Efficient Computation of K-fold Cross-validation Error for Linear Models

Shi Binbin

September 10, 2017

For ridge regression, the coefficients are found by minimizing the squared-error and L_2 regularization term:

$$\text{minimize}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2 \tag{1}$$

The solution to ridge regression is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

The estimate of $\mathbf{y}$ is:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{3}$$

We can also write $\hat{\mathbf{y}}$ as:

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} \tag{4}$$

where $\mathbf{S}$ is a smoother matrix of $\mathbf{y}$: $\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$.

The leave-one-out cross-validation error of linear regression can be computed efficiently:

$$\text{LOOCV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{y}^{-i}(x_i)]^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}\right]^2 \tag{5}$$

where S_{ii} is the ith diagonal element of $\mathbf{S}$.

The GCV approximation of the leave-one-out cross-validation is:

$$\text{GCV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - \text{trace}(\mathbf{S})/N}\right]^2 \tag{6}$$

where $\text{trace}(\mathbf{S})$ is the effective number of parameters.

For k-fold cross-validation, the cross-validation error is:

$$\text{CV}(\hat{f}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k}[y_{ki} - \hat{f}^{-k}(\mathbf{x}_{ki})]^2 \tag{7}$$

where N_k is the number of test samples in the kth part of the dataset. $(\mathbf{x}_{ki}, y_{ki})$ is the ith sample in the kth part of the dataset. $\hat{f}^{-k}(\mathbf{x}_{ki}$ is the fitted function on the dataset with the kth part removed.

The smoother matrix of the training samples is:

$$\mathbf{S}_k = \mathbf{X}_k \mathbf{A}^{-1} \mathbf{X}_k^T \tag{8}$$

where $\mathbf{A} = \mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}$

The estimate of kth part of the test samples by the function fitted on the full dataset is:

$$\hat{\mathbf{f}}(\mathbf{X}_k) = \mathbf{X}_k \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y} \tag{9}$$

Denote the fitted function with the kth part removed by $\hat{\mathbf{f}}^{-k}(\mathbf{X}_k)$.

$$\begin{align} \hat{\mathbf{f}}^{-k}(\mathbf{X}_k) &= \mathbf{X}_k (\mathbf{X}^T \mathbf{X} - \mathbf{X}_k^T \mathbf{X}_k + \lambda \mathbf{I})^{-1} (\mathbf{X}^T \mathbf{y} - \mathbf{X}_k^T \mathbf{y}_k) \tag{10} \\ &= \mathbf{X}_k (\mathbf{A} - \mathbf{X}_k^T \mathbf{X}_k)^{-1} (\mathbf{X}^T \mathbf{y} - \mathbf{X}_k^T \mathbf{y}_k) \tag{11} \end{align}$$

Following the properties of inverse of a block matrix:

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{12}$$

we can separate $\mathbf{X}_k$ from $(\mathbf{A} - \mathbf{X}_k^T \mathbf{X}_k)^{-1}$:

$$\begin{align} (\mathbf{A} - \mathbf{X}_k^T \mathbf{X}_k)^{-1} &= \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_k(\mathbf{I} - \mathbf{X}_k \mathbf{A} \mathbf{X}_k^T)^{-1} \mathbf{X}_k^T \mathbf{A}^{-1} \tag{13} \\ &= \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_k(\mathbf{I} - \mathbf{S}_k)^{-1} \mathbf{X}_k^T \mathbf{A}^{-1} \tag{14} \end{align}$$

Plugging in $(\mathbf{A} - \mathbf{X}_k^T \mathbf{X}_k)^{-1}$ into the calculation of $\hat{\mathbf{f}}^{-k}(\mathbf{X}_k)$:

$$\begin{aligned} \hat{\mathbf{f}}^{-k}(\mathbf{X}_k) &= \mathbf{X}_k[\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_k(\mathbf{I} - \mathbf{S}_k)^{-1}\mathbf{X}_k^T \mathbf{A}^{-1}](\mathbf{X}^T \mathbf{y} - \mathbf{X}_k^T \mathbf{y}_k) \\ &= \mathbf{X}_k \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y} \\ &\quad + \mathbf{X}_k \mathbf{A}^{-1} \mathbf{X}_k^T \mathbf{y}_k \\ &\quad + \mathbf{X}_k \mathbf{A}^{-1} \mathbf{X}_k (\mathbf{I} - \mathbf{S}_k)^{-1} \mathbf{X}_k^T \mathbf{A}^{-1} \mathbf{X}^T \mathbf{y} \\ &\quad + \mathbf{X}_k \mathbf{A}^{-1} \mathbf{X}_k (\mathbf{I} - \mathbf{S}_k)^{-1} \mathbf{X}_k^T \mathbf{A}^{-1} \mathbf{X}_k^T \mathbf{y}_k \\ &= \hat{\mathbf{f}}(\mathbf{X}_k) + \mathbf{S}_k \mathbf{y}_k + \mathbf{S}_k (\mathbf{I} - \mathbf{S}_k)^{-1} \hat{\mathbf{f}}(\mathbf{X}_k) + \mathbf{S}_k (\mathbf{I} - \mathbf{S}_k)^{-1} \mathbf{S}_k \\ &= [\mathbf{I} + \mathbf{S}_k (\mathbf{I} - \mathbf{S}_k)^{-1}][\hat{\mathbf{f}}(\mathbf{X}_k) - \mathbf{y}_k] + \mathbf{y}_k \end{aligned}$$

Then the cross-validated residual on the test samples can be written as:

$$\mathbf{y}_k - \hat{\mathbf{f}}^{-k}(\mathbf{X}_k) = [\mathbf{I} + \mathbf{S}_k (\mathbf{I} - \mathbf{S}_k)^{-1}][\mathbf{y}_k - \hat{\mathbf{f}}(\mathbf{X}_k)] \tag{15}$$

The cross-validated squared error is:

$$\frac{1}{N_k} ||\mathbf{y}_k - \hat{\mathbf{f}}^{-k}(\mathbf{X}_k)||_2^2 = [\mathbf{y}_k - \hat{\mathbf{f}}(\mathbf{X}_k)]^T \mathbf{B}_k^T \mathbf{B}_k [\mathbf{y}_k - \hat{\mathbf{f}}(\mathbf{X}_k)] \tag{16}$$

where $\mathbf{B}_k = \mathbf{I} + \mathbf{S}_k(\mathbf{I} - \mathbf{S}_k)^{-1}$.

$\mathbf{S}_k$ can be approximated by only considering the diagonal elements. Then the cross-validation error on the *k*th part can be approximated:

$$\frac{1}{N_k}||\mathbf{y}_k - \hat{\mathbf{f}}^{-k}(\mathbf{X}_k)||_2^2 \approx \frac{1}{N_k}\sum_{i=1}^{N_k}\left[\frac{y_{ki} - \hat{f}(\mathbf{x}_{ki})}{1 - S_{ki}}\right]^2 \tag{17}$$

where S_{ki} is the *i*th diagonal element of $\mathbf{S}_k$.