Chapter 2: Statistical Learning – Full Notes

1. What Is Statistical Learning?

Statistical learning refers to a set of tools for modeling and understanding complex datasets. It provides methods to estimate the relationship between a set of input variables$X = (X1, X2, \dots, Xp)$ and an output variable Y.

This relationship is typically modeled as:

$$Y = f(X) + \varepsilon$$

Where:

- f is an **unknown function** that relates the inputs to the output.
- ε is a **random error term** that accounts for **unexplained variation** in Y, even if we knew the true function f.

2. Why Estimate f?

There are two main reasons to estimate f:

1. Prediction:

- Use $\hat{f}\ (X)$ to predict future values of Y based on observed values of X.
- The accuracy of prediction depends on the **error**:

$$Y - \hat{f}\ (X) = f(X) + \varepsilon - \hat{f}\ (X)$$

This has two components:

  - **Reducible error**: due to the difference between $f(X)$ and $\hat{f}(X)$. Can be reduced by improving $\hat{f}(X)$.
  - **Irreducible error**: due to ε, which captures noise and unmeasured variables. Cannot be reduced.

2. Inference:

- Understand the relationship between X and Y
- For example, which variables are important? How do they affect the output? What happens to Y if we change X_j?

3. How Do We Estimate f?

Two main approaches:

A. Parametric Methods:

- Simplify the problem by assuming a specific form for $f(X)$, such as a linear model:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

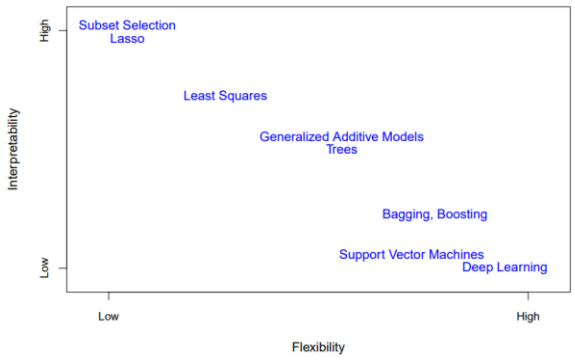
- Estimate the parameters β using methods like least squares.
- **Advantages**: Simple, interpretable.
- **Disadvantages**: Risk of model misspecification (high bias).

B. Non-Parametric Methods:

- Do not assume a fixed functional form. They try to estimate f more flexibly.
- Examples: K-Nearest Neighbors (KNN), decision trees.
- **Advantages**: Capture complex relationships.
- **Disadvantages**: Require more data, can overfit (high variance).

4. The Trade-Off Between Prediction Accuracy and Model Interpretability

- **Flexible models** (e.g. KNN) can fit training data closely, but may overfit and be hard to interpret.
- **Less flexible models** (e.g. linear regression) are more interpretable but may underfit the data.
- The goal is to find a **balance** that minimizes prediction error.



5. Supervised vs. Unsupervised Learning

Supervised Learning:

- Output variable Y is observed.

  Goal: Predict or estimate Y from X.

- Examples: Regression, classification.

Unsupervised Learning:

- No observed output Y.
- Goal: Discover structure in the data.
- Examples: Clustering, Principal Component Analysis (PCA).

6. Regression vs. Classification

- **Regression**: Predict a **quantitative** output (e.g., income, temperature).
- **Classification**: Predict a **qualitative** output (e.g., spam vs. not spam, disease vs. no disease).

7. Assessing Model Accuracy

For Regression:

Use **Mean Squared Error (MSE)**:

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

Lower MSE indicates better predictive accuracy.

For Classification:

Use **Classification Error Rate**:

$$\text{Error Rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y_i})$$

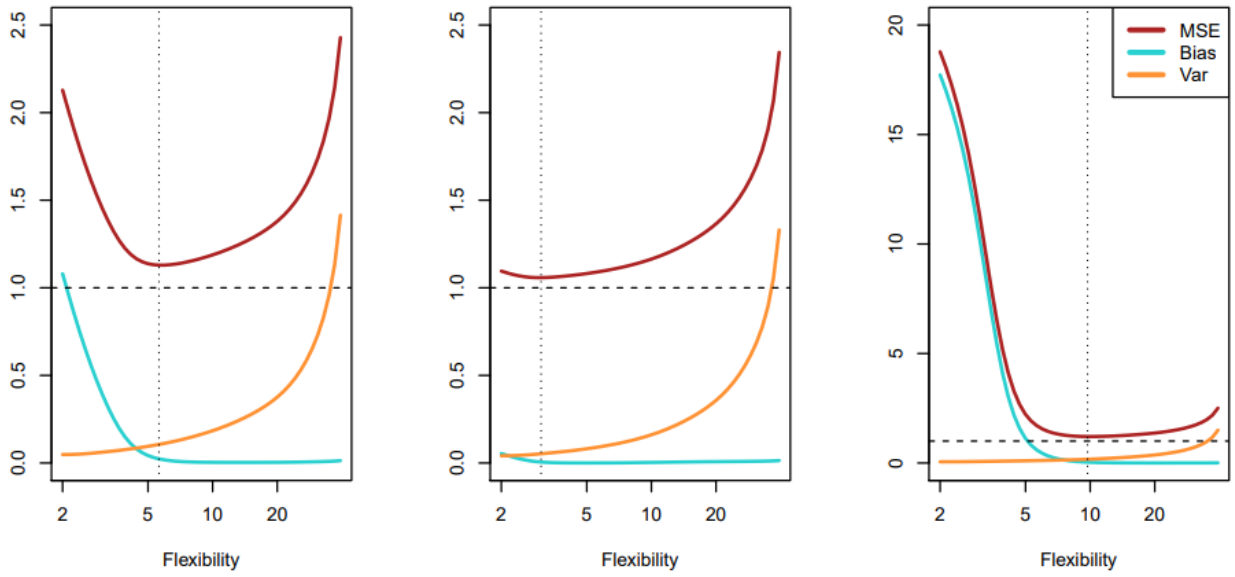
Where $I(\quad)$ is the indicator function (1 if the condition is true, 0 otherwise).

8. The Bias-Variance Trade-Off

Test MSE can be decomposed into:

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

- **Bias**: Error due to oversimplified assumptions.
- **Variance**: Error due to model's sensitivity to training data.
- **Irreducible Error**: Variance of ε, cannot be eliminated.



9. The Classification Setting

The Bayes Classifier:

- Theoretical ideal classifier.
- Assigns observation xxx to the class with the highest conditional probability:

$$\hat{y}=\arg\max_k P\left(Y=k \mid X=x\right)$$

K-Nearest Neighbors (KNN):

- For a new input x_0, find the k closest points in training data.
- Predict the most common class among those neighbors.

Advantages:

- Non-parametric, simple.

Disadvantages:

- Sensitive to choice of k, suffers in high dimensions (curse of dimensionality).