STATS305A - Lecture 12

John Duchi
Scribed by Michael Howes

10/28/21

Contents

1 Announcements

- Etude 2 due today 5pm.
- No class next Tuesday.

2 Model Selection and prediction

2.1 Motivation

Up to this point we've treated the model $Y = X\beta + \varepsilon$ as "god-give". This is a bit inaccurate. In real life we will typically have data and no model and have to figure it out and select a model. When selecting a model we have two desiderata:

- Identify important features thhat are relating x to our response y.
- Pure predictive accuracy: how well can we predict y from x?

These two are intertwinned. We don't always have to choose one over the other.

2.2 Bias/Variance Decomposition

Suppose we are in a setting where $y = f(x) + \varepsilon$ and $\mathbb{E}[\varepsilon|x] = 0$. This is equivalent to having $f(x) = \mathbb{E}[Y|X = x]$ since if $\varepsilon = y - f(x)$, then

$$\mathbb{E}[\varepsilon|x] = \mathbb{E}[y|x] - f(x).$$

Thus $\mathbb{E}[\varepsilon|x] = 0$ if and only if $f(x) = \mathbb{E}[y|x]$. Define $\sigma^2(x) = \mathbb{E}[\varepsilon^2|x]$ which is the conditional variance of ε.

Our goal is to fit a predictor $\widehat{f}$ using a sample $\{(x_i, y_i)\}_{i=1}^n$. Note that if we think of the sample of $\{(x_i, y_i)\}_{i=1}^n$ as random, then the predictor $\widehat{f}$ is random (like how $\widehat{\beta}$ is random in the linear model). Thus we can take the expectation of quantities involving $\widehat{f}$ over all samples $\{(x_i, y_i)\}_{i=1}^n$. This idea will be used many times over the course of this lecture.

Definition 1. If we have a predictor $\widehat{f}$ of a model $y = f(x) + \varepsilon$, then we define the *in-sample (MSE) risk* of $\widehat{f}$ to be

$$R_{in}(\widehat{f}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(x_i) - f(x_i))^2\right],$$

where the above expectation is taken over all samples $\{(x_i, y_i)\}_{i=1}^n$ with x_i fixed. (That is we fix x and calculate $\widehat{f}$ using different samples (x, y), we then calculate the quantity $\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(x_i) - f(x_i))^2$ and take the expectation over all samples (x, y).)

Aside 1. Sometimes the in-sample risk is called the $L^2(P_n)$ risk. This is because R_{in} is the expectation of the L^2 norm error of $\widehat{f} - f$ with respect to the distribution

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{x_i}.$$

Definition 2. Sometimes the insample risk is defined with respect to a fresh sample $\{Y_i^*\}_{i=1}^n$ where

$$Y_i^* = \text{ a new sample of } Y_i = f(x_i) + \varepsilon_i^*,$$

where ε_i^* is an independent copy of ε_i. We then define

$$R_{in}^*(\widehat{f}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^* - \widehat{f}(x_i)\right)^2\right],$$

where here the expectation is over both $Y_1, \ldots, Y_n$ (used to calculate $\widehat{f}$) and over $Y_1^*, \ldots, Y_n^*$ (used to calculate $(Y_i^* - \widehat{f}(x_i))^2$).

Note that

$$\begin{aligned}
R_{in}^*(\widehat{f}) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \widehat{f}(x_i))^2\right] && (1)\\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - f(x_i))^2\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(x_i) - f(x_i))^2\right] && (2)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sigma^2(x_i) + R_{in}(\widehat{f}). && (3)
\end{aligned}$$

We call $\frac{1}{n}\sum_{i=1}^{n}\sigma^2(x_i)$ the irreducible error.

Now suppose that we have a function $g : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is the space X lives in. Note that g is different to $\widehat{f}$. The predictor $\widehat{f}$ is something that the depend on the sample (x, y) used to fit $\widehat{f}$. The function g is simply a function. It is a way of taking an X and producing a number. With this in mind we define

Definition 3. Given a function $g : \mathcal{X} \to \mathbb{R}$, the *(MSE) out of sample risk* of g is

$$R_{out}(g) = \mathbb{E}[(Y - g(X))^2] = \int_{\mathcal{X}} \mathbb{E}[(Y - g(x))^2 | X = x]p(x)dx.$$

Here the expectation is over both Y and X (hence out of sample - we are allowing X to change).

Note that

$$\begin{aligned}
R_{out}(g) &= \mathbb{E}\left[(Y - f(X) + f(X) + g(X))^2\right]\\
&= \mathbb{E}\left[(Y - f(X))^2\right] + \mathbb{E}[(f(X) - g(X))^2] + 2\mathbb{E}\left[(Y - f(X))(f(X) - g(X))\right]\\
&= \mathbb{E}[\sigma^2(X)] + \mathbb{E}[(f(X) - g(X))^2].
\end{aligned}$$

We again call $\mathbb{E}[\sigma^2(X)]$ the irreducible error and we could call $\mathbb{E}[(f(X) - g(X))^2]$ the error in mean prediction (this last term is just a term John used - he said that there isn't really a term in literature for it).

In the out of sample risk we average over all the X's we could possible draw. In the in sample we fix the value x_i and average over all possible y_i. Note that if our data if i.i.d., then

$$R_{out}(g) = \mathbb{E}[(g(X_{n+1}) - Y_{n+1})^2].$$