Introduction to Cross Validation[1]

Lecture 1: Overview

Chi-Kuang Yeh[2]
McGill University and University of Waterloo

November 26, 2024

[1] https://chikuang.github.io/course/directstudy/
[2] https://chikuang.github.io/

Today's Agenda

Today (Lec 1):

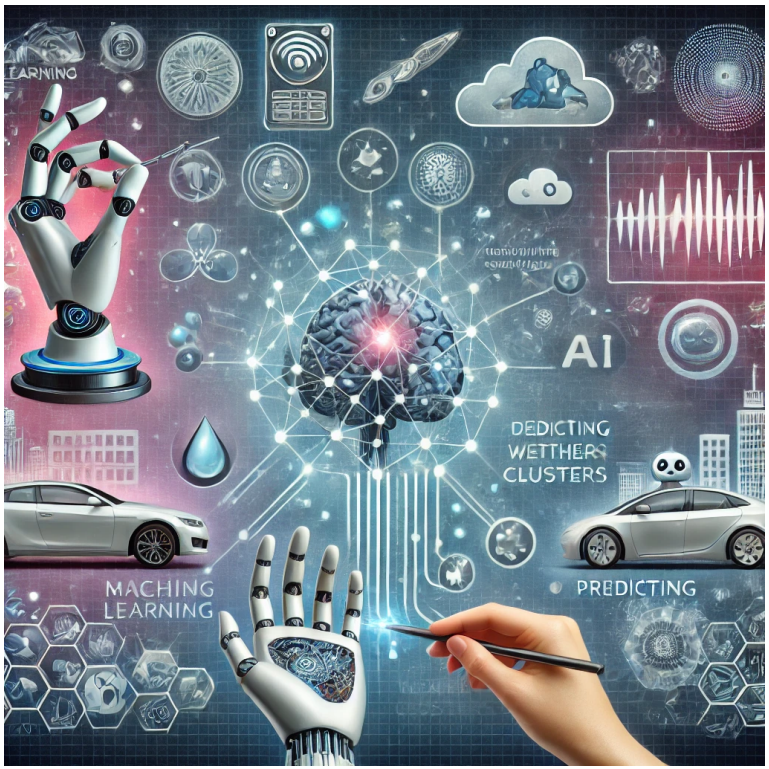
- Background
- Cross-Validation
- Application

Note: This lecture is based on the book by Hastie et al. (2009), and James et al. (2013).

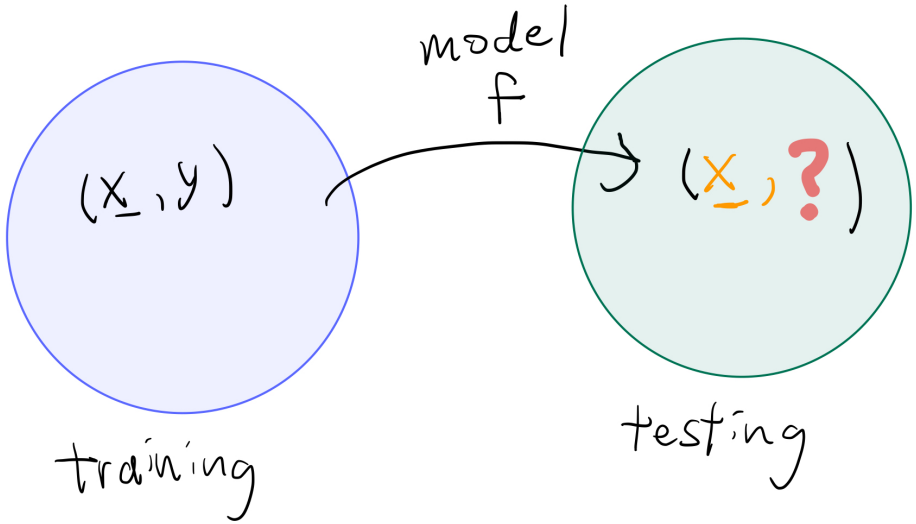
Machine Learning Models

Left: What machine learning can do

Right: Model/Methods



Training and testing/validating sets



Questions

How do we choose between different models $f_1, \ldots, f_m$?

Setup

Suppose we have a **supervised learning** model $f(X) \to Y$.

Denote the training set by $\mathcal{T} = \{(X_i, Y_i)\}_{i=1}^{N_{train}}$.

How to choose between the models $f_1, \ldots, f_m$?

Ideal: $(X, Y) \sim F_{X,Y}$.

Define the generalization error (Population error) as

$$\mathrm{Err}(f) = E_{X,Y}[(Y - f(X))^2]$$

Choose f by

$$\arg \min_{f \in \{f_1, \ldots, f_m\}} \mathrm{Err}(f)$$

Suppose we have a **supervised learning** model $f(X) \to Y$.

Denote the training set by $\mathcal{T} = \{(X_i, Y_i)\}_{i=1}^{N_{train}}$.

How to choose between the models $f_1, \ldots, f_m$?

Ideal: $(X, Y) \sim F_{X,Y}$ (**unknown**).

Define the generalization error (Population error) as

$$\mathrm{Err}(f) = E_{X,Y}[(Y - f(X))^2]$$

Choose f by

$$\arg \min_{f \in \{f_1, \ldots, f_m\}} \mathrm{Err}(f)$$

What to do if we do not know about F_{X_Y}?

Let $V := \{(X_i, Y_i)\}_{i=1}^{N_{Val}}$ be the validating/testing set.

$$\mathrm{Err}(f) = E[(Y - f(X))^2] \approx \frac{1}{N_{Val}} \sum_{i=1}^{N_{Val}} (Y_i - f(X_i))^2 =: \mathrm{err}_{Val}(f)$$

When $N_{Val} \to \infty$, $\mathrm{err}_{Val}(f) \to \mathrm{Err}(f)$.

We can then do

$$\arg \min_{f \in \{f_1, \ldots, f_m\}} \mathrm{err}_{Val}(f)$$

If N_{val} is large, we can have good estimate of $\mathrm{Err}(f)$.

But actually, we may only have *small* validating set.

Problems with simple Train-Test Split:

- Splitting 50-50 wastes data that could improve the model.
- Splitting 80-20 may leave too little test data for reliable evaluation.

But actually, we may only have *small* validating set.

Problems with simple Train-Test Split:

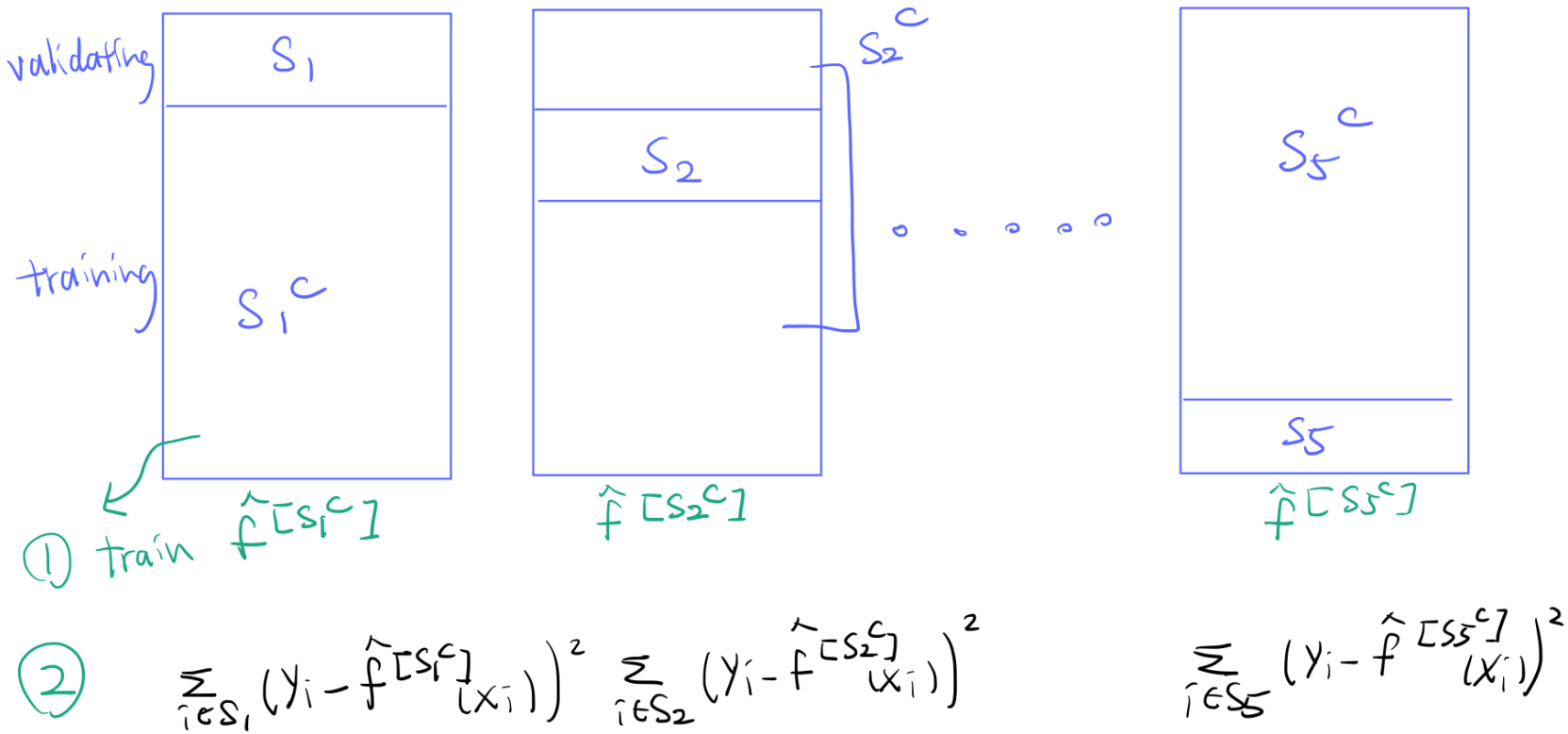
- Splitting 50-50 wastes data that could improve the model.
- Splitting 80-20 may leave too little test data for reliable evaluation.

What we can do? Cross-validation!

Cross-validation uses **all** data efficiently for training and testing.

What is Cross-Validation?

Suppose there are 5 folds ($K = 5$).



Put together, we have

$$\mathrm{err}_{cv}(f) = \frac{1}{N} \sum_{k=1}^{5} \sum_{i \in S_k} (y_i - \hat{f}^{[s_k^C]}(x_i))^2.$$

Then cross-validation is to find

$$f^* = \arg \min_{f \in \{f_1, \ldots, f_m\}} \mathrm{err}_{cv}(f).$$

K-Fold Cross-Validation

Let $\mathcal{D} = \{X_i, Y_i\}_{i=1}^N$ be our data.

1. Split the data into K approximately equal sizes parts/fold K
2. For each $k = 1, 2, \ldots, K$, repeat the following steps:
   (i) Leave the kth fold S_k from the data $\mathcal{D}$, and denote the remaining data as S_k^C. We fit the model to S_k^C and denote the corresponding model we obtained by $\hat{f}^{[S_k^C]}$
   (ii) Calculate the total prediction error on the fitted model $\hat{f}^{[S_k^C]}$ on the left-out fold S_k

$$\mathrm{err}_{cv,k}(f) = \sum_{i \in S_k} L(Y_i, \hat{f}^{[S_k^C]}(X_i)).$$

3. The **CV estimate** of prediction error is

$$\mathrm{err}_{cv}(f) = \frac{1}{N} \sum_{k=1}^{K} \mathrm{err}_{cv,k}(f).$$

So if we have M models, $f_1, f_2, \ldots, f_M$, we can use cross-validation to select the best model by computing the cross-validation error for each model $\mathrm{err}_{cv}(f_1), \mathrm{err}_{cv}(f_2), \ldots \mathrm{err}_{cv}(f_M)$

Then the best model is

$$f^* = \arg \min_{f \in \{f_1, \ldots, f_M\}} \mathrm{err}_{cv}(f).$$

There are two many use of the K-fold CV

1. Tune hyperparameters

   e.g., In linear regression: $y = \beta_0 + \sum_{j=1}^{L} \beta_k x^k + \varepsilon$, L is the hyperparameter here.

   M1. $y = \beta_0 + \beta_1 x_1 + \varepsilon$

   M2. $y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \varepsilon$

   M3. $y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

2. To better evaluate the performance of a model

The number of folds depends on the data size.

Discussion

What would you consider when choosing K?

[3] Hastie et al. (2009).

Discussion

What would you consider when choosing K?

The Bias-Variance Decomposition.

$$\begin{array}{lcll} \text{Generalization error} & = & \text{variance} & + \text{bias} & + \text{irreducible error} \\ \mathbb{E}_{\mathcal{T}}[(y - f(x;\mathcal{T}))^2] & = & \mathbb{V}ar(x) & + \mathbb{B}\text{ias}^2(x) & + \varepsilon^2. \end{array}$$

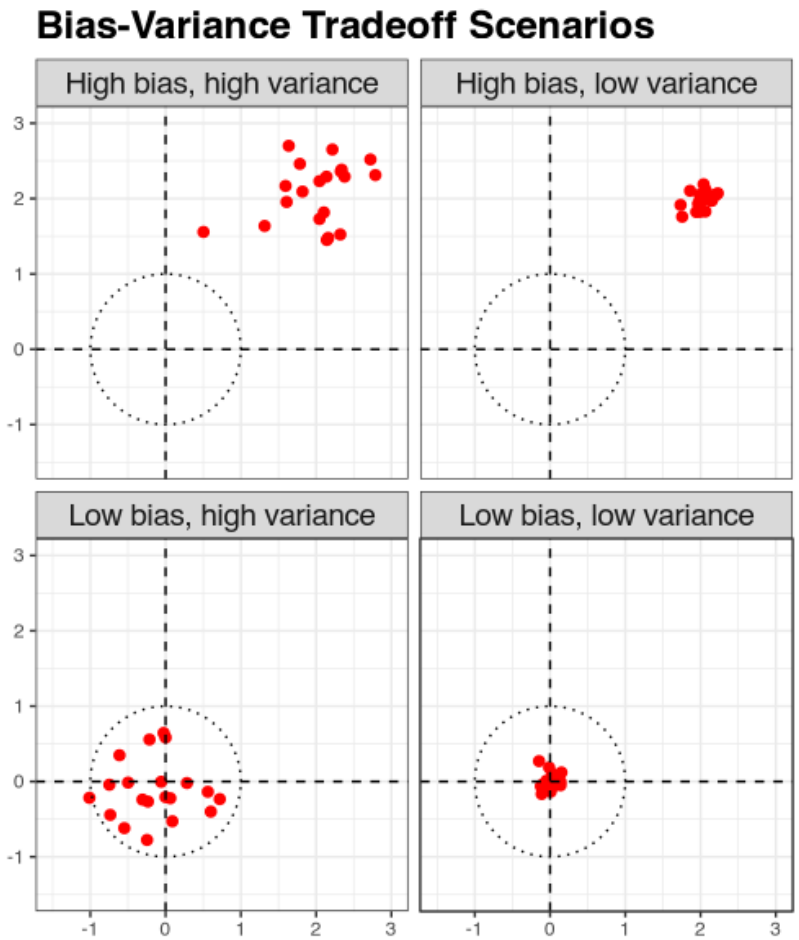
See Section 7.3, Equation (7.9) in [3].

[3] Hastie et al. (2009).

Pop-up quiz

Q: What are the characteristic a good model f should have?

1. Low bias, high variance
2. High bias, low variance
3. Low bias, low variance
4. High bias, high variance
5. None of above



- Bias (Systematic error): The difference between predicted values and the true target.
- Variance (Sensitivity to data changes): How much predictions change with new data.
- Goal: Minimize both bias and variance.

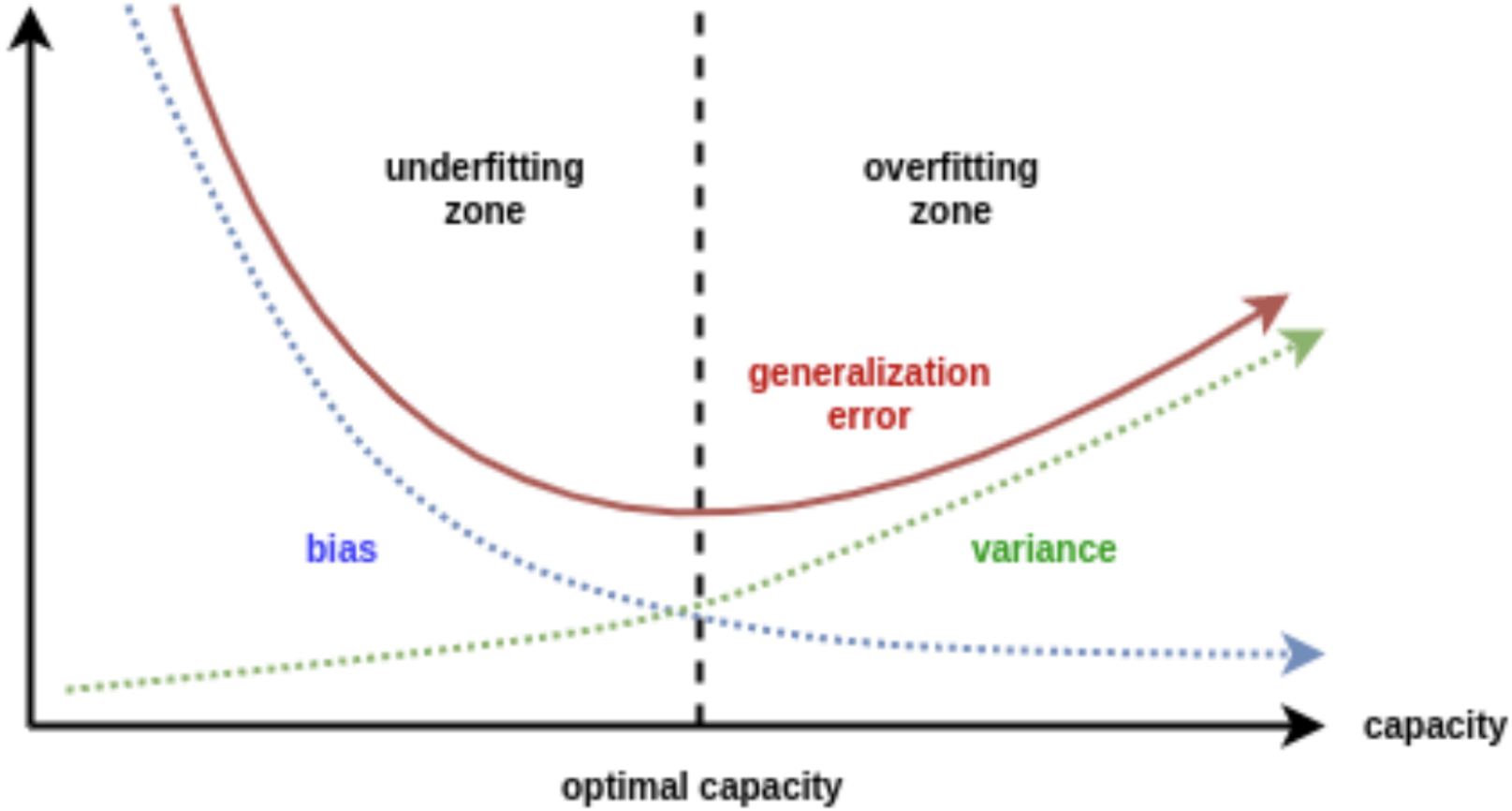
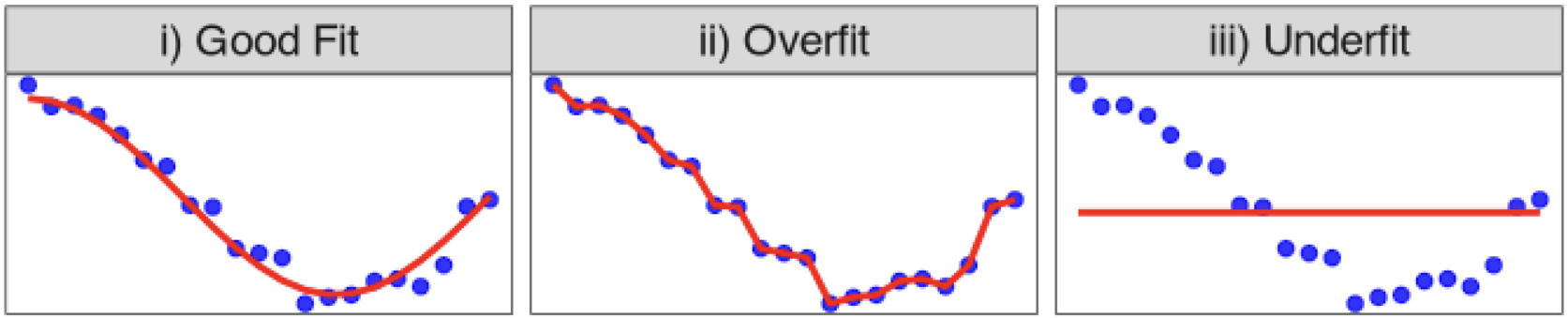
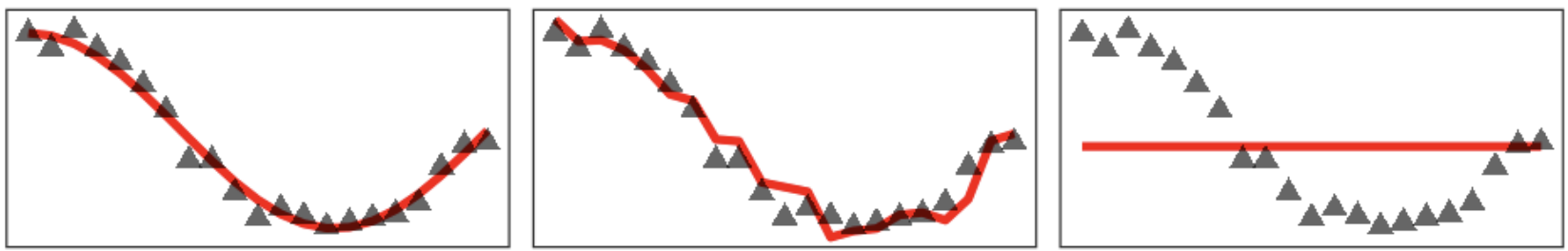


Figure 1: Picture borrowed from [4].

[4] https://djsaunde.wordpress.com/2017/07/17/the-bias-variance-tradeoff/



With new data (X, Y) from the same distribution:



Choice of Fold K

Bias-Variance Tradeoff

The choice of K is a tradeoff between bias and variance.

Choice of Fold K

Bias-Variance Tradeoff

The choice of K is a tradeoff between bias and variance.

Q: What values of $K, 2 \leq K \leq N$ should we use?

Choice of Fold K

Bias-Variance Tradeoff

The choice of K is a tradeoff between bias and variance.

Q: What values of $K, 2 \leq K \leq N$ should we use?

- Large K: high variance, but small bias.
- Small K: low variance, but high bias.

Bias decreases as K increases.

Example: Stock market

We look at the dataset in **Smarket** package in R.

It contains the daily percentage returns for the S&P 500 stock index between 2001 and 2005.

$N = 1250$

Table 1: First Few Rows of Smarket Dataset

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
2001	0.381	-0.192	-2.624	-1.055	5.010	1.1913	0.959	Up
2001	0.959	0.381	-0.192	-2.624	-1.055	1.2965	1.032	Up
2001	1.032	0.959	0.381	-0.192	-2.624	1.4112	-0.623	Down
2001	-0.623	1.032	0.959	0.381	-0.192	1.2760	0.614	Up
2001	0.614	-0.623	1.032	0.959	0.381	1.2057	0.213	Up
2001	0.213	0.614	-0.623	1.032	0.959	1.3491	1.392	Up

Leave-One-Out Cross-Validation (LOOCV)

- When $K = N$, the size of the training data, it is leave-one-out cross validation.
- Instead of creating two subsets of comparable size, a single observation (x_i, y_i) is used for the validation set and the remaining observations make up the training set.
- Repeat this for each observation and get the average.

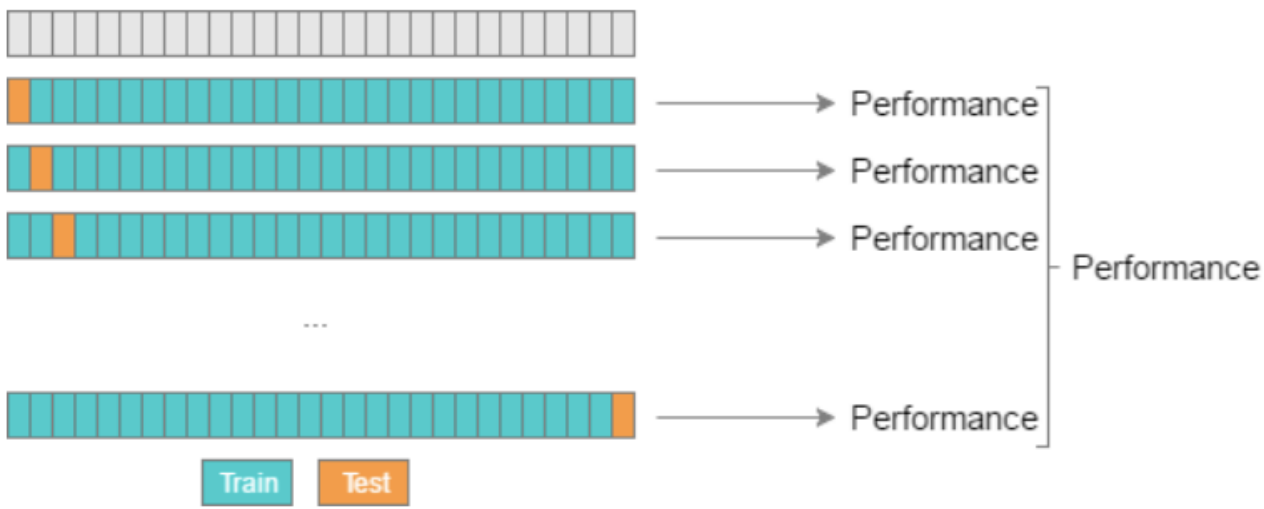
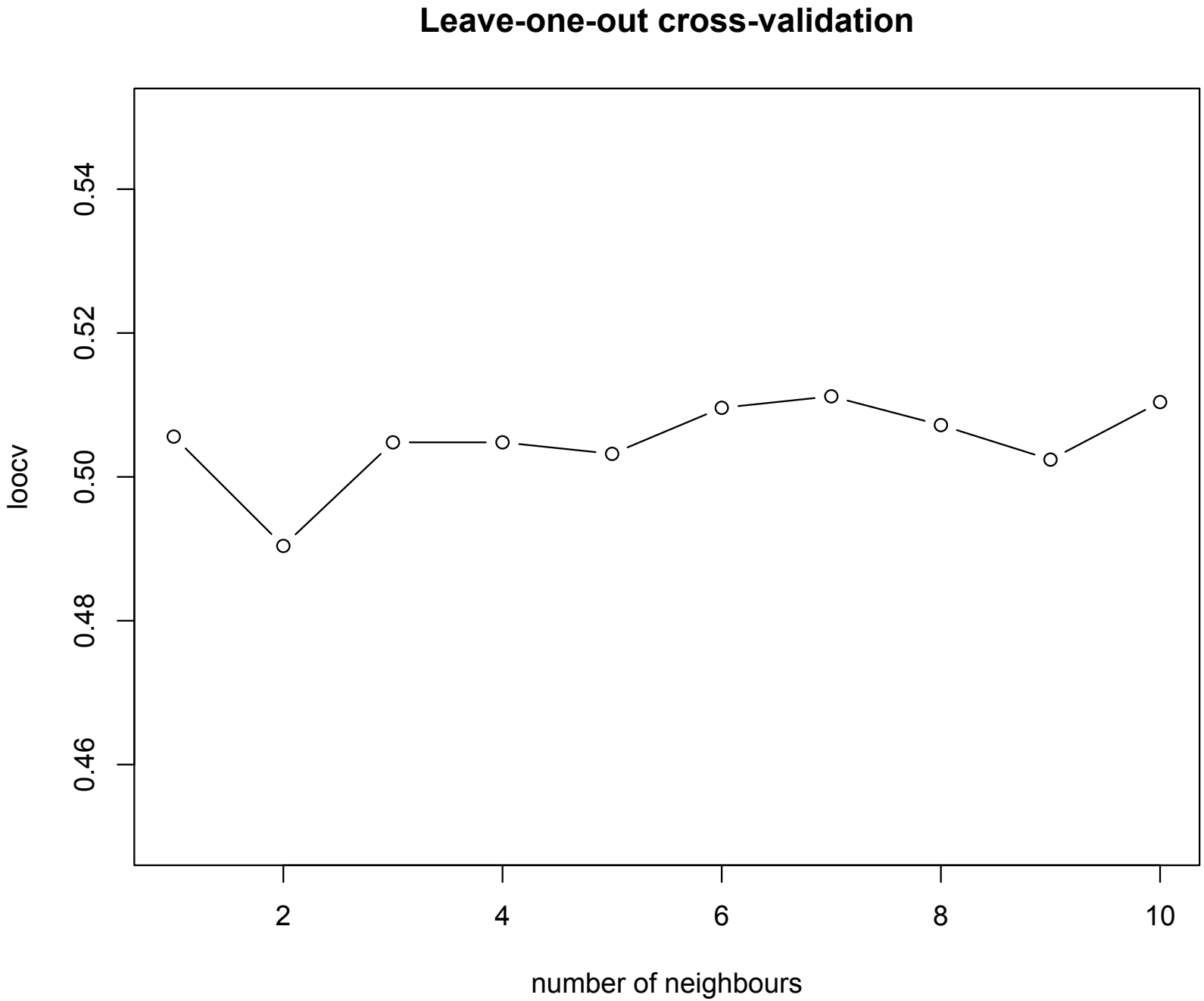


Figure 2: Illustration for Leave one out CV.


Leave-one-out cross-validation
loocv
0.46
0.48
0.50
0.52
0.54
2
4
6
8
10
number of neighbours

5-fold CV

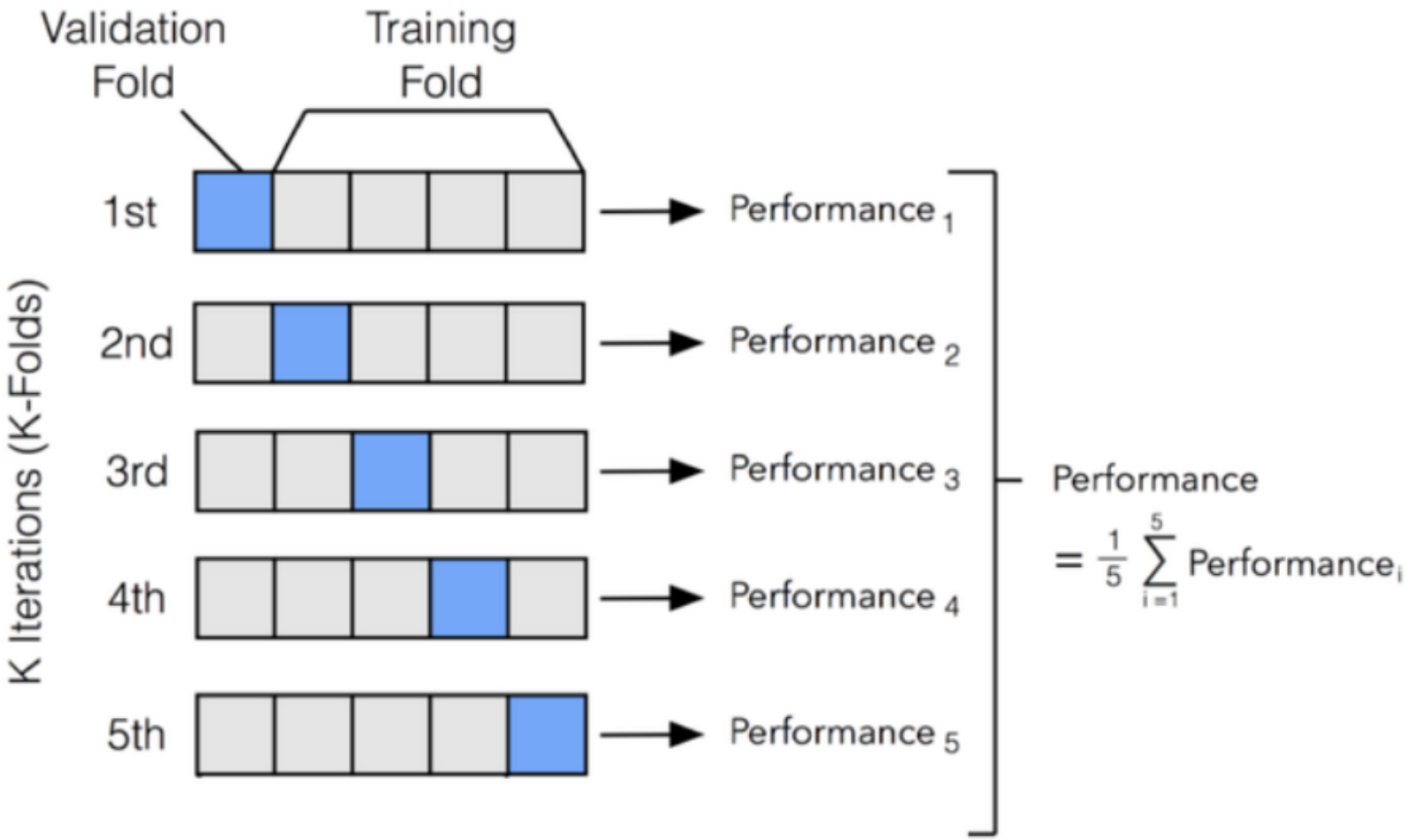
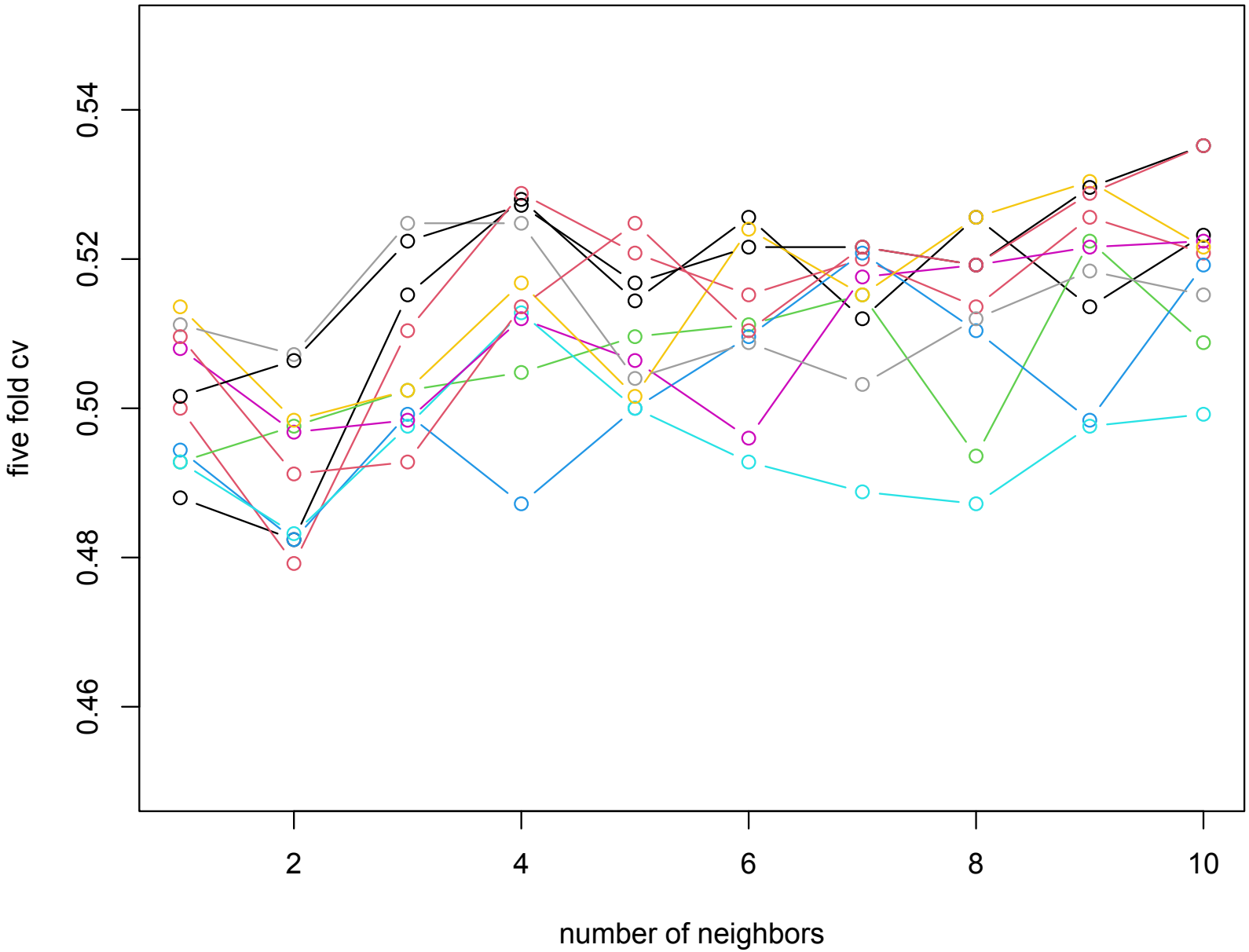


Figure 4: 5 fold CV.


five fold cross-validation
five fold cv
0.46
0.48
0.50
0.52
0.54
2
4
6
8
10
number of neighbors

Take Home Messages

What is CV?

- A method to estimate prediction error using all data efficiently.

Why K-fold?:

- Balances bias and variance effectively.

LOOCV:

- Special case with $K = N$, unbiased but expensive.

Practical Tips:

- $K = 5$ or $K = 10$ is common and (usually) works well.
- Use CV to tune hyperparameters and compare models.

Note:

Cross-validation can be applied in various contexts!

Tutorial

We will be using the `sklearn` package in Python to demonstrate how to use cross-validation to choose the best hyperparameter on Friday's lab.

Find a lab partner!

Reading Section 7.10 in Hastie et al. (2009).

References

1. Hastie, T., Tibshirani R.J. and Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

2. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York.

3. James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023). *Introduction to Statistical Learning with Applications in Python*. Springer, New York.

4. Bates, S., Hastie, T. and Tibshirani, R. (2023). Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of American Statistical Association*, 119, 1434–1445.