IEOR 265 – Lecture 4
Cross-Validation

1 Comparison of Different High-Dimensional Estimators

Suppose we consider the three estimators defined as

$$\begin{aligned}
\hat{\beta}_1 &= \arg\min\{\|Y - X\beta\|_2 \mid \|\beta\|_1 \leq 1\} \\
\hat{\beta}_2 &= \arg\min\{\|Y - X\beta\|_2 \mid \|\beta\|_2 \leq 1\} \\
\hat{\beta}_\infty &= \arg\min\{\|Y - X\beta\|_2 \mid \|\beta\|_\infty \leq 1\}.
\end{aligned}$$

Pictorially in two-dimensions, only the $\hat{\beta}_1$ estimator leads to sparsity. Furthermore, we can compare these three estimators using the expected estimation error from the M^* bound. Specifically, we have

$$\begin{aligned}
\mathbb{E}\|\hat{\beta}_1 - \beta\| &\leq C \cdot \sqrt{\frac{\log p}{n}} \\
\mathbb{E}\|\hat{\beta}_1 - \beta\| &\leq C \cdot \sqrt{\frac{p}{n}} \\
\mathbb{E}\|\hat{\beta}_1 - \beta\| &\leq C \cdot \frac{p}{\sqrt{n}}.
\end{aligned}$$

In high-dimensions (i.e., when p is large relative to n), only the $\hat{\beta}_1$ estimator has small error. This is because its error has a logarithmic dependence on dimension p, whereas the other estimators have either a square-root $\sqrt{p}$ or linear p dependence of dimension.

2 Bias-Variance Tradeoff

Consider the case of parametric regression with $\beta \in \mathbb{R}$, and suppose that we would like to analyze the expectation of the squared loss of the difference between a estimate $\hat{\beta}$ and the true parameter β. In particular, we have that

$$\begin{aligned}
\mathbb{E}\big((\hat{\beta} - \beta)^2\big) &= \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}) + \mathbb{E}(\hat{\beta}) - \beta)^2\big) \\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big) + 2\mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)(\hat{\beta} - \mathbb{E}(\hat{\beta})\big) \\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big) + 2(\mathbb{E}(\hat{\beta}) - \beta)(\mathbb{E}(\hat{\beta}) - \mathbb{E}(\hat{\beta})) \\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big).
\end{aligned}$$

The term $\mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big)$ is clearly the variance of the estimate $\hat{\beta}$. The other term $\mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big)$ measures how far away the "best" estimate is from the true value, and it is common to define bias($\hat{\beta}$) $= \mathbb{E}\big(\mathbb{E}(\hat{\beta}) - \beta\big)$. With this notation, we have that

$$\mathbb{E}\big((\hat{\beta} - \beta)^2\big) = (\text{bias}(\hat{\beta}))^2 + \text{var}(\hat{\beta}).$$

This equation states that the expected estimation error (as measured by the squared loss) is equal to the bias-squared plus the variance, and in fact there is a tradeoff between these two aspects in an estimate.
It is worth making three comments.

1. The first is that if bias($\hat{\beta}$) $= \mathbb{E}\big(\mathbb{E}(\hat{\beta}) - \beta\big) = 0$, then the estimate $\hat{\beta}$ is said to be *unbiased*.
2. Second, this bias-variance tradeoff exists for vector-valued parameters $\beta \in \mathbb{R}^p$, for non-parametric estimates, and other models.
3. Lastly, the term *overfit* is sometimes used to refer to an model with low bias but extremely high variance.

3 Bias and Variance of OLS

Recall that the OLS estimate of a linear model $y = x'\beta + \epsilon$ is given by

$$\hat{\beta} = (X'X)^{-1}(X'Y).$$

We begin by computing the expectation of the OLS estimate

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}((X'X)^{-1}X'Y)\\ &= \mathbb{E}((X'X)^{-1}X'(X\beta + \underline{\epsilon}))\\ &= \mathbb{E}(\beta + (X'X)^{-1}X'\underline{\epsilon})\\ &= \beta + \mathbb{E}(\mathbb{E}[(X'X)^{-1}X'\underline{\epsilon}|X])\\ &= \beta.\end{aligned}$$

As a result, we conclude that bias($\hat{\beta}$) $= 0$.
Now assume that the x_i are independent and identically distributed (i.i.d.) random variables, with mean zero $\mathbb{E}(x_i) = 0$ and have covariance matrix Σ that is **assumed to be invertible**. By the weak law of large numbers (wlln), we have that $\frac{1}{n}X'X \xrightarrow{p} \Sigma$ and that $\frac{1}{n}X'Y \xrightarrow{p} \Sigma\beta$. Using the Continuous Mapping Theorem, we have that $\hat{\beta} \xrightarrow{p} \beta$. When an estimate converges in probability to its true value, we say that the estimate is *consistent*. Thus, the OLS estimate is consistent under the model we have described; however, consistency does not hold if Σ is not invertible.
We next determine the asymptotic variance of the OLS estimate. By the Central Limit Theorem (CLT), we have that $\frac{1}{n}X'X = \Sigma + O_p(\frac{1}{\sqrt{n}})$ and $\frac{1}{n}X'Y = \Sigma\beta + O_p(\frac{1}{\sqrt{n}})$. Applying the Continuous Mapping Theorem gives that $\hat{\beta} = \beta + O_p(\frac{1}{\sqrt{n}})$. Since we know that

$$\mathbb{E}((\hat{\beta} - \beta)^2) = (\text{bias}(\hat{\beta}))^2 + \text{var}(\hat{\beta}),$$

we must have that var($\hat{\beta}$) $= O_p(1/n)$, since we showed above that bias($\hat{\beta}$) $= 0$.

4 Cross-Validation

Cross-validation is a data-driven approach that is used to choose tuning parameters for regression. The choice of bandwidth h is an example of a tuning parameter that needs to be chosen in order to use LLR. The basic idea of cross-validation is to split the data into two parts. The first part of data is used to compute different estimates (where the difference is due to different tuning parameter values), and the second part of data is used to compute a measure of the quality of the estimate. The tuning parameter that has the best computing measure of quality is selected, and then that particular value for the tuning parameter is used to compute an estimate using all of the data. Cross-validation is closely related to the jackknife and bootstrap methods, which are more general. We will not discuss those methods here.

4.1 Leave k-Out Cross-Validation

We can describe this method as an algorithm.

```
data   : $(x_i, y_i)$ for $i = 1, \dots, n$ (measurements)
input  : $\lambda_j$ for $j = 1, \dots, z$ (tuning parameters)
input  : $R$ (repetition count)
input  : $k$ (leave-out size)
output: $\lambda^*$ (cross-validation selected tuning parameter)
for $j \leftarrow 1$ to $z$ do
    set $e_j \leftarrow 0$;
end
for $r \leftarrow 1$ to $R$ do
    set $\mathcal{V}$ to be $k$ randomly picked indices from $\mathcal{I} = \{1, \dots, n\}$;
    for $j \leftarrow 1$ to $z$ do
        fit model using $\lambda_j$ and $(x_i, y_i)$ for $i \in \mathcal{I} \setminus \mathcal{V}$;
        compute cross-validation error $e_j \leftarrow e_j + \sum_{i \in \mathcal{V}} (y_i - \hat{y}_i)^2$;
    end
end
set $\lambda^* \leftarrow \lambda_j$ for $j := \arg\min e_j$;
```

4.2 k-Fold Cross-Validation

We can describe this method as an algorithm.

```
data   : $(x_i, y_i)$ for $i = 1, \ldots, n$ (measurements)
input  : $\lambda_j$ for $j = 1, \ldots, z$ (tuning parameters)
input  : $k$ (block sizes)
output: $\lambda^*$ (cross-validation selected tuning parameter)
for $j \leftarrow 1$ to $k$ do
    set $e_j \leftarrow 0$;
end
partition $\mathcal{I} = \{1, \ldots, n\}$ into $k$ randomly chosen subsets $\mathcal{V}_r$ for $r = 1, \ldots, k$;
for $r \leftarrow 1$ to $k$ do
    for $j \leftarrow 1$ to $z$ do
        fit model using $\lambda_j$ and $(x_i, y_i)$ for $i \in \mathcal{I} \setminus \mathcal{V}_r$;
        compute cross-validation error $e_j \leftarrow e_j + \sum_{i \in \mathcal{V}_r} (y_i - \hat{y}_i)^2$;
    end
end
set $\lambda^* = \lambda_j$ for $j := \arg\min e_j$;
```

4.3 Notes on Cross-Validation

There are a few important points to mention:

1. The first point is that the cross-validation error is an estimate of prediction error, which is defined as
$$\mathbb{E}((\hat{y} - y)^2).$$
One issue with cross-validation error (and this issue is shared by jackknife and bootstrap as well), is that these estimates of prediction error must necessarily be biased lower. The intuition is that we are trying to estimate prediction error using data we have seen, but the true prediction error involves data we have not seen.

2. The second point is related, and it is that the typical use of cross-validation is heuristic in nature. In particular, the consistency of an estimate is affected by the use of cross-validation. There are different cross-validation algorithms, and some algorithms can "destroy" the consistency of an estimator. Because these issues are usually ignored in practice, it is important to remember that cross-validation is usually used in a heuristic manner.

3. The last point is that we can never eliminate the need for tuning parameters. Even though cross-validation allows us to pick a λ^* in a data-driven manner, we have introduced new tuning parameters such as k. The reason that cross-validation is considered to be a data-driven approach to choosing tuning parameters is that estimates are usually less sensitive to the choice of cross-validation tuning parameters, though this is not always true.