Linear Regression Models

Segment 4 – Model Diagnostics

Topic 3 – Bias and Variance

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

1. The Bias-Variance Decomposition

2. Bias and Variance of Linear Regression

The Bias-Variance Decomposition

The Bias-Variance Decomposition

- For this lecture, we will consider the true population relationship $Y = f(X) + \varepsilon$ with the assumptions:

The Bias-Variance Decomposition

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- For this lecture, we will consider the true population relationship $Y = f(X) + \varepsilon$ with the assumptions:

$$E\left[\varepsilon \mid X\right] = 0,$$

The Bias-Variance Decomposition

- For this lecture, we will consider the true population relationship $Y = f(X) + \varepsilon$ with the assumptions:

$$E\left[\varepsilon \mid X\right] = 0,$$
$$\text{Var}\left[\varepsilon \mid X\right] = \sigma^2.$$

The Bias-Variance Decomposition

- For this lecture, we will consider the true population relationship $Y = f(X) + \varepsilon$ with the assumptions:

$$E[\varepsilon \mid X] = 0,$$
$$\text{Var}[\varepsilon \mid X] = \sigma^2.$$

- This implies $f(X) = E[Y \mid X]$.

The Bias-Variance Decomposition

- For this lecture, we will consider the true population relationship $Y = f(X) + \varepsilon$ with the assumptions:

$$E[\varepsilon \mid X] = 0,$$
$$\text{Var}[\varepsilon \mid X] = \sigma^2.$$

- This implies $f(X) = E[Y \mid X]$.
- Suppose we build a model $\hat{f}$ using the dataset $(\mathbf{X}, \mathbf{y})$:

The Bias-Variance Decomposition

- For this lecture, we will consider the true population relationship $Y = f(X) + \varepsilon$ with the assumptions:

$$E\left[\varepsilon \mid X\right] = 0,$$
$$\text{Var}\left[\varepsilon \mid X\right] = \sigma^2.$$

- This implies $f(X) = E\left[Y \mid X\right]$.
- Suppose we build a model $\hat{f}$ using the dataset $(\mathbf{X}, \mathbf{y})$: the prediction error for an unseen data X can be shown to be

The Bias-Variance Decomposition

- For this lecture, we will consider the true population relationship $Y = f(X) + \varepsilon$ with the assumptions:

$$E[\varepsilon \mid X] = 0,$$
$$\text{Var}[\varepsilon \mid X] = \sigma^2.$$

- This implies $f(X) = E[Y \mid X]$.
- Suppose we build a model $\hat{f}$ using the dataset $(\mathbf{X}, \mathbf{y})$: the prediction error for an unseen data X can be shown to be

$$\sigma^2 + \left(\hat{f}(X) - f(X)\right)^2.$$

The Bias-Variance Decomposition–Continued

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is,

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is, we fix the design matrix $\mathbf{X}$,

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is, we fix the design matrix $\mathbf{X}$, generate the true response vector values $\mathbf{y}$ several times,

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is, we fix the design matrix $\mathbf{X}$, generate the true response vector values $\mathbf{y}$ several times, create a model $\hat{f}$ using each dataset $(\mathbf{X}, \mathbf{y})$,

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is, we fix the design matrix $\mathbf{X}$, generate the true response vector values $\mathbf{y}$ several times, create a model $\hat{f}$ using each dataset $(\mathbf{X}, \mathbf{y})$, and calculate the quantity in the previous step for each model.

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is, we fix the design matrix $\mathbf{X}$, generate the true response vector values $\mathbf{y}$ several times, create a model $\hat{f}$ using each dataset $(\mathbf{X}, \mathbf{y})$, and calculate the quantity in the previous step for each model.
- How close $\hat{f}(X)$ is to $f(X) = E\left[Y \mid X\right]$ represents the bias.

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is, we fix the design matrix $\mathbf{X}$, generate the true response vector values $\mathbf{y}$ several times, create a model $\hat{f}$ using each dataset $(\mathbf{X}, \mathbf{y})$, and calculate the quantity in the previous step for each model.
- How close $\hat{f}(X)$ is to $f(X) = E[Y \mid X]$ represents the bias.
- How much $\hat{f}(X)$ varies w.r.t. each dataset represents the variance.

The Bias-Variance Decomposition–Continued

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is, we fix the design matrix $\mathbf{X}$, generate the true response vector values $\mathbf{y}$ several times, create a model $\hat{f}$ using each dataset $(\mathbf{X}, \mathbf{y})$, and calculate the quantity in the previous step for each model.
- How close $\hat{f}(X)$ is to $f(X) = E[Y \mid X]$ represents the bias.
- How much $\hat{f}(X)$ varies w.r.t. each dataset represents the variance.
- A more general result called the *bias-variance* decomposition shows how prediction error changes w.r.t. the training data:

The Bias-Variance Decomposition–Continued

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- We will focus on the term $\left(\hat{f}(X) - f(X)\right)^2$.
- To assess how good the model $\hat{f}$ is, we fix the design matrix $\mathbf{X}$, generate the true response vector values $\mathbf{y}$ several times, create a model $\hat{f}$ using each dataset $(\mathbf{X}, \mathbf{y})$, and calculate the quantity in the previous step for each model.
- How close $\hat{f}(X)$ is to $f(X) = E\left[Y \mid X\right]$ represents the bias.
- How much $\hat{f}(X)$ varies w.r.t. each dataset represents the variance.
- A more general result called the *bias-variance* decomposition shows how prediction error changes w.r.t. the training data:

$$\text{prediction error} = \text{irreducible error } \sigma^2 + \text{bias}^2 + \text{variance}.$$

Bias and Variance of Linear Regression

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$,

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.
- We see that $f(X_1, X_2) = 1 + X_1 + 2X_2$.

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.
- We see that $f(X_1, X_2) = 1 + X_1 + 2X_2$.
- We will investigate three models:

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.
- We see that $f(X_1, X_2) = 1 + X_1 + 2X_2$.
- We will investigate three models: (1) Y~X_1,

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.
- We see that $f(X_1, X_2) = 1 + X_1 + 2X_2$.
- We will investigate three models: (1) `Y~X_1`, (2) `Y~X_1+X_2`,

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.
- We see that $f(X_1, X_2) = 1 + X_1 + 2X_2$.
- We will investigate three models: (1) `Y~X_1`, (2) `Y~X_1+X_2`, (3) `Y~X_1+X_2+...+I(X_1^4)+I(X_2^4)`.

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.
- We see that $f(X_1, X_2) = 1 + X_1 + 2X_2$.
- We will investigate three models: (1) `Y~X_1`, (2) `Y~X_1+X_2`, (3) `Y~X_1+X_2+...+I(X_1^4)+I(X_2^4)`.
- Which model will have a higher bias?

Bias and Variance of Linear Regression

- We consider the population model $Y = 1 + X_1 + 2X_2 + \varepsilon$, where X_1 and X_2 are independent and standard normal random variables (mean 0, standard deviation 1), and $\varepsilon \sim N(0, 5)$.
- We see that $f(X_1, X_2) = 1 + X_1 + 2X_2$.
- We will investigate three models: (1) `Y~X_1`, (2) `Y~X_1+X_2`, (3) `Y~X_1+X_2+...+I(X_1^4)+I(X_2^4)`.
- Which model will have a higher bias?
- Which model will have a higher variance?

Summary

Summary

- Describe the general equation for bias-variance decomposition.

Summary

- Describe the general equation for bias-variance decomposition.
- Describe bias and variance in the context of linear regression