What is "statistical learning"?

Statistical learning refers to a set of tools for modeling and understanding complex datasets.

Example: Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

Example: Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.

Example: Identify the risk factors for prostate cancer, based on clinical and demographic variables.

Example: Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product. The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media. Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales. In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

In this setting, the advertising budgets are input variables while sales is an output variable.

The input variables (predictors, independent variables, features, or sometimes just variables) are typically denoted using the symbol X, with a subscript to distinguish them.

X_1: TV budget, X_2: radio budget, X_3: newspaper budget

The output variable (response or dependent variable) is typically denoted using the symbol Y.

Y: sales

In general, suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, ..., X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, ..., X_p)$, which can be written in the very general form

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

some fixed but unknown function

random error term, independent of X, $E\epsilon = 0$

f represents the systematic information that X provides about Y.

Statistical learning **basically refers to a set of approaches for estimating f.**

Why Estimate f?

There are two main reasons that we may wish to estimate f:

1. Prediction
2. Inference

Prediction

- In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained.
- We can predict Y using
$$\hat{Y} = \hat{f}(X),$$
where $\hat{f}$ represents our estimate for f , and $\hat{Y}$ represents the resulting prediction for Y.
- In this setting, $\hat{f}$ is often treated as a black box, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.
- Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed. Then,

$$\text{Var}(\varepsilon) \leq E(Y - \hat{Y})^2 = E[f(X) + \varepsilon - \hat{f}(X)]^2$$
$$= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{reducible error}} + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible error}}$$

Inference

- We are often interested in understanding the relationship between X and Y , or more specifically, understanding how Y changes as a function of $X_1, ..., X_p$.
- Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form.