MONASH University

ETC3250

Business Analytics

Week 6
Model assessment and selection

10 April 2018

Outline

Week	Topic	Chapter	Lecturer
1	Introduction	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression	3	Souhaib
4	Classification	4	Souhaib
5	Clustering	10	Souhaib
	Semester break		
6	Model selection and resampling methods	5	Souhaib
7	Dimension reduction	6,10	Souhaib
8	Advanced regression	6	Souhaib
9	Advanced regression	6	Souhaib
10	Advanced classification	9	Souhaib
11	Tree-based methods	8	Souhaib
12	Project presentation		Souhaib

The model selection problem

- Which predictors should we choose?
- How do we choose the df for a spline?
- How do we choose the number of neighbours K in KNN?
- We need a way of **comparing** multiple competing models
- If there are a limited number of predictors, we can study **all possible models**. Otherwise we need a **search strategy** to explore some potential models.

The model selection procedure

1. **Model generation**

   Generate a set of candidate model structures among which the best one is to be selected.

   If applicable, estimate the parameters of each candidate model (*parametric identification*).

2. **Model assessment/validation**

   Evaluate the model's performance by computing the validation error.

3. **Model selection**

   Select the final model structure in the set that has been proposed by model generation and assessed by model validation. We typically select the model structure that minimizes the validation error.

Model selection procedure $\equiv$ *Structural identification* procedure

Training vs Test MSEs

Suppose we want to compute a regression function $\hat{f}$ from some **training data**, $Tr = \{x_i, y_i\}_1^n$.

Training Mean Squared Error

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}(x_i)]^2$$

We compute test MSE using **test data** $Te = \{\tilde{x}_i, \tilde{y}_i\}_1^m$

Test Mean Squared Error

$$\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{f}(\tilde{x}_i)]^2$$

Training vs Test MSEs

Let $\hat{\beta}$ denote the OLS estimate using the training data, $Tr = \{x_i, y_i\}_1^n$:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

then

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}' x_i]^2\right] \leq \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}' \tilde{x}_i]^2\right].$$

Training vs Test MSEs

1. $\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right]$

2. $A = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}'x_i]^2$ and $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2$ where $\tilde{\beta}'$ is the OLS estimate using the test data, $\{\tilde{x}_i, \tilde{y}_i\}_1^n$.
   - A and B have the same distribution, so $\mathrm{E}[A] = \mathrm{E}[B]$.

3. $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2 \leq \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2$ by OLS property.

4. $\mathrm{E}[A] = \mathrm{E}[B] \leq \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2]$

$$\Longrightarrow \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}'x_i]^2\right] \leq \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right].$$

Training vs Test MSEs

1. $\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right]$

2. $A = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}' x_i]^2$ and $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2$ where $\tilde{\beta}'$ is the OLS estimate using the test data, $\{\tilde{x}_i, \tilde{y}_i\}_1^n$.

   - A and B have the same distribution, so $\mathrm{E}[A] = \mathrm{E}[B]$.

3. $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2 \leq \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2$ by OLS property.

4. $\mathrm{E}[A] = \mathrm{E}[B] \leq \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2]$

$$\implies \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}' x_i]^2\right] \leq \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right].$$

Training vs Test MSEs

1. $\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right]$
2. $A = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}' x_i]^2$ and $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2$ where $\tilde{\beta}'$ is the OLS estimate using the test data, $\{\tilde{x}_i, \tilde{y}_i\}_1^n$.
   - A and B have the same distribution, so $\mathrm{E}[A] = \mathrm{E}[B]$.
3. $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2 \leq \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2$ by OLS property.
4. $\mathrm{E}[A] = \mathrm{E}[B] \leq \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2]$

$$\Longrightarrow \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}' x_i]^2\right] \leq \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right].$$

Training vs Test MSEs

1. $\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right]$
2. $A = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}'x_i]^2$ and $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2$ where $\tilde{\beta}'$ is the OLS estimate using the test data, $\{\tilde{x}_i, \tilde{y}_i\}_1^n$.
   - A and B have the same distribution, so $\mathrm{E}[A] = \mathrm{E}[B]$.
3. $B = \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \tilde{\beta}'\tilde{x}_i]^2 \leq \frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2$ by OLS property.
4. $\mathrm{E}[A] = \mathrm{E}[B] \leq \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2]$

$$\Longrightarrow \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\beta}'x_i]^2\right] \leq \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}[\tilde{y}_i - \hat{\beta}'\tilde{x}_i]^2\right].$$

More on Training vs Test MSEs

We want our model to have **small** expected test error for a **new random point** (x_0, y_0) where both x_0 and y_0 are random.

An easier task would be to produce predictions at the **same** values of the predictor variables as before, but with **different** noises. That is we fit our model using samples from

$$y = x\beta + \varepsilon,$$

and predict for samples from

$$y^{'} = x\beta + \varepsilon^{'}$$

where ε and $\varepsilon^{'}$ are independent but identically distributed.

We want to see if the coefficients estimated using (x_i, y_i) produce good predictions for $(x_i, y_i^{'})$.
$\rightarrow$ If the model can't predict well any more, it has just memorize the noise (only the noise has changed).

More on Training vs Test MSEs

Compare $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i^{'}-\hat{m}_i]^2\right]$ with $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{m}_i]^2\right]$ where $\hat{m}_i = x_i\hat{\beta}$.

- y_i and $\hat{m}_i$ are dependent random variables since $\hat{m}_i$ depends notably on y_i
- $y_i^{'}$ and $\hat{m}_i$ are independent random variables

More on Training vs Test MSEs

$$
\begin{aligned}
\mathrm{E}\left[(y_i - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i - \hat{m}_i) + (\mathrm{E}[y_i - \hat{m}_i])^2 \\
&= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{m}_i) - 2\mathrm{Cov}(y_i, \hat{m}_i) + (\mathrm{E}[y_i] - \mathrm{E}[\hat{m}_i])^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[(y_i^{'} - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i^{'} - \hat{m}_i) + (\mathrm{E}[y_i^{'} - \hat{m}_i])^2 \\
&= \mathrm{Var}(y_i^{'}) + \mathrm{Var}(\hat{m}_i) - 2\mathrm{Cov}(y_i^{'}, \hat{m}_i) + (\mathrm{E}[y_i^{'}] - \mathrm{E}[\hat{m}_i])^2
\end{aligned}
$$

- y_i is independent of $y_i^{'}$ but has the same distribution: $\mathrm{E}[y_i] = \mathrm{E}[y_i^{'}]$ and $\mathrm{Var}(y_i) = \mathrm{Var}(y_i^{'})$
- $\mathrm{Cov}(y_i^{'}, \hat{m}_i) = 0$

$$
\begin{aligned}
\mathrm{E}\left[(y_i^{'} - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{m}_i) + (\mathrm{E}[y_i] - \mathrm{E}[\hat{m}_i])^2 \\
&= \mathrm{E}\left[(y_i - \hat{m}_i)^2\right] + 2\mathrm{Cov}(y_i, \hat{m}_i)
\end{aligned}
$$

More on Training vs Test MSEs

$$\begin{aligned}\mathrm{E}\left[(y_i - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i - \hat{m}_i) + (\mathrm{E}[y_i - \hat{m}_i])^2 \\ &= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{m}_i) - 2\mathrm{Cov}(y_i, \hat{m}_i) + (\mathrm{E}[y_i] - \mathrm{E}[\hat{m}_i])^2\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left[(y_i^{'} - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i^{'} - \hat{m}_i) + (\mathrm{E}[y_i^{'} - \hat{m}_i])^2 \\ &= \mathrm{Var}(y_i^{'}) + \mathrm{Var}(\hat{m}_i) - 2\mathrm{Cov}(y_i^{'}, \hat{m}_i) + (\mathrm{E}[y_i^{'}] - \mathrm{E}[\hat{m}_i])^2\end{aligned}$$

- y_i is independent of $y_i^{'}$ but has the same distribution: $\mathrm{E}[y_i] = \mathrm{E}[y_i^{'}]$ and $\mathrm{Var}(y_i) = \mathrm{Var}(y_i^{'})$
- $\mathrm{Cov}(y_i^{'}, \hat{m}_i) = 0$

$$\begin{aligned}\mathrm{E}\left[(y_i^{'} - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{m}_i) + (\mathrm{E}[y_i] - \mathrm{E}[\hat{m}_i])^2 \\ &= \mathrm{E}\left[(y_i - \hat{m}_i)^2\right] + 2\mathrm{Cov}(y_i, \hat{m}_i)\end{aligned}$$

More on Training vs Test MSEs

$$\mathrm{E}\left[(y_i - \hat{m}_i)^2\right] = \mathrm{Var}(y_i - \hat{m}_i) + (\mathrm{E}[y_i - \hat{m}_i])^2$$
$$= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{m}_i) - 2\mathrm{Cov}(y_i, \hat{m}_i) + (\mathrm{E}[y_i] - \mathrm{E}[\hat{m}_i])^2$$

$$\mathrm{E}\left[(y_i^{\prime} - \hat{m}_i)^2\right] = \mathrm{Var}(y_i^{\prime} - \hat{m}_i) + (\mathrm{E}[y_i^{\prime} - \hat{m}_i])^2$$
$$= \mathrm{Var}(y_i^{\prime}) + \mathrm{Var}(\hat{m}_i) - 2\mathrm{Cov}(y_i^{\prime}, \hat{m}_i) + (\mathrm{E}[y_i^{\prime}] - \mathrm{E}[\hat{m}_i])^2$$

- y_i is independent of $y_i^{\prime}$ but has the same distribution: $\mathrm{E}[y_i] = \mathrm{E}[y_i^{\prime}]$ and $\mathrm{Var}(y_i) = \mathrm{Var}(y_i^{\prime})$
- $\mathrm{Cov}(y_i^{\prime}, \hat{m}_i) = 0$

$$\mathrm{E}\left[(y_i^{\prime} - \hat{m}_i)^2\right] = \mathrm{Var}(y_i) + \mathrm{Var}(\hat{m}_i) + (\mathrm{E}[y_i] - \mathrm{E}[\hat{m}_i])^2$$
$$= \mathrm{E}\left[(y_i - \hat{m}_i)^2\right] + 2\mathrm{Cov}(y_i, \hat{m}_i)$$

More on Training vs Test MSEs

$$\begin{aligned} \mathrm{E}\left[(y_i - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i - \hat{m}_i) + (\mathrm{E}[y_i - \hat{m}_i])^2 \\ &= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{m}_i) - 2\mathrm{Cov}(y_i, \hat{m}_i) + (\mathrm{E}[y_i] - \mathrm{E}[\hat{m}_i])^2 \end{aligned}$$

$$\begin{aligned} \mathrm{E}\left[(y_i' - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i' - \hat{m}_i) + (\mathrm{E}[y_i' - \hat{m}_i])^2 \\ &= \mathrm{Var}(y_i') + \mathrm{Var}(\hat{m}_i) - 2\mathrm{Cov}(y_i', \hat{m}_i) + (\mathrm{E}[y_i'] - \mathrm{E}[\hat{m}_i])^2 \end{aligned}$$

- y_i is independent of y_i' but has the same distribution: $\mathrm{E}[y_i] = \mathrm{E}[y_i']$ and $\mathrm{Var}(y_i) = \mathrm{Var}(y_i')$
- $\mathrm{Cov}(y_i', \hat{m}_i) = 0$

$$\begin{aligned} \mathrm{E}\left[(y_i' - \hat{m}_i)^2\right] &= \mathrm{Var}(y_i) + \mathrm{Var}(\hat{m}_i) + (\mathrm{E}[y_i] - \mathrm{E}[\hat{m}_i])^2 \\ &= \mathrm{E}\left[(y_i - \hat{m}_i)^2\right] + 2\mathrm{Cov}(y_i, \hat{m}_i) \end{aligned}$$

Model assessment/validation

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{m}_i]^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_i]^2\right] + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{m}_i)$$

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{m}_i]^2\right] \approx \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_i]^2 + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{m}_i)$$

The **optimism** is the amount by which the training error systematically under-estimates the expected test error.

- One way to estimate test error is to **estimate the optimism** and then **add it to the training error**, e.g. AIC and BIC.
- Another way is to **directly estimate the test error** using resampling methods, e.g. validation set, cross-validation and bootstrap.

Linear model assessment

For the linear model, since $\text{Cov}(y_i, \hat{m}_i) = \sigma^2 H_{ii}$, we have

$$\frac{2}{n}\sum_{i=1}^{n}\text{Cov}(y_i, \hat{m}_i) = \frac{2}{n}\sigma^2\ \text{tr}(H) = \frac{2}{n}\sigma^2(p+1),$$

where we used $\text{tr}(H) = p + 1$.

$$\text{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{m}_i]^2\right] \approx \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_i]^2 + \frac{2}{n}\sigma^2(p+1)$$

- Here, the optimism grows with σ^2 and p. It shrinks with n.

Residual Sum of Squares

Residual Sum of Squares (or SSE)

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Minimizing RSS will always choose the model with the most predictors.

Estimated residual variance

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n - p - 1}$$

where $p =$ no. predictors.
Minimizing $\hat{\sigma}^2$ works quite well for choosing predictors (but better methods to follow).

The R^2 statistic

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

The R^2 gives the proportion of variance explained, and is independent of the scale of y.

However . . .

- R^2 does not allow for “degrees of freedom”.
- Adding *any* variable tends to increase the value of R^2, even if that variable is irrelevant.

To overcome this problem, we can use *adjusted* R^2:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Maximizing $\bar{R}^2$ is equivalent to minimizing $\hat{\sigma}^2$.

The $\bar{R}^2$ statistic

Minimizing $\hat{\sigma}^2$, what does that translate to?

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p-1} = \mathrm{MSE}\frac{n}{n-p-1} = \mathrm{MSE}\frac{1}{1-(p+1)/n}$$

The binomial theorem gives $(1-x)^{-1} = 1 + x + x^2 + \ldots$, and we truncate the series at first order. For a fixed k, the approximation becomes exact as $n \to \infty$.

$$\hat{\sigma}^2 \approx \mathrm{MSE}\left(1 + \frac{p+1}{n}\right) = \mathrm{MSE} + \mathrm{MSE}\frac{p+1}{n} \textit{ vs } \mathrm{MSE} + 2\sigma^2\frac{(p+1)}{n}$$

$\rightarrow$ even for the right model where MSE is a consistent estimator of σ^2, the penalty is half as big as what it should be. $\bar{R}^2$ is better than R^2 but it is still not going to work very well.

Akaike's Information Criterion

$$\mathrm{AIC} = -2\log(L) + 2(p+1)$$

where L is the likelihood and k is the number of predictors in the model.

- This is a penalized **likelihood** approach.
- *Minimizing* the AIC gives the best model for prediction.
- AIC penalizes more heavily than $\bar{R}^2$.
- Minimizing the AIC is asymptotically equivalent to minimizing MSE via leave-one-out cross-validation.

Corrected AIC

For small values of n, the AIC tends to select too many predictors, and so a bias-corrected version of the AIC has been developed.

$$\mathrm{AIC_C} = \mathrm{AIC} + \frac{2(p+2)(p+3)}{n-p-1}$$

As with the AIC, the $\mathrm{AIC_C}$ should be minimized.

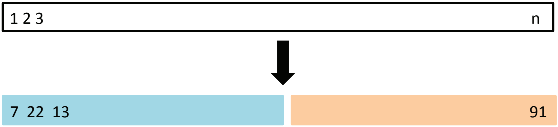
Schwartz Bayesian Information Criterion

$$\text{BIC} = -2\log(L) + (p+1)\log(n)$$

where L is the likelihood and k is the number of predictors in the model.

- *Minimizing* the BIC gives the best model for prediction.
- BIC penalizes more heavily than AIC
- Also called SBIC and SC.
- Minimizing BIC is asymptotically equivalent to leave-v-out cross-validation when $v = n[1 - 1/(log(n) - 1)]$.

The validation set



1 2 3 n

7 22 13 91

$$\{x_i, y_i\}_1^n \implies Tr = \{x_i, y_i\}_1^k \text{ and } Val = \{x_i, y_i\}_{k+1}^n$$

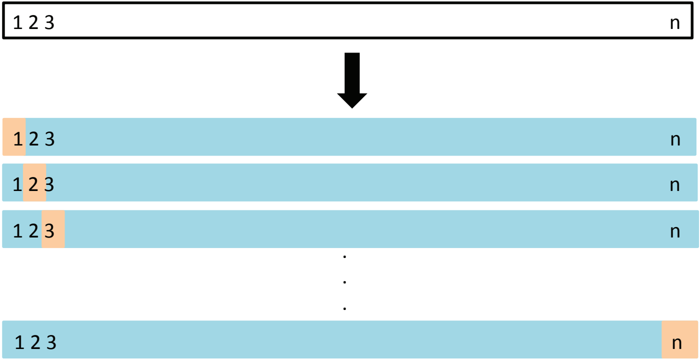
Training Mean Squared Error

$$\frac{1}{k}\sum_{i=1}^{k}[y_i - \hat{f}(x_i)]^2$$

Validation Mean Squared Error

$$\frac{1}{n-k}\sum_{i=k+1}^{n}[y_i - \hat{f}(x_i)]^2$$

Leave one-out cross-validation



LOO Cross-validation

Leave-one-out cross-validation (LOOCV) for regression can be carried out using the following steps.

- Remove observation i from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation.
- Repeat step 1 for $i = 1, \ldots, n$.
- Compute the MSE from $\{e_1^*, \ldots, e_n^*\}$. We shall call this the CV.

LOOCV vs validation set

- LOOCV has less bias
  - We repeatedly fit the statistical learning method using training data that contains $n-1$ obs., i.e. almost all the data set is used
- LOOCV produces a less variable MSE
  - The validation approach produces different MSE when applied repeatedly due to randomness in the splitting process, while performing LOOCV multiple times will always yield the same results, because we split based on 1 obs. each time
- LOOCV is (usually) computationally intensive
  - We fit each model n times!

LOOCV for linear models

Fitted values

$$\hat{\boldsymbol{Y}} = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the “hat matrix”.

Leave-one-out residuals

Let $h_1, \dots, h_n$ be the diagonal values of $\boldsymbol{H}$, then the cross-validation statistic is

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

where e_i is the residual obtained from fitting the model to all n observations.

Linear model selection procedures

Different strategies to search for the best model: Best subsets, Forward stepwise, Backwards stepwise, ...

Best subsets regression

- Fit all possible regression models using one or more of the predictors.
- Choose the best model based on CV (or an asymptotic equivalent: AIC, AICc).

Warning!

- If there are a large number of predictors, this is not possible.

  For example, 44 predictors leads to 18 trillion possible models!

Linear model selection procedures

Backwards stepwise regression

- Start with a model containing all variables.
- Try subtracting one variable at a time. Keep the model if it has lower CV, AICc or AIC.
- Iterate until no further improvement.

Notes:

- Stepwise regression is not guaranteed to lead to the best possible model.
- If you are trying several different models, use the CV, AICc or AIC value to select between them.