Statistical Learning

https://github.com/ggorr/Machine-Learning/tree/master/ISLR

5 Resampling Methods

- 5.1 Cross-Validation
- 5.2 The Bootstrap

What is resampling method?

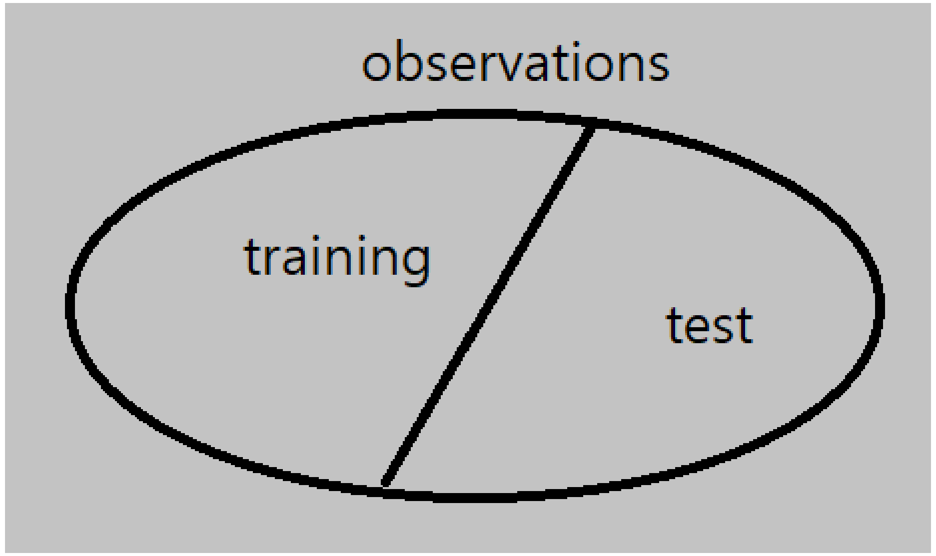
- Drawing samples from a training set
- Fitting a model on each sample

5.1 Cross-Validation

- 5.1.1 The Validation Set Approach
- 5.1.2 Leave-One-Out Cross-Validation
- 5.1.3 k-Fold Cross-Validation
- 5.1.4 Bias-Variance Trade-Off for k-Fold Cross-Validation
- 5.1.5 Cross-Validation on Classification Problems

What is cross-validation

- Holding out some observations from the fitting process
- Computing test error rates from the holding out observations

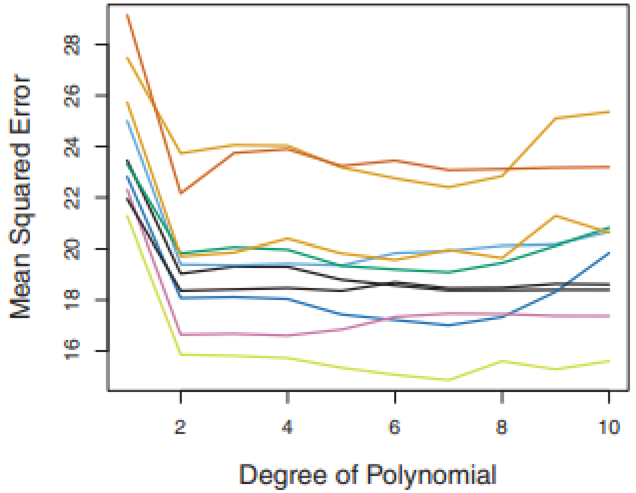


5.1.1 The Validation Set Approach

- Validation set approach
  - Random splitting
    - A training set – 50%
    - A validation set(= hold-out set) – 50%
  - Training and computing test error rate

The Validation Set Approach

- Test errors are highly variable
  - high variance
- Test errors are overestimated
  - because training uses the half of data set



5.1.2 Leave-One-Out Cross-Validation

- Leave-one-out cross-validation (LOOCV)
  - Training on observations except one
  - Computing test error for leave one observation
  - Averaging test errors

Applying LOOCV

- Observations: $(x_1, y_1), \dots, (x_n, y_n)$
- Fitting the model on $(x_2, y_2), \dots, (x_n, y_n)$
- Let

$$\mathrm{MSE}_1 = (y_1 - \hat{y}_i)^2$$

- Similarly, compute $\mathrm{MSE}_2, \dots, \mathrm{MSE}_n$
- Averaging test errors

$$\mathrm{CV}_{(n)} = \frac{1}{n} \sum \mathrm{MSE}_i$$

Linear or polynomial regression

- Interesting formula

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

  where h_i is the leverage, i.e.

$$h_i = i\text{-th diagonal entry of the hat matrix } X(X^T X)^{-1} X^T$$

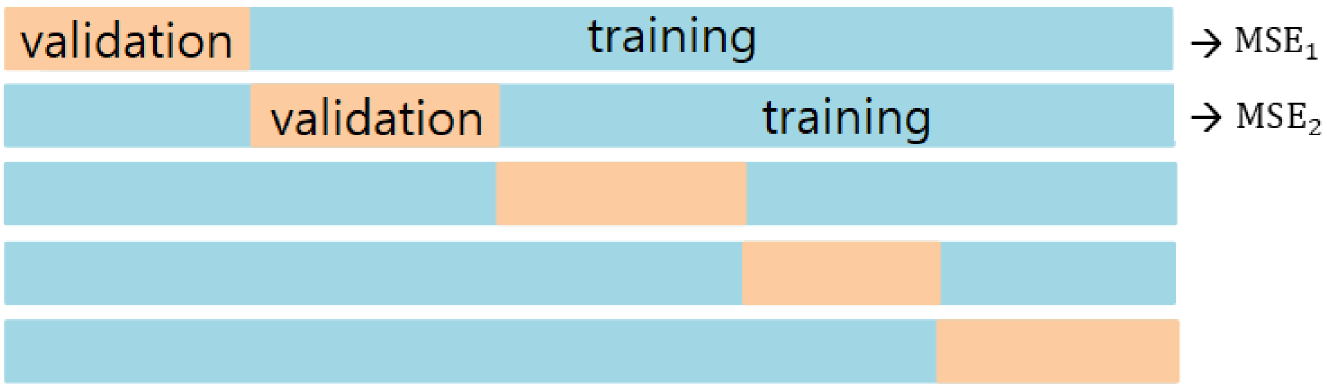
LOOCV

- Computational cost is high

5.1.3 k-Fold Cross-Validation

- Dividing observations into k groups (or folds)
  - The first fold is treated as a validation set
  - Fitting on the remaining $k-1$ folds
  - Averaging test errors

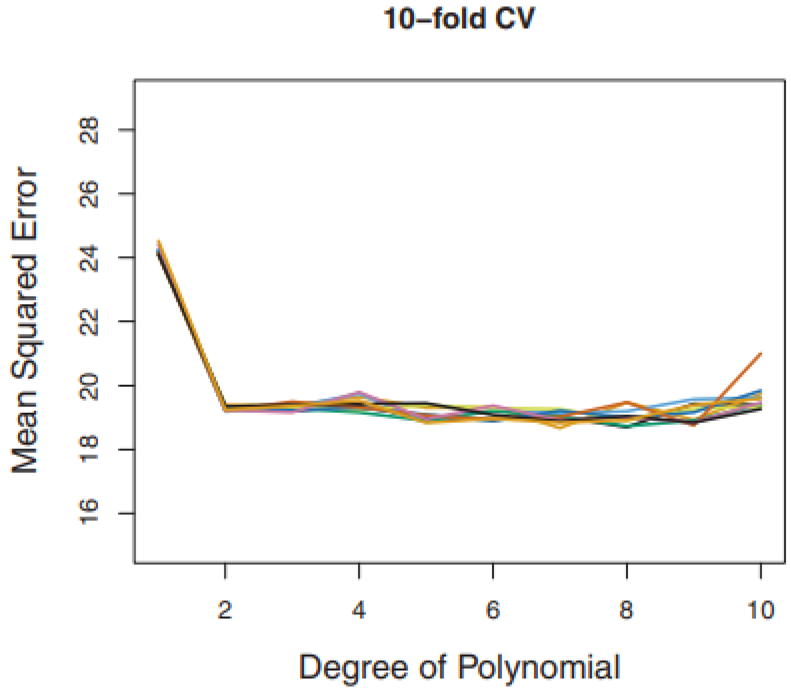
$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum \mathrm{MSE}_i$$



5-folds CV

k-fold CV

- Computational cost is not high
- Test errors are not highly variable



5.1.4 Bias-Variance Trade-Off for k-Fold Cross-Validation

- Bias of test error

$$\text{Validation set approach} \geq k\text{-Fold CV} \geq \text{LOOCV}$$

- Variance of test error

$$k\text{-Fold CV} \leq \text{LOOCV}$$

5.1.5 Cross-Validation on Classification Problems

- LOOCV

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum \mathrm{Err}_i$$

where $\mathrm{Err}_i = I(y_i \neq \hat{y}_i)$

- k-fold CV

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum \mathrm{Err}_i$$

where $\mathrm{Err}_i = \sum_{j \in \mathrm{Validation}_i} I(y_j \neq \hat{y}_j)$

5.2 The Bootstrap

- Bootstrap
  - Repeatedly sampling from the original data set with replacement (恢复提取)

  - Quantifying the uncertainty of an estimator
    - Example - estimating the standard error of coefficient

Example: Invest

- X, Y: returns(收益) of two financial assets
- Invest ratio: α, $1-\alpha$
- Minimize the risk or the variance $\mathrm{Var}(\alpha X + (1-\alpha)Y)$
  - Solution

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

    where σ_X^2 =Var(X), σ_Y^2 =Var(Y), and $\sigma_{XY} = \mathrm{Cov}(X, Y)$

  - From observation, estimating σ_X, σ_Y and σ_{XY} and computing

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

- Estimating the variance of α
  - Repeatedly sampling from the original data set with replacement
  - Computing the variance of $\hat{\alpha}$

$$\widehat{\mathrm{Ver}}(\hat{\alpha}) = \frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2$$