Statistical Learning

Maryclare Griffin

2024-09-05

This material is based on Chapter 2 of Introduction to Statistical Learning (ISL) and Chapter 2 of Elements of Statistical Learning (ESL). We will tend to follow ISL more closely, and look to ESL for occasional additional higher level material.

For the next several lectures, we will focus on **supervised** learning where we observe inputs $\boldsymbol{X}$ and an output $\boldsymbol{y}$. We will assume that there is some statistical model relating pairs of inputs and outputs,

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

where f is a fixed but unknown function of $X = X_1, \ldots, X_p$ and ϵ is a mean zero random error term which is independent of X. The random error may reflect random variation in the measurement of Y or variation associated with additional unobserved variables that account for variation of observed output y_i corresponding to the same inputs x_i.

We will be interested in estimating f. We will refer to the estimated f as $\hat{f}$. Generally, the goal of estimating f is either:

- **Prediction**: Predicting the output Y given certain inputs, leveraging the fact that ϵ is mean zero. A predicted value of Y will be denoted as $\hat{Y} = \hat{f}(X)$.
- **Inference**: Understanding the association between Y and $X_1, \ldots, X_p$.

When we are interested in **prediction**, our goal is to minimize

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

specifically the reducible error $E[(f(X) - \hat{f}(X))^2]$. We refer to $E[(f(X) - \hat{f}(X))]2$ as the reducible error because improving our estimate $\hat{f}$ of f can make it smaller. We refer to $\text{Var}[\epsilon]$ as the irreducible error because is inherent to the data and cannot be reduced. When making predictions, the form of $\hat{f}(X)$ is a secondary concern and **black box** methods are popular.

When we are interested in **inference**, understanding the form of $\hat{f}$, is more important because we may want to understand phenomena that correspond to specific properties of f, e.g.:

- Which predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are associated with the response $\boldsymbol{y}$
- Which predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ have an approximately linear association with the response $\boldsymbol{y}$
- The relationship between each predictor $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ and the response $\boldsymbol{y}$

There is often a trade-off between methods that are best for **prediction** versus **inference**, for instance linear models can be very useful for **inference** because they are interpretable, but tend to provide poorer predictions.

This class will cover many methods for estimating linear and/or nonlinear functions f from n observations, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i = (x_{i1}, \ldots, x_{ip})^\top$. The data we use to estimate f will sometimes be called the **training data**. Methods for estimating f can often be described as **parametric** or **non-parametric**. We will consider both, and we note that sometimes there is some overlap between the two.

Parametric methods require (i) choice of a functional form for $f(X)$ that reduces the problem of estimating an entire p-dimensional function to a problem of estimating a smaller number of parameters and a (ii) choice of a procedure for finding values of the parameters that yield $\hat{f}(X)$ satisfying $\hat{f}(X) \approx Y$.

- For instance, ordinary least squares (i) assumes linearity of $f(X)$, i.e. $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, and reduces the problem of estimating $f(X)$ to the problem of estimating $\beta_0, \beta_1, \ldots, \beta_p$ and (ii) estimates $\beta_0, \beta_1, \ldots, \beta_p$ by minimizing the sum of squared errors $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2$ with respect to $\beta_0, \beta_1, \ldots, \beta_p$.

A problem with **parametric** methods is that the choice of a functional form for $f(X)$ may be incorrect. We can protect against this by choosing more flexible functional forms for $f(X)$, but more flexible functional forms for $f(X)$ will depend on more parameters and tend to **overfit** the data by including too much information from the irreducible part of Y - the noise - in the estimate of $\hat{f}(X)$.

Non-Parametric methods do not make explicit assumptions about the functional form for $f(X)$. This is a great thing! But it is expensive in terms of data. Non-parametric tend to require much more data, i.e. larger values of n, to produce accurate estimates of $f(X)$. Additionally, the flexibility of non-parametric approaches can make it difficult to perform inference, as they tend to be less interpretable. Specifically, it can be harder to explain how changes in an individual predictor $\boldsymbol{x}_j$ are associated with changes in the response $\boldsymbol{y}$ based on a non-parametric estimate.

The mention of **overfitting** suggests a measure of how well a model fits, i.e. a quantification of how close the estimate $\hat{f}(X)$ is to the observed response Y. One of the most commonly used measures is **mean squared error (MSE)**,

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

Above, the MSE is defined as a function of the training data, i.e. the n observations of x_i and y_i that we fit the model to, and will sometimes be called the **training MSE**. We make this distinction because often - especially when we are interested in prediction - we are not interested in how well the model fits the data used to estimate $\hat{f}$, but rather how well the model fits data we haven't seen before. Letting (x_0, y_0) refer to an arbitrary data point that has not been seen before, we define the **test MSE** as

$$(y_0 - \hat{f}(x_0))^2.$$

We can imagine averaging this over all possible observations that we haven't seen before but may in the future. This quantity measures how well an estimate $\hat{f}$ of f performs when used to fit data that we have not seen before.

Having defined training and test MSE allows us to better define **overfitting**. A model that is **overfitting** will have a very low training MSE and a very low test MSE.

Unfortunately, we cannot compute the test MSE as defined, because the test MSE depends on data that we have not seen before. Fortunately, we can approximate it by repeatedly defining training data as a subset of our observed data, and defining test data as the remainder that was not used to fit the model. This procedure describes cross-validation, and we will return to it in future lectures.

A concept that is closely related to **overfitting** that also helps us understand differences between **parametric** and **non-parametric** methods is the bias-variance trade off. Let $E[(y_0 - \hat{f}(x_0))^2]$ refer to the expected squared error over unobserved data points. It can be decomposed as:

$$E[(y_0 - \hat{f}(x_0))^2] = \underbrace{E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2]}_{\text{Var}[\hat{f}(x_0)]} + \underbrace{E[\hat{f}(x_0) - f(x_0)]^2}_{\text{Squared Bias}[\hat{f}(x_0)]} + \text{Var}[\epsilon].$$

We can think of the first term - $\text{Var}[\hat{f}(x_0)]$ - as quantifying how much our estimate $\hat{f}$ is expected to change if fit to a different training dataset. Estimates $\hat{f}$ that **overfit** the data will be expected to change a lot and have high $\text{Var}[\hat{f}(x_0)]$. Because they are more flexible, **non-parametric** methods tend to be associated with higher variance.

We can think of the second term - $\text{Bias}[\hat{f}(x_0)]^2$ - as quantifying how accurate our estimate $\hat{f}$ is. Because they are constrained in ways that may not reflect the true structure of $f(X)$, **parametric** methods tend to be associated with bias.

In general, more flexible methods are associated with higher variance and lower bias. This phenomenon is called the **bias-variance tradeoff**. Finding the best balance of the two will require being able to toggle how flexible a method or estimate $\hat{f}$ is and find a happy medium. This is also something we will discuss in future lectures.