1 StatLearn Theoretical exercise 2

Let X be a p-dimensional stochastic variable and Y a scalar stochastic variable. Let some mapping $f : \mathbb{R}^p \to \mathbb{R}$ be given, and assume given some other mapping $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. We think of f as a predictor function used for predicting samples of Y from samples of X, and we think of L as a loss function used for assessing the quality of approximation of Y by $f(X)$. We define $\mathrm{EPE}(f) = EL(Y, f(X))$ as the expected prediction error. This expression depends explicitly on f and implicitly on the joint distribution of (X, Y), and cannot be calculated exactly in practice.

Considering observation pairs (x_i, y_i), $i \leq n$, and assuming given an estimator $\hat{f}(x_i)$ of $f(x_i)$ based on (x_i, y_i), we may define $\overline{\mathrm{err}} = \frac{1}{n}\sum_{i=1}^n L(y_i, \hat{f}(x_i))$ as the training error. This expression depends solely on the observation pairs (x_i, y_i), and can be calculated exactly in practice.

Now keep the observations of the independent variables (x_i) fixed, and let Y_i denote independent samples with Y_i having the conditional distribution of Y given $X = x_i$. We then define the stochastic training error by $\overline{\mathrm{err}}_s = \frac{1}{n}\sum_{i=1}^n L(Y_i, \hat{f}_s(x_i))$, where $\hat{f}_s(x_i)$ is the estimator of $f(x_i)$ based on the fixed independent variables (x_i) and the stochastic responses (Y_i). This expression depends on the observations (x_i) and (Y_i), and can be calculated exactly in practice. Furthermore, we may also define $\mathrm{Err_{in}} = \frac{1}{n}\sum_{i=1}^n EL(Y_i', \hat{f}_s(x_i))$ as the expected in-sample error, where (Y_i') are independent samples from the conditional distribution of Y given $X = x_i$, and $\hat{f}_s(x_i)$ remains the estimator of $f(x_i)$ based on (x_i) and (Y_i). This expression depends on the observations (x_i) and on the conditional distribution of Y given X, and cannot be calculated exactly in practice. The expected optimism is then $\mathrm{eop} = \mathrm{Err_{in}} - E\overline{\mathrm{err}}_s$. This expression depends on the observations (x_i) and on the conditional distribution of Y given X, and cannot be calculated exactly in practice.

Note that all of $\overline{\mathrm{err}}$, $\overline{\mathrm{err}}_s$, $\mathrm{Err_{in}}$ and eop depend only on the fitted values $\hat{f}_s(x_i)$ and not on any full estimate of f. This distinction will allow us extra flexibility in the following.

Exercise 1.1. *Show that with L being the squared error loss,* $\mathrm{eop} = \frac{2}{n}\sum_{i=1}^n \mathrm{cov}(\hat{f}_s(x_i), Y_i)$.

Solution. Plugging in the expression for the loss function and expanding terms, we obtain

$$\begin{aligned}
\mathrm{eop} &= \mathrm{Err_{in}} - E\overline{\mathrm{err}}_s \\
&= \frac{1}{n}\sum_{i=1}^n E((Y_i' - \hat{f}_s(x_i))^2 - (Y_i - \hat{f}_s(x_i))^2) \\
&= \frac{1}{n}\sum_{i=1}^n E((Y_i')^2 - 2Y_i'\hat{f}_s(x_i) + \hat{f}_s(x_i)^2 - (Y_i^2 - 2Y_i\hat{f}_s(x_i) + \hat{f}_s(x_i)^2)) \\
&= \frac{1}{n}\sum_{i=1}^n E(Y_i')^2 - 2EY_i'\hat{f}_s(x_i) - EY_i^2 + 2EY_i\hat{f}_s(x_i).
\end{aligned}$$

Now, Y_i' and Y_i has the same distribution, namely the conditional distribution of Y given $X = x_i$, and therefore the moments and second moments are equal. Furthermore, (Y_i') is independent of (Y_i), in particular independent of $\hat{f}_s(x_i)$, and we therefore find

$$\begin{aligned}
\text{eop} &= \frac{1}{n}\sum_{i=1}^n 2EY_i\hat{f}_s(x_i) - 2EY_i'\hat{f}_s(x_i) \\
&= \frac{1}{n}\sum_{i=1}^n 2EY_i\hat{f}_s(x_i) - 2(EY_i')(E\hat{f}_s(x_i)) \\
&= \frac{1}{n}\sum_{i=1}^n 2EY_i\hat{f}_s(x_i) - 2(EY_i)(E\hat{f}_s(x_i)) \\
&= \frac{2}{n}\sum_{i=1}^n \operatorname{cov}(\hat{f}_s(x_i), Y_i),
\end{aligned}$$

as required. □

Now consider a $n \times n$ matrix $\mathbf{S}$, understood to be depending only on the observations (x_i). Let $\hat{\mathbf{f}}$ denote the vector of fitted values, $\hat{\mathbf{f}}_i = \hat{f}_s(x_i)$. Assume that $\hat{\mathbf{f}} = \mathbf{SY}$, where $\mathbf{Y}$ is the n-dimensional vector whose i'th entry is Y_i. Assume that the conditional variance of Y given $X = x$ does not depend on x, and let σ^2 denote the common value of the conditional variance.

Exercise 1.2. *Show that* $\sum_{i=1}^n \operatorname{cov}(\hat{f}_s(x_i), Y_i) = \sigma^2 \operatorname{tr} \mathbf{S}$.

Solution. Using linearity of the covariance, we find

$$\begin{aligned}
\sum_{i=1}^n \operatorname{cov}(\hat{f}_s(x_i), Y_i) &= \sum_{i=1}^n \operatorname{cov}((\mathbf{SY})_i, Y_i) \\
&= \sum_{i=1}^n \operatorname{cov}\left(\sum_{j=1}^n \mathbf{S}_{ij}Y_j, Y_i\right) \\
&= \sum_{i=1}^n\sum_{j=1}^n \mathbf{S}_{ij}\operatorname{cov}(Y_j, Y_i) \\
&= \sum_{i=1}^n \mathbf{S}_{ii}VY_i \\
&= \sigma^2\sum_{i=1}^n \mathbf{S}_{ii} \\
&= \sigma^2 \operatorname{tr} \mathbf{S}.
\end{aligned}$$

□

Exercise 1.3. *Let $\hat{\sigma}^2$ denote an unbiased estimator of σ^2. Define* $\hat{\text{Err}}_{\text{in}} = \overline{\text{err}}_s + \frac{2}{n}(\text{tr}\,\mathbf{S})\hat{\sigma}^2$. *Show that the mean of* $\hat{\text{Err}}_{\text{in}}$ *is* Err_{in}.

Solution. Combining our previous results of Exercise 1.1 and Exercise 1.2, we have

$$\begin{aligned} E\hat{\text{Err}}_{\text{in}} &= E\overline{\text{err}}_s + \frac{2}{n}(\text{tr}\,\mathbf{S})E\hat{\sigma}^2 \\ &= \text{Err}_{\text{in}} - \text{eop} + \frac{2}{n}(\text{tr}\,\mathbf{S})\sigma^2 \\ &= \text{Err}_{\text{in}} - \frac{2}{n}\sum_{i=1}^{n}\text{cov}(\hat{f}_s(x_i), Y_i) + \frac{2}{n}(\text{tr}\,\mathbf{S})\sigma^2 \\ &= \text{Err}_{\text{in}}. \end{aligned}$$

□

Exercise 1.3 shows that for prediction methods where the fitted values are a linear function of the responses given the design matrix, we have a simple estimator for the in-sample error, which is often of interest to us.

Next, we define the generalization error as $\text{Err} = EL(Y', \hat{f}(X'))$, where $\hat{f}$ is the predictor function estimate based on (X_i) and (Y_i) and take interest in estimating Err. To this end, we assume that the diagonal of $\mathbf{S}$ does not contain any ones, and define

$$\hat{f}^{-i}(x_i) = \sum_{j\neq i}\frac{\mathbf{S}_{ij}}{1-\mathbf{S}_{ii}}y_j,$$

and think of $\hat{f}^{-i}(x_i)$ as the fitted value at x_i for the data set excluding x_i. Note that $\hat{f}^{-i}(x_i)$ is this the predicted value of the response for a point outside of the data set, and thus requires an prediction methodology for obtaining predicted values outside of our observed independent variables, for example through a full estimate of f. We then define the leave-one-out cross-validation estimator of Err as

$$\hat{\text{Err}} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}^{-i}(x_i))$$

Exercise 1.4. *Show that* $y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1-\mathbf{S}_{ii}}$.

Solution. This follows as

$$\begin{aligned}
y_i - \hat{f}^{-i}(x_i) &= y_i - \sum_{j \neq i} \frac{\mathbf{S}_{ij}}{1 - \mathbf{S}_{ii}} y_j \\
&= \frac{1}{1 - \mathbf{S}_{ii}} \left(y_i(1 - \mathbf{S}_{ii}) - \sum_{j \neq i} \mathbf{S}_{ij} y_j \right) \\
&= \frac{1}{1 - \mathbf{S}_{ii}} \left(y_i - \sum_{j=1}^{n} \mathbf{S}_{ij} y_j \right) \\
&= \frac{y_i - \hat{f}(x_i)}{1 - \mathbf{S}_{ii}}.
\end{aligned}$$

□

Exercise 1.5. *Explain why the above result may be used to compute* $\hat{\mathrm{Err}}$ *with squared error loss efficiently using the diagonal elements of* $\mathbf{S}$*.*

Solution. We find

$$\begin{aligned}
\hat{\mathrm{Err}} &= \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}^{-i}(x_i)) \\
&= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}^{-i}(x_i))^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}(x_i)}{1 - \mathbf{S}_{ii}} \right)^2,
\end{aligned}$$

which allows fast computation given $\hat{f}$ and the diagonal of $\mathbf{S}$, removing the need to calculate each $\hat{f}^{-i}(x_i)$ separately. □

Exercise 1.6. *Show that the assumption* $\hat{f}^{-i}(x_i) = (1 - \mathbf{S}_{ii})^{-1} \sum_{j \neq i} \mathbf{S}_{ij} Y_j$ *is equivalent to assuming that the fit at* x_i*,* $\hat{f}_s(x_i)$*, based on the reduced data set excluding the i'th observation pair, is the same as the fit at* x_i *based the data set exchanging the i'th observation pair with* $(x_i, \hat{f}_s^{-i}(x_i))$*.*

Solution. The property that the fit at x_i, $\hat{f}_s(x_i)$, based on the reduced data set excluding the i'th observation pair is the same as the fit at x_i based the data set exchanging the i'th observation pair with $(x_i, \hat{f}_s^{-i}(x_i))$ may, be formalized as:

$$(\mathbf{S}(\mathbf{Y} - (Y_i - \hat{f}^{-i}(x_i))e_i))_i = \hat{f}^{-i}(x_i),$$

where e_i denotes the unit vector in the i'th direction. The left-hand side is

$$\begin{aligned}
(\mathbf{S}(\mathbf{x})(\mathbf{Y} - (Y_i - \hat{f}^{-i}(x_i))e_i))_i &= \sum_{j=1}^{n} \mathbf{S}_{ij}(\mathbf{Y} - (Y_i - \hat{f}^{-i}(x_i))e_i)_i \\
&= \sum_{j=1}^{n} \mathbf{S}_{ij}\mathbf{Y}_j - (Y_i - \hat{f}_s^{-i}(x_i))\mathbf{S}_{ii} \\
&= \mathbf{S}_{ii} f_s^{-i}(x_i) + \sum_{j \neq i} \mathbf{S}_{ij}\mathbf{Y}_j,
\end{aligned}$$

and so the requirement is that $(1 - \mathbf{S}_{ii})\hat{f}_s^{-i}(x_i) = \sum_{j\neq i} \mathbf{S}_{ij}\mathbf{Y}_j$, which yields the result. $\square$

Exercise 1.7. *Show that least squares regression and ridge regression linear smoothers satisfying the regularity criterion on leave-one-out estimates. Show that the k-nearest neighbor method satisfies the regularity criterion if the leave-one-out estimates are based on the $(k-1)$-nearest neighbor method.*

Solution. For the least squares regression, we are given a response $y \in \mathbb{R}^n$ and a design matrix $\mathbf{X}$ of full column rank, and the estimate of the prediction function is then obtained as $\hat{f}(x) = x^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y$. The estimate of the prediction function for the reduced data set is then $\hat{f}^{-i}(x) = x^t((\mathbf{X}_{-i})^t\mathbf{X}_{-i})^{-1}(\mathbf{X}_{-i})^t y_{-i}$, where $\mathbf{X}_{-i}$ is the $(n-1)\times p$ matrix obtained by removing the i'th row of $\mathbf{X}$, and y_{-i} is the $(n-1)$-dimensional vector obtained by removing the i'th entry of y. We have

$$\begin{aligned}
& \|y - (y_i - \hat{f}^{-1}(x_i))e_i - \mathbf{X}\beta\|_2^2 \\
= \; & \|y_{-i} - \mathbf{X}_{-i}\beta\|_2^2 + (y_i - (y_i - \hat{f}^{-1}(x_i)) - x_i^t\beta)^2 \\
= \; & \|y_{-i} - \mathbf{X}_{-i}\beta\|_2^2 + (\hat{f}^{-1}(x_i) - x_i^t\beta)^2.
\end{aligned}$$

The first term is minimized for $\beta_{-i} = (\mathbf{X}_{-i})^t\mathbf{X}_{-i})^{-1}(\mathbf{X}_{-i})^t y_{-i}$, and this argument incidentally also minimizes the second term, yielding the value zero. Therefore, we conclude

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|y - (y_i - \hat{f}^{-1}(x_i))e_i - \mathbf{X}\beta\|_2^2 = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|y_{-i} - \mathbf{X}_{-i}\beta\|_2^2.$$

As both these argument minima are solutions to ordinary least squares problems, we conclude

$$(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(y - (y_i - \hat{f}^{-1}(x_i))e_i) = ((\mathbf{X}_{-i})^t\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^t y_{-i},$$

and in particular the fitted values, obtained by multiplying by x_i^t on the right, coincide. The left-hand side becomes the fitted value at x_i for the full data set with y_i exchanged by $f^{-i}(x_i)$, and the right-hand side becomes the fitted value at x_i for the reduced dataset. This proves the result in the least squares regression case.

Next, we consider the ridge regression case. Let $\lambda \geq 0$ be given. Here, the estimate of the prediction function is $\hat{f}(x) = x^t(\mathbf{X}^t\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t y$, and the corresponding estimate of the prediction function for the reduced data set is $\hat{f}^{-i}(x) = x^t((\mathbf{X}_{-i})^t\mathbf{X}_{-i} + \lambda I_p)^{-1}(\mathbf{X}_{-i})^t y_{-i}$. As before, we find

$$\begin{aligned} & \|y - (y_i - \hat{f}^{-1}(x_i))e_i - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 \\ = \; & \|y_{-i} - \mathbf{X}_{-i}\beta\|_2^2 + (y_i - (y_i - \hat{f}^{-1}(x_i)) - x_i^t\beta)^2 + \lambda\|\beta\|_2^2 \\ = \; & \|y_{-i} - \mathbf{X}_{-i}\beta\|_2^2 + \lambda\|\beta\|_2^2 + (\hat{f}^{-1}(x_i) - x_i^t\beta)^2. \end{aligned}$$

The first two terms are minimized for $\beta_{-i} = (\mathbf{X}_{-i})^t\mathbf{X}_{-i} + \lambda I_p)^{-1}(\mathbf{X}_{-i})^t y_{-i}$, and as in the least squares case, this argument also minimizes the second term, and so we obtain

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|y - (y_i - \hat{f}^{-1}(x_i))e_i - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|y_{-i} - \mathbf{X}_{-i}\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

and therefore

$$(\mathbf{X}^t\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t(y - (y_i - \hat{f}^{-1}(x_i))e_i) = ((\mathbf{X}_{-i})^t\mathbf{X}_{-i} + \lambda I_p)^{-1}\mathbf{X}_{-i}^t y_{-i}.$$

As in the least squares case, this shows that the fitted value at x_i for the full data set with y_i exchanged by $f^{-i}(x_i)$ and the fitted value at x_i for the reduced data set match, as desired.

Finally, we consider the k-nearest neighbor method. As before, we are given an n-dimensional response vector y and an $n \times p$ design matrix $\mathbf{X}$, and the estimate of the prediction function is $\hat{f}(x) = \frac{1}{k}\sum_{j=1}^n 1_{(x_j\in N(x))}y_j$, where $N(x)$ is a neighborhood of x containing k points. Note that in contrast to the case for least squares and ridge regression, $\hat{f}$ is not linear, although the fitted values $\hat{\mathbf{f}}$ is a linear function of y. We base the estimate of the prediction function for the reduced data set on the $(k-1)$-nearest neighbor method, and put $\hat{f}^{-i}(x) = \frac{1}{k-1}\sum_{j\neq i} 1_{(x_j\in N^{-i}(x))}y_j$, where $N^{-i}(x)$ are the neighborhoods for the reduced data set containing $k-1$ points each.

We wish to calculate the fitted value at x_i for the full data set with y_i exchanged by $f^{-i}(x_i)$. As the reduced data set does not contain x_i, we have $N^{-i}(x_i) = N(x_i) \setminus \{x_i\}$, and so

$$\begin{aligned} & \frac{1}{k}\sum_{j\neq i} 1_{(x_j\in N(x_i))}y_j + \frac{1}{k}1_{(x_i\in N(x_i))}f^{-i}(x_i) \\ = \; & \frac{1}{k}\sum_{j\neq i} 1_{(x_j\in N(x_i))}y_j + \frac{1}{k}\left(\frac{1}{k-1}\sum_{j\neq i} 1_{(x_j\in N^{-i}(x_i))}y_j\right) \\ = \; & \frac{1}{k}\sum_{j\neq i} y_j\left(1_{(x_j\in N^{-i}(x_i))} + \frac{1}{k-1}1_{(x_j\in N^{-i}(x_i))}\right) \\ = \; & \frac{1}{k}\left(1 + \frac{1}{k-1}\right)\sum_{j\neq i} 1_{(x_j\in N^{-i}(x_i))}y_j \\ = \; & \frac{1}{k-1}\sum_{j\neq i} 1_{(x_j\in N^{-i}(x_i))}y_j, \end{aligned}$$

which is $f^{-i}(x_i)$. $\square$

We now define

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{f}(x_i)}{1 - \frac{1}{n}\mathrm{tr}\,\mathbf{S}}\right)^2,$$

and call GCV the generalized cross-validation estimator.

Exercise 1.8. *Show that*

$$\left(\frac{Y_i - \hat{f}(x_i)}{1 - \frac{1}{n}\mathrm{tr}\,\mathbf{S}}\right)^2 \approx (Y_i - \hat{f}(x_i))^2(1 + \tfrac{2}{n}\mathrm{tr}\,\mathbf{S}),$$

and use this to obtain an approximate relation between GCV *and* $\hat{\mathrm{Err}}_{\mathrm{in}}$.

Solution. Put $f(x) = (1-x)^{-2}$, we then have $f'(x) = 2(1-x)^{-3}$, so that in particular, $f(1) = 1$ and $f'(1) = 2$, yielding the first-order Taylor approximation $(1-x)^{-2} \approx 1 + 2x$, and so

$$\left(\frac{Y_i - \hat{f}(x_i)}{1 - \frac{1}{n}\mathrm{tr}\,\mathbf{S}}\right)^2 = (Y_i - \hat{f}(x_i))^2(1 - \tfrac{1}{n}\mathrm{tr}\,\mathbf{S})^{-2} \approx (Y_i - \hat{f}(x_i))^2(1 + \tfrac{2}{n}\mathrm{tr}\,\mathbf{S}).$$

The estimate of the in-sample error considered previously was $\hat{\mathrm{Err}}_{\mathrm{in}} = \overline{\mathrm{err}}_s + \frac{2}{n}(\mathrm{tr}\,\mathbf{S})\hat{\sigma}^2$. We therefore obtain

$$\begin{aligned}
\mathrm{GCV} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{f}(x_i)}{1 - \frac{1}{n}\mathrm{tr}\,\mathbf{S}}\right)^2 \\
&\approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2(1 + \tfrac{2}{n}\mathrm{tr}\,\mathbf{S}) \\
&= (1 + \tfrac{2}{n}\mathrm{tr}\,\mathbf{S})\overline{\mathrm{err}}_s \\
&= \left(1 + \frac{\hat{\mathrm{Err}}_{\mathrm{in}} - \overline{\mathrm{err}}_s}{\hat{\sigma}^2}\right)\overline{\mathrm{err}}_s
\end{aligned}$$

If we further assume that we estimate $\hat{\sigma}^2$ with the stochastic training error, that is, $\hat{\sigma}^2 = \overline{\mathrm{err}}_s$, we find $\mathrm{GCV} = \hat{\mathrm{Err}}_{\mathrm{in}}$. $\square$