UNIVERSITY OF ARKANSAS

DASC 4113 Machine Learning

Lecture 5
Ukash Nakarmi

Resampling

Learning Objectives

In this class, we will learn how to:

Quantify the variability of Test Error of Learning Model

Resampling Methods

- Tools to repeatedly sample and fit models to training data.
- Allows us to understand the behavior of our fitted model.

For example: We can create a new subset of training data, fit linear regression model, then observe how the fitted models are similar/different.

Two Approaches:

Cross-Validation

Bootstrap

Cross -Validation

Key Idea:

- The training error can be very different (underestimate) the test error.
- If large test data is available, we can measure the performance of the fitted model (test-error) using test data. (Not usually the case)
- Cross-Validation technique hold-out some subset of training example and use it for estimating test error.

1. The Validation Set Approach
2. Leave-One-Out-Cross-Validation(LOOCV)
3. k-fold Cross-Validation

The Validation Set Approach

Divide the available data set into Training Set and Validation Set

Challenges:

- Validation estimate of test error can be very highly sensitive to specific training set and validation set.
- Reduces the number of data samples in the training set.

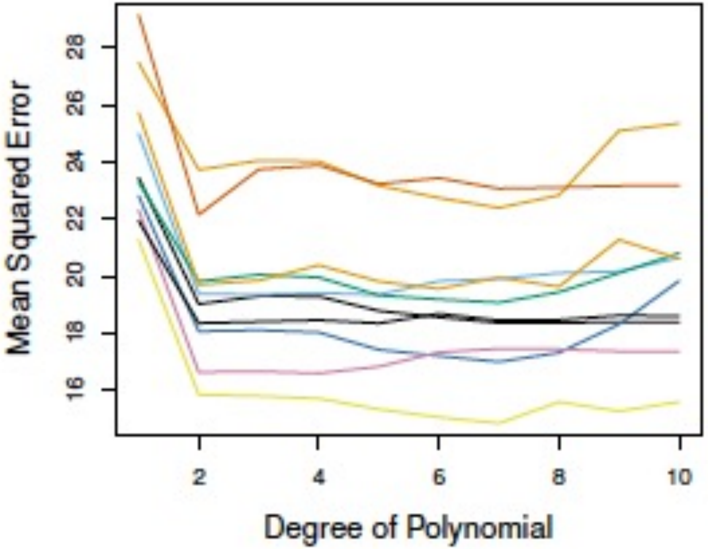


Fig: 5.2. MSE on validation set on different random split for MPG vs Horsepower data.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \cdots \beta_{N+1} X_1^N$$

Y : MPG

X_1: Horsepower

Leave One Out Cross Validation (LOOCV)

Given n data samples $\{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$:

We do n model fits such that at i^{th} fit:

- Validation Set is a single data sample (x_i, y_i),
- Training Set is rest n-1 data samples,

Then:

Test Set Error is calculated as average of n test error.

For example, if MSE is error metric:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i .$$

- Could be computationally expensive because we would need to fit the model n times.
- (For Linear Regression, computational cost do not increase. For other methods, potentially increases computational cost.)

K-Fold Cross Validation

Given n data samples divide the data into k groups of approximately equal size.

We do k model fits such that at i^{th} fit:

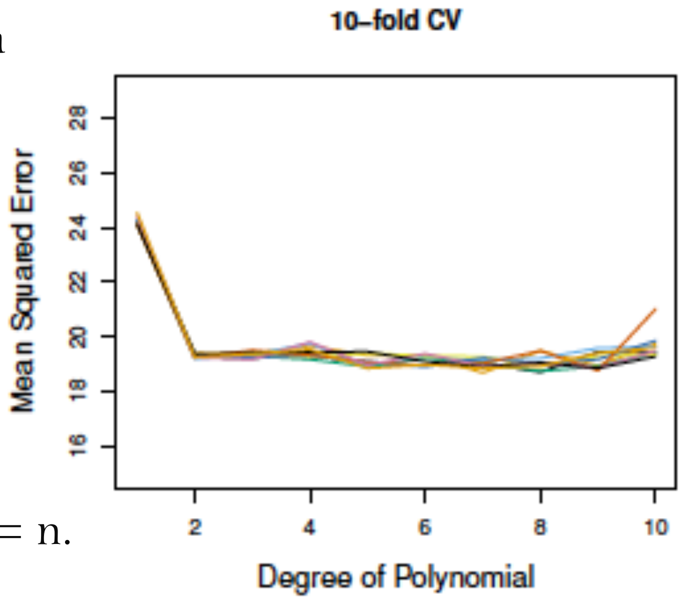
- Validation Set is i^{th} group of data
- Training Set is rest k-1 group,

Then:

Test Set Error is calculated as average of k test error.

For example, if MSE is error metric:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i.$$



- LOOCV is a special case of k-fold validation when k = n.
- Computationally cheaper then LOOCV.
- Typically, k = 5, 10.
- LOOCV test error has better bias reduction, but higher variance compared to k-fold.
- k-fold is preferred not only due to computational adv, but also it reduces variance in test error.

Cross Validation

- These Cross Validation Techniques could be used for any Statistical Learning Methods even though we used Linear Regression as an Example.

General Form:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i,$$

- Note: Cross validation do not necessarily improve the model itself, but it allows us to more precisely quantify the model model's variability in test error.

The Bootstrap

- Alternative approach to resample that data for quantifying the variability of statistical model.
- Unlike cross-validation, bootstrapping resamples using replacement.
- Resampling methods are Not limited to variability of Error.

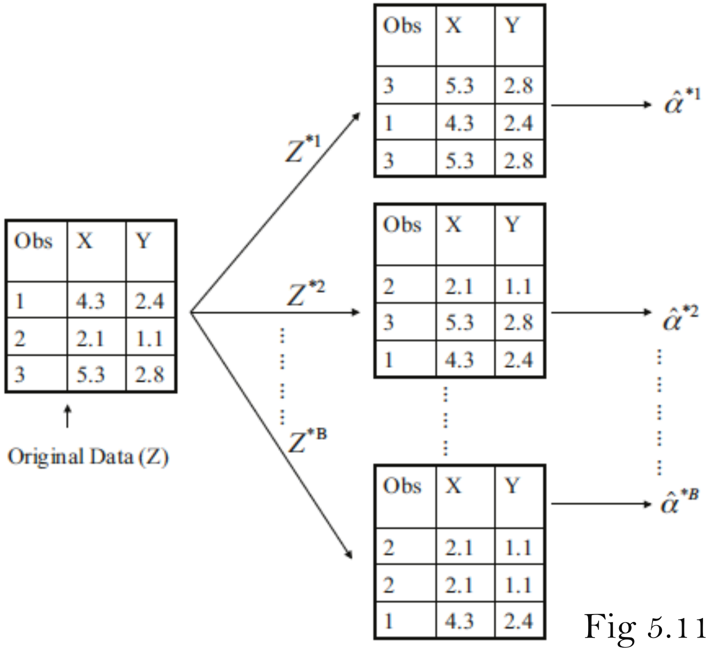


Fig 5.11. Resampling with Replacement

The Bootstrap

- Example:

Scenario:

- We have some amount of money and two investment options.
- We want to invest fraction of money(α) on money on option 1 and $(1-\alpha)$ on option 2.
- Option 1 has return rate of X and option 2 has return rate of Y

Then: Return ~ $\alpha X + (1-\alpha) Y$

There are variability associated with Return rate X and .

So, We want to minimize the the risk or Var $(\alpha X + (1-\alpha) Y)$

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

$$\sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y), \text{ and } \sigma_{XY} = \mathrm{Cov}(X, Y)$$

The Bootstrap

- Example:

Scenario:

- We have some amount of money and two investment options.
- We want to invest fraction of money(α) on money on option 1 and $(1-\alpha)$ on option 2
- Option 1 has return rate of X and option 2 has return rate of Y

Then: Return ~ αX + $(1-\alpha)$ Y

There are variability associated with Return rate X and .

So, We want to minimize the the risk or Var (αX + $(1-\alpha)$ Y)

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

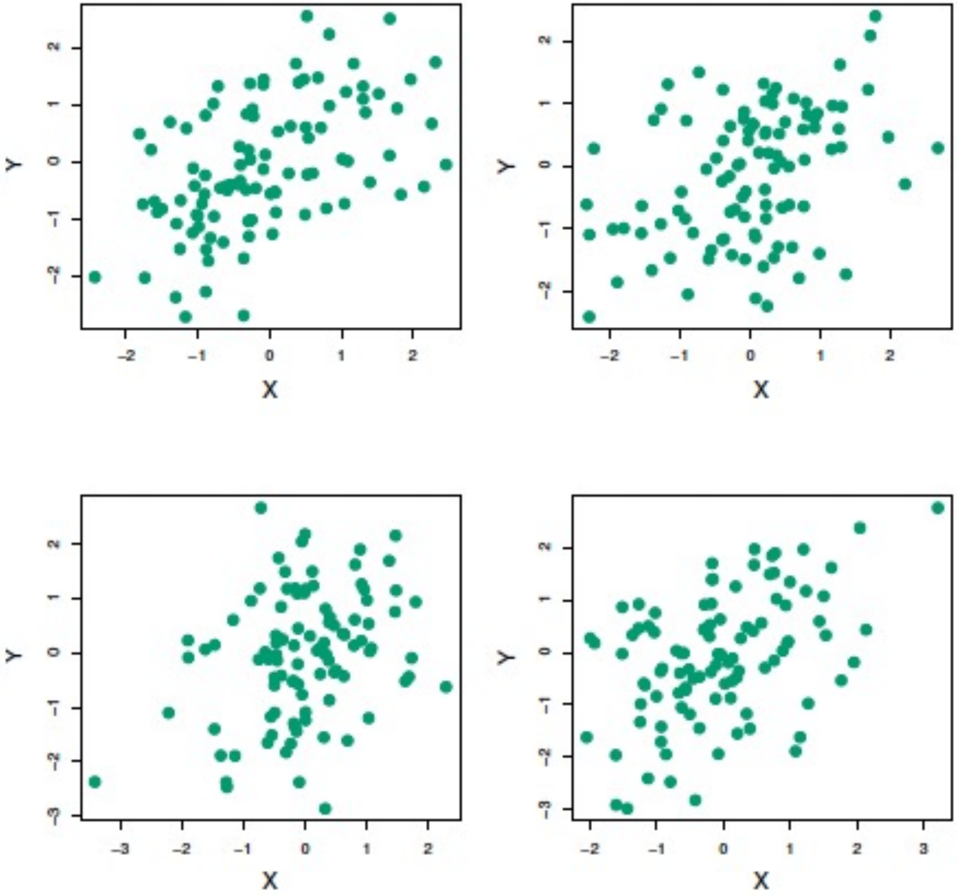
$$\sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y), \text{ and } \sigma_{XY} = \mathrm{Cov}(X, Y)$$

The Bootstrap

Fig: 5.9
4-Dataset with 100 samples each generated from replacement.
Each set with different α.

$\alpha = 0.576$

$\alpha = 0.532$



$\alpha = 0.657$

$\alpha = 0.651$

The Bootstrap

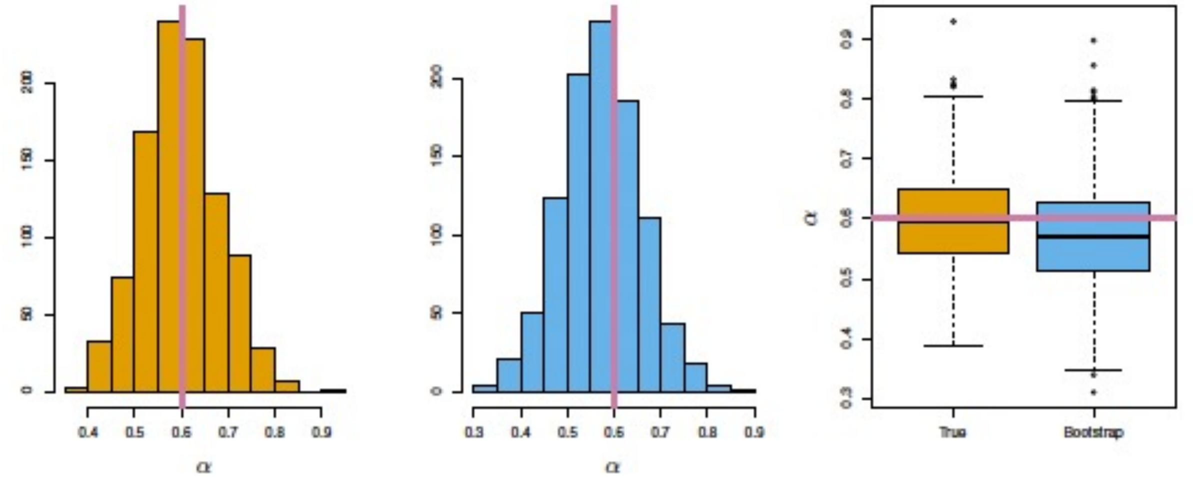


Fig: 5.10, Left ; Estimation from Bootstrapping in Population, Center: ~ bootstrapping in 1 given dataset, Right: Box plot

Since we do re-sampling with replacement, we can generate many dataset. (Say 1000) and get many estimates $\alpha_1, \alpha_2 \dots \alpha_{1000}$

Then:

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

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

Takeaway: Estimation of $\hat{\alpha}$ (0.5996) from bootstrapping in 1 sample dataset is close ~ Estimation of $\hat{\alpha}$ (0.6) from bootstrapping in actual data (population)

Take Away Points

- It is important to quantify the variability of test error of Learning Models.

- Two ways to do that are :
- Cross-Validation -> Resampling without replacement
- Bootstrapping -> With Replacement