Introduction to Statistical Learning

November 20, 2020

Statistical Learning

What is statistical learning?

We want to predict some output variable (also called response, dependent variable) Y based on some input variables (also called predictors, independent variables, features) X. We suppose there exists a relationship of the form $Y = f(X) + \varepsilon$. f is some unknown function of X and $\varepsilon$is a random error term, independent from X and of mean zero. f represents the *systematic information* that X provides about Y. The ε are the difference between the observations and the true underlying relationship between X and Y, which is usually unknown. Statistical learning refers to a set of approaches for estimating f.

Why estimate f?

Prediction We can predict Y using $\hat{Y} = \hat{f}(X)$ where $\hat{f}$ represents an estimate for f. Accuracy of the resulting prediction depends on the *reducible error* ($\hat{f}$ cannot be a perfect estimate of f) and the *irreducible error* (Y is also a function of some error term ε that cannot be predicted using X). The average of the squared distances between the observations Y and the predictions $\hat{Y}$ is thus given by: $E(Y - \hat{Y}) = \underbrace{E[f(X) - \hat{f}(X)]^2}_{Reducible} + \underbrace{Var[\varepsilon]}_{Irreducible}$. Statistical learning: estimating f and minimizing the reducible error.

Assessing Model Accuracy