1 Chapter 2

Statistical learning involves tools and techniques for modeling and understanding data. It is fundamental to tasks in business analytics and spans two key goals: *inference* and *prediction*.

1.1 Inference and Prediction

- **Inference:** Understand the relationship between predictors X and response Y. For example, determining which predictors significantly affect Y and the nature of this relationship.
- **Prediction:** Focus on accurately predicting Y based on new X values. This requires estimating $f(X)$ in the model:

  $$Y = f(X) + \epsilon,$$

  where ϵ is the irreducible error, assumed to have mean zero and constant variance.

1.2 Bias-Variance Tradeoff

The expected test mean squared error (MSE) for a prediction model is given by:

$$E\big[(Y - \hat{f}(X))^2\big] = Bias^2(\hat{f}(X)) + Var(\hat{f}(X)) + \sigma^2,$$

where:

- **Bias:** Error introduced by approximating a complex reality with a simpler model.
- **Variance:** Error due to model sensitivity to training data variations.
- σ^2: Irreducible error.

This tradeoff guides the selection of model complexity.

2 Supervised vs. Unsupervised Learning

2.1 Supervised Learning

Supervised learning methods involve a set of predictors X and a response Y. Examples include:

- **Regression:** Predicting quantitative responses.
- **Classification:** Assigning categories to observations.

2.2 Unsupervised Learning

Unsupervised learning methods analyze data without labeled responses, identifying patterns or structures. Examples:

- **Clustering:** Grouping observations into clusters.
- **Dimensionality Reduction:** Simplifying data representations.

3 Model Accuracy

Model accuracy is assessed using metrics like Mean Squared Error (MSE) in regression:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

In classification, accuracy is measured by the error rate:

$$ErrorRate = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i),$$

where $I(\cdot)$ is an indicator function.

3.1 Choosing Models

Key considerations include:

- **Flexibility:** Flexible models fit data closely but risk overfitting.
- **Interpretability:** Simple models are easier to understand but may underfit.

Cross-validation techniques are used to estimate test MSE and select models with optimal bias-variance tradeoffs.