bias(x_{test}) := f_0(x_{test}) - E[f(x_{test}, \hat{\beta})]$$

Conclusion: We can see that when the model is linear and there is no under-modeling, there is no bias:

$$Bias(x_{test}) = 0$$

.

Variance of linear model

Recall that ϵ_i are independent for different samples, with $E[\epsilon_i] = 0$ and $E[\epsilon_i^2] = \sigma_\epsilon^2$.

Then,

$$Cov(\epsilon_i, \epsilon_j) = \{0, \quad i \neq j\sigma_\epsilon^2 \quad , quad i = j$$

so the variance matrix is

$$Var(\epsilon) = \sigma_\epsilon^2 I$$

Also recall from our discussion of bias,

$$\hat{\beta} = (A^T A)^{-1} A^T y = (A^T A)^{-1} A^T (A\beta^0 + \epsilon) = \beta^0 + (A^T A)^{-1} A^T \epsilon$$

Then we can compute the variance of the *parameters* in the linear model:

$$\begin{aligned} E[(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)^T] &= (A^T A)^{-1} A^T Var(\epsilon) A (A^T A)^{-1} \\ &= \sigma_\epsilon^2 (A^T A)^{-1} A^T A (A^T A)^{-1} \\ &= \sigma_\epsilon^2 (A^T A)^{-1} \end{aligned}$$

We can also compute the variance of the *estimate* in the linear model. First, recall from our discussion of bias,

$$E[f(x_{test}, \hat{\beta})] = \phi(x_{test})^T E[\hat{\beta}] = \phi(x_{test})^T \beta^0 = f_0(x_{test})$$

Also note the following trick: if $\mathbf{a}$ is a non-random vector and $\mathbf{z}$ is a random vector, then

$$E[\mathbf{a}^T \mathbf{z}]^2 = E[\mathbf{a}^T \mathbf{z}\mathbf{z}^T \mathbf{a}] = \mathbf{a}^T E[\mathbf{z}\mathbf{z}^T]\mathbf{a}$$

Then the variance of the estimate of the linear model (when there is no under-modeling) is:

$$\begin{aligned} Var(x_{test}) &= E[f(x_{test}, \hat{\beta}) - E[f(x_{test}, \hat{\beta})]]^2 \\ &= \phi(x_{test})^T E[(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)^T] \phi(x_{test}) \\ &= \sigma_\epsilon^2 \phi(x_{test})^T (A^T A)^{-1} \phi(x_{test}) \end{aligned}$$

Let us assume that the test point x_{test} is distributed identically to the training data:

- Training data is $\mathbf{x}_i, i = 1, \ldots, N$
- $\mathbf{x}_{test} = \mathbf{x}_i$ with probability $\frac{1}{N}$

Since the rows of A are $\phi(\mathbf{x}_i)^T$, then

$$A^T A = \sum_{i=1}^{N} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T$$

We will use a trick: for random vectors $\mathbf{u}, \mathbf{v}$, $E[\mathbf{u}^T\mathbf{v}] = Tr(E[\mathbf{v}\mathbf{u}^T])$, where $Tr(A) = \sum_i A_{ii}$ is the sum of diagonals of A.

Then the variance averaged over x_{test} is:

$$\begin{aligned} E[Var(x_{test})] &= \sigma_\epsilon^2 E[\phi(x_{test})^T (A^T A)^{-1} \phi(x_{test})] \\ &= \sigma_\epsilon^2 Tr\left(E[\phi(x_{test})\phi(x_{test})^T](A^T A)^{-1}\right) \\ &= \frac{\sigma_\epsilon^2}{N} Tr\left(\sum_i \phi(x_i)\phi(x_i)^T](A^T A)^{-1}\right) \\ &= \frac{\sigma_\epsilon^2}{N} Tr\left((A^T A)(A^T A)^{-1}\right) \\ &= \frac{\sigma_\epsilon^2}{N} Tr\left(I_p\right) \\ &= \frac{\sigma_\epsilon^2 p}{N} \end{aligned}$$

The average variance increases with the number of parameters p, and decreases with the number of samples used for training N, as long as the test point is distributed like the training data.

Summary of results for linear models

Suppose the model class is linear with N samples and p parameters.

Result 1: Uniqueness of coefficient estimate

When $N < p$, the least squares estimate of the coefficients is not unique.

Result 2: Bias of estimate of target variable

When $N \geq p$ and the least squares estimate of the coefficients is unique, *and* there is no under-modeling, then the esimate of the target variable is unbiased:

$$E[f(x_{test}, \hat{\beta})] = f_0(x_{test})$$

Result 3: Variance of estimate of target variable

When $N \geq p$ and the least squares estimate of the coefficients is unique, *and* the test point is drawn from the same distribution of the trainng data, then variance increases linearly with the number of parameters and inversely with the number of samples used for training:

$$Var = \frac{p}{N}\sigma_\epsilon^2$$