Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 5: Resampling Methods (ISL ch. 5)

Instructor: Ma, Jun

Renmin University of China

October 21, 2021

Cross-validation and bootstrap

- These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
  - Cross validation: estimate the test error to evaluate its performance (model assessment) and select the appropriate level of flexibility (model selection).
  - Bootstrap: standard error of parameter estimates (measure of estimation accuracy).
- They are computationally expensive: fitting the same method multiple times.

Test error and training error

- Training data: $(Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n)$.
- The test error (regression MSE or misclassification error rate) is the average error of predicting the response Y_0 to a new input vector X_0. (Y_0, X_0) was not used in training the method.
- The training error is the average error of applying the method to the observations used in its training.

- Regression: Y is quantitative and (Te: test; Tr: training)

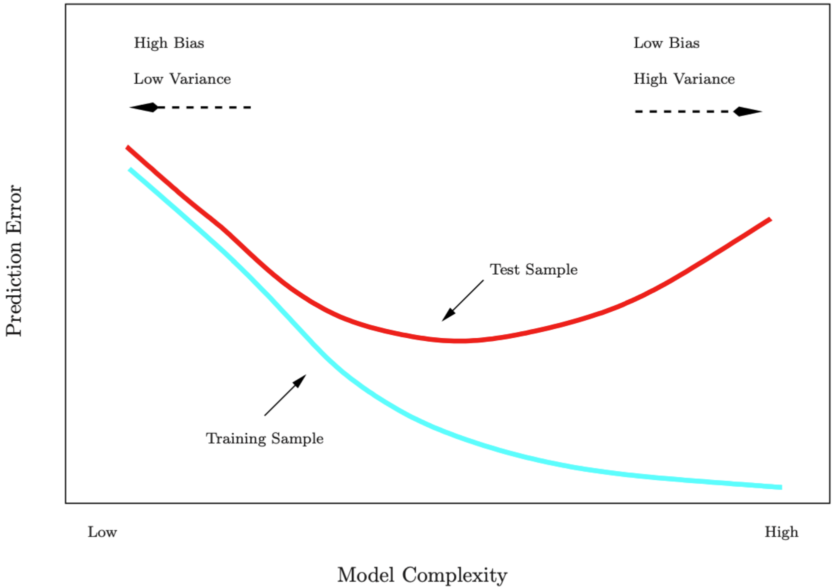
$$\begin{aligned}\mathrm{MSE}_{\mathsf{Te}} &= \mathrm{E}\left[\left(Y_0 - \hat{f}\,(X_0)\right)^2\right] \\ \mathrm{MSE}_{\mathsf{Tr}} &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}\,(X_i)\right)^2 .\end{aligned}$$

- Classification: Y is categorical and (Te: test; Tr: training)

$$\begin{aligned}\mathrm{Err}_{\mathsf{Te}} &= \Pr\left[Y_0 \neq \hat{f}\,(X_0)\right] \\ \mathrm{Err}_{\mathsf{Tr}} &= \frac{1}{n}\sum_{i=1}^{n} 1\left(Y_i \neq \hat{f}\,(X_i)\right) .\end{aligned}$$

- $\hat{f}$ depends on the training data.

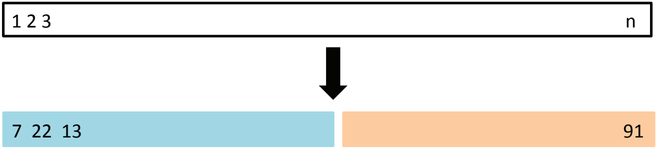
- Given a data set, the use of a particular method is warranted if it results in a low test error.
- Which method in the ML toolbox results in the lowest test error depends on the underlying data generating mechanism.
- The test error can be estimated if a large test data is available. Unfortunately, this is usually not the case. In contrast, the training error can be easily calculated.
- The training error can dramatically underestimate the test error.


High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
Prediction Error
Low
High
Model Complexity

- Some methods make a mathematical adjustment to the training error rate in order to estimate the test error rate.
    - Think of the adjusted R^2 in the regression context.
- A class of methods estimates the test error by holding out a subset of the training observations from the fitting process, and then applying the method to those held out observations.

Validation-set approach

- Randomly divide the data points into two parts: a training set and a validation set.
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error.

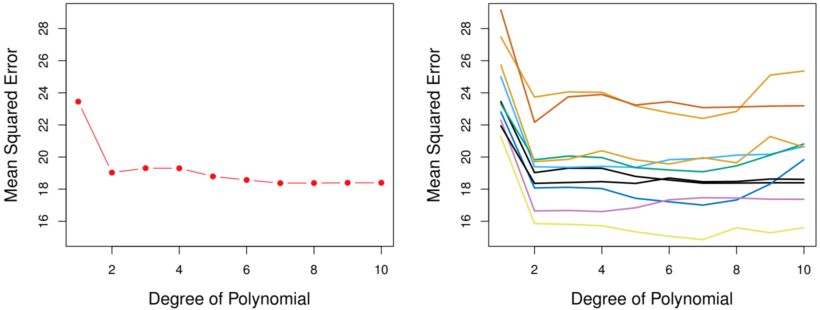


ISL Figure 5.1

- A random splitting into two halves: left part is training set, right part is validation set.
- Possible $\binom{n}{n/2}$ splits. The validation-set approach randomly selects one out of these splits.

Example: the `Auto` data

- There appears to be a non-linear relationship between `mpg` and `horsepower`. A model that predicts mpg using `horsepower` and `horsepower`^{2} gives better in-sample fit (training error) than a model that uses only a linear term. We compare linear versus higher-order polynomial terms in a linear regression.
- We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.



ISL Figure 5.2

- Left: single split; right: multiple splits.

Drawbacks of the validation set approach

- The validation estimate of the test error can be highly variable, depending on how to split the sample into training and validation sets.
- Only half of the observations (the training set) are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

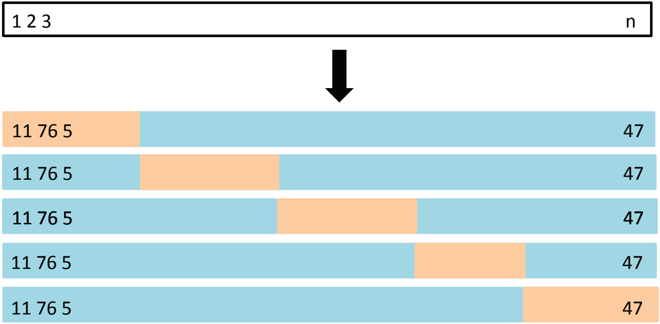
K-fold Cross-validation (CV)

- Randomly divide the data into K (approximately equal-sized) parts: $C_1, C_2,..., C_K$ with $\cup_{k=1}^{K} C_k = \{1, 2, ..., n\}$, where C_k denote the indices of observations in part k.
- Suppose that C_k has n_k observations so that $\sum_{k=1}^{K} n_k = n$. If n can be divided by K, then $n_k = n/K$.
- Use observations in $\cup_{j \neq k} C_j$ to predict $\{Y_i : i \in C_k\}$.
- This is done in turn for each $k = 1, 2, ..., K$, and compute:

$$\text{MSE}_k = \frac{1}{n_k} \sum_{i \in C_k} \left(Y_i - \hat{f}_{-k}(X_i)\right)^2 \text{ or } \text{Err}_k = \frac{1}{n_k} \sum_{i \in C_k} 1\left(Y_i \neq \hat{f}_{-k}(X_i)\right),$$

  where $\widehat{f_{-k}}$ uses observations in $\cup_{j \neq k} C_j$.
- The K-fold CV estimate of the test error:

$$\text{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k \text{ or } \text{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \text{Err}_k.$$



ISL Figure 5.5

- n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (beige), and the remainder as a training set (blue).
- The test error is estimated by averaging the five resulting MSE estimates.

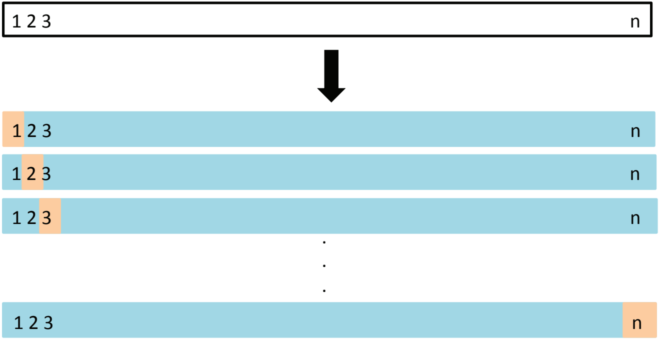
Leave-one out cross-validation (LOOCV)

- LOOCV is a special case of K-fold CV that sets $K = n$.
- A single observation (X_k, Y_k) is used for the validation set, and the remaining observations $(X_1, Y_1), ..., (X_{k-1}, Y_{k-1}), (X_{k+1}, Y_{k+1}), ..., (X_n, Y_n)$ make up the training set to form the predictor/classifier $\hat{f}_{-k}$.
- The method is fit on the $n - 1$ training observations and a prediction $\hat{f}_{-k}(X_k)$ is made for the excluded observation. Then,

  $$\mathrm{MSE}_k = \left(Y_k - \hat{f}_{-k}(X_i)\right)^2 \text{ or } \mathrm{Err}_k = 1\left(Y_k \neq \hat{f}_{-k}(X_k)\right)$$

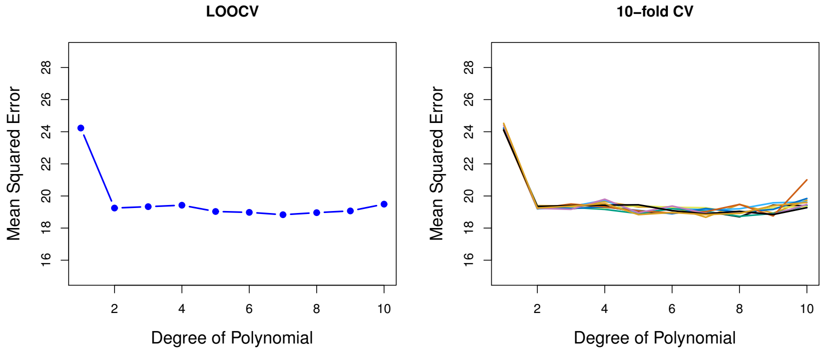
  and do it in turn for $k = 1, 2, ..., n$ to get $\mathrm{MSE}_1, ..., \mathrm{MSE}_n$ (or $\mathrm{Err}_1, ..., \mathrm{Err}_n$).
- The LOOCV estimate for the test error is the average of these n test error estimates:

  $$\mathrm{CV}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i \text{ or } \mathrm{CV}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i.$$



ISL Figure 5.3

- A set of n data points is repeatedly split into a training set (blue) containing all but one observation, and a validation set that contains only that observation (beige).
- The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.



ISL Figure 5.4

- Performing LOOCV multiple times yields the same results: no randomness from training/validation splits.
- Left: The LOOCV error curve. Right: 10-fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.

LOOCV for linear regressions

- In linear regressions, we do not need to fit the model n times, since

$$\mathrm{CV}_n = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - h_i} \right)^2,$$

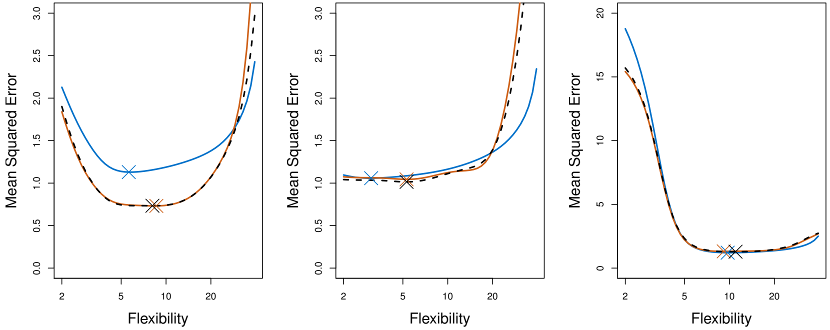
  where $\hat{Y}_i$ is the i-th fitted value from the regression using all data and h_i is the i-th leverage statistic.
- The training MSE is just $n^{-1} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$.

LOOCV versus K-fold CV

- LOOCV incurs way more computational burden: fit the model n times. Usually n is very large.
- LOOCV and K-fold CV are different estimators ($\hat{\theta}$) of the unknown test error (θ). We consider and compare the estimation MSE $\mathrm{E}\left[\left(\hat{\theta}-\theta\right)^2\right]$ which is decomposed into

$$\mathrm{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = \underbrace{\left(\mathrm{E}\left[\hat{\theta}\right]-\theta\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}\left[\hat{\theta}\right]}_{\text{variance}}.$$

- In LOOCV, the estimates from each fold are highly correlated (use the same $n-2$ data points) and hence their average can have high variance.
- For K-fold CV, we are averaging the outputs of K fitted models that are somewhat less correlated, since the overlap between the training sets is smaller.
- LOOCV: high variance, low bias; K-fold CV: lower variance but higher bias. A popular choice is $K = 5$ or 10.



ISL Figure 5.6

- Blue: true test MSE; black dashed: LOOCV; 10-fold CV: orange; Crosses: minimum. Estimation of the minimizer is more accurate than estimation of the test MSE.
- We perform CV on a number of ML methods to find the one that results in the lowest test error. So we are more interested in the minimizer.

The bootstrap

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide a bootstrap standard error $\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)$ of a regression coefficient $\hat{\beta}_1$. A 95% confidence interval is given by $\left[\hat{\beta}_1 - 2\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)\right]$. However, in this case, one typically uses conventional standard errors which are easy to compute.
- Bootstrap is computationally burdensome.

- Many estimators $\hat{\theta}_n$ (n denotes the sample size) have the property $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \sim \mathrm{N}\left(0, \sigma^2\right)$and equivalently $\hat{\theta}_n \sim \mathrm{N}\left(\theta, \sigma^2/n\right)$, approximately, when n is very large.
- Conventional methods estimate σ^2 using the analogue principle (i.e., replace population moments/unknown quantities in σ^2 by their sample moments/estimates) and then construct the standard error $\hat{\sigma}/\sqrt{n}$.
- Conventional methods require knowledge of the expression (formula) of σ^2, which can be very complicated in many contexts.
- Bootstrap is a computation-intensive approach that does not requires knowledge of the expression of σ^2.

A simple example

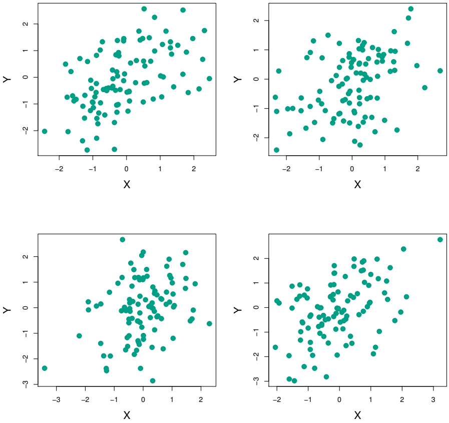
- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y.
- We wish to choose α to minimize the variance of our investment $\operatorname{Var}\left[\alpha X + (1-\alpha) Y\right]$.
- One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

  where $\sigma_X^2 = \operatorname{Var}[X]$, $\sigma_Y^2 = \operatorname{Var}[Y]$ and $\sigma_{XY} = \operatorname{Cov}[X, Y]$.
- Let $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$ and $\hat{\sigma}_{XY}$ denote the sample variance/covariances. We can then estimate α using

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

- It can be shown that $\sqrt{n}\,(\hat{\alpha} - \alpha) \sim \mathrm{N}\left(0, \sigma^2\right)$ and $\hat{\alpha} \sim \mathrm{N}\left(\alpha, \sigma^2/n\right)$, when n is very large, for some $\sigma^2 > 0$ with a complicated expression.



ISL Figure 5.9

- Each panel displays 100 simulated returns for investments X and Y, with $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$, $\sigma_{XY} = 0.5$ and $\alpha = 0.6$. The resulting estimates for α are 0.576, 0.532, 0.657, and 0.651.
- We can repeat the process of simulating 100 paired observations more times.

- To estimate the standard deviation of $\hat{\alpha}$, we repeated the process of simulating 100 paired observations of X and Y , and estimating α 1000 times to get $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_{1000}$.
- $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_{1000}$ are independent observations of $\hat{\alpha}$ in the simulation context.
- The mean over all 1000 estimates for α is

  $$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996,$$

  which is very close to $\alpha = 0.6$ and the sample standard deviation of the estimates from the repeated samples is

  $$\sqrt{\frac{1}{1000-1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083.$$

- 0.083 should be an accurate estimate of the population standard deviation of $\hat{\alpha}$, which should be approximately $\sigma/\sqrt{n}$, when n is large.

- The procedure outlined above cannot be applied. We have only one sample (data set) and we cannot generate new samples from the original population, which is unknown.
- The bootstrap approach allows us to use a computer to mimic the process of obtaining new samples.
- Rather than repeatedly obtaining independent samples from the population, we instead repeatedly sample observations from the original data set with replacement.
- Each of these bootstrap samples is created by sampling with replacement, and is the same size as our original sample. Some observations may appear more than once and some not at all.

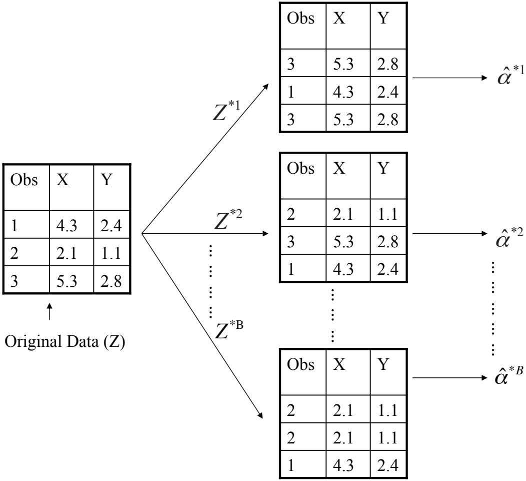
- This procedure is repeated B times for some large value of B (e.g., $B = 1000$), in order to produce B different bootstrap samples.
- Each bootstrap data set is used to obtain an estimate of α: $\hat{\alpha}^{*1}$, $\hat{\alpha}^{*2}$, ..., $\hat{\alpha}^{*B}$.
- We estimate the standard error of these bootstrap estimates using the formula

$$\mathrm{SE}_B\left(\hat{\alpha}\right) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*}\right)^2},$$

where $\bar{\hat{\alpha}}^{*} = B^{-1}\sum_{r=1}^{B}\hat{\alpha}^{*r}$.

- In our numerical example, $\mathrm{SE}_B\left(\hat{\alpha}\right) = 0.087$.
- A feasible 95% confidence interval for α is $[\hat{\alpha} - 2\mathrm{SE}_B\left(\hat{\alpha}\right), \hat{\alpha} + 2\mathrm{SE}_B\left(\hat{\alpha}\right)]$.
- Indeed, it can be shown that $\mathrm{SE}_B\left(\hat{\alpha}\right)/\left(\sigma/\sqrt{n}\right)$ converges to 1, in probabilistic sense.

Example with just 3 observations



Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original Data (Z)

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

ISL Figure 5.11