2.1 What Is Statistical Learning?

Zongyi Liu

2023-05-10

Prologue

炎正中微，大盜移國。九縣飆回，三精霧塞。人厭淫詐，神思反德。

光武誕命，靈貺自甄。沈幾先物，深略緯文。尋邑百萬，貔虎為群。

長轂雷野，高鋒彗雲。英威既振，新都自焚。虔劉庸代，紛紜梁趙。

三河未澄，四關重擾。神旌乃顧，遞行天討。金湯失險，車書共道。

靈慶既啟，人謀咸贊。明明廟謨，赳赳雄斷。於赫有命，系隆我漢。

2.1 What is Statistical Learning?

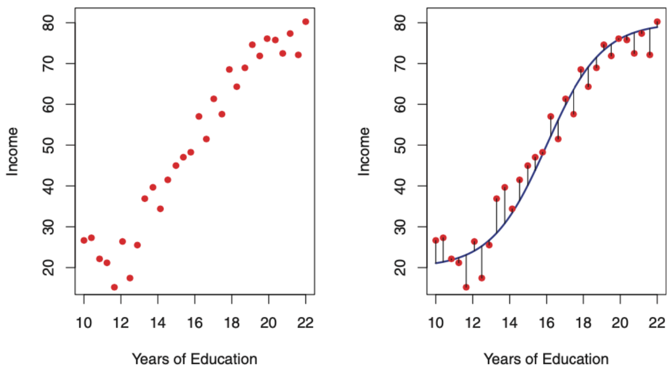
We want to predict one variable with other information we have, thus, we would use some kind of techniques. There are two types of variables:

- **Input variables**, or predictors, independent variables
- **Output variables**, or response, dependent variables

Suppose we have a set of X, and we want to get Y, we need:

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

Here f represents the systematic information that X provides about Y. In plots, we can see as:



2.1.1 Why Estimate f?

This books gives two reasons that we should estimate f . The first one is prediction:

$$\hat{Y} = \hat{f}(X),$$

Here, $\hat{Y}$ represents the resulting prediction for Y, and f-hat represents the resulting prediction for f.

The accuracy of Y-hat depends on two quantities, **reducible error** and **irreducible error**. The aim of this book is to minimize the reducible error.

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \quad , \end{aligned}$$

$E(Y-\hat{Y})^2$ represents the average, or the expected value, of the squared difference between the predicted and actual values of Y.

The second reason is inference, from which we want to know how Y changes as a function of X_1 to X_p. We want to answer those questions:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

2.1.2 How Do We Estimate f?

There are two ways to estimate f, the first one is the parametric method

Parametric Methods

The parametric methods involves a two-step model-based approach

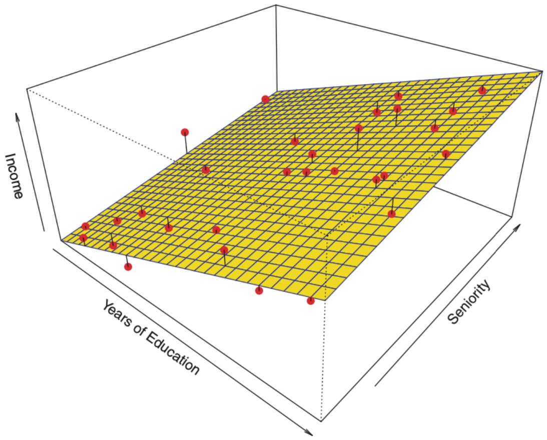
1. First is to make assumptions about the function form, and the most simple one is the linear form

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

2. After the model is selected, we would like to use it and to find the values of those parameters (beta 1 to p)

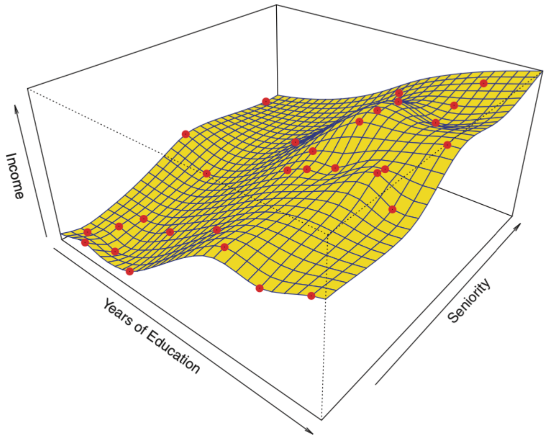
$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

The most common approach to fit the model is the ordinary least square method (OLS). In plots we can see it as



Non-parametric Method

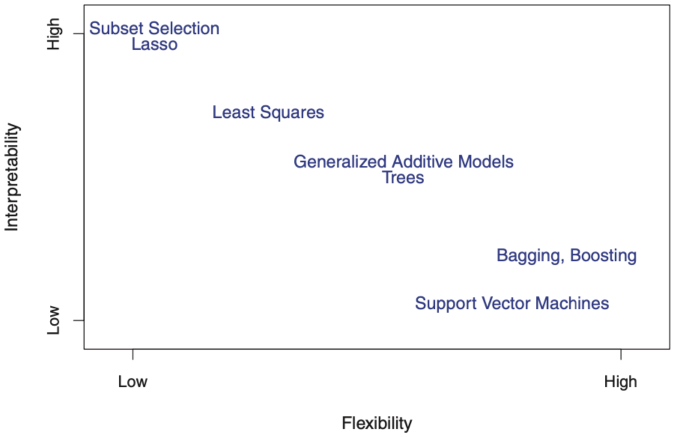
There is another form to estimate f, which is non-parametric method, it doesn't have specific function, but has a smooth spline to estimate the f.



2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

There are two characteristics of the model:

- **Flexibility**: refers to how well the model fits the data set. For example, linear model is not flexible, but it's good for inference
- **Interpretability**: refers to how well the model can be used for inference. For example, splines are very accurate, but they are hard to interpret

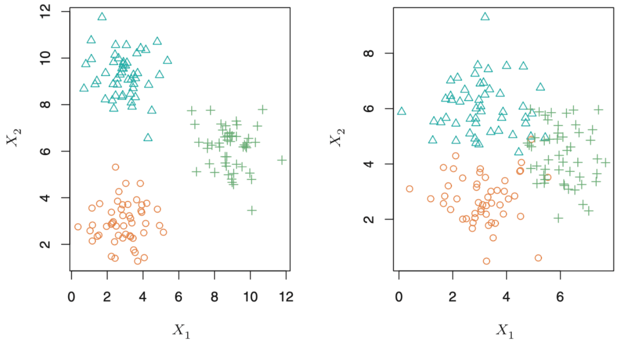


2.1.4 Supervised and Unsupervised Learning

There are two types of statistical learning, supervised and unsupervised learning

- **Supervised Learning**: For each observation of the predictor measurement(s) xi, i = 1,...,n there is an associated response measurement yi.
- **Unsupervised Learning**: every observation i = 1,...,n, we observe a vector of measurements xi but no associated response yi.

One way to solve the issue that each predictor is not associated with response is by **cluster analysis**, or **clustering**.



2.1.5 Regression and Classification

There are two types of variables, and we would face two problems separately:

- **Quantitative (or numerical)**: regression problem
- **Qualitative (or categorical)**: classification problem