第七次作业

Ext 7.1

Ex. 7.1 Derive the estimate of in-sample error (7.24).

In summary, we have the important relation

$$\mathrm{E}_{\mathbf{y}}(\mathrm{Err_{in}}) = \mathrm{E}_{\mathbf{y}}(\overline{\mathrm{err}}) + \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i). \tag{7.22}$$

This expression simplifies if $\hat{y}_i$ is obtained by a linear fit with d inputs or basis functions. For example,

$$\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i) = d\sigma_\varepsilon^2 \tag{7.23}$$

for the additive error model $Y = f(X) + \varepsilon$, and so

$$\mathrm{E}_{\mathbf{y}}(\mathrm{Err_{in}}) = \mathrm{E}_{\mathbf{y}}(\overline{\mathrm{err}}) + 2\cdot\frac{d}{N}\sigma_\varepsilon^2. \tag{7.24}$$

Notice that from (7.22) we just need to show that:

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = d\sigma_\epsilon^2$$

We use trace to simplify the equation(using the truth that $\hat{y} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$):

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = tr(Cov(\hat{y}, y)) = tr(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma_\epsilon^2) = tr(I_d)\sigma_\epsilon^2 = d\sigma_\epsilon^2$$

Ext 7.3

Ex. 7.3 Let $\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$ be a linear smoothing of $\mathbf{y}$.

(a) If S_{ii} is the ith diagonal element of $\mathbf{S}$, show that for $\mathbf{S}$ arising from least squares projections and cubic smoothing splines, the cross-validated residual can be written as

$$y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}. \tag{7.64}$$

(b) Use this result to show that $|y_i - \hat{f}^{-i}(x_i)| \geq |y_i - \hat{f}(x_i)|$.

(c) Find general conditions on any smoother $\mathbf{S}$ to make result (7.64) hold.

Firstly we know that $\hat{f} = Sy$ is a linear smoothing of y, so we know that:

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{\Omega})^{-1}\mathbf{X}^T.$$

(a)

We write $\hat{f}^{-i}(x_i)$ first. We use X_{-i} to notation for input X without the i-th row.

$$\begin{aligned}\hat{f}^{-i}(x_i) &= x_i^T(\mathbf{X}_{-i}^T\mathbf{X}_{-i} + \lambda\mathbf{\Omega})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}\\ &= x_i^T(\mathbf{X}^T\mathbf{X} - x_ix_i^T + \lambda\mathbf{\Omega})^{-1}(\mathbf{X}^T\mathbf{y} - x_iy_i).\end{aligned}$$

And use the Woodbury matrix Identity that we see $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{\Omega}$ as $\mathbf{A}$. so

$$(\mathbf{A} - x_ix_i^T)^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}x_ix_i^T\mathbf{A}^{-1}}{1 - x_i^T\mathbf{A}^{-1}x_i}.$$

Then we use

$$
\begin{aligned}
\hat{f}^{-1}(x_i) &= x_i^T \left(\mathbf{A}^{-1} + \frac{\mathbf{A}^{-1} x_i x_i^T \mathbf{A}^{-1}}{1 - x_i^T \mathbf{A}^{-1} x_i}\right) (\mathbf{X}^T \mathbf{y} - x_i y_i) \\
&= \left(x_i^T \mathbf{A}^{-1} + \frac{S_{ii} x_i^T \mathbf{A}^{-1}}{1 - S_{ii}}\right) (\mathbf{X}^T \mathbf{y} - x_i y_i) \\
&= x_i^T \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y} - x_i^T \mathbf{A}^{-1} x_i y_i + \frac{S_{ii} x_i^T \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y}}{1 - S_{ii}} - \frac{S_{ii} x_i^T \mathbf{A}^{-1} x_i y_i}{1 - S_{ii}} \\
&= \hat{f}(x_i) - y_i S_{ii} + \frac{S_{ii} \hat{f}(x_i)}{1 - S_{ii}} - \frac{y_i S_{ii}^2}{1 - S_{ii}} \\
&= \frac{\hat{f}(x_i) - y_i S_{ii}}{1 - S_{ii}}.
\end{aligned}
$$

Then we get (7.64) easily.

$$
y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}.
$$

(b)

From this part in textbook:

5.4.1 Degrees of Freedom and Smoother Matrices

We have not yet indicated how λ is chosen for the smoothing spline. Later in this chapter we describe automatic methods using techniques such as cross-validation. In this section we discuss intuitive ways of prespecifying the amount of smoothing.

A smoothing spline with prechosen λ is an example of a *linear smoother* (as in linear operator). This is because the estimated parameters in (5.12) are a linear combination of the y_i. Denote by $\hat{\mathbf{f}}$ the N-vector of fitted values $\hat{f}(x_i)$ at the training predictors x_i. Then

$$\begin{aligned} \hat{\mathbf{f}} &= \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y} \\ &= \mathbf{S}_\lambda\mathbf{y}. \end{aligned} \tag{5.14}$$

Again the fit is linear in $\mathbf{y}$, and the finite linear operator $\mathbf{S}_\lambda$ is known as the *smoother matrix*. One consequence of this linearity is that the recipe for producing $\hat{\mathbf{f}}$ from $\mathbf{y}$ does not depend on $\mathbf{y}$ itself; $\mathbf{S}_\lambda$ depends only on the x_i and λ.

Linear operators are familiar in more traditional least squares fitting as well. Suppose $\mathbf{B}_\xi$ is a $N \times M$ matrix of M cubic-spline basis functions evaluated at the N training points x_i, with knot sequence ξ, and $M \ll N$. Then the vector of fitted spline values is given by

$$\begin{aligned} \hat{\mathbf{f}} &= \mathbf{B}_\xi(\mathbf{B}_\xi^T\mathbf{B}_\xi)^{-1}\mathbf{B}_\xi^T\mathbf{y} \\ &= \mathbf{H}_\xi\mathbf{y}. \end{aligned} \tag{5.15}$$

Here the linear operator $\mathbf{H}_\xi$ is a projection operator, also known as the *hat* matrix in statistics. There are some important similarities and differences between $\mathbf{H}_\xi$ and $\mathbf{S}_\lambda$:

- Both are symmetric, positive semidefinite matrices.
- $\mathbf{H}_\xi\mathbf{H}_\xi = \mathbf{H}_\xi$ (idempotent), while $\mathbf{S}_\lambda\mathbf{S}_\lambda \preceq \mathbf{S}_\lambda$, meaning that the right-hand side exceeds the left-hand side by a positive semidefinite matrix. This is a consequence of the *shrinking* nature of $\mathbf{S}_\lambda$, which we discuss further below.
- $\mathbf{H}_\xi$ has rank M, while $\mathbf{S}_\lambda$ has rank N.

We can get that $S^2 = S^\top S \preceq S$

where $\preceq$ denotes that the components are less than, so that there are $\forall i, (S^2)_{ii} \leq S_{ii}$

Considering the expansion $S^2 = S^\top S$ it is easy to see that the actual results are

$$(S^2)_{ii} = \sum_j (S_{ij})^2 = \sum_{i \neq j} (S_{ij})^2 + (S_{ii})^2$$

Thus,

$$0 \leq \sum_{i \neq j} (S_{ij})^2 + (S_{ii})^2 \leq S_{ii}$$

And the other side of the inequality can be obtained simply by reductio ad absurdum, assuming that $S_{ii} \geq 1$
then we have:

$$\sum_{i \neq j} (S_{ij})^2 + (S_{ii})^2 \geq (S_{ii})^2 \geq S_{ii}$$

Contradicts the conclusion above. So it can only be $S_{ii} \leq 1$

(c)

For general linear smoother $\hat{f} = \mathbf{S}\mathbf{y}$, if $\mathbf{S}$ only depends on $\mathbf{X}$ and other tuning parameters (i.e, independent of y.

To see that, note that if we replace y_i with $\hat{f}^{-i}(x_i)$ and denote the new vector by $\mathbf{y}', \mathbf{S}$ is not changed. Thus we have

$$\begin{aligned}\hat{f}^{-i}(x_i) &= (\mathbf{S}\mathbf{y}')i \\ &= \sum i \neq j S_{ij} \mathbf{y} j' + S ii \hat{f}^{-i}(x_i) \\ &= \hat{f}(x_i) - S_{ii} y_i + S_{ii} \hat{f}^{-i}(x_i),\end{aligned}$$

therefore we obtain (1).

Ext 7.4

Ex. 7.4 Consider the in-sample prediction error (7.18) and the training error $\overline{\text{err}}$ in the case of squared-error loss:

$$\begin{aligned}\text{Err}_{\text{in}} &= \frac{1}{N}\sum_{i=1}^{N}\text{E}_{Y^0}(Y_i^0-\hat{f}(x_i))^2\\ \overline{\text{err}} &= \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{f}(x_i))^2.\end{aligned}$$

Add and subtract $f(x_i)$ and $\text{E}\hat{f}(x_i)$ in each expression and expand. Hence establish that the average optimism in the training error is

$$\frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i),$$

as given in (7.21).

Firstly we define the following notation:

$$\begin{aligned}A_i &= E_{Y^0}(Y_i^0-f(x_i))^2\\ B_i &= E_{Y^0}(f(x_i)-E\hat{y}_i)^2=(f(x_i)-E\hat{y}_i)^2\\ C_i &= E_{Y^0}(E\hat{y}_i-\hat{y}_i)^2=(E\hat{y}_i-\hat{y}_i)^2\\ D_i &= 2E_{Y^0}(Y_i^0-f(x_i))(f(x_i)-E\hat{y}_i)\\ E_i &= 2E_{Y^0}(Y_i^0-f(x_i))(E\hat{y}_i-\hat{y}_i)\\ F_i &= 2E_{Y^0}(f(x_i)-E\hat{y}_i)(E\hat{y}_i-\hat{y}_i)=2(f(x_i)-E\hat{y}_i)(E\hat{y}_i-\hat{y}_i)\end{aligned}$$

and

$$\begin{aligned}\text{Gi} &= (y_i-f(x_i))^2\\ H_i &= 2(y_i-f(x_i))(f(x_i)-E\hat{y}_i)\\ J_i &= 2(y_i-f(x_i))(E\hat{y}_i-\hat{y}_i).\end{aligned}$$

Then we can rewrite the two type of error.

$$
\begin{aligned}
Err_{in} &= \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}\left(Y_i^0 - f(x_i) + f(x_i) - E\hat{y}_i + E\hat{y}_i - \hat{y}_i\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} A_i + B_i + C_i + D_i + E_i + F_i,
\end{aligned}
$$

$$
\begin{aligned}
\overline{err} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i) + f(x_i) - E\hat{y}_i + E\hat{y}_i - \hat{y}_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} G_i + B_i + C_i + H_i + J_i + F_i,
\end{aligned}
$$

Therefore we get

$$
\begin{aligned}
E_{\mathbf{y}}(\mathrm{op}) &= E_{\mathbf{y}}(\mathrm{Err_{in}} - \overline{\mathrm{err}}) \\
&= \frac{1}{N}\sum_{i=1}^{N} E_{\mathbf{y}}[(A_i - G_i) + (D_i - H_i) + (E_i - J_i)] \\
&= -\frac{2}{N}\sum_{i=1}^{N} J_i \\
&= -\frac{2}{N}\sum_{i=1}^{N} E_{\mathbf{y}}(y_i - f(x_i))(E\hat{y}_i - \hat{y}_i) \\
&= \frac{2}{N}\sum_{i=1}^{N}[E_{\mathbf{y}}(y_i\hat{y}_i) - E_{\mathbf{y}}y_i E_{\mathbf{y}}\hat{y}_i] \\
&= 2\mathrm{Cov}(y_i, \hat{y}_i).
\end{aligned}
$$

Ext 7.5

Ex. 7.5 For a linear smoother $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, show that

$$\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) = \mathrm{trace}(\mathbf{S})\sigma_\varepsilon^2, \tag{7.65}$$

which justifies its use as the effective number of parameters.

This problem is quite same as the Ex7.1, and we use the same method to get this:

$$\begin{aligned}
\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) &= \mathrm{trace}(\mathrm{Cov}(\hat{\mathbf{y}}, \mathbf{y})) \\
&= \mathrm{trace}(\mathrm{Cov}(\mathbf{S}\mathbf{y}, \mathbf{y})) \\
&= \mathrm{trace}(\mathbf{S}\mathrm{Cov}(\mathbf{y}, \mathbf{y})) \\
&= \mathrm{trace}(\mathbf{S}\mathrm{Var}(\mathbf{y})) \\
&= \mathrm{trace}(\mathbf{S})\sigma_\epsilon^2.
\end{aligned}$$