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).

Training and test

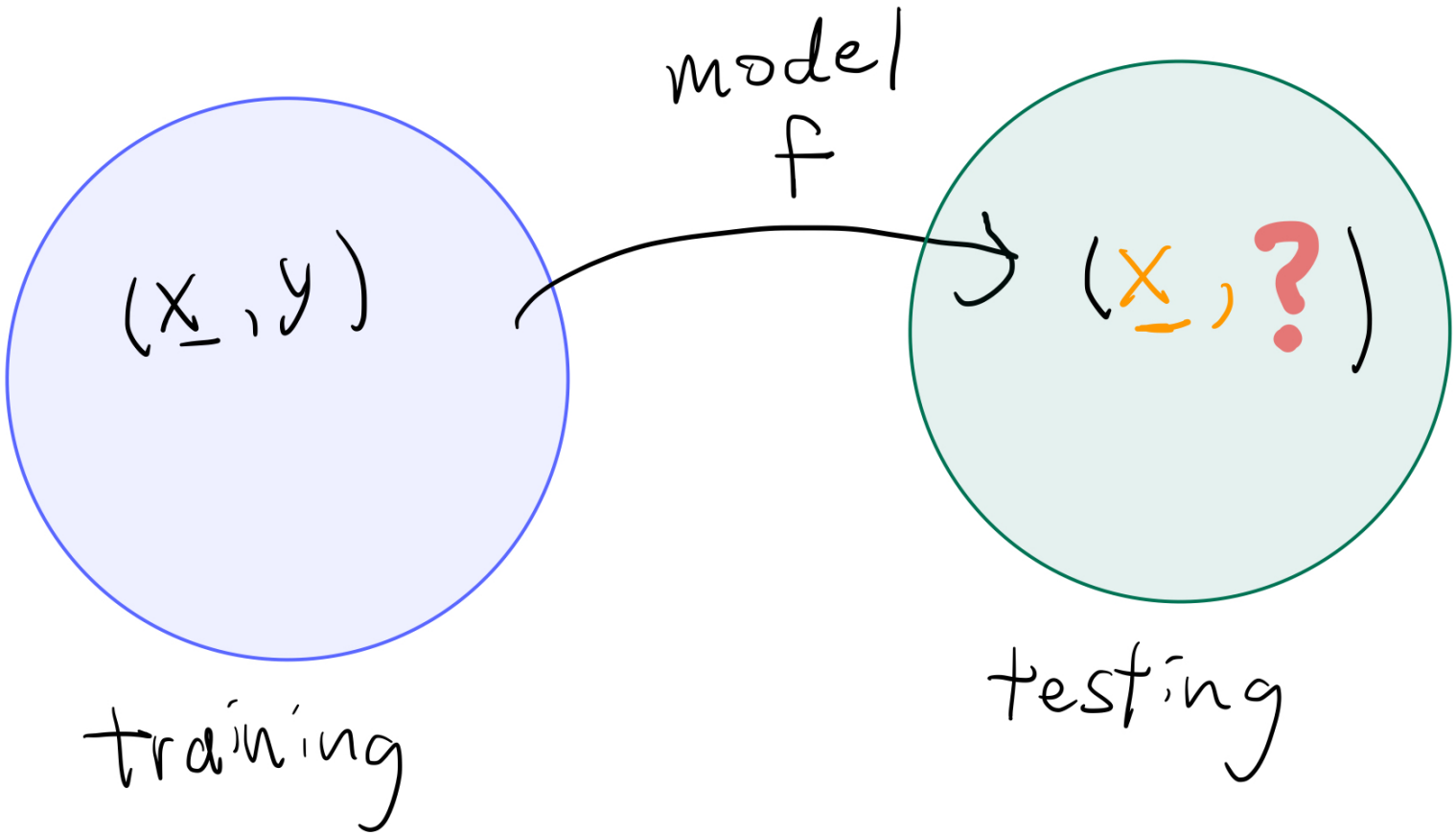


Figure 1: Training and testing sets.

Data split

- Train a model $\hat{f}$ on the training set.
- Use the $\hat{f}$ with the features to predict the outcomes, $\hat{y}$
- Difference between y and $\hat{y}$.

	fea. 1	fea 2	$\cdots$	fea. k	outcome	
X_1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,k}$	y_1	Train
X_2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,k}$	y_2	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
X_r	$X_{r,1}$	$X_{r,2}$	$\cdots$	$X_{r,k}$	y_r	Train
X_{r+1}	$X_{r+1,1}$	$X_{r+1,2}$	$\cdots$	$X_{r+1,k}$	y_{r+1}	Test
X_{r+2}	$X_{r+2,1}$	$X_{r+2,2}$	$\cdots$	$X_{r+2,k}$	y_{r+2}	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
X_N	$X_{N,1}$	$X_{N,2}$	$\cdots$	$X_{N,k}$	y_N	Test

Figure 2: Data split.

Evaluation

There are different metrics to evaluation the prediction. For instance

If y is continuous variable, we can use Mean-Square error.

If y is categorical, we can use the error rate.

Bias-Variance Tradeoff

If we have a complex model, we may have many different tuning parameters.

Bias-Variance Tradeoff

$$\mathrm{MSE}[y, \hat{y}] \propto \mathbb{V}ar(\hat{y}) + \mathbb{B}\mathrm{ias}^2(\hat{y})$$

Pop-up quiz

Q: What are the characterisitic a good estimate $\hat{y}$ should have?

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

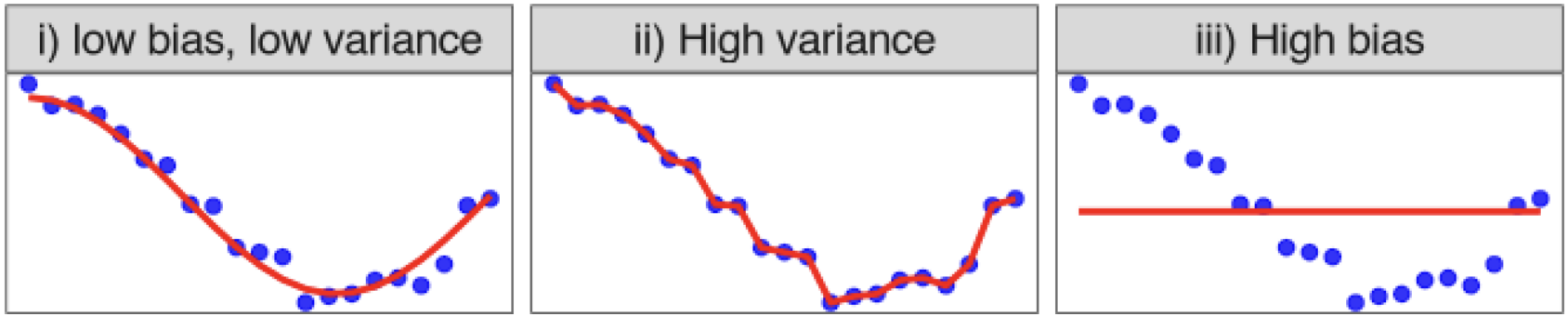


Figure 3: Illustraton of Bias and Variance.

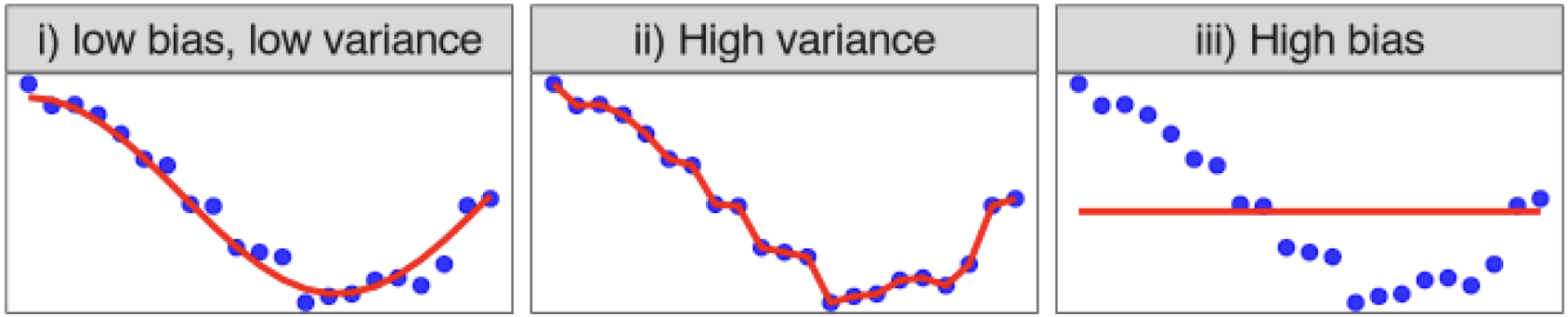


Figure 3: Illustraton of Bias and Variance.

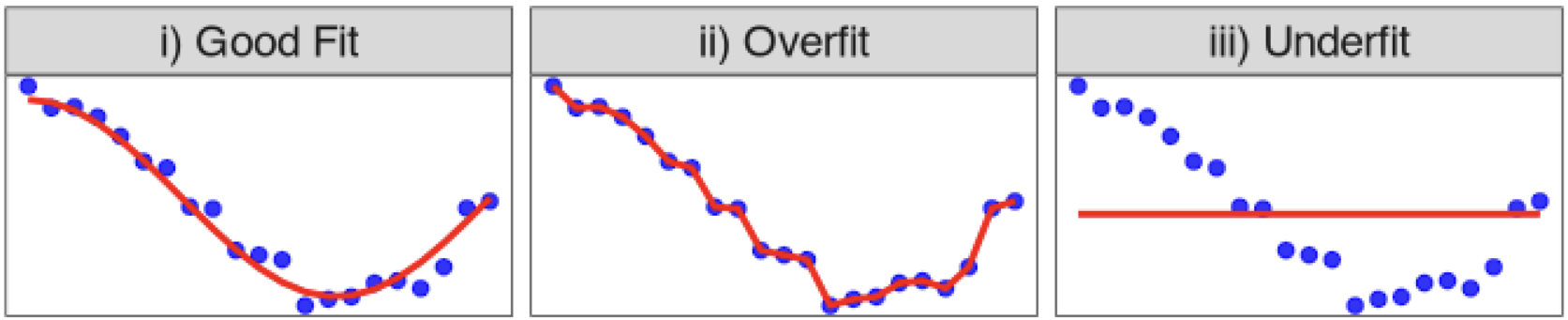
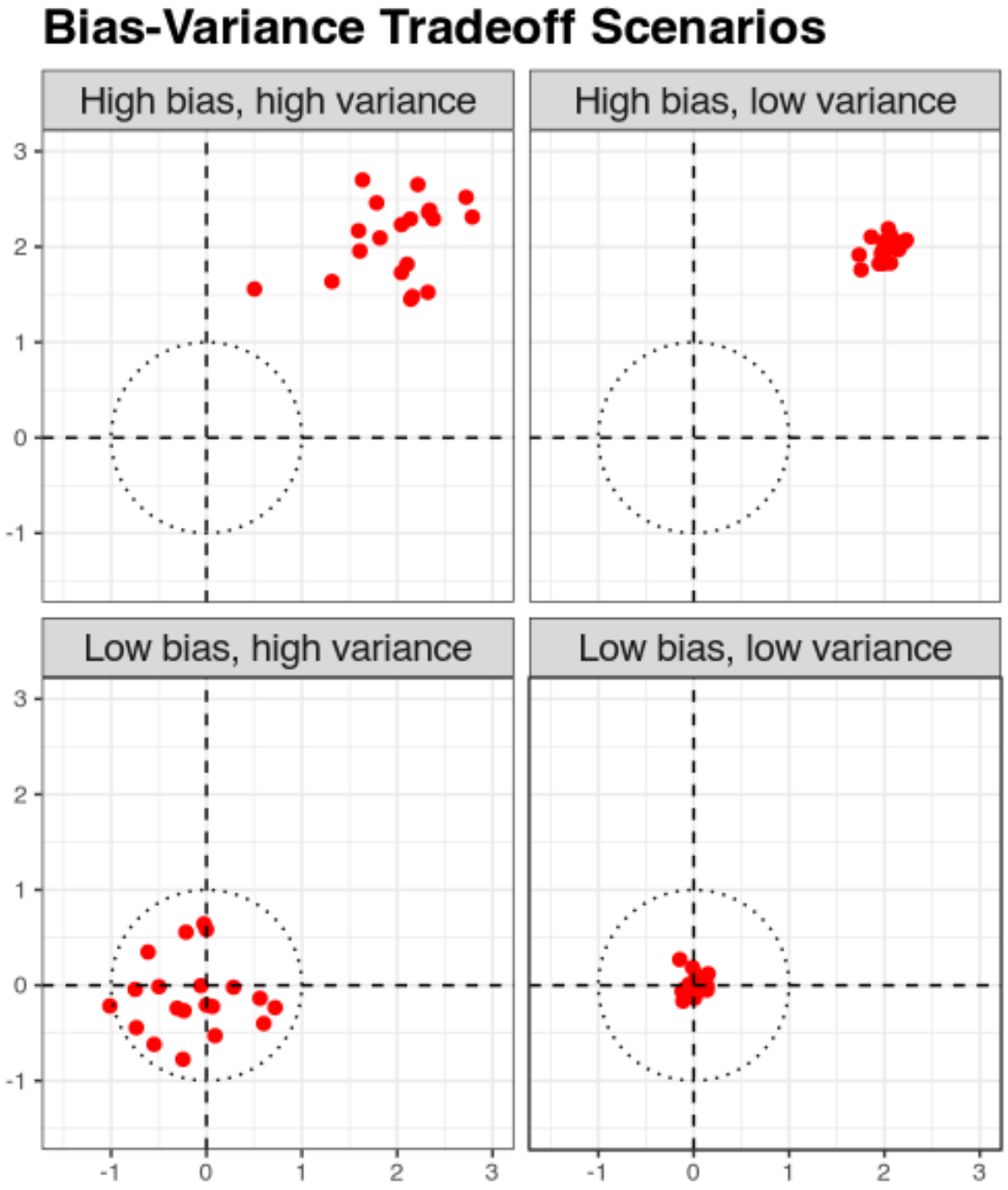


Figure 4: Illustraton of Bias and Variance.


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

What is cross validation?

Cross validation (CV) is a widely used technique to estimate **prediction error**.

One would like to think cross-validation estimates the **prediction error** for the model at hand, fit to the training data

Background: Cross-Validation

- To avoid overfitting, we have separated the data into training and test data
- If you split 50-50 into training and test data, you are wasting 50% of the training data which might have improved the model.
- If you split 80-20, depending on the data set, there may be insufficient test data to evaluate the fit well
- Cross validation is a technique that wastes less training data and can test on 100% of the data

Pop-up quiz

Q: If you are given an arbitrary data set? What is the split ratio between training and testing data you would choose?

1. 50-50
2. 90-10
3. 80-20
4. 70-30
5. Depends
6. Unsure

Prediction

Suppose we have a training data and testing data

$$\mathrm{Err} = \mathbb{E}[L(Y, \hat{f}(X))],$$

the average generalization error when the method $\hat{f}(X)$ is applied to an independent test sample from the joint distribution of X and Y.

Notation

Let the training set that consist N samples as

$$\mathcal{T} = \{(X_i, Y_i)\}_{i=1}^{N},$$

where

- $X_i \in \mathbb{R}^P$ are the features.
- $Y_i \in \mathbb{R}$ are the responses.

We train the model, and use some *prediction rule* to determine $\hat{Y} = \hat{f}_{\mathcal{T}}(X)$.

K-fold cross validation I

There are two many use of the K-fold CV

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

In short, yes the number of folds depends on the data size.

- The cross-validation approach involves randomly dividing the set of observations into K groups, or folds, of approximately equal size.
- The first fold is treated as a validation set and the method is fit on the remaining $K - 1$ folds
- Repeat step 2 for K times; each time, a different group of observations is treated as a validation set.

K-fold cross validation II

K-fold CV uses part of the data to fit the model and a different part to test it.

1. Split the data into K roughly equal sizes parts K $= 5$.
2. For each $k = 1, 2, \ldots, K$, repeat the following steps:
   (i) Leave the kth fold $\mathcal{T}_k$ (i.e., part) from the data $\mathcal{T}$, and denote the remaining data as $\mathcal{T}_{-k}$. We fit the model to $\mathcal{T}_{-k}$ and denote the corresponding model we obtained by $\hat{f}_{\mathcal{T}_{-k}}$
   (ii) Calculate the total prediction error on the fitted model $\hat{f}_{\mathcal{T}_{-k}}$ on the left-out fold $\mathcal{T}_k$

$$cv_k = \sum_{i \in \mathcal{T}_k} L(Y_i, \hat{f}_{\mathcal{T}_{-k}}(X_i)).$$

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

$$\widehat{\mathrm{Err}}_{cv} = \frac{1}{K} \sum_{k=1}^{K} cv_k.$$

K-fold cross validation III

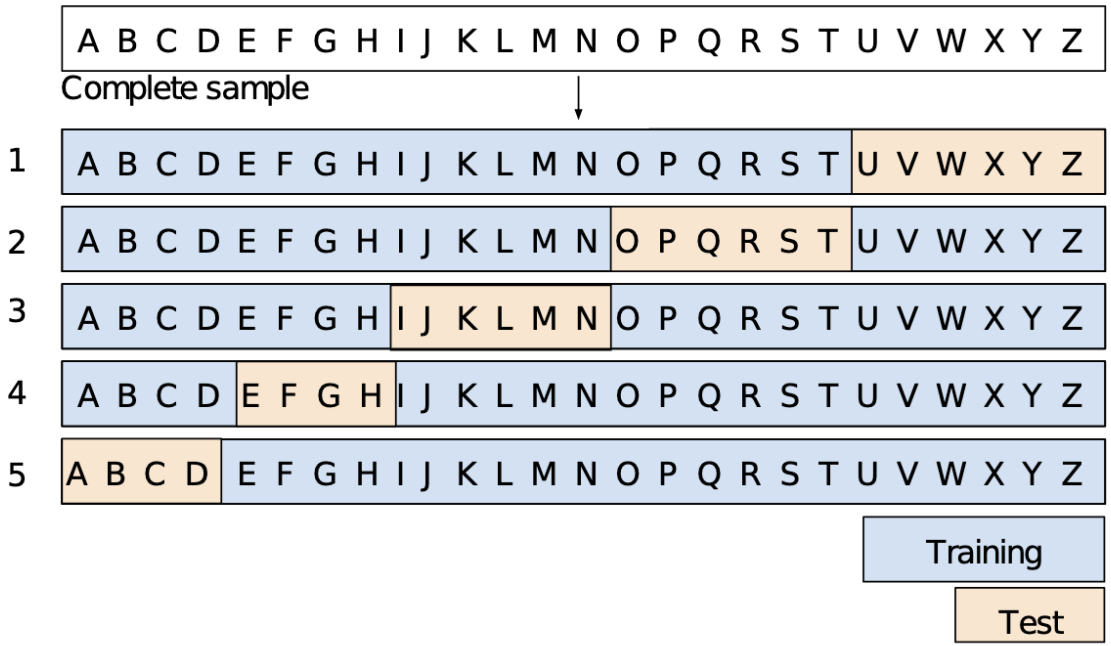


Figure 5: Example of the CV procedure with $K = 5$ folds. Picture borrowed from [4].

[4] https://www.math.mcgill.ca/yyang/slides/cv/cv.pdf

So if we have M models/prediction rules $f_1, f_2, \ldots, f_M$, we can use cross-validation to select the best model by computing the cross-validation error for each model
$\widehat{\mathrm{Err}}_{cv}(\hat{f}_1), \widehat{\mathrm{Err}}_{cv}(\hat{f}_2), \ldots \widehat{\mathrm{Err}}_{cv}(\hat{f}_M)$

How to choose the fold K?

It is interesting to wonder about what quantity K-fold CV estimates.

Bias-Variance Tradeoff
The choice of K is a tradeoff between bias and variance.

If we have our fitted model $\hat{f}$, we then have

$$\begin{aligned}
\mathrm{MSE}(x_0) &= \mathbb{E}_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\
&= \mathbb{E}_{\mathcal{T}}[\hat{y}_0 - \mathbb{E}_{\mathcal{T}}(\hat{y}_0)]^2 + \{\mathbb{E}_{\mathcal{T}}(\hat{y}_0 - f(x_0)\}^2 \\
&= \mathbb{V}ar_{\mathcal{T}}(\hat{y}_0) + \mathbb{B}\mathrm{ias}^2(\hat{y}_0)
\end{aligned}$$

- An extreme case $K = N$: the CV estimator is approximately unbiased for the true (expected) prediction error, but can have high variance because the N "training sets" T_{-i} are so similar to one another. The computational burden is also considerable, requiring N applications of the learning method.
- On the other hand, with $K = 5$ say, cross-validation has lower variance. But bias could be a problem, depending on how the performance of the learning method varies with the size of the training set.

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.

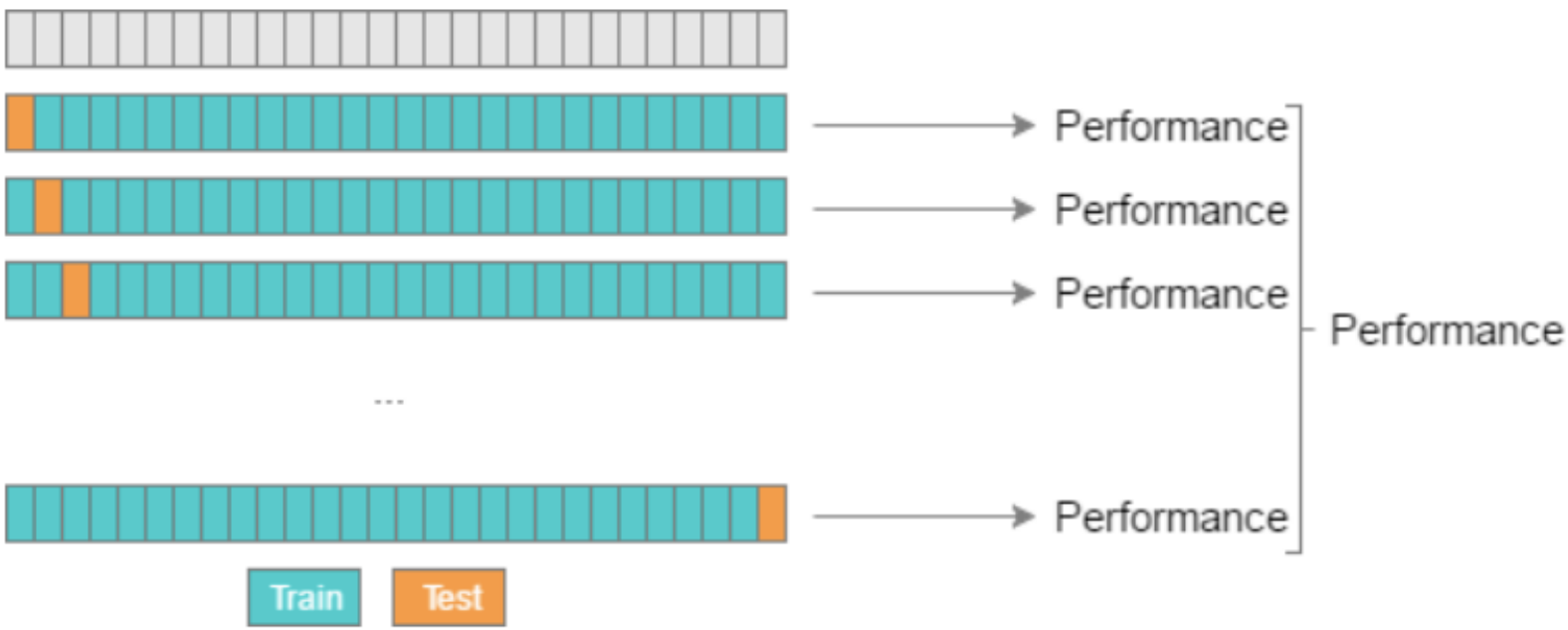
$N = 1250$

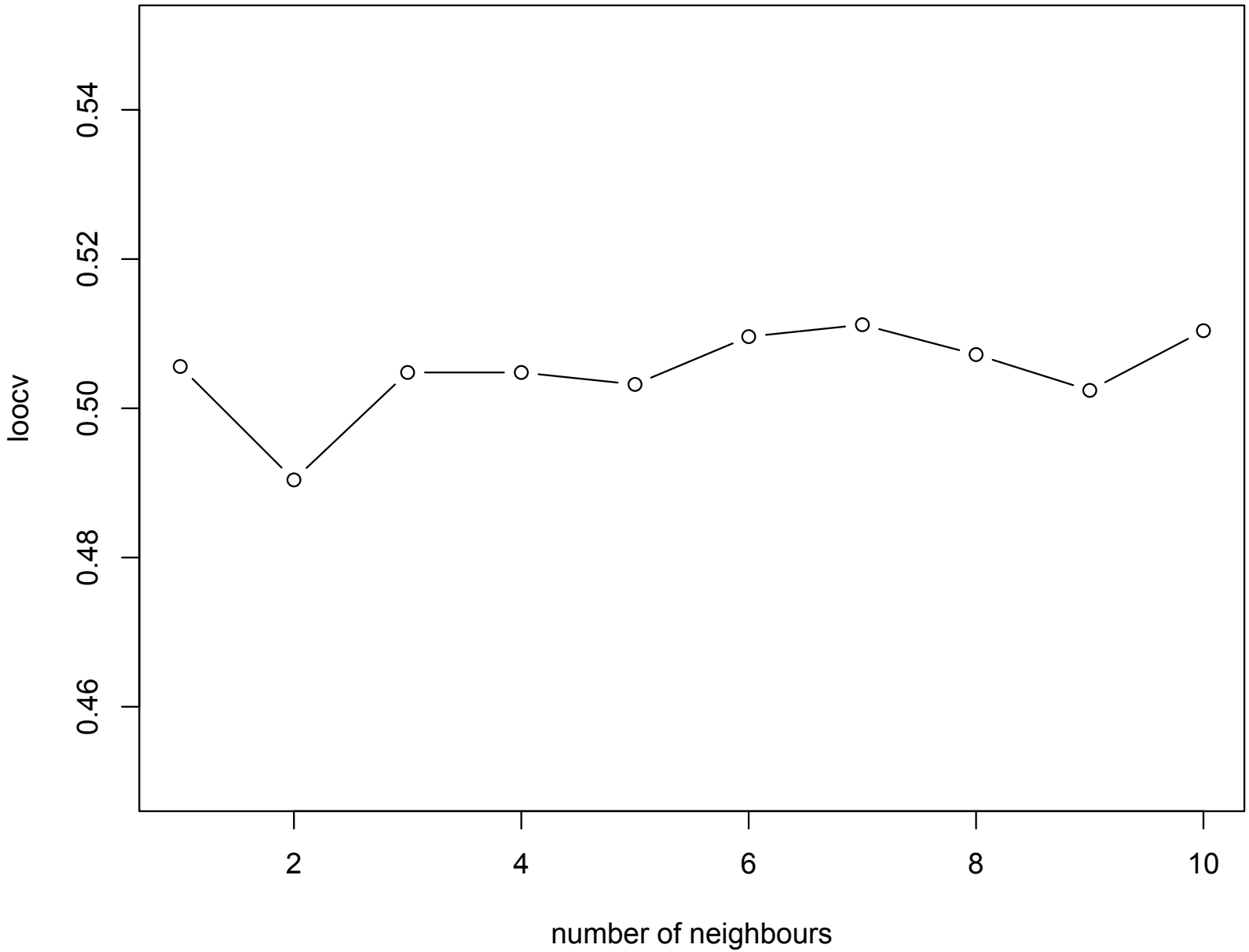
: First Few Rows of Smarket Dataset

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

- 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.




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

- Partition data in five random subsets of roughly equal size
- For each of five subsets:
  1. Use this subset as validation data, the remainder as training data
  2. Fit the model using the training data
  3. Evaluate the performance on the validation data
- Average over five evaluation results

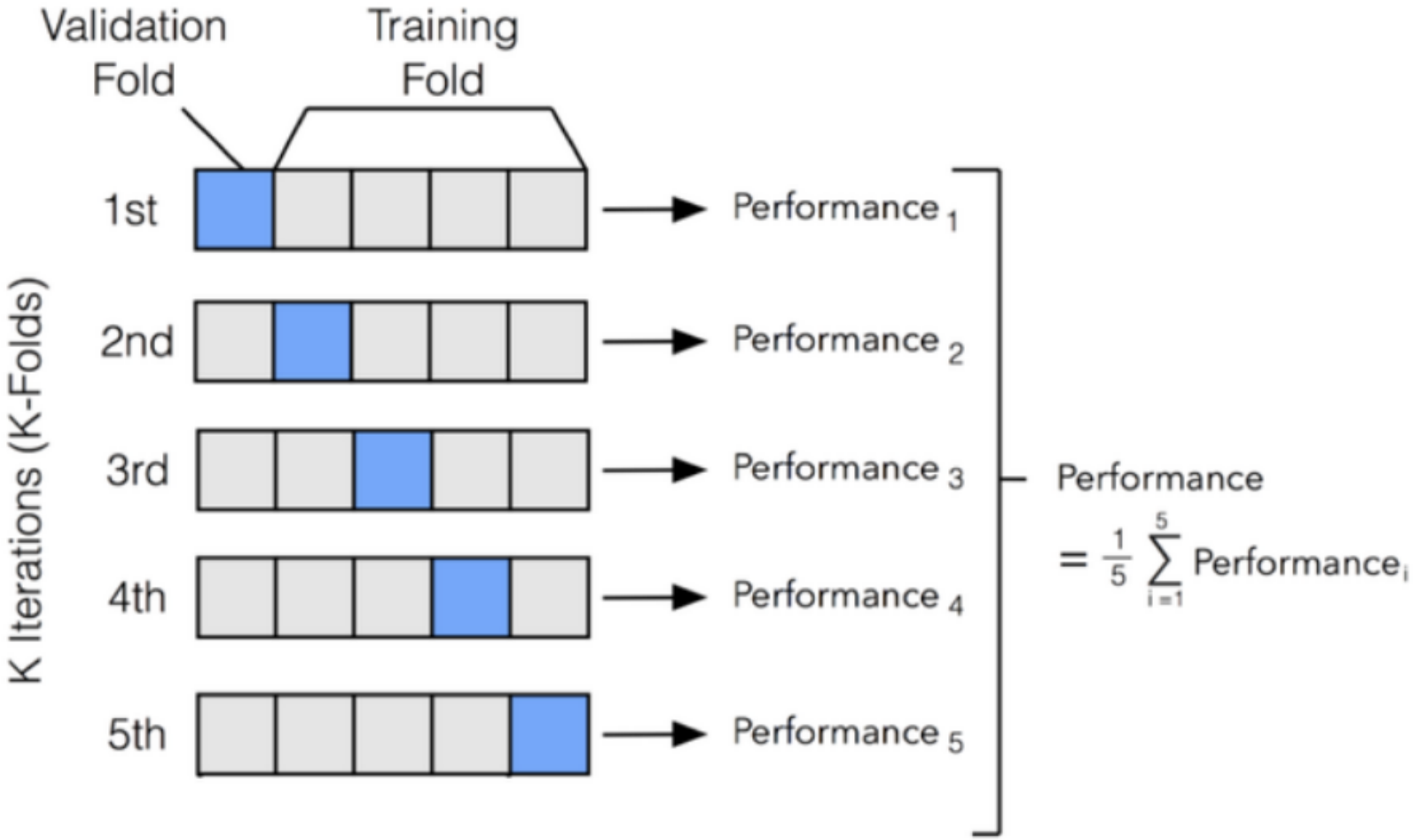
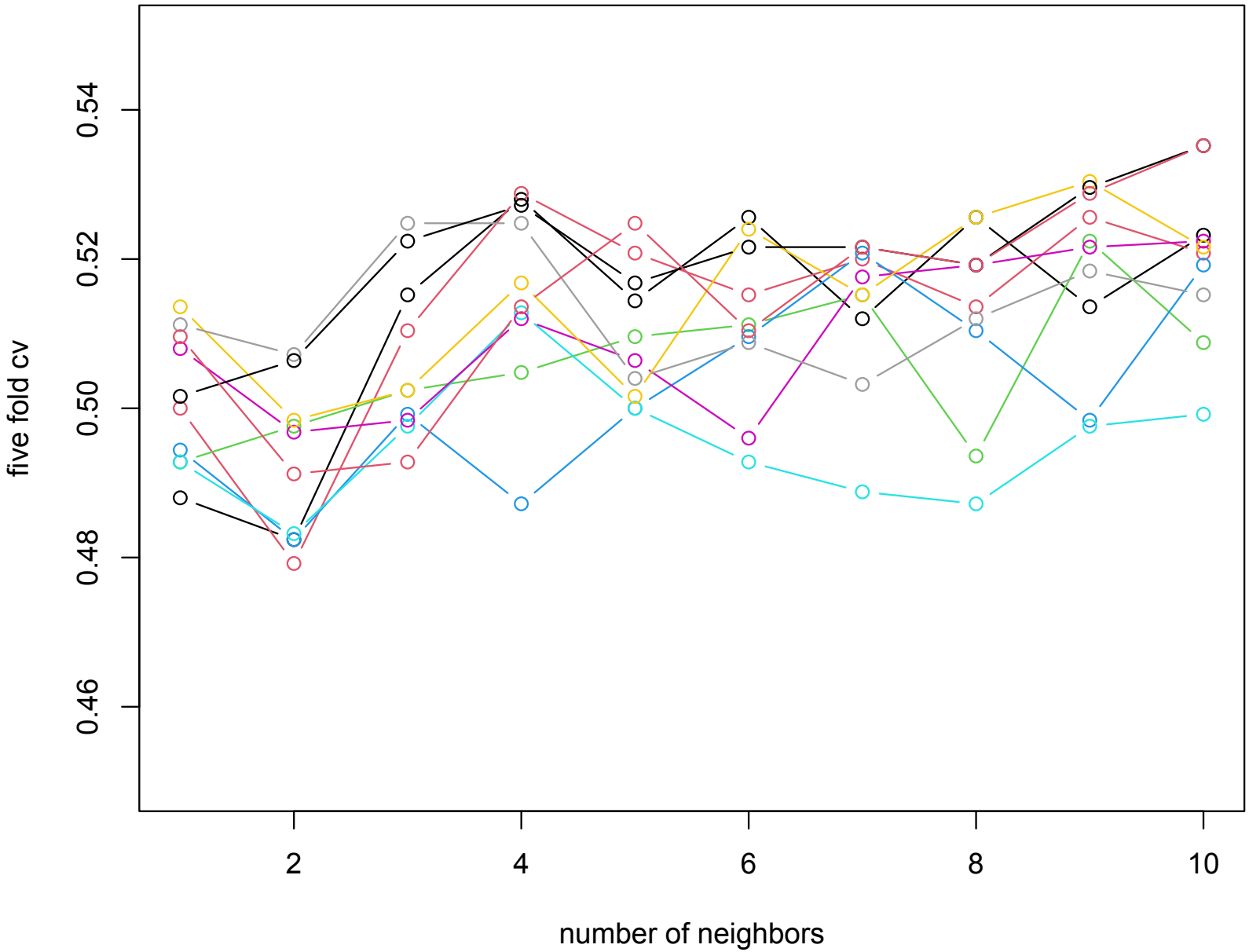


Figure 8: 5 fold CV.


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

Cross-Validation

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

Cross-Validation

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

- As always, it is a bias-variance tradeoff.

Cross-Validation

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

- As always, it is a bias-variance tradeoff.
- It turns out that larger values of K yield estimates with higher variance, but smaller bias.
  - The smaller the number of folds, the smaller the size of each training set. CV overestimates the test error for the model fit on the entire data set. Thus, the CV-estimate of prediction error is always biased high, and bias decreases as K increases.

Discussion

What would you consider when choosing K?

Discussion

What would you consider when choosing K?

1. Run time
2. Sample size
3. ...

Cross validation may be used to choose the hyper parameter

- Linear Regression (No hyper parameter)
- LASSO Regression
- Ridge Regression
- Random Forest
- Gradient Boosting
- Neural Networks

Many of the models have something called **Hyperparameter**.

Some models have hyperparameters that need to be tuned, some have only a few, some have more.

For example, in LASSO regression, there is only 1 hyperparameter

whereas in Random Forest, the hyperparameters are

1. max depth
2. min sample split
3. max terminal node
4. min sample leaf
5. ...

CV can help in choosing the hyperparameter(s).

Take home message

- Cross-validation is a widely used technique to estimate prediction error.
- K-fold CV is a popular choice for cross-validation. The value of K influences the bias-variance tradeoff.
  - LOOCVL is a special case of K-fold CV when $K = N$.
  - 5- or 10-fold CV are common chocies
- CV can be used to choose the hyperparameter(s) for a model.

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.