IST 5535: Machine Learning Algorithms and Applications

Langtao Chen, Spring 2021

5. Resampling Methods

OUTLINE

- (I) Resampling methods
  1. Cross-validation
     - I. K-fold cross validation
     - II. Leave-one-out cross validation
  2. Bootstrap

- (II) Implementing resampling methods in R
  1. Using caret package for quick experimentation
  2. Directly implement the logic for more detailed control

AGENDA

- Introduction to Resampling Methods
- Using Caret Package
- Repeated K-Fold Cross-Validation
- Bootstrap

Resampling Methods

- Resampling methods involve drawing samples from a training set and refitting a model on each sample.

- The objective of resampling is to obtain additional information about the fitted model.

- In this section, we'll discuss two most commonly used resampling methods:
  - Cross-validation
  - Bootstrap

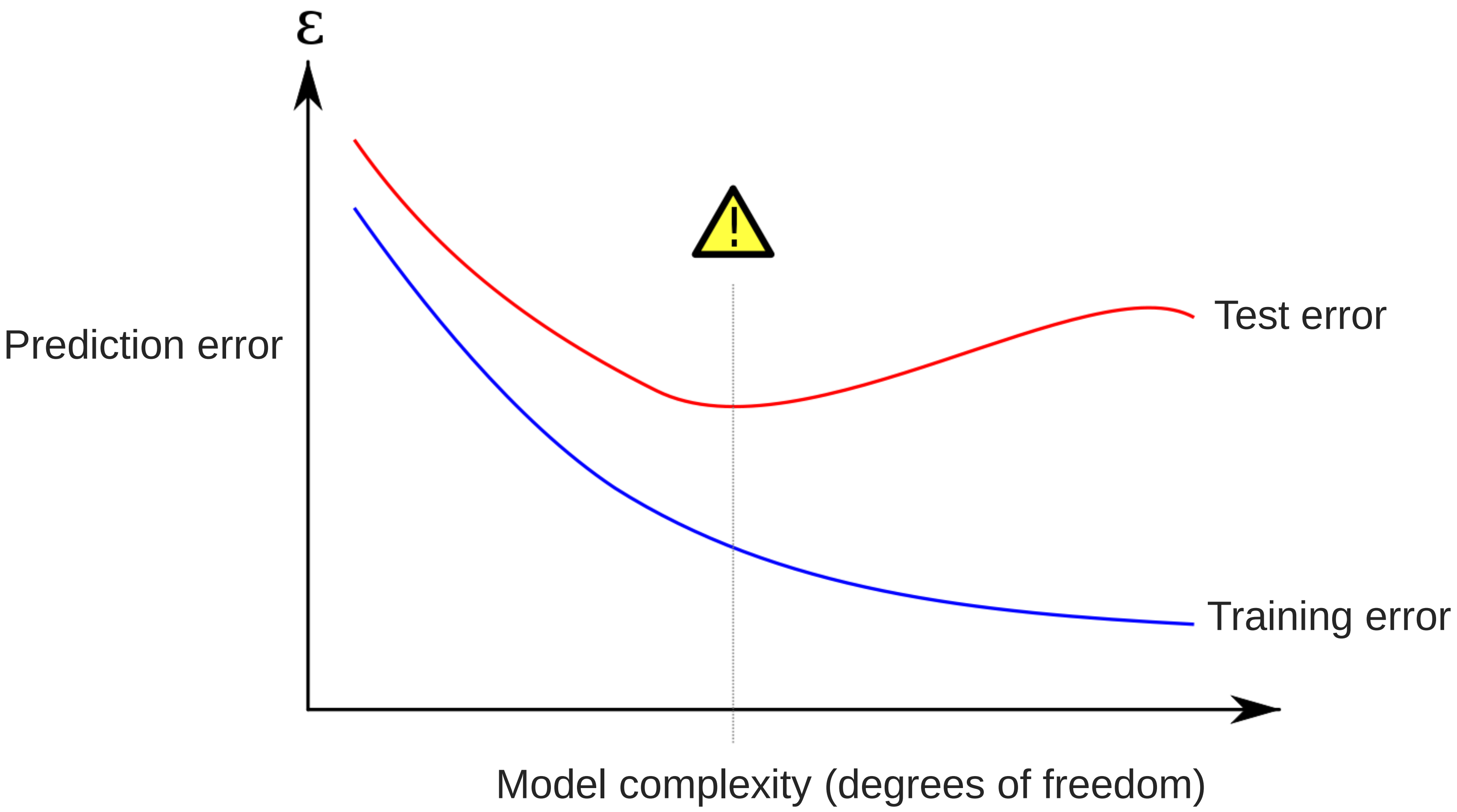
Training Error vs. Test Error

- Test error is the average error that results from using a statistical learning method to predict the response on a new observation—a measurement that was not used in training the method.

- Training error can be easily calculated by applying the statistical learning method to the observations used in its training.

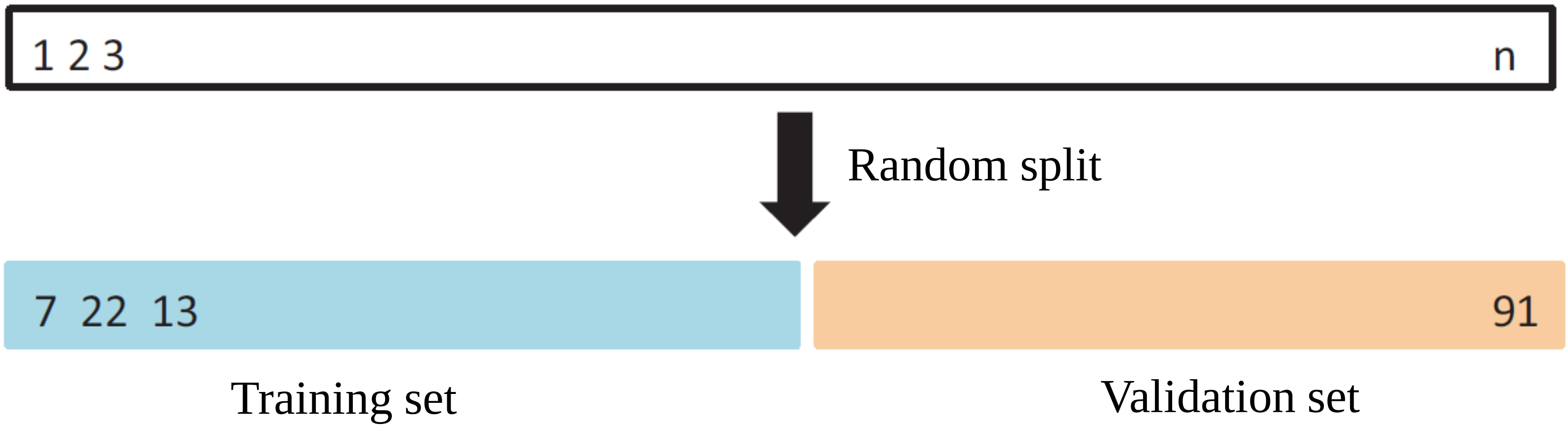
- The training error rate often is quite different from the test error rate, and in particular the former can dramatically <u>underestimate</u> the latter.

Training Error vs. Test Error



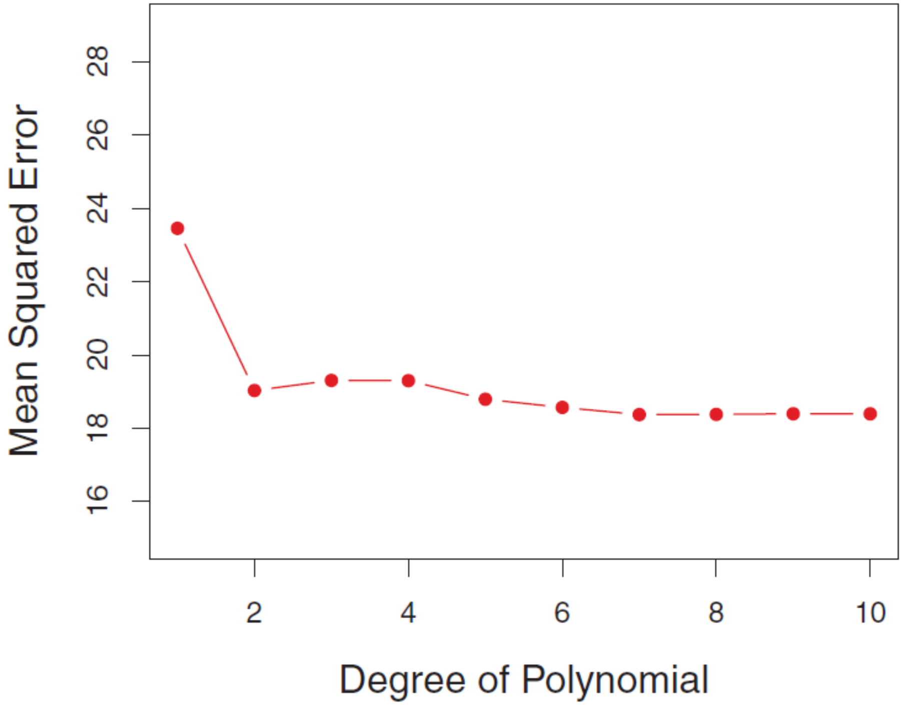
Validation Set Approach

- When we don't have a large designated test set, what can we do?
- Randomly divide the available set of observations into two parts, a *training set* and a *validation set* or *hold-out 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 rate provides an estimate of the test error rate.

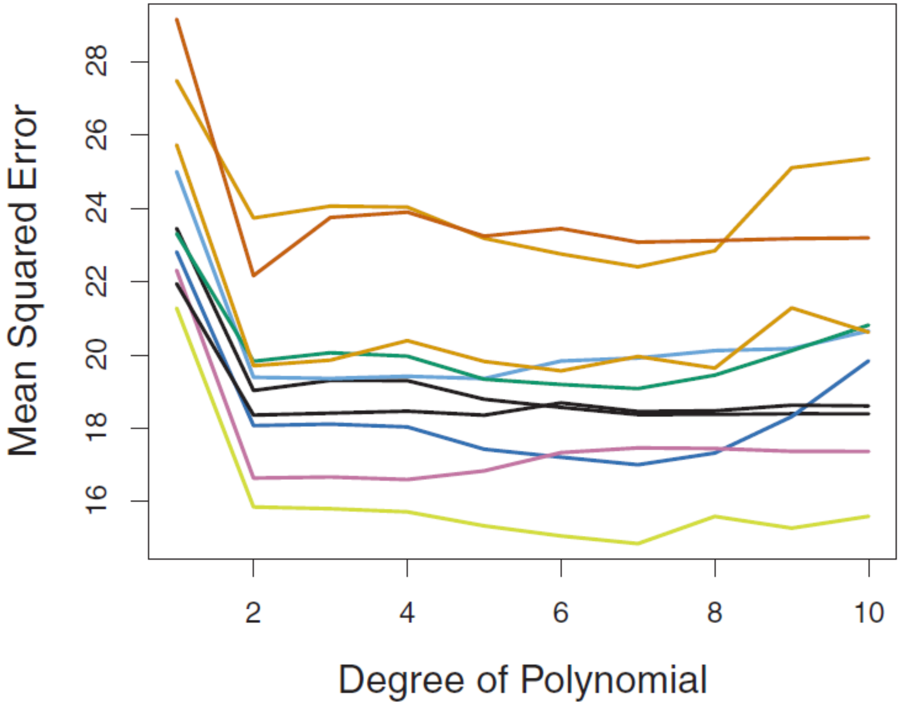


Validation Set Approach on Auto Dataset

- Predict mpg using polynomial functions of horsepower



A single random split



Repeat the process ten times

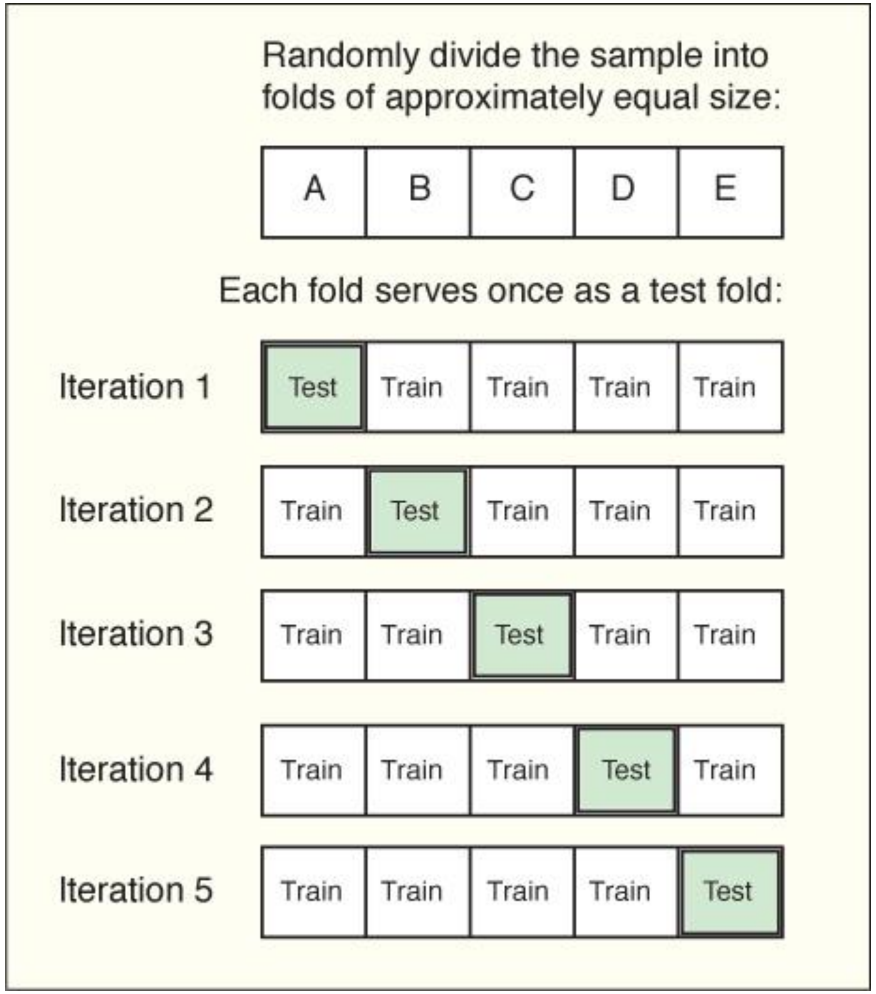
Summary of Validation Set Approach

- Advantages
  - Simple and easy to implement

- Disadvantages
  - The estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
  - Only a subset of the observations (included in the training set) are used to fit the model.
  - This suggests that the validation set error tends to overestimate the test error for the model fit on the entire data set.

K-Fold Cross-Validation

A 5-fold cross validation



In practice, k = 5 or 10.
Magical k?

Cross-Validation Error Rate:

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

Leave-One-Out Cross-Validation (LOOCV)

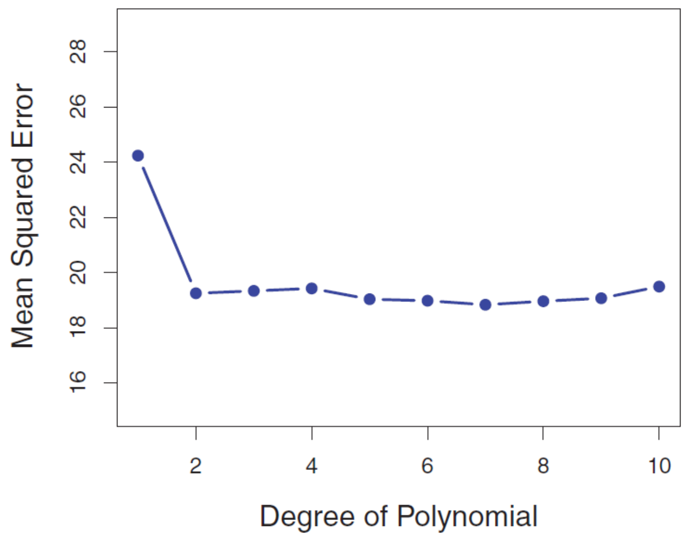
- Set k=n, it is called n-fold or leave-one-out cross-validation (LOOCV)
- Each instance in turn is left out, and the model is trained on all remaining instances.

- Advantages
  - Greatest possible amount of data is used for training.
  - Tends to have lower bias than k-fold cross-validation.
  - The procedure is deterministic: no random sampling is involved, obtain the same result each time.
- Disadvantages
  - Computationally expensive
  - Nonstratified sample (only one instance in the validation/test set) => May lead to poor performance
  - Tends to have higher variance than k-fold cross-validation (bias-variance tradeoff).

LOOCV vs. k-Fold CV and Validation Set

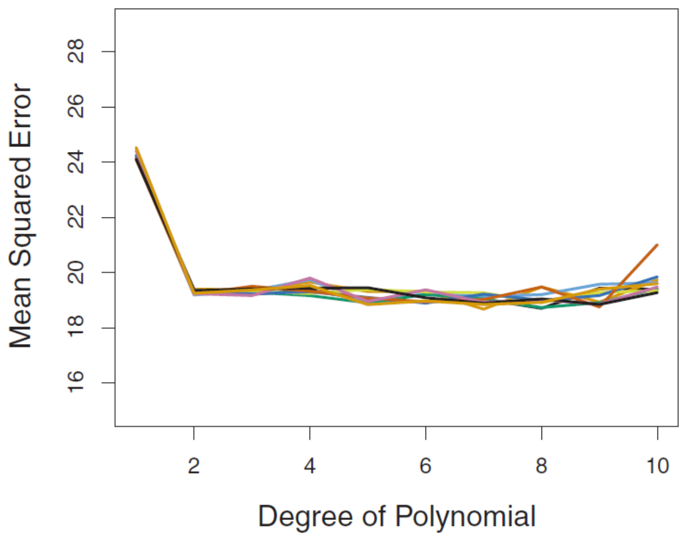
- Predict mpg using polynomial functions of horsepower

LOOCV



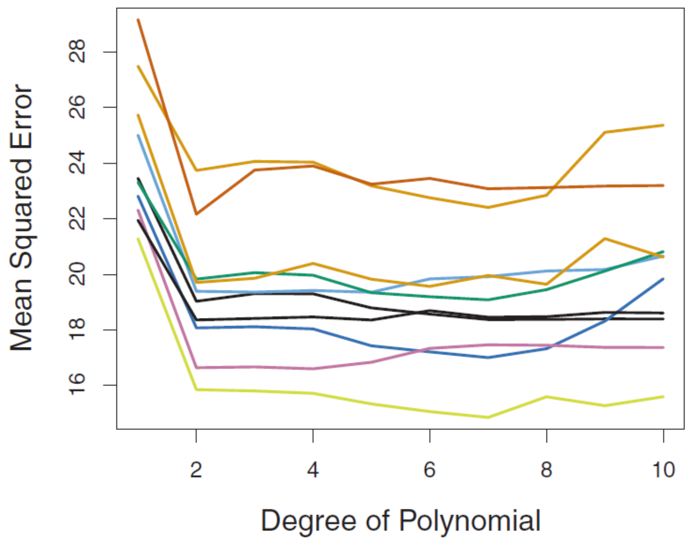
deterministic

K-fold CV (repeated)



low variability

Validation Set (repeated)



high variability

Why k = 5 or 10?

- Computational advantage compared with a large k or $k = n$ (LOOCV).

- Bia-variance trade-off
  - If k is too small, a large portion of the data is not used to train the model. The estimate of prediction error tends to be biased upward.
  - If k is too large, the bias can be reduced. However, the estimated prediction error tends to have a large variance.
  - k = 5 or 10 provides a good trade-off between bias and variance.

Cross-Validation on Classification Problems

- So far, we have discussed cross-validation in regression setting.

- The procedure would be similar in classification setting where the outcome is qualitative, except that we use misclassification rate to quantify test error.
  - K-Fold Cross-Validation

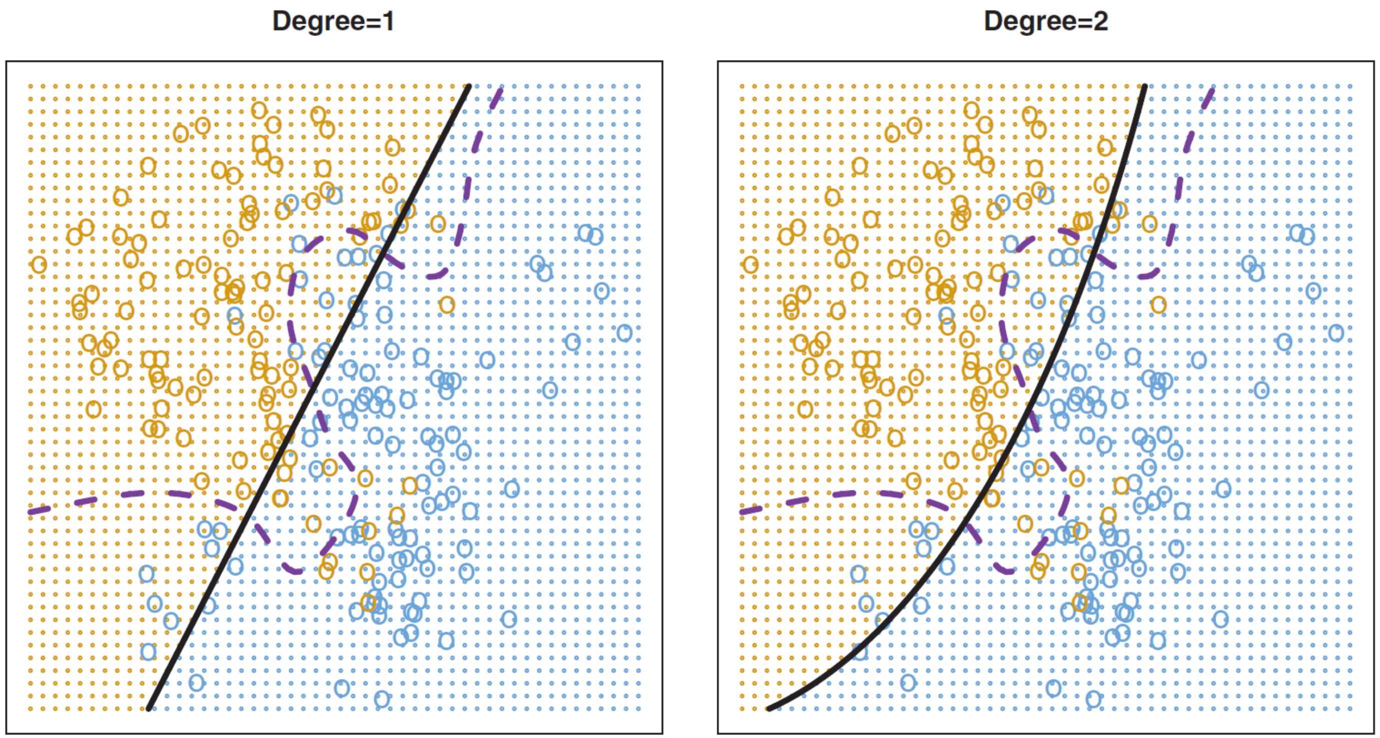
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i \qquad \text{where } Err_i = I(y_j \neq \hat{y}_j)$$

  - LOOCV

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i$$

Use CV to Select the Best Model (or Tune Hyperparameters)

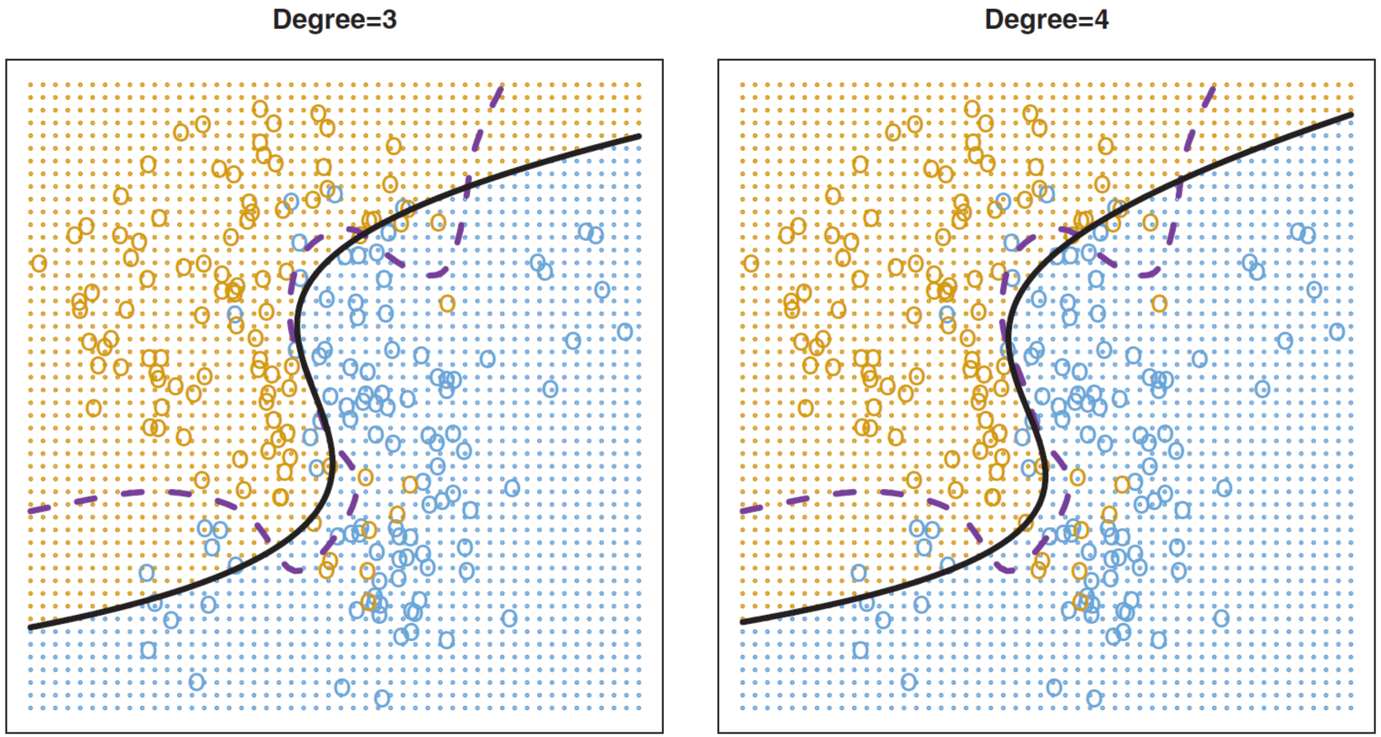
- Select the order of polynomial in logistic regression



Linear Logit
Test Error: 0.201

Quadratic Logit
Test Error: 0.197

Purple dashed: Bayes decision boundary
Black: polynomial logit



Purple dashed: Bayes decision boundary
Black: polynomial logit

Cubic Logit
Test Error: 0.160

Quartic Logit
Test Error: 0.162

Overall Process of Cross Validation for Hyperparameter Tuning

- Some machine learning algorithms have hyperparameters that cannot be directly estimated from the data.
- Cross-validation provides a simple way to tune parameters.

```
Define a grid of parameter values
for each parameter value do
    for each cross-validation iteration do
        Hold-out specification samples
        [Optional] Pre-process the data
        Fit the model on the remainder
        Predict the hold-out samples
    end
    Calculate the average performance across all iterations
end
Determine the optimal parameter value
Fit the final model to all training data using the optimal parameter value
```

Nested Cross-Validation

- Cross-validation can be used to only tune hyperparemeters, or estimate performance of the model.
- It can also be used in a nested structure, to both tune hypermarameters and estimate performance.

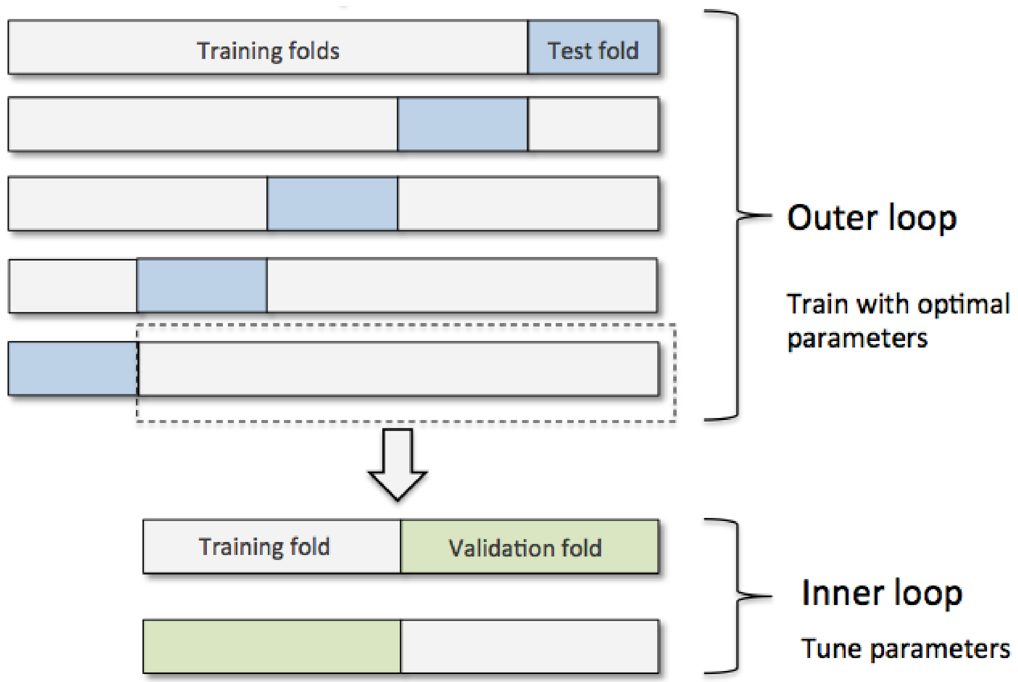


Image: https://sebastianraschka.com/faq/docs/evaluate-a-model.html

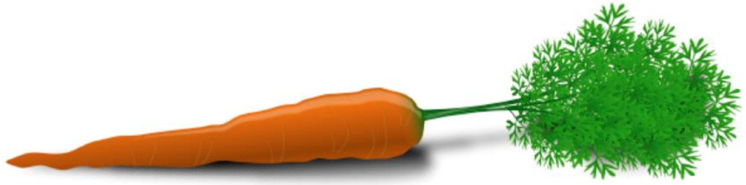
- Use inner loop to tune hyperparameters
- Use outer loop to estimate performance

AGENDA

- Introduction to Resampling Methods
- Using Caret Package
- Repeated K-Fold Cross-Validation
- Bootstrap

Use caret R Package

- caret = classification and regression training
- The caret package is a set of functions that attempt to streamline the process for creating predictive models.
- The package contains tools for:
  - data splitting
  - pre-processing
  - feature selection
  - model tuning using resampling
  - variable importance estimation
  - ……



- To learn more, visit http://topepo.github.io/caret/index.html

Validation Set (Simple Split)

- A single 80/20% split of the corolla data

```
# Read data file
df <- read.csv("Telco-Customer-Churn.csv")

# Use caret package
library(caret)

# Data partition
set.seed(1234)
trainIndex <- createDataPartition(df$Churn, p = .8, list = FALSE)
head(trainIndex)

train_data <- df[ trainIndex,]
test_data  <- df[-trainIndex,]
```

Advanced Modeling Training/Tuning

- Use caret::train() to tune model parameters

```
Define a grid of parameter values
for each parameter value do
    for each cross-validation iteration do
        Hold-out specification samples
        [Optional] Pre-process the data
        Fit the model on the remainder
        Predict the hold-out samples
    end
    Calculate the average performance across all iterations
end
Determine the optimal parameter value
Fit the final model to all training data using the optimal parameter value
```

Why the final model is fitted to all training data in the final step?

K-Fold Cross Validation

- Use 5-fold Cross-Validation

```
fitControl <- trainControl(method = "cv",number = 5)

set.seed(123)
svmRadial_fit <- train(Churn ~ ., data = train_data[-1],
                     trControl = fitControl,
                      method = "svmRadial")
print(svmRadial_fit)
```

The train() method in caret only support a few performance measures (overall accuracy and kappa for classification, RMSE, MAE, and R^2 for regression). If you need other measures, you can implement the k-fold cross-validation by your own code.

Example

- Customer Churn Analysis

AGENDA

- Introduction to Resampling Methods
- Using Caret Package
- Repeated K-Fold Cross-Validation
- Bootstrap

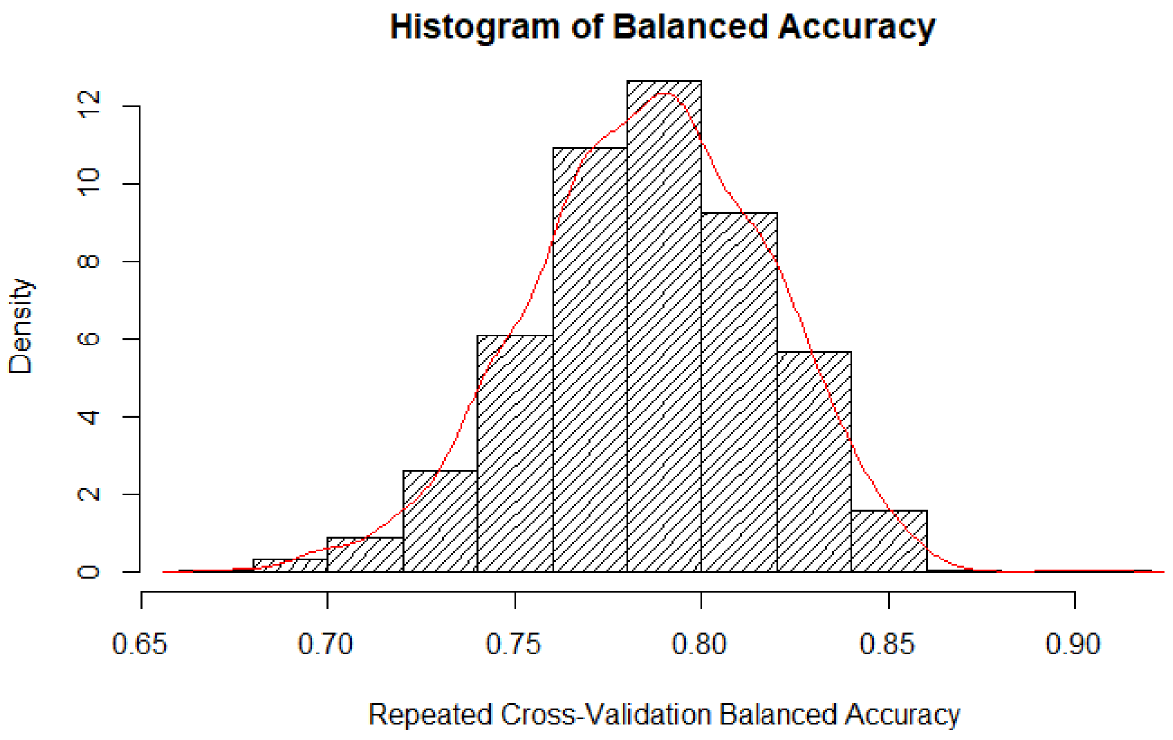
Extension on k-Fold Cross-Validation

- Repeated k-Fold Cross-Validation
  - Alleviate the random effect due to resampling

- Repeated Stratified k-Fold Cross-Validation
  - Alleviate the random effect due to resampling
  - Make sure the folds are balanced, in order to provide a more accurate performance estimate

Repeated K-Fold Cross Validation

- K-fold cross validation does not provide a robust estimate of mean performance.
- We can repeat the k-fold cross validation multiple times to better estimate performance.



How to Implement Repeated K-Fold Cross Validation

- Method 1: Using caret, you can set the train control as:

```
trainControl(method = "repeatedcv",
             number = 5,
             repeats = 200)
```

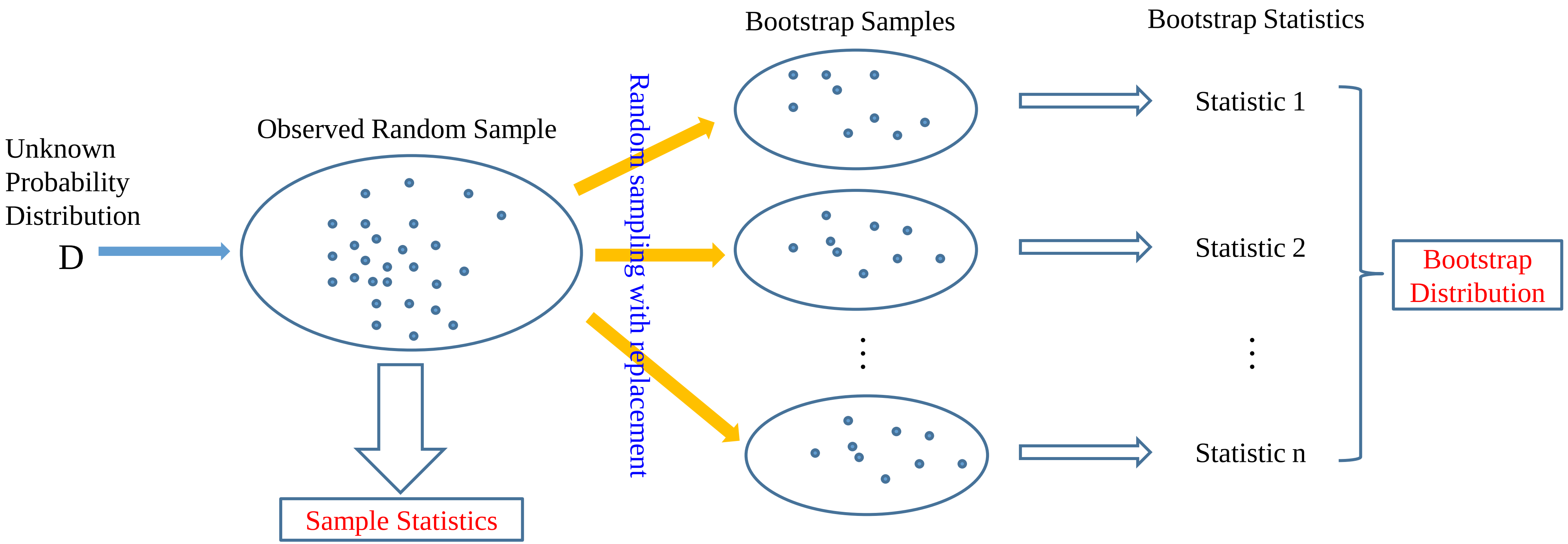
- Method 2: Directly implement the logic by your own code.
  - Refer to Example: Titanic Survival Analysis

Agenda

- Introduction to Resampling Methods
- Using Caret Package
- Repeated K-Fold Cross-Validation
- Bootstrap

Bootstrap

- Bootstrap provides a general way for quantifying uncertainty of a statistical method based on random sampling with replacement.



Bootstrap distribution is usually closer to true distribution than sample statistics.

Implementing Bootstrap

- Refer to Example: Bootstrap

Summary of Bootstrap

- Bootstrap is based on the law of large numbers: if we sample enough times, we can approximate the true population distribution.

- The number of bootstrap samples should be large (e.g., 1000).

- Bootstrap can easily derive standard errors and confidence intervals for complicated statistics. Hypothesis testing can be very simple.

- Bootstrap works for small sample.

Recap: Outline

- (I) Resampling methods
  1. Cross-validation
     - I. K-fold cross validation
     - II. Leave-one-out cross validation
  2. Bootstrap

- (II) Implementing resampling methods in R
  1. Using caret package for quick experimentation
  2. Directly implement the logic for more detailed control

Q & A