Bibliographic Notes


Exercises

Ex. 7.1 Derive the estimate of in-sample error (7.20).

Ex. 7.2 For 0–1 loss with \( Y \in \{0, 1\} \) and \( \Pr(Y = 1|x_0) = f(x_0) \), show that

\[
\text{Err}(x_0) = \Pr(Y \neq \hat{G}(x_0)|X = x_0) \\
= \text{Err}_B(x_0) + |2f(x_0) - 1|\Pr(\hat{G}(x_0) \neq G(x_0)|X = x_0),
\]

(7.56)

where \( \hat{G}(x) = I(\hat{f}(x) > \frac{1}{2}) \), \( G(x) = I(f(x) > \frac{1}{2}) \) is the Bayes classifier, and \( \text{Err}_B(x_0) = \Pr(Y \neq G(x_0)|X = x_0) \), the irreducible Bayes error at \( x_0 \).

Using the approximation \( \hat{f}(x_0) \sim N(E\hat{f}(x_0), \text{Var}(\hat{f}(x_0)) \), show that

\[
\Pr(\hat{G}(x_0) \neq G(x_0)|X = x_0) \approx \Phi\left(\frac{\text{sign}(\frac{1}{2} - f(x_0))(E\hat{f}(x_0) - \frac{1}{2})}{\sqrt{\text{Var}(\hat{f}(x_0))}}\right).
\]

(7.57)

In the above,

\[
\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-t^2/2)dt,
\]

the cumulative Gaussian distribution function. This is an increasing function, with value 0 at \( t = -\infty \) and value 1 at \( t = +\infty \).
We can think of $\text{sign}(\frac{1}{2} - f(x_0))(E\hat{f}(x_0) - \frac{1}{2})$ as a kind of boundary-bias term, as it depends on the true $f(x_0)$ only through which side of the boundary ($\frac{1}{2}$) that it lies. Notice also that the bias and variance combine in a multiplicative rather than additive fashion. If $E\hat{f}(x_0)$ is on the same side of $\frac{1}{2}$ as $f(x_0)$, then the bias is negative, and decreasing the variance will decrease the misclassification error. On the other hand, if $E\hat{f}(x_0)$ is on the opposite side of $\frac{1}{2}$ to $f(x_0)$, then the bias is positive and it pays to increase the variance! Such an increase will improve the chance that $\hat{f}(x_0)$ falls on the correct side of $\frac{1}{2}$ (Friedman 1997).

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

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

$$y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}.$$  \hspace{1cm} (7.58)

(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 $S$ to make result (7.58) hold.

**Ex. 7.4** Consider the in-sample prediction error (7.15) and the training error $\text{err}$ in the case of squared-error loss:

$$\text{Err}_{\text{in}} = \frac{1}{N} \sum_{i=1}^{N} E_{Y_{\text{new}}}(Y_{i}^{\text{new}} - \hat{f}(x_i))^2$$

$$\text{err} = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2.$$  

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

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

as given in (7.17).

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

$$\sum_{i=1}^{N} \text{Cov}(\hat{y}_i, y_i) = \text{trace}(S)\sigma^2,$$  \hspace{1cm} (7.59)

which justifies its use as the effective number of parameters.
**Ex. 7.6** Use the approximation $1/(1-x)^2 \approx 1+2x$ to expose the relationship between $C_p$/AIC (7.22) and GCV (7.46), the main difference being the model used to estimate the noise variance $\sigma^2$.  

**Ex. 7.7** Show that the set of functions $\{I(\sin(\alpha x) > 0)\}$ can shatter the following points on the line:

$$z^1 = 10^{-1}, \ldots, z^\ell = 10^{-\ell}, \quad (7.60)$$

for any $\ell$. Hence the VC dimension of the class $\{I(\sin(\alpha x) > 0)\}$ is infinite.  

**Ex. 7.8** For the prostate data of Chapter 3, carry out a best-subset linear regression analysis, as in Table 3.3 (third column from left). Compute the AIC, BIC, five- and tenfold cross-validation, and bootstrap.632 estimates of prediction error. Discuss the results.