Statistical Learning

Jack Krebsbach

Sep 4

0.1 Overview of Statistical Learning

0.1.1 Regression

The goal of regression is to model the relationship between a set of *predictors*, also called features, or independent variables and *response*, also called the dependent variables.

Consider the predictors

$$\vec{\mathbf{X}} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_p}).$$

Notice we have p predictors. Each column is a feature of our data set. We try and find a model to predict $\mathbf{Y}$.

The predicted value that we hope to be close to $\mathbf{Y}$ is

$$\hat{\mathbf{Y}} = \hat{f}(\mathbf{X}).$$

We can fully describe the value of $\mathbf{Y}$ by introducing the error term ϵ.

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

We assume that ε is a random error term that is independent of $\vec{X}$ and has zero mean: $\mathrm{E}(\varepsilon) = 0$. That is the mean is zero.

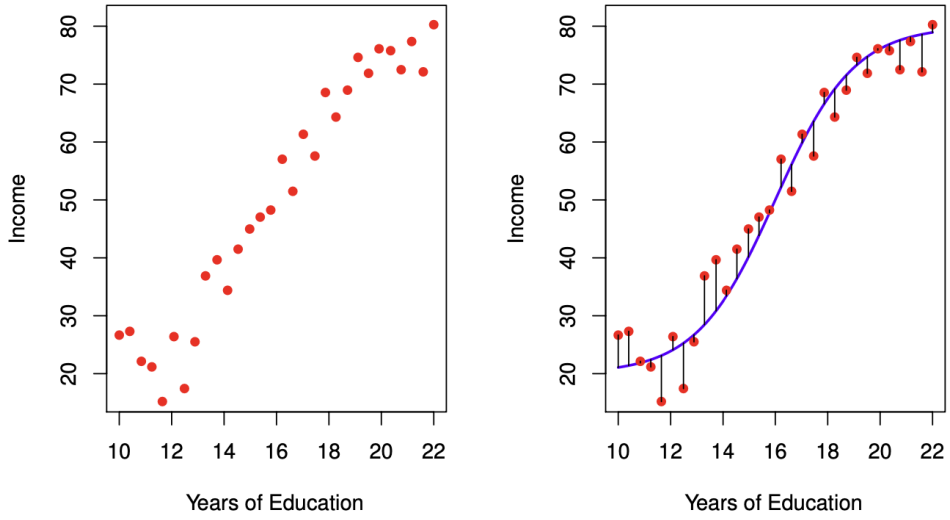

Figure 1: regression

Generally, we are interested in two types of regression tasks:

- **Inference** (model simplicity and interpretability): Which variables are important to the response and in which way?
- **Prediction** (model accuracy): What is the response at a given location?

In this prediction task we can break up the error at a particular $\vec{\mathbf{X}}$ into reducible error and irreducible error.

$$\mathrm{E}_{\mathbf{Y}}\left[(\mathbf{Y}-\hat{\mathbf{Y}})^2 \mid \vec{\mathbf{X}}, \hat{f}\right] = \underbrace{[f(\vec{\mathbf{X}})-\hat{f}(\vec{\mathbf{X}})]^2}_{\text{reducible error}} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{irreducible error}}$$

This quantity represents the average, or expected value of the squared difference between the predicted and actual value of $\mathbf{Y}$. In this conditional expectation we are looking at a particular value of $\vec{\mathbf{X}}$ and a particular $\hat{f}$.

Methods for estimating f can be divided into the following two categories:

- **Parametric** Suppose f has a particular functional form, such as linear:

  $$f(\vec{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

  Then the problem of estimating f reduces to estimating the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ (called parameters of the model), based on a set of observations.

- **Nonparametric**: No explicit assumptions about the functional form of f; often relies on local approximation (globally any shape can be produced).

We can work to improve the reducible error (by varying $\hat{f}$) but not the irreducible error because it is independent of the choice of $\hat{f}$. The focus is thus on the reducible error but minimizing it can be very challenging because we do not directly observe $f(\vec{X})$ but only a corrupted version of it, $Y = f(\vec{X}) + \varepsilon$. Another reason is that the space of possible functions $\hat{f}$ is extremely large.

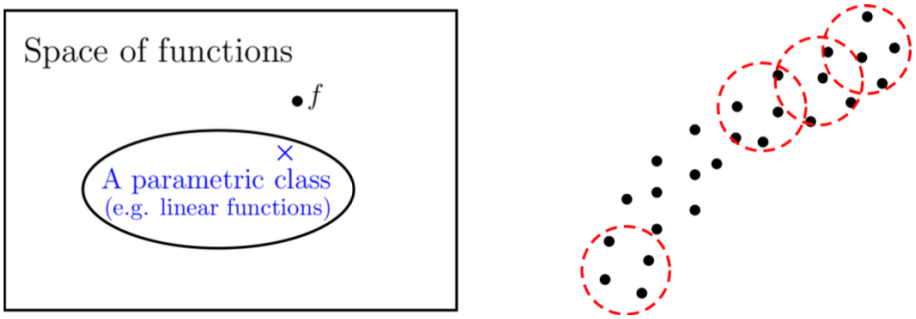

Figure 2: parametric-vs-non-parametric

Parametric models (especially linear) tend to be more restrictive (but faster to compute and easier to interpret). In contrast, nonparametric methods are more flexible (but more complicated in computing and harder to interpret)

To measure the error we can find the *Mean Squared Error*.

$$\mathrm{MSE}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}\left(\vec{x}_i\right)\right)^2$$

Which is our empirical estimate for the *Population Mean Squared Error* - the mean squared error. This is the average error over all values of $\vec{X}$.

$$\mathrm{E}_{\vec{X},Y}\left[(Y-\hat{Y})^2\right] = \mathrm{E}_{\vec{X}}\left[\mathrm{E}_Y(Y-\hat{Y})^2 \mid \vec{X}\right]$$

Let us decompose the expected squared error at a given $\vec{\mathbf{x}}$.

$$\begin{aligned}
\mathrm{E}\left[\left(y_0 - \hat{f}\left(\vec{x}_0\right)\right)^2\right] &= \mathrm{E}\left[\left(\varepsilon_0 + f\left(\vec{x}_0\right) - \hat{f}\left(\vec{x}_0\right)\right)^2\right] \\
&= \mathrm{E}\left[\left(\varepsilon_0 + f\left(\vec{x}_0\right) - \mathrm{E}\left(\hat{f}\left(\vec{x}_0\right)\right) + \mathrm{E}\left(\hat{f}\left(\vec{x}_0\right)\right) - \hat{f}\left(\vec{x}_0\right)\right)^2\right] \\
&= \underbrace{\mathrm{Var}\left(\epsilon_0\right)}_{\text{irreducible}} + \underbrace{\mathrm{Bias}\left(\hat{f}\left(\vec{x}_0\right)\right)^2 + \mathrm{Var}\left(\hat{f}\left(\vec{x}_0\right)\right)}_{\text{reducible error}}
\end{aligned}$$

- $\mathrm{Bias}\left(\hat{f}\left(\vec{X}_0\right)\right)^2$: On average (over all possible training sets), how much does $\hat{f}\left(x_0\right)$ differ from $f\left(x_0\right)$? This can be understood as the error introduced by approximating a real world problem.
- $\mathrm{Var}\left(\hat{f}\left(\vec{x}_0\right)\right)$: How much does $\hat{f}\left(x_0\right)$ vary from sample to sample? If we used a different training set how much would $\hat{f}$ differ?