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-given". 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 that 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 In sample and out of sample risk

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 squared 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

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

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 depends 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 have our next definition.

Definition 3. Given a function $g : \mathcal{X} \to \mathbb{R}$, the *out of sample (MSE) 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 $\mathbb{E}[(f(X) - g(X)^2)]$).

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 is i.i.d., then

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

The quantity $R_{out}(\widehat{f})$ is a number but it is random since it depends on the sample $(x_i, y_i)_{i=1}^n$ used to fit $\widehat{f}$ and the sample $(x_i, y_i)_{i=1}^n$ is random. Thus $\mathbb{E}[R_{out}(\widehat{f})]$ is our expected out of sample risk over all samples used to fit $\widehat{f}$.

2.3 Bias/variance decomposition

Let $\widehat{f}$ be any estimator of f. For any $x \in \mathcal{X}$ we can write

$$\begin{aligned} \mathbb{E}[(\widehat{f}(x) - f(x))^2] &= \mathbb{E}\left[\left(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)] + \mathbb{E}[\widehat{f}(x)] - f(x)\right)^2\right] \\ &= \mathbb{E}[\left(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)]\right)^2] + \left(\mathbb{E}[\widehat{f}(x)] - f(x)\right)^2 \\ &= \mathrm{Var}(\widehat{f}(x)) + \mathrm{Bias}(\widehat{f}(x))^2, \end{aligned}$$

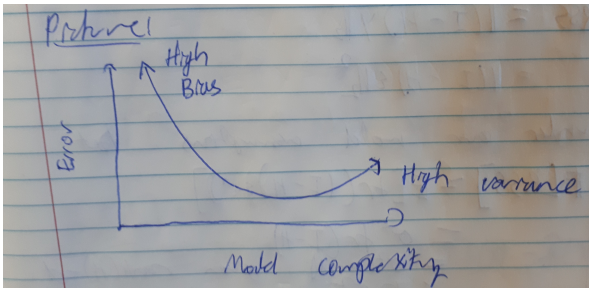
where, as in $R_{in}(\widehat{f})$, the above expectation is taken with respect to the samples used to fit $\widehat{f}$. Thus we have

$$R_{in}(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Bias}(\widehat{f}(x_i))^2 + \mathrm{Var}(\widehat{f}(x_i))\right),$$

and

$$\begin{aligned} \mathbb{E}[R_{out}(\widehat{f})] &= \mathbb{E}[\sigma^2(X)] + \mathbb{E}[\mathrm{Bias}(\widehat{f}(X))^2] + \mathbb{E}[\mathrm{Var}(\widehat{f}(X))] \\ &= \text{irreducible error} \ + \ \text{expected bias squared} \ + \ \text{expected variance}. \end{aligned}$$

This decomposition means that in practice we often see the following curve when we plot error against model complexity. The error initally goes down as complexity increases but then increases. This is because intially the model has high bias when the model is simple but once the model is very complicated it has high variance.



Although one word of warning: We sometimes don't know how to properly measure model complexity. This means that when we use a proxy for model complexity we might not see the above picture. This is sometimes the case with machine learning algorithms where the error continues to go down even as the number of parameters is getting very very high. The complexity of the model may not be as high as the number of parameters suggests.

2.4 Comparing models

The above analysis gives us the idea that we should choose a model that trades optimally between bias and variance. Unfortunately we can't do this exactly since we can only approximate the bias and variance of a model. We thus will have to estimate the prediction error. We can then use this estimate to choose a model. There is a challenge though, the natural quantity

$$\frac{1}{n}RSS = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i\widehat{Y}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{f}(x_i)\right)^2,$$

is biased downwards since we chose $\widehat{f}$ to minimize RSS. Thus $\frac{1}{n}RSS$ will tend to be smaller than R^*_{in} since in R^*_{in} we have a new independent sample of Y's.

Example 1. Consider for concreteness the linear model $Y = X\beta + \varepsilon$ where $\varepsilon \sim (0, \sigma^2 I_n)$ and X is full rank with $\text{rank}(X) = d$. Let $H = X(X^TX)^{-1}X^T$, we know that

$$RSS = \|(I-H)Y\|_2^2 = \|(I-H)\varepsilon\|_2^2.$$

Thus we have

$$\begin{aligned}\mathbb{E}[RSS] &= \mathbb{E}[\text{tr}((I-H)\varepsilon\varepsilon^T)] \\ &= \sigma^2\,\text{tr}(I-H) \\ &= \sigma^2(n-d) \\ &< \sigma^2 n.\end{aligned}$$

Thus $\mathbb{E}\left[\frac{1}{n}RSS\right] = \frac{n-d}{n}\sigma^2 < \sigma^2$ which is a problem since the irreducible error is σ^2 so we will always have $R^*_{in} \geq \sigma^2$.

There are two approaches to overcome this issue:

(a) Penalized risk estimates.

(b) Validation.

We will discuss the first approach today.

2.5 Estimating risk

How do we estimate the risk of a model? Let's think about the sum of square errors. Note that

$$\begin{aligned}\mathbb{E}\left[\left(Y_i - \widehat{Y}_i\right)^2\right] &= \mathbb{E}\left[\left(Y_i - f(x_i) + f(x_i) - \widehat{Y}_i\right)^2\right] \\ &= \sigma_i^2 + 2\mathbb{E}\left[(Y_i - f(x_i))(f(x_i) - \widehat{Y}_i)\right] + \mathbb{E}\left[\left(f(x_i) - \widehat{f}(x_i)\right)^2\right] \\ &= \sigma_i^2 + \mathbb{E}\left[(f(x_i) - \widehat{f}(x_i))^2\right] - 2\,\text{Cov}(Y_i, \widehat{Y}_i)\end{aligned}$$

The last equality holds because $f(x_i) = \mathbb{E}[Y_i]$ and so for any constant c_i

$$\mathbb{E}\left[(Y_i - f(x_i))(\widehat{Y}_i - c_i)\right] = \mathbb{E}\left[(Y_i - f(x_i))(\widehat{Y}_i - \mathbb{E}[\widehat{Y}_i])\right] = \mathrm{Cov}(Y_i, \widehat{Y}_i).$$

Thus we have

$$\begin{aligned}
\mathbb{E}[RSS] &= \mathbb{E}\left[\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2\right] \\
&= \sum_{i=1}^n \sigma_i^2 + \sum_{i=1}^n \mathbb{E}\left[(f(x_i) - \widehat{f}(x_i))^2\right] - 2\,\mathrm{Cov}(Y, \widehat{Y}) \\
&= \sum_{i=1}^n \sigma_i^2 + nR_{in}(\widehat{f}) - 2\,\mathrm{Cov}(Y, \widehat{Y}).
\end{aligned}$$

Equivalently,

$$\begin{aligned}
R^*_{in}(\widehat{f}) &= R_{in}(\widehat{f}) + \frac{1}{n}\sum_{i=1}^n \sigma_i^2 \\
&= \frac{1}{n}\mathbb{E}[RSS] + \frac{2}{n}\,\mathrm{Cov}(Y, \widehat{Y}).
\end{aligned}$$

The quantity $\mathrm{Cov}(Y, \widehat{Y})$ is called the "effective degrees of freedom". The upshot of these calculations is that if we can estimate $\frac{2}{n}\,\mathrm{Cov}(Y, \widehat{Y})$, then we can estimate the in-sample risk of $\widehat{f}$.

Example 2. Let's calculate the $\mathrm{Cov}(Y, \widehat{Y})$ in some special cases. Suppose we have a "true" linear model $Y = X\beta + \varepsilon$ with $\varepsilon \sim (0, \sigma^2 I_n)$ and we fit $\widehat{Y} = HY$. Then

$$\begin{aligned}
\mathrm{Cov}(Y, \widehat{Y}) &= \mathrm{Cov}(\varepsilon, H\varepsilon) \\
&= \mathbb{E}[\varepsilon^T H \varepsilon] \\
&= \sigma^2\,\mathrm{tr}(H) \\
&= \sigma^2 d,
\end{aligned}$$

if H has rank d. Suppose know that we don't know that Y comes from a linear model but we have the more general model $Y_i = f(x_i) + \varepsilon_i$. Suppose we still fit $\widehat{f}(x) = x^T\widehat{\beta}$ where $\widehat{\beta} = (X^TX)^{-1}X^TY$. We then still have $\widehat{Y} = HY$ and so

$$\begin{aligned}
\mathrm{Cov}(Y, \widehat{Y}) &= \mathrm{Cov}(\varepsilon, HY) \\
&= \mathrm{Cov}(\varepsilon, H[f(x_i) + \varepsilon_i]_{i=1}^n) \\
&= \mathrm{Cov}(\varepsilon, H\varepsilon) \\
&= \sum_{i=1}^n \sigma_i^2 H_{ii},
\end{aligned}$$

where $\sigma_i^2 = \mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i)$. If we have homoscedastic noise, then $\sigma_i^2 = \sigma$ for all i and

$$\mathrm{Cov}(Y, \widehat{Y}) = \sigma^2 d,$$

where again d is the rank of H. These two examples give motivation for calling $\mathrm{Cov}(Y, \widehat{Y})$ the effective degrees of freedom.

Definition 4. *Mallow's C_p statistic* is the statistic.

$$C_p := \frac{1}{n}RSS + 2\frac{\sigma^2}{n}\operatorname{rank}(X) = \frac{1}{n}RSS + 2\frac{\sigma^2}{n}p,$$

where $p = \operatorname{rank}(X)$.

Definition 5. *Mallow's method* is the following.

- Given a number of possible models, calculate C_p for each model.
- Choose the model that minimizes C_p

Unfortunately we do not know σ^2 and so to use Mallow's method we have to estimate σ^2. Some options are

(a) If we have a pair $\left(Y_i^{(1)}, Y_i^{(2)}\right)$ for each x_i, then $\mathbb{E}\left[\left(Y_i^{(1)} - Y_i^{(2)}\right)^2\right] = 2\sigma$ and so we can use

$$\widehat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i^{(1)} - Y_i^{(2)}\right)^2.$$

(b) If we do not have paired data, then we can find x's that are close to each other. We can then pair these x's and use the above estimator.

(c) We can also use the most complicated model to estimate σ^2. Suppose that we have $Z \in \mathbb{R}^{n\times q}$ where q is very big. Then we can define

$$\widehat{\gamma} = (Z^TZ)^{-1}Z^TY \in \mathbb{R}^q,$$

and use the estimator

$$\widehat{\sigma}^2 = \frac{1}{n-q}\sum_{i=1}^{n}(Y_i - Z_i^T\widehat{\gamma})^2.$$