Leave-one-out cross validation for linear models

ETC3250 Sunday, 14 August 2016

The leave-one-out cross-validation (LOOCV) statistic is given by

$$\text{CV} = \frac{1}{N}\sum_{i=1}^{N} e_{[i]}^2,$$

where $e_{[i]} = y_i - \hat{y}_{[i]}$, $y_1, \dots, y_N$ are the observations, and $\hat{y}_{[i]}$ is the predicted value obtained when the model is estimated with the ith case deleted. It turns out that for linear models, we do not actually have to estimate the model N times, once for each omitted case. Instead, CV can be computed after estimating the model once on the complete data set.

Suppose we have a linear regression $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the "hat-matrix". It has this name because it is used to compute $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$. If the diagonal values of $\boldsymbol{H}$ are denoted by $h_1, \dots, h_N$, then the leave-one-out cross-validation statistic can be computed using

$$\text{CV} = \frac{1}{N}\sum_{i=1}^{N}[e_i/(1-h_i)]^2,$$

where $e_i = y_i - \hat{y}_i$ and $\hat{y}_i$ is the predicted value obtained when the model is estimated with all data included.

Proof[1]

Let $\boldsymbol{X}_{[i]}$ and $\boldsymbol{Y}_{[i]}$ be similar to $\boldsymbol{X}$ and $\boldsymbol{Y}$ but with the ith row deleted in each case. Let $\boldsymbol{x}_i'$ be the ith row of $\boldsymbol{X}$ and let

$$\hat{\boldsymbol{\beta}}_{[i]} = (\boldsymbol{X}_{[i]}'\boldsymbol{X}_{[i]})^{-1}\boldsymbol{X}_{[i]}'\boldsymbol{Y}_{[i]}$$

be the estimate of $\boldsymbol{\beta}$ without the ith case. Then $e_{[i]} = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{[i]}$.

Now $\boldsymbol{X}_{[i]}'\boldsymbol{X}_{[i]} = (\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i')$ and $\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i = h_i$. So by the Sherman–Morrison–Woodbury formula[2],

$$(\boldsymbol{X}_{[i]}'\boldsymbol{X}_{[i]})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-h_i}.$$

Also note that $\boldsymbol{X}_{[i]}'\boldsymbol{Y}_{[i]} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{x}y_i$. Therefore

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{[i]} &= \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}}{1-h_i}\right](\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{x}_i y_i) \\
&= \hat{\boldsymbol{\beta}} - \left[\frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i}{1-h_i}\right]\left[y_i(1-h_i) - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + h_i y_i\right] \\
&= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i/(1-h_i)
\end{aligned}$$

Thus

$$\begin{aligned}
e_{[i]} &= y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{[i]} \\
&= y_i - \boldsymbol{x}_i'\left[\hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i/(1-h_i)\right] \\
&= e_i + h_i e_i/(1-h_i) \\
&= e_i/(1-h_i),
\end{aligned}$$

and the result follows.

References

Seber, G. A. F. and A. J. Lee (2003). *Linear Regression Analysis*. 2nd. John Wiley & Sons.

[1] Credit to Rob J Hyndman (adapted from Seber and Lee, 2003)

[2] `https://en.wikipedia.org/wiki/Sherman-Morrison_formula`