MODEL BUILDING 2

ETHAN LEVIEN

Contents

1. Linear models with features

In the context of interactions, we saw how a model can be extended by defining a new predictor $X_3 = X_1X_2$. The more general idea that we can define a new predictor which is a function of the other predictors allows us to develop very complex and flexible models which nonetheless can be analyzed within linear regression framework. Here, we will formalize this, beginning with the case of a single predictor.

For a single prediction, consider the model

$$y = f(X) + \epsilon \tag{1}$$

A simple example would be $f(x) = b + ax + a_2x^2$. This is simply a linear model if we define

$$X_1 = X, \quad X_2 = X_1^2. \tag{2}$$

More generally, a trick is to select a series of **basis** function, $\phi_1, \dots, \phi_m$ and express f as a combination of them:

$$f(X) = \sum_{i=1}^{m} a_i \phi_i(X) \tag{3}$$

The function $\phi_i(x)$ are also called **features**.

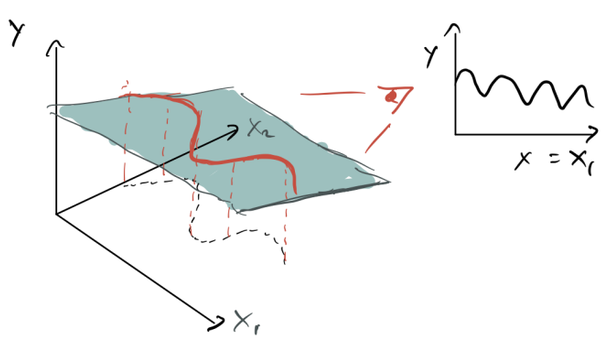

Figure 1. An illustration of how a nonlinear dependence on our predictor can be incorporated into the linear modeling framework by adding a feature.

Which functions ϕ should we use? The answer of course depends on the problem at hand. For example, we might know something about the physics of the data we are modeling. In some cases, we may select the ϕ so that the parameters a_i have clear interpretations (as is the case in linear regression model). The following illustrates such an example.

Date: April 2022.

Example 1. *Mauna Kea Data*

Exercise 1: *Killing a tumor*

1.1. **Orthogonality.** It is often to our advantage to select basis function ϕ which contain very different "information". We know from before that ideally our predictor variables should be uncorrelated, otherwise $\hat{a}_i$ will be very correlated in the sample distribution. Thus, ideally, we should select ϕ_i so that the correlation coefficient between $\phi_i(X)$ and $\phi_j(X)$ is very small. When

$$\mathbb{E}[\phi_i(X)\phi_j(X)] = \mathbb{E}[\phi_i(X)]\mathbb{E}[\phi_j(X)] \tag{4}$$

we say that ϕ_i and ϕ_j are **orthogonal**s with respect to the distribution of X. We won't get too deep into this, but it's important to understand when you go looking for basis functions. One can show for example, that the basis function $\phi_i(x) = \sin(2\pi x)$ are orthogonal with respect to

$$X \sim \text{Uniform}(-1, 1). \tag{5}$$

2. Direct assessment of model predictions

We have seen that any model has deficiencies and that **expanding our model always increases** R^2. A natural question is: Why not make our model as complex as possible? That is, why not add as many variables as we can and nonlinear terms? A simple answer could be given with the example of a linear regression: We can't draw a line through only two points. Similarly, we can't find a unique plane in d dimensions that goes through $< d$ points. This suggests we must not have more parameters in our model than data. Moreover, as the number of parameters in our model approach the amount of data, we become unable to resolve the parameters (the sample distributions become too wide) and interpret them in a meaningful way.

However, if our goal is make predictions with a model, there is much more to the story. Even before we have nearly as many parameters as data points, our model looses its power. This is not something we can see based on the behavior of error between the model and the data, such as R^2, which always decreases as we expand our model. Instead, we need to look at a **direct assessment of out-of-sample predictive power**. Ideally, we could look at the error between the model and the predictors for new X values. One way to access this is to break our data up into two subsets, a **training** (denoted $Y_{\text{train},i}$) and **test** set (denoted $Y_{\text{test},i}$) . We fit the model using only the training set, and then see how well our model can predict the values in the test set. The following examples reveals what we can learn from this:

Example 2. *Fitting polynomial data*

2.1. **Bais-variance tradeoff.** To make sense of the results in the previous example, let's define a few things. We define the training error as

$$\epsilon^2_{\text{train}} = \mathbb{E}[(\hat{Y}_i - Y_{\text{train},i})^2] \tag{6}$$

where the average is taken over different realizations of our data and $\hat{Y}_i$ is our prediction of $\mathbb{E}[Y|X]$. Not the relationship between R^2 and the training error:

$$R^2 \approx 1 - \frac{\epsilon^2_{\text{train}}}{\text{var}(Y_{\text{train},i})}. \tag{7}$$

Similarly, we define the test error as

$$\epsilon^2_{\text{test}} = \mathbb{E}[(\hat{Y}_i - Y_{\text{test},i})^2] \tag{8}$$

It can be shown that

$$\epsilon^2_{\text{test}} = \underbrace{(\mathbb{E}[\hat{Y}_i] - y)^2}_{=\text{bias}} + \underbrace{\text{var}(\hat{y})}_{=\text{variance}} + \sigma^2_\epsilon \tag{9}$$

The **bias** results from the fact that our model will systematically under or over estimate the y values. For example, if we try to fit an exponentially decaying curve with a straight line, different data sets will give consistent result, but on average they will overestimate the middle of the data and underestimate the ends.

The **variance** variation between our model predictions and the data between different datasets from the same model. For example, if we interpolate every single-point, then a different set of points will cause our curve to change in ways that differ from the data.

Roughly speaking, a biased model will give us consistent but incorrect results, while a low bias high variance model will be correct on average, but our predictions will vary a lot from data set to data set and therefore will not be reliable. A key conceptual point to understand is that the **variance arrises from a model being too complex, while bias arrises from a model not being complex enough.**

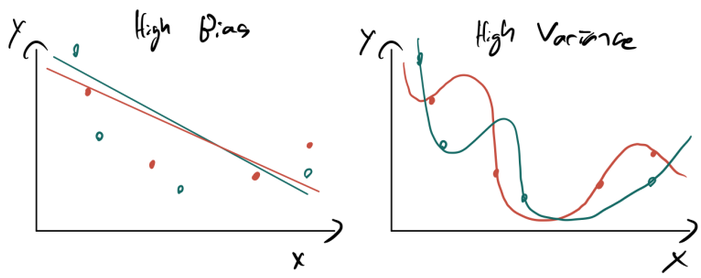

FIGURE 2. Bias an variance