Chapter 1. Linear regression Model part 2

January 10, 2007

1 Other Cross-validation methods

1.1 the inconsistency of delete-one-out CV

One important issue is that whether our selection is correct as $n \to \infty$. (This is the so called asymptotic theory). Consistency means if n tends to infinity, our statistical inference should tend to be correct. The delete-one-out CV tends to select a larger number of predictors than the true model theoretically even with very large n. It is also said that delete-one-out CV is not consistent in linear regression model selection

1.2 delete-m-out CV

The reason that delete-one-out CV is not consistent is that the validation set is too small (we have only one observation in the validation set)

Each time we split the whole data set $\{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)\}$ into two parts training set and validation set. The latter contains m observations, denoted by V_i; the former contains $n-m$ observations, dented by T_i. We estimate a model based on T_i, denote the estimators by $\hat{\beta}^i$; and calculate the prediction error for data set V_i, denoted the prediction errors by err_i:

$$err_i = \sum_{(X_j, Y_j) \in V_i} \{Y_j - X_j^\top \hat{\beta}^i\}^2$$

Note that there are C_n^m possible splitting cases. The delete-m-out CV is defined as

$$CV_m = (C_n^m)^{-1} \sum_{i=1}^{C_n^m} err_i$$

It is proved that if $m \to \infty$ as $n \to \infty$, then the delete-m-out CV is consistent, i.e., under some conditions, we can select correct model when sample size tends to infinity .

1.3 M-fold CV

Partition the whole data into M sets with roughly equal sizes. Use $M-1$ sets to estimate the model, and the remainder to validate the model, denote the prediction error by err_i. swap the validation sets and repeat the same procedure. Then the m-fold CV is defined as

$$MCV = \frac{1}{M}\sum_{i=1}^{M} err_i$$

1.4 Generalized CV

The calculation of delete-one-out or delete-m-out CV is still not easy. Consider the fitted value

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

where $\hat{\beta}$ is based on the whole data set. Let

$$S_n = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

Then, one can prove that the delete-one-observation is

$$CV = n^{-1}\sum_{i=1}^{n}\frac{(Y_i - X_i^\top\hat{\beta})^2}{(1-S_n(i,i))^2}$$

Based on this, Craven and Wahba (1979) proposed to consider the so called generalized cross-validation

$$GCV = \frac{n^{-1}\sum_{i=1}^{n}(Y_i - X_i^\top\hat{\beta})^2}{\{1-tr(S_n)/n\}^2}$$

Example 1.1 *For the same data above* **(data)** *Our candidate models are*

$$\begin{aligned}
&\textit{model 0} \quad Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 + \varepsilon\\
&\textit{model 1} \quad Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \varepsilon\\
&\textit{model 2} \quad Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_5\mathbf{x}_5 + \varepsilon\\
&\textit{model 3} \quad Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 + \varepsilon\\
&\textit{model 4} \quad Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 + \varepsilon\\
&\textit{model 5} \quad Y = \beta_0 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 + \varepsilon
\end{aligned}$$

The GCV values for the above model are respectively

$$GCV0 = 0.2646481, GCV1 = 0.25096, GCV2 = 0.9335248,$$

$$GCV3 = 0.2879277, GCV4 = 1.211334, GCV5 = 0.3392096$$

Thus model 1 is selected (and variable $\mathbf{x}_5$ *is deleted)*

R code for the calculation **(code)**

1.5 Other model selection criterion

Akaike's Information Criterion (AIC): suppose a linear regression model has p covariates (predictors), then its AIC

$$AIC = n \log(\hat{\sigma}^2) + 2p$$

where $\hat{\sigma}^2 = RSS/(n - p - 1)$. In a set of model candidates, AIC prefers the model with smallest AIC.

Schwarz's Bayesian Information Criterion (BIC)

$$BIC = n \log(\hat{\sigma}^2) + p \log n$$

In a set of model candidates, AIC prefers the model with smallest AIC.

Example 1.2 *For the same data and model candidates. The AIC values for the above model are respectively*

$$AIC_0 = -16.58708, AIC_1 = -19.64923, AIC_2 = 6.624244,$$

$$AIC_3 = -16.90092, AIC_4 = 11.83445, AIC_5 = -13.62274.$$

thus, model 1 is selected by AIC (and variable $\mathbf{x}_5$ *is deleted)*

$$BIC_0 = -11.60842, BIC_1 = -15.66630, BIC_2 = 10.60717,$$

$$BIC_3 = -12.91799, BIC_4 = 15.81738, BIC_5 = -9.63981.$$

thus, model 1 is selected by BIC (and variable $\mathbf{x}_5$ *is deleted)*

R code for the calculation **(code)**

2 Some theories of the distributions for LSE of linear regression model

Lemma 2.1 *If $\xi \sim N(b, \Sigma)$. Then for any constant matrix A,*

$$A\xi + c \sim N(Ab + c, A\Sigma A^\top).$$

Recall that our estimator of $\beta = (\beta_0, \beta_1, ..., \beta_p)^\top$ in the model

$$Y = \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p + \varepsilon$$

is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \mathcal{E}) = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E}$$

Therefore,

$$\begin{aligned} \hat{\beta} \sim N(\beta, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 I\{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\}^\top\} &= N(\beta, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2\} \\ &= N(\beta, (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2). \end{aligned}$$

In other words

$$\hat{\beta} - \beta \sim N(\beta, (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2).$$

If $E\mathcal{E} = 0, Var(\mathcal{E}) = \sigma^2 I$, but does not follow normal distribution, then, UNDER SOME CONDITIONS, the distribution of

$$\hat{\beta} - \beta$$

can still be approximated by $N(\beta, (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2)$ if n is large enough.