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

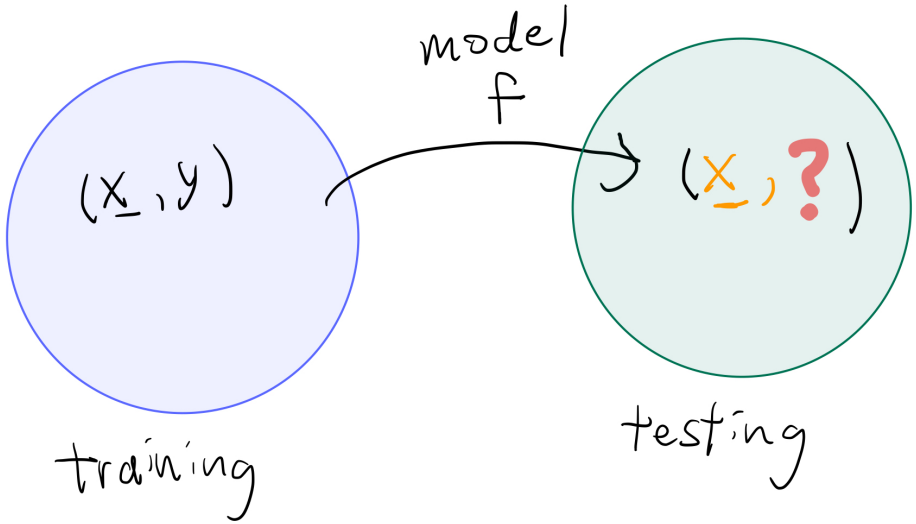
Today (Lec 1):

- Background
- Cross-validation
- Example

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

Machine learning models

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 $f(X) \to Y$.

$\mathcal{T} = \{X_i, Y_i\}_{i=1}^{N_{train}}$

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

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

Define the generlization error (Population error) as

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

Choose f by

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

How to test the model using generalised error?

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

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

When $N_{Val} \to \infty$, $err_{Val}(f) \to Err(f)$.

We can then do

$$\arg \min_{f \in \{f_1, \ldots, f_m\}} err_{Val}(f)$$

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

But actually, we have *small* validating set.

What we can do? Cross-validation!

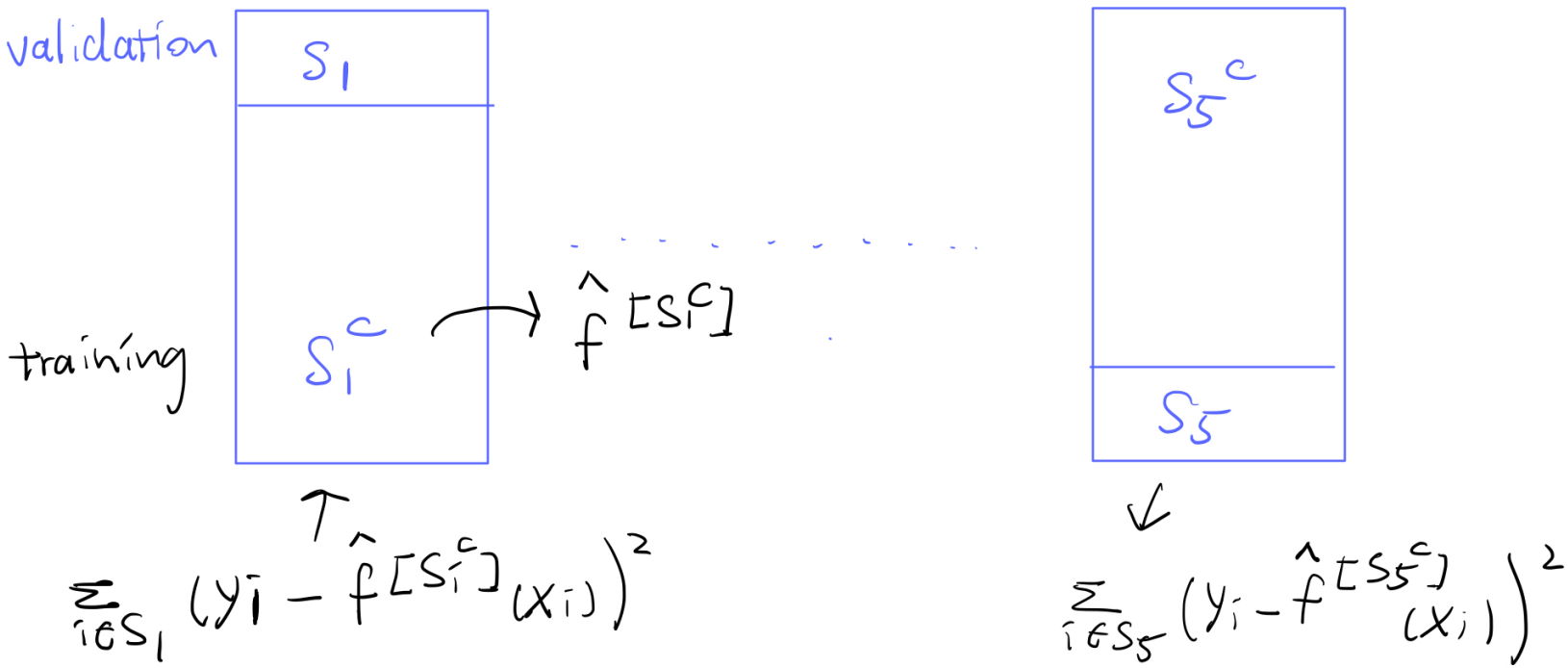
Problems with Simple Train-Test Split:

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

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

What is Cross-Validation?

Suppose $K = 5$.



Put together, we have

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

Then cross-validation is to find

$$\hat{f} = \arg \min_{f \in \{f_1, \ldots, f_m\}} err_{cv}(f)$$

Understanding K-Fold Cross-Validation I

There are two many use of the K-fold CV

1. Tune hyperparameters
2. To better evaluate the performance of a model

The number of folds depends on the data size.

Understanding K-Fold Cross-Validation II

Let $\mathcal{T} = \{X_i, Y_i\}_{i=1}^N$ be the training 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 (i.e., part) from the data $\mathcal{T}$, and denote the remaining data as S_k^C. We fit the model to S_k 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

$$cv_k = \sum_{i \in S_k} L(Y_i, \hat{f}^{S_k^C}(X_i)).$$

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

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

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 $\mathrm{err}_{cv}(\hat{f}_1), \mathrm{err}_{cv}(\hat{f}_2), \ldots \mathrm{err}_{cv}(\hat{f}_M)$

Discussion

What would you consider when choosing K?

Note

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

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

Key Concepts: Bias and Variance

Bias

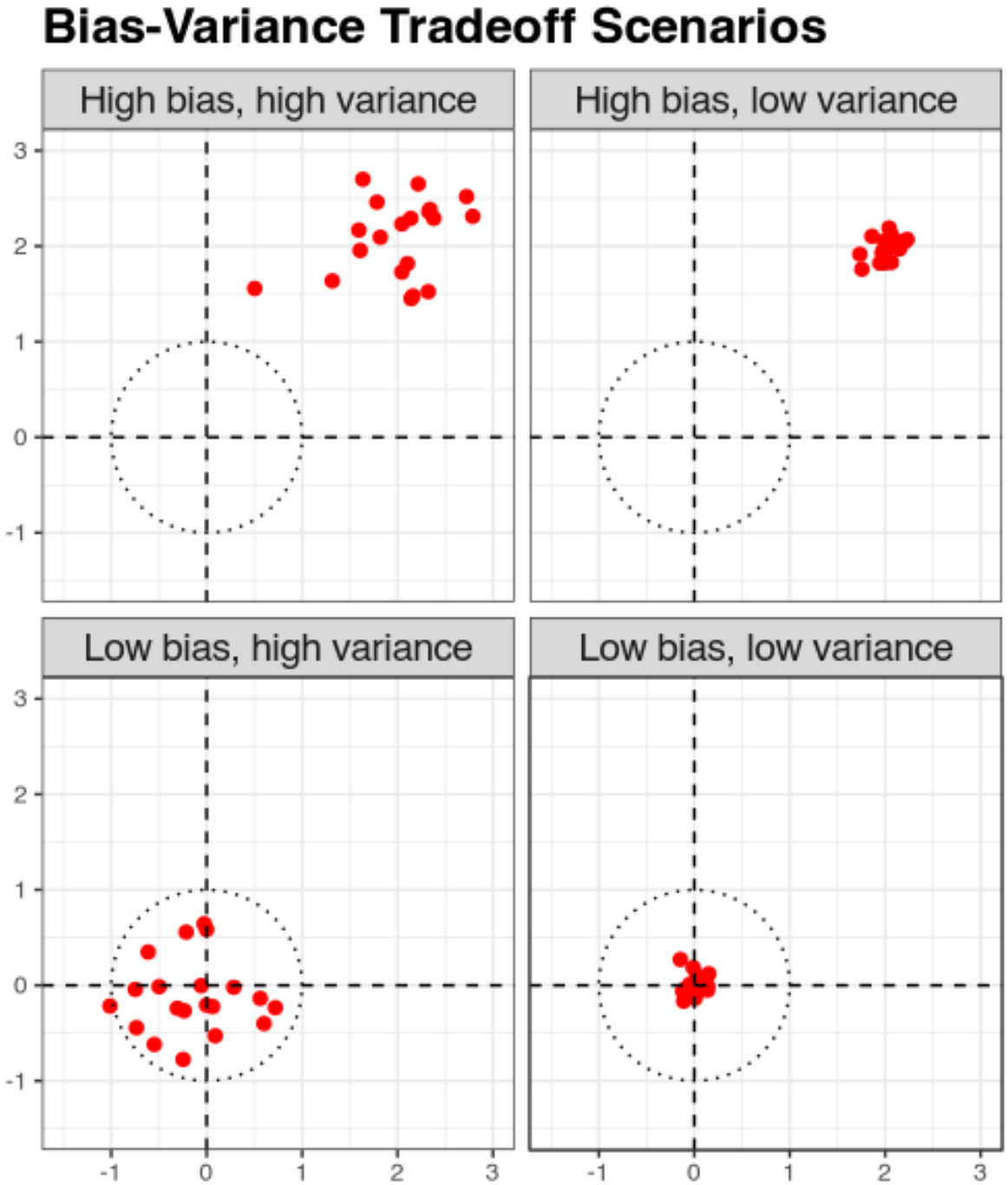
- Systematic error: The difference between predicted values and the true target.
- Example: A model missing the target entirely.

Variance

- Sensitivity to data changes: How much predictions change with new data.
- Example: Hitting different spots on the target each time.

Goal

- Minimize both bias and variance.


Bias-Variance Tradeoff Scenarios
High bias, high variance
High bias, low variance
Low bias, high variance
Low bias, low variance
3
2
1
0
-1
-1
0
1
2
3

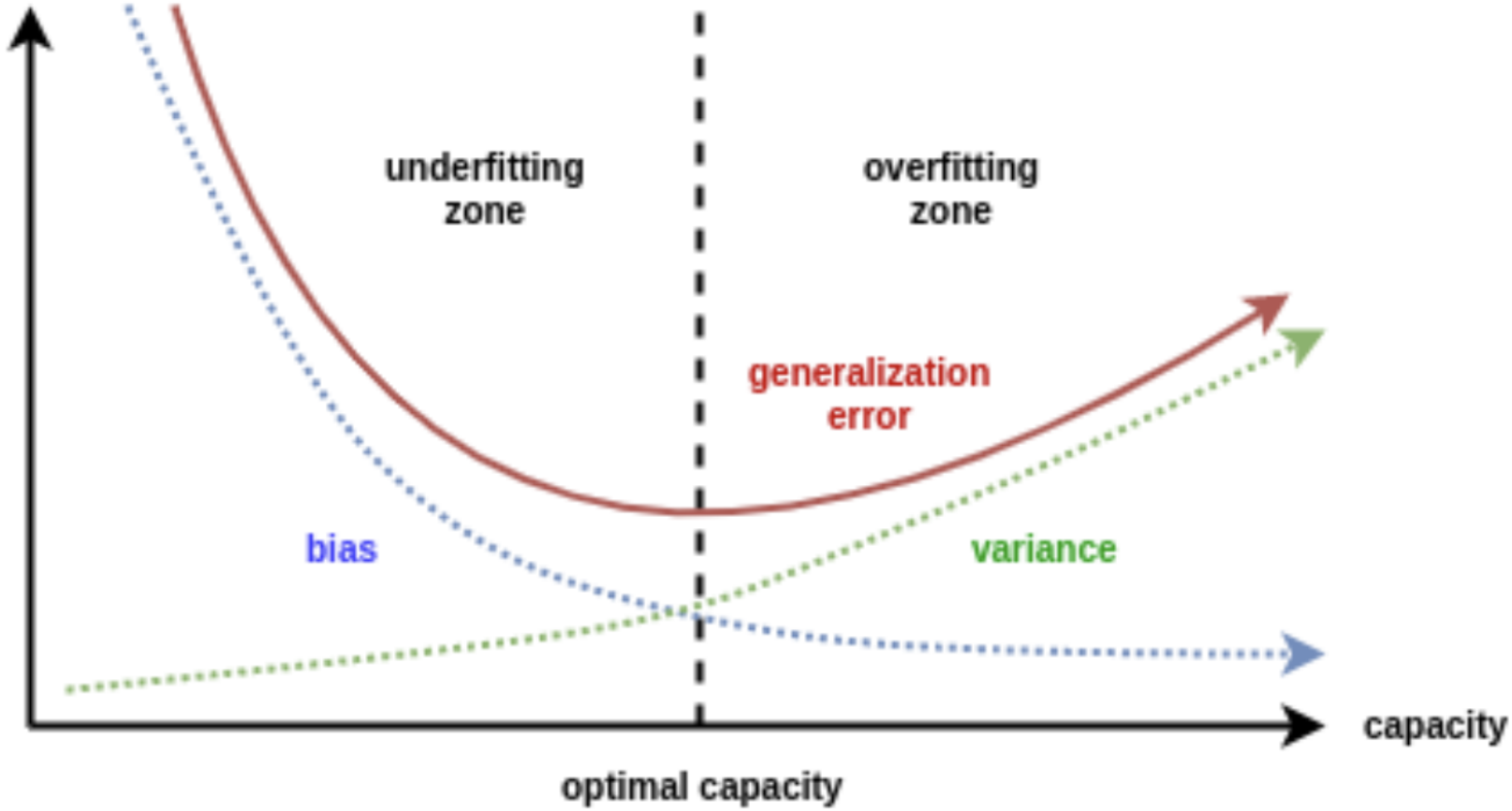
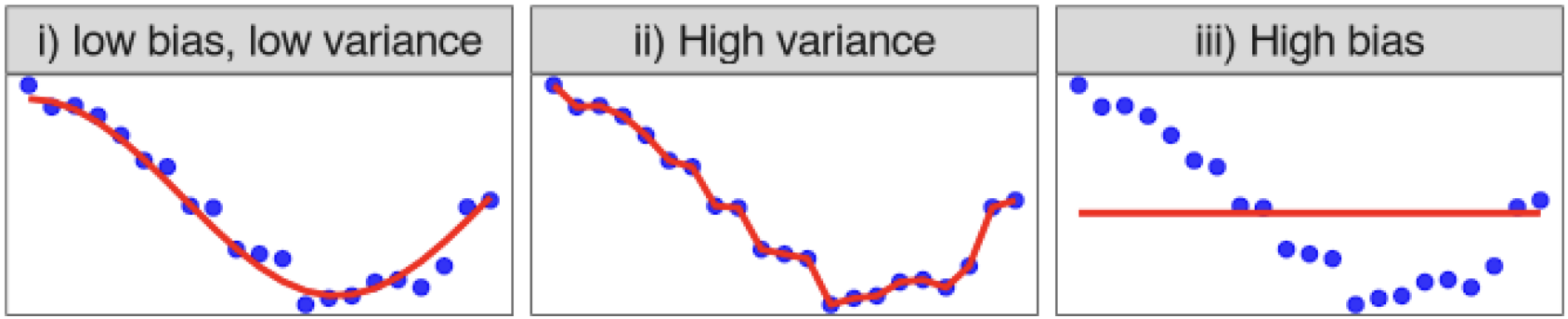
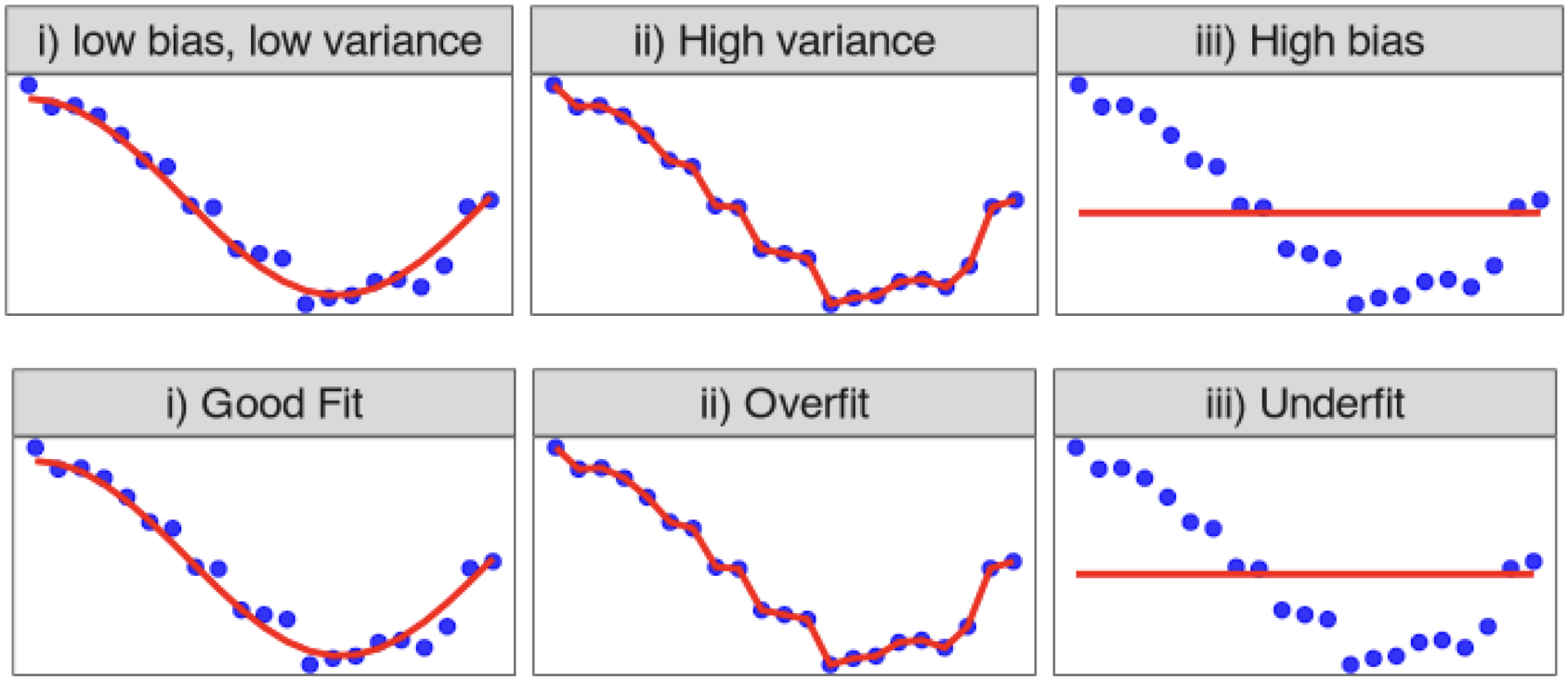


Figure 1: Picture borrowed from [3].

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


i) low bias, low variance
ii) High variance
iii) High bias


i) low bias, low variance
ii) High variance
iii) High bias
i) Good Fit
ii) Overfit
iii) Underfit

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.

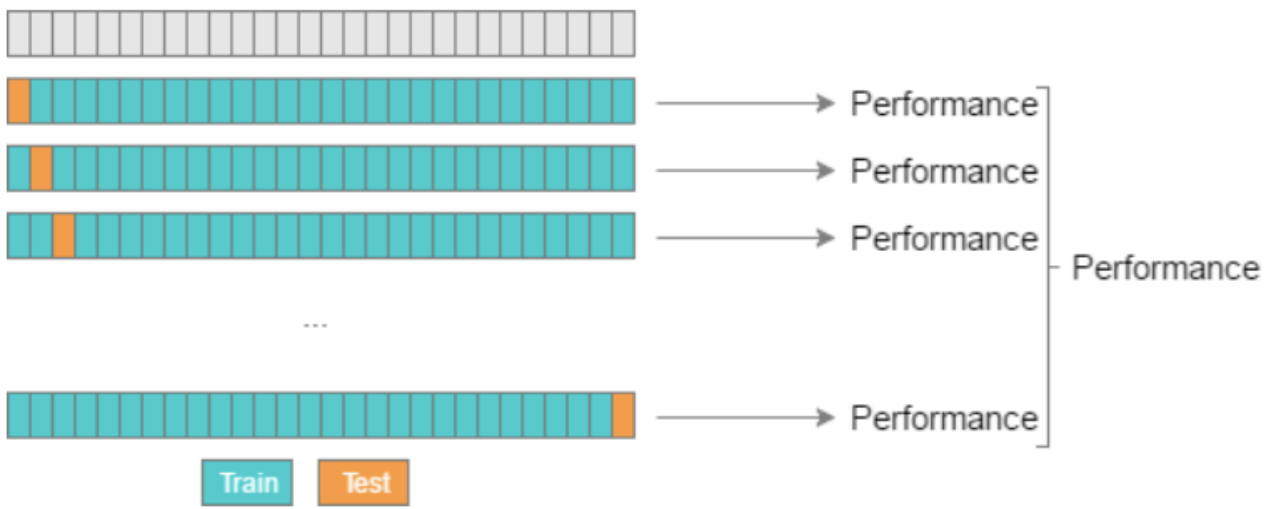
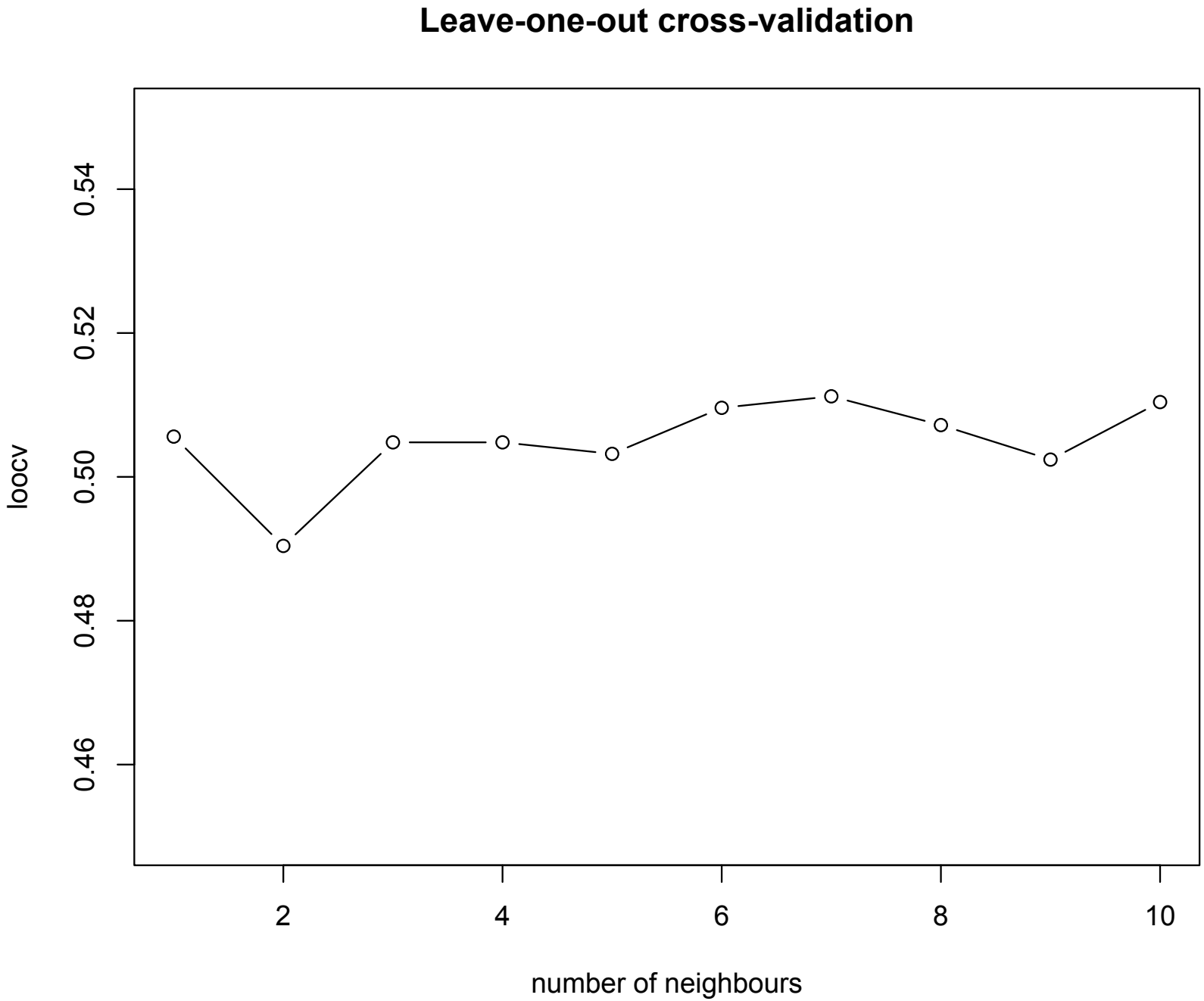


Figure 2: Illustration for Leave one out CV.


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

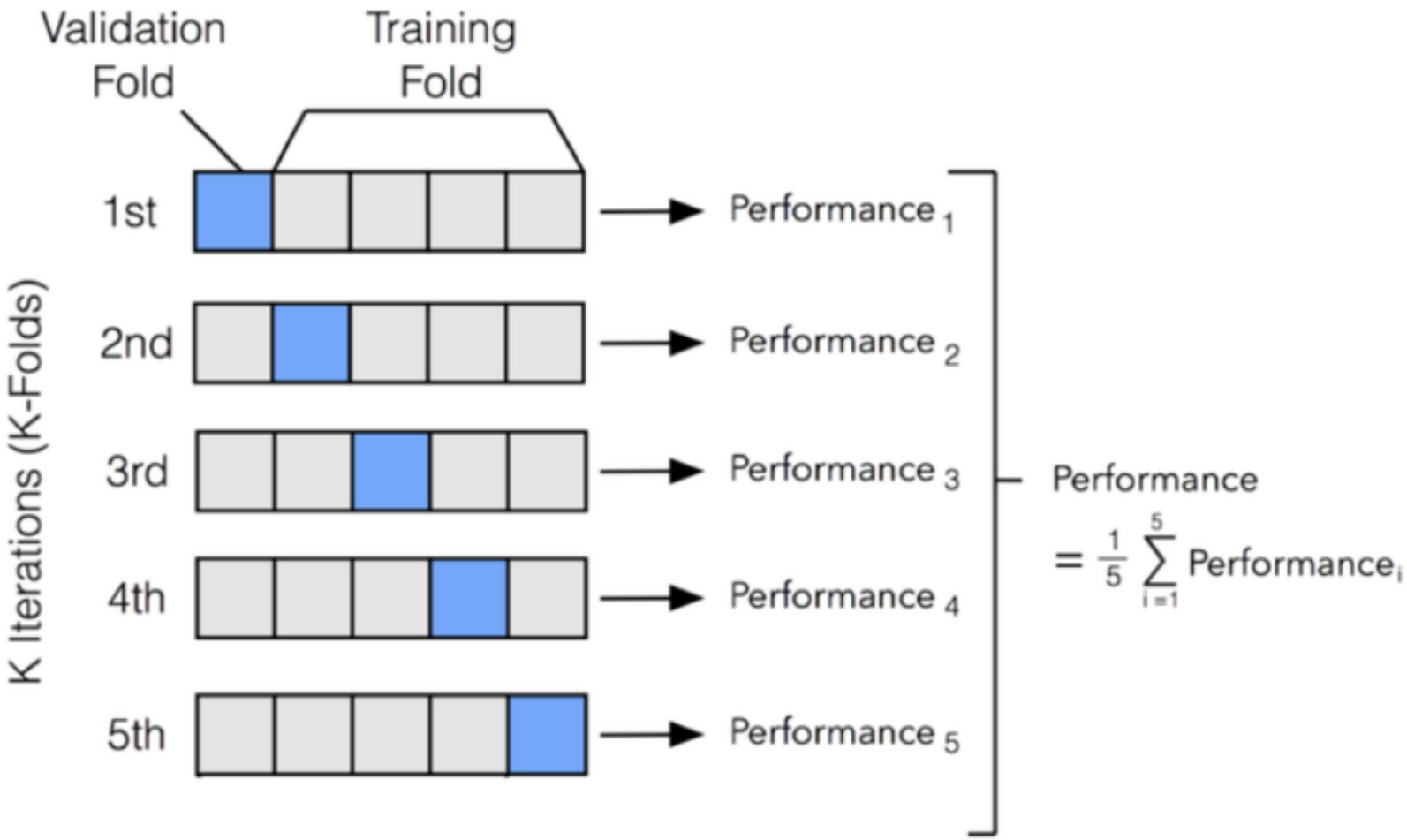
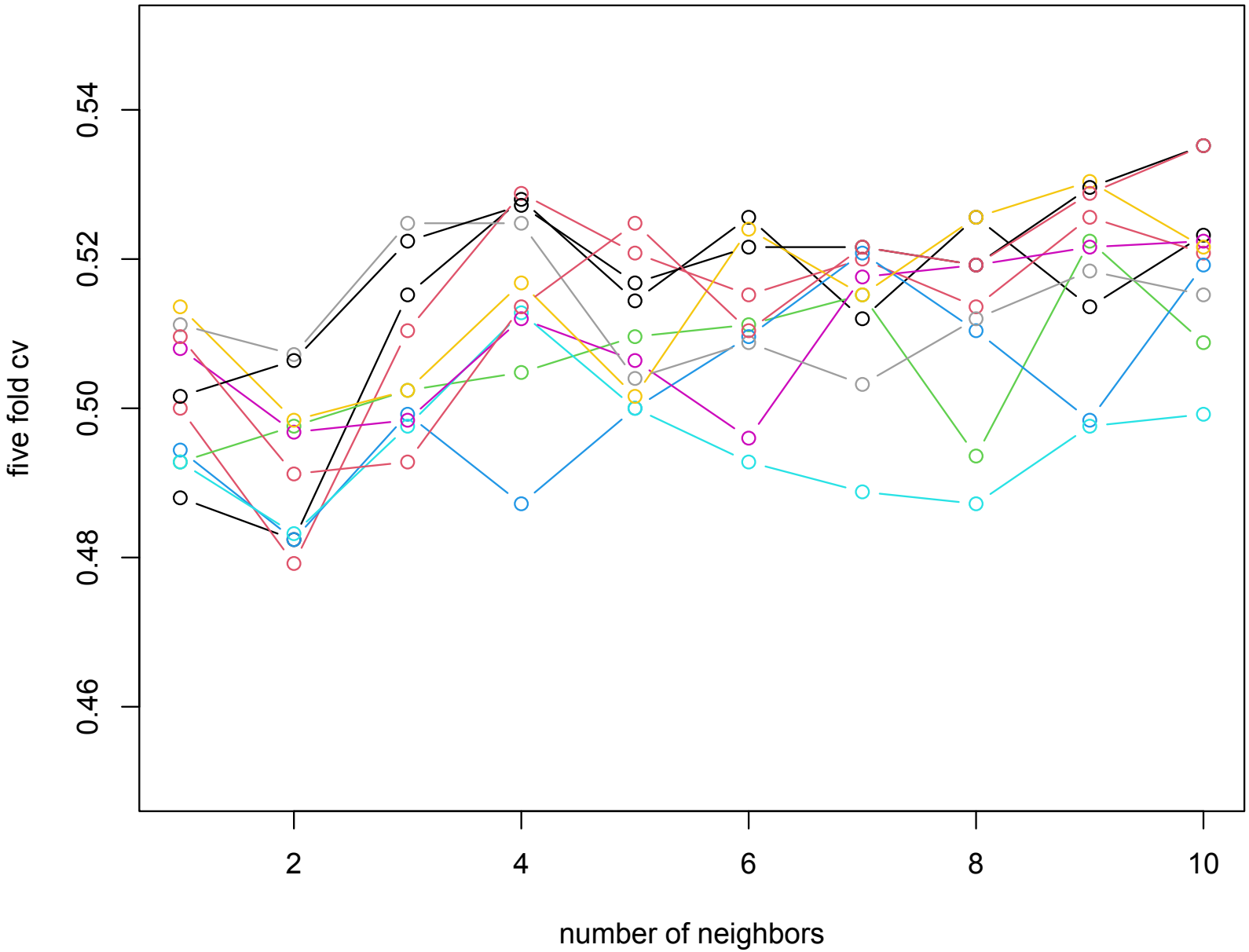


Figure 4: 5 fold CV.


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 message

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 works well.
- Use CV to tune hyperparameters and compare models.

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 the lab partner!

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

Reference

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.