Chapter 2

0.1 What Is Statistical Learning?

Statistical learning involves tools for modeling and understanding the relationship between input variables $X_1, X_2, \ldots, X_p$ and an output Y. The primary goal is to estimate the relationship $Y = f(X) + \varepsilon$, where f is an unknown deterministic function, and ε is a random error term with mean zero.

0.1.1 Why Estimate f?

The two main objectives are:

- **Prediction:** Predict Y for unseen inputs using $\hat{f}(X)$, where $\hat{f}$ is the estimate of f.
- **Inference:** Understand how Y changes as a function of X.

The error in prediction can be decomposed into:

$$E\left[(Y - \hat{f}(X))^2\right] = [f(X) - \hat{f}(X)]^2 + Var(\varepsilon).$$

0.1.2 Parametric vs Non-Parametric Methods

- **Parametric:** Assume a specific form for f, such as linear models, and estimate parameters. Simplicity but risk of misspecification.
- **Non-Parametric:** No explicit assumption on f. Flexible but requires more data to avoid overfitting.

0.1.3 The Trade-Off Between Accuracy and Interpretability

Methods range from simple (e.g., linear regression) to complex (e.g., support vector machines, boosting). Simpler models are more interpretable but less flexible.

0.2 Supervised vs Unsupervised Learning

- **Supervised Learning:** Both predictors and responses are observed (X, Y).
- **Unsupervised Learning:** Only predictors (X) are observed; goal is to find structure (e.g., clustering).

0.3 Assessing Model Accuracy

0.3.1 Bias-Variance Trade-Off

The expected test error can be decomposed as:

$$E\left[(y_0 - \hat{f}(x_0))^2\right] = Bias(\hat{f}(x_0))^2 + Var(\hat{f}(x_0)) + Var(\varepsilon).$$

There is a trade-off between bias (error from model assumptions) and variance (sensitivity to training data).

0.3.2 Training vs Test Error

Training error:

$$MSE_{train} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2$$

Test error measures model performance on unseen data and often exhibits a U-shape as model flexibility increases.

0.4 K-Nearest Neighbors (KNN)

KNN classifies a test observation x_0 by identifying the K closest training points and assigning the majority class:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j),$$

where N_0 is the neighborhood of the K closest points.

0.5 Bayes Classifier

The Bayes classifier assigns an observation to the class with the highest conditional probability:

$$Class = \arg\max_j P(Y = j | X = x_0).$$

This classifier achieves the minimum possible test error, called the Bayes error rate:

$$1 - E \left[\max_j P(Y = j | X) \right].$$