Statistical Machine Learning

Exercise sheet 5

Exercise 5.1 (Leave-one-out cross-validation for linear smoothers) In this exercise we consider linear smoothers, i.e., models $\widehat{f}$ for which the fitted values verify $\widehat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y}$, where $\mathbf{S}$ is an $n \times n$ matrix whose values only depend on the inputs $\boldsymbol{x}_i, \dots, x_n$. We have already encountered several linear smoothers (can you cite some?). The goal of this exercise is to derive a fast way of computing the leave-one-out (or n-fold) cross-validation (CV) error for linear smoothers under a *regularity assumption.* The leave-one-out CV error is

$$\mathrm{CV}(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\}^2,$$

where $\widehat{f}^{-i}$ denote the model fitted to the original training sample with the ith observation $(y_i, \boldsymbol{x}_i)$ removed.

(a) Assume that the leave-ith-out fit at $\boldsymbol{x}_i$ is given by

$$\widehat{f}^{-i}(\boldsymbol{x}_i) = \sum_{j \neq i} \frac{\mathbf{S}_{ij}}{1 - \mathbf{S}_{ii}} y_j. \tag{1}$$

With this regularity assumption, show that

$$y_i - \widehat{f}^{-i}(\boldsymbol{x}_i) = \frac{y - \widehat{f}(\boldsymbol{x}_i)}{1 - \mathbf{S}_{ii}}. \tag{2}$$

(b) Explain why (2) may be used to compute the CV error more efficiently.

(c) Show that the regularity assumption given in (1) is equivalent to assuming that the fit at $\boldsymbol{x}_i$, based on the *reduced data set* that excludes the ith observation pair, is the same as the fit at $\boldsymbol{x}_i$, based on the *adapted data set* that replaces the ith observation pair with $(\boldsymbol{x}_i, \widehat{f}^{-i}(\boldsymbol{x}_i))$, where $\widehat{f}^{-i}(\boldsymbol{x}_i)$ is the fit at $\boldsymbol{x}_i$ based on the reduced data set.

Hint: Start by arguing that the statement above can be formalized as

$$\Big(\mathbf{S}[\boldsymbol{y} - \{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\}\boldsymbol{e}_i]\Big)_i = \widehat{f}^{-i}(\boldsymbol{x}_i),$$

where $\boldsymbol{e}_i$ is the unit vector in the ith direction.

(d) Argue that the least squares and ridge regression estimators satisfy the regularity assumption given in (c) and hence (a).

Exercise 5.2 (Generalized cross-validation and C_p) Define the generalized cross-validation estimator (GCV) as

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1 - \mathrm{tr}(\mathbf{S})/n}\right\}^2.$$

Show that

$$\left\{\frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1 - \mathrm{tr}(\mathbf{S})/n}\right\}^2 \approx \{y_i - \widehat{f}(\boldsymbol{x}_i)\}^2\{1 + 2\mathrm{tr}(\mathbf{S})/n\},$$

and use this to obtain an approximate relation between GCV and C_p.

Exercise 5.3 (Practical: Cross-validation) This exercise continues on from Exercise 4.2.

(a) Write your own code to perform K-fold cross-validation and estimate the standard error of the cross-validation error estimate.

(b) Use your code and the one-standard-error rule to choose the optimal value of λ for the ridge and lasso estimators on the bodyfat dataset.