Notes on The Elements of Statistical Learning

Chapter 1

Chapter 2 Overview of Supervised Learning

1.1. Linear Models

Given a vector of inputs $X^T = (X_1, X_2, ..., X_p)$, we predict the output Y via the model

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j \tag{1}$$

The term $\hat{\beta}_0$ is the intercept( ), also known as the bias( ) in machine learning. Often it is convenient to include the constant variable 1 in X, include $\hat{\beta}_0$ in the vector of coefficients $\hat{\beta}$, and then write the linear model in vector form as an inner product $\hat{Y} = X^T \hat{\beta}$.

Viewed as a function over the p-dimensional input space, $f(X) = X^T \beta$ is linear, and the gradient $f\ (X) = \beta$ is a vector in input space that points in the steepest uphill direction.

The most popular is the method to fit the linear model to a set of training data is least squares. In this approach, we pick the coefficients β to minimize the residual sum of squares

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \end{aligned} \tag{2}$$

Note.

$$\sum_{i=1}^{N} \alpha_i \beta_i = (\alpha_1, \alpha_2, \cdots, \alpha_n)^T (\beta_1, \beta_2, \cdots, \beta_n) \tag{3}$$