Lecture 3, Part I: Error and Validation

Text references: Chapter 3 in Shalizi

Review/Introductory Remark

Suppose X represents different covariates and Y is our response variable, and suppose that we are interested in the relationships between X and Y. Here are three *different* but related concepts:

1. association
2. causation
3. prediction

Association does not imply causation. Do you remember when you would be able to make causal statements?

Furthermore, a strong association between X and Y does not necessarily mean that X is an important predictor in the model. Why? Give an example.

In Lecture 2, Parts III and IV, we discussed causal inference, and the pitfalls of interpreting relationships or the output of a multiple regression program as casual. In this lecture, we are going to take a purely *predictive viewpoint* of regression. We start by describing the different sources of errors in prediction and how to estimate error in practice.

The Predictive Viewpoint. Set-Up.

Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ we have two goals:

estimation: Find an estimate $\hat{r}(x)$ of the regression function $r(x)$.

prediction: Given a *new* X, predict Y; we use $\hat{Y} = \hat{r}(X)$ as the prediction.

Note the following (what is random? what is fixed?):

Q: How well does the the function $\hat{r}$, constructed on the training set/sample, predict our new test point (X, Y)?

To measure this, we define the **expected test error** (which is often referred to as **prediction error** or **prediction risk**):

where the expectation is over all that is random (training set, and test point).

Q: Why would we want to measure this? Here are two major reasons:

1. *Model assessment:* sometimes we would just like to know how well we can predict a future observation, in absolute terms
2. *Model selection:* often we will want to choose between different fitting functions; this could mean choosing between two different model classes entirely (e.g., linear regression versus some other fixed method) or choosing an underlying tuning parameter for a single method (e.g., choosing k in k-nearest neighbors). How to do this? A common way is to compare prediction errors (which are in practice *estimated* from data).

Note that there are two estimation problems:

The Bias-Variance Decomposition

First some intuition

The expected test error or prediction risk R above has an important property: it decomposes into informative parts. But, before we go through the mathematics of the decomposition, let's think about a few points:

1. Can we ever predict Y from X with zero prediction error? Likely, no. Even if our function $\widehat{r}$ happened to capture the ideal relation underlying X, Y, i.e., the true regression function $r(X) = \mathbb{E}(Y|X)$, we would still incur error, due to noise. We call this *irreducible error*; it can represent

2. What happens if our fitted function $\widehat{r}$ belongs to a model class that is far from the true regression function r? E.g., we may choose to fit a linear model in a setting where the true relationship is far from linear? We'll often refer to this as a *misspecified model*. As a result, we encounter error, what we call *estimation bias*; it represents

3. What happens if our fitted (random) function $\widehat{r}$ is itself quite variable? In other words, for different training samples of size n, we end up constructing substantially different functions $\widehat{r}$? This is another source of

error, that we'll call *estimation variance*; it represents

Point Estimation Revisited. The Bias-Variance Decomposition of MSE

Before we return to regression, first recall the problem of **point estimation**, i.e. the problem of providing a single "best guess" of some fixed quantity of interest. More formally let $Y_1, \ldots, Y_n$ be n IID data points from some distribution. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $Y_1, \ldots, Y_n$ (what is fixed? what is random?):

The bias of a point estimator is defined by:

The distribution of $\widehat{\theta}_n$ is called the **sampling distribution**. Think about this figure:

The quality of a point estimate is sometimes assessed by the **mean squared error**, or MSE, defined by:

The MSE can be written as:

Now compare **point estimation** with regression and **curve estimation**. Think about this figure and compare with the one above!

The Bias-Variance Decomposition of the MSE in Regression

If X and Y are random variables, recall the rule of iterated expectations

$$\mathbb{E}[g(X,Y)] = \mathbb{E}[\mathbb{E}[g(x,Y)|X=x]],$$

where the inner expectation is taken with respect to $Y|X$ and the outer one is taken with respect to the marginal distribution of X. Throughout the following section, we use this rule, conditioning on X to obtain the risk function $R(x)$ at $X = x$.

Hence, for our expected test error or prediction error, we have

Now look at $R(x)$, the *prediction error conditional on* $X = x$ for some fixed value of x:

The first term is just the *irreducible error*. The second term can be further decomposed just as the MSE of a point estimator (NOTE: $\hat{r}(x)$ is a *random variable* itself, calculated at a fixed value of x) into a bias component and a variance component at x. Therefore, altogether,

which is called the **bias-variance decomposition**.

Typical trend: *underfitting* means high bias and low variance, *overfitting* means low bias but high variance. E.g., think about k in k-nearest-neighbors regression: relatively speaking, how do the bias and variance behave for small k, and for large k? [See R Demo 1 and HW Set 1, Problem 1]

Finally, we have for the prediction error R:

and R_{av} is called the **average prediction risk**.

<u>To summarize:</u> We wish to know R, the prediction risk. R_{av} provides an excellent approximation, but R_{av} is not a quantity that we can readily calculate empirically because we do not know $R(X_i)$. Let us next explore why it is challenging to calculate R.

The Optimism of the Training Error

Define the **training error**

What wrong with the training error or expected training error? (As before, we will condition on $x_1, \ldots, x_n$ in our discussion. If needed, one can always derive unconditional expectations by the "law of total expectations", and unconditional variances by the "law of total variance".)

We might guess that $\hat{R}_{\text{training}}$ estimates the prediction error (R) well but this is not true. The reason is that we used the observed pairs (x_i, Y_i) to obtain $\hat{Y}_i = \hat{r}(x_i)$. As a consequence Y_i *and* $\hat{Y}_i$ *are correlated*. Typically $\hat{Y}_i$ "predicts" Y_i better than it predicts a new Y at the same x_i. Let us explore this formally. Let $\overline{r}_i = \mathbb{E}(\hat{r}(x_i))$ and compute

$$\begin{aligned}
\mathbb{E}(Y_i - \hat{Y}_i)^2 &= \mathbb{E}(Y_i - r(x_i) + r(x_i) - \overline{r}(x_i) + \overline{r}(x_i) - \hat{Y}_i)^2 \\
&= \sigma^2 + \mathbb{E}(r(x_i) - \overline{r}(x_i))^2 + \mathbb{V}(\hat{r}(x_i)) - 2\mathsf{Cov}(\hat{Y}_i, Y_i).
\end{aligned}$$

Note: this time the cross-product involving the 1st and 3rd terms is not 0 because $\mathsf{Cov}(\hat{Y}_i, Y_i) \neq 0$. This is because Y_i is a particular observation from which we calculated $\hat{Y}_i$, hence the two terms are correlated. This introduces a bias into the estimate of risk:

Typically, $\mathsf{Cov}(\hat{Y}_i, Y_i) > 0$ and so $\hat{R}_{\text{training}}$ underestimates the risk. Later, we shall see how to estimate the prediction risk.