DATASCI 347 Machine Learning

Lecture 7: Cross-Validation

Ruoxuan Xiong

Suggested reading: ISL Chapter 5

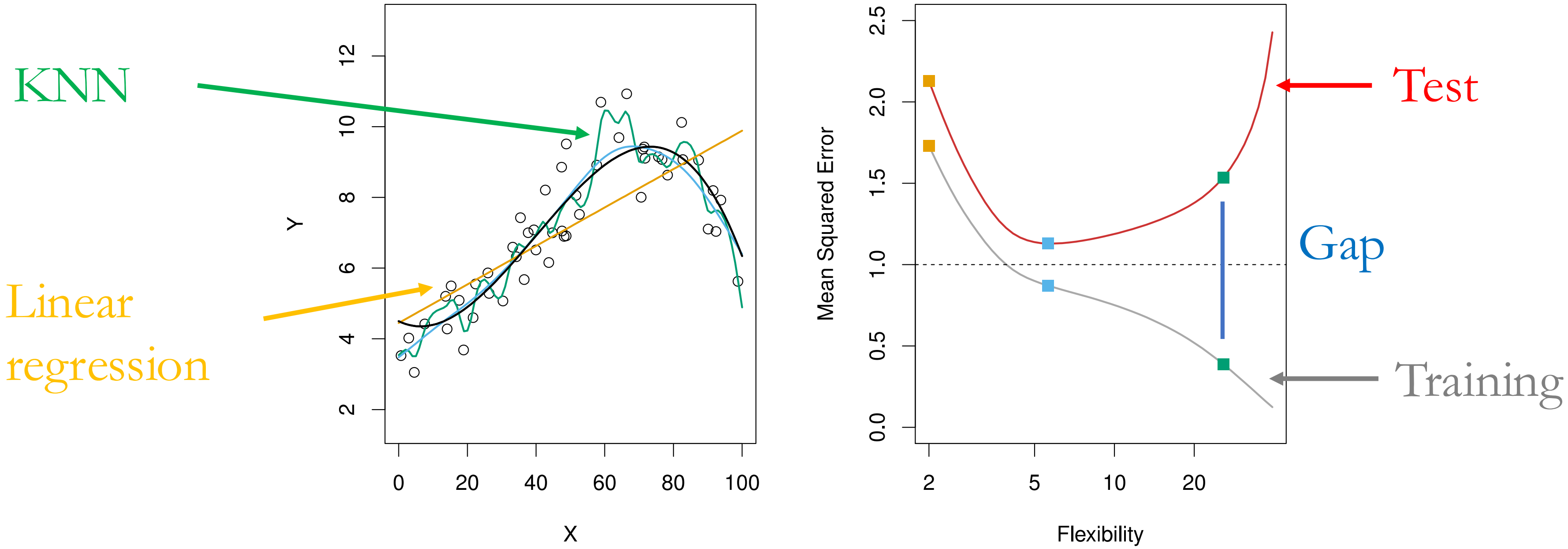
EMORY

Lecture plan

- Cross validation

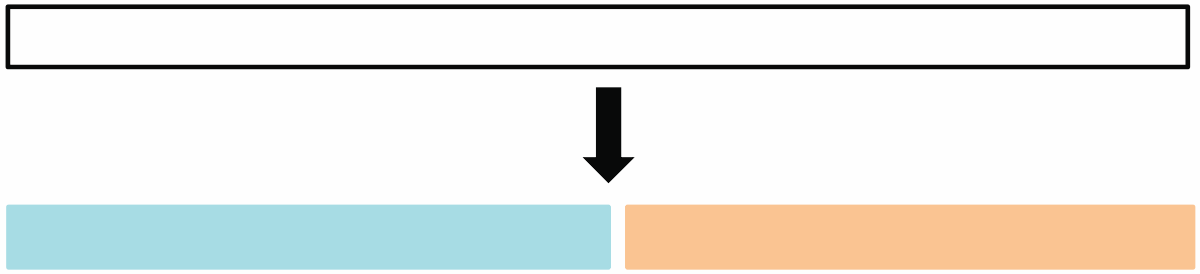
Motivation

- **Supervised learning**: Minimize test error
  - However, we only have access to the training error
  - There is often a gap between them
- **Illustration**: Suppose we know what f is (the black curve)
  - We generate data according to f as simulated data (in circles)



Validation set approach

- **Goal of validation set approach**: Using the training data set alone, find out the test error as closely as possible
- **A first attempt**:
  - Randomly split the data in two parts
  - Train the method in the first part
  - Compute the error on the second part

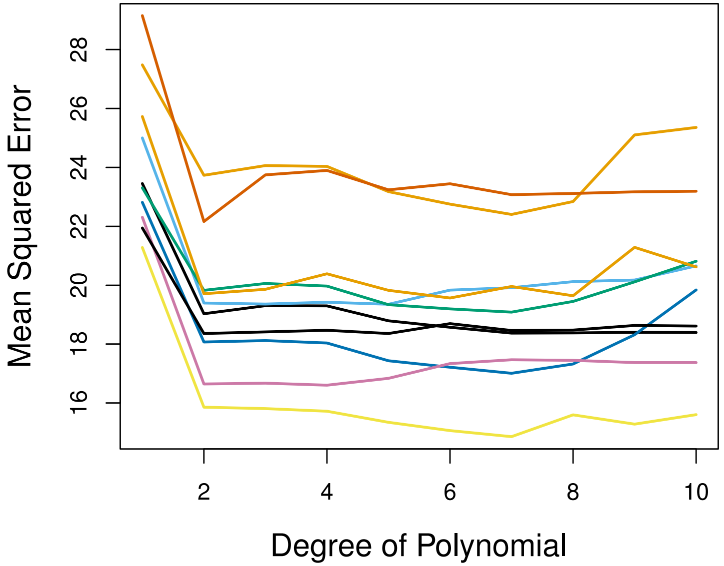


Example

- Estimate miles per gallon (mpg) from engine horsepower
  - Auto data: horsepower, gas mileage, and other information for 392 vehicles
- **Simple linear regression**
  - $\text{mpg} = \beta_0 + \beta_1 \text{horsepower}$
- **Multiple linear regression with polynomial features**
  - $\text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \beta_2 \text{horsepower}^2$
  - $\text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \beta_2 \text{horsepower}^2 + \beta_3 \text{horsepower}^3$
- **Which polynomial is the right relationship?**
  - **Resampling**
    - Partition 392 samples into two sets with equal size
    - One is the training set and the other one is the validation set

Example

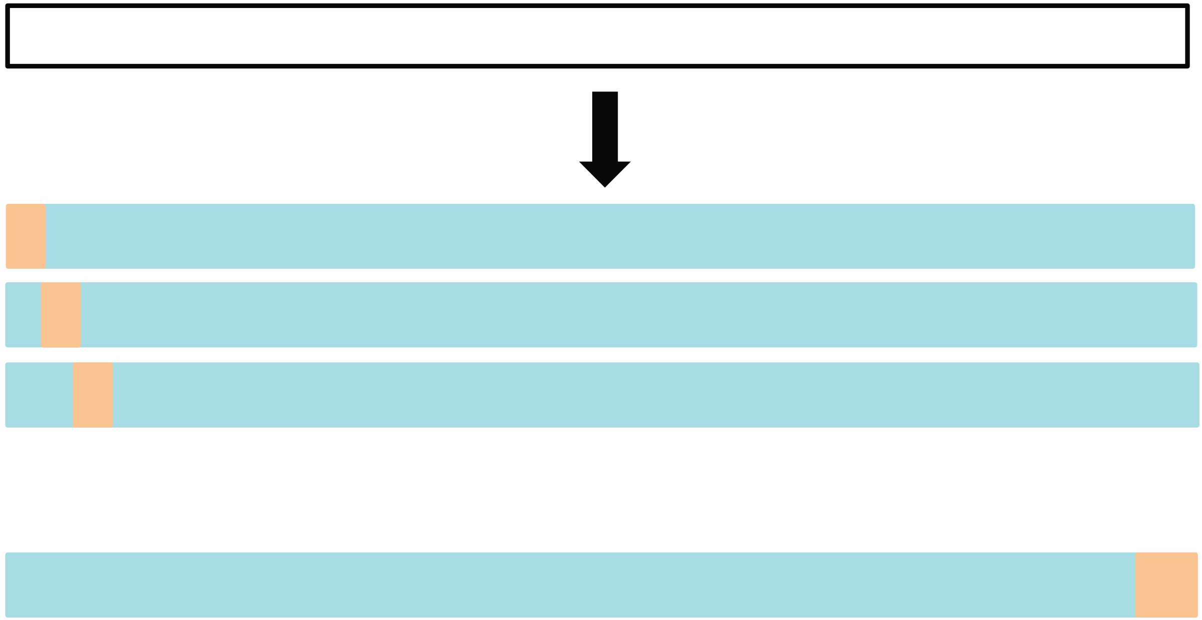
- Estimate miles per gallon (mpg) from engine horsepower



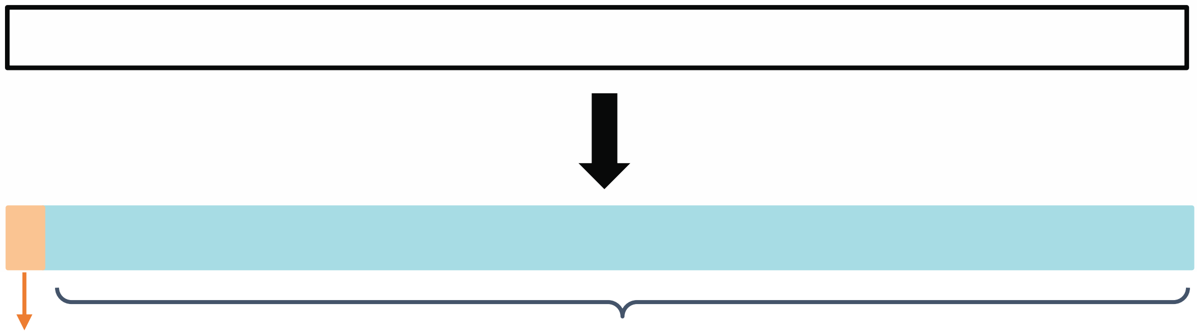
- Each line is the result with a different random split of the data into two parts
- Every split yields a different estimate of the error ☹

Leave one out cross-validation

- Leave one out cross-validation (split the data into n *folds*)
- For every $i = 1, \cdots, n$,
  - Train the model on every point except i
  - Compute the test error on the hold-out point
  - Average over all n points



Leave-one-out cross-validation

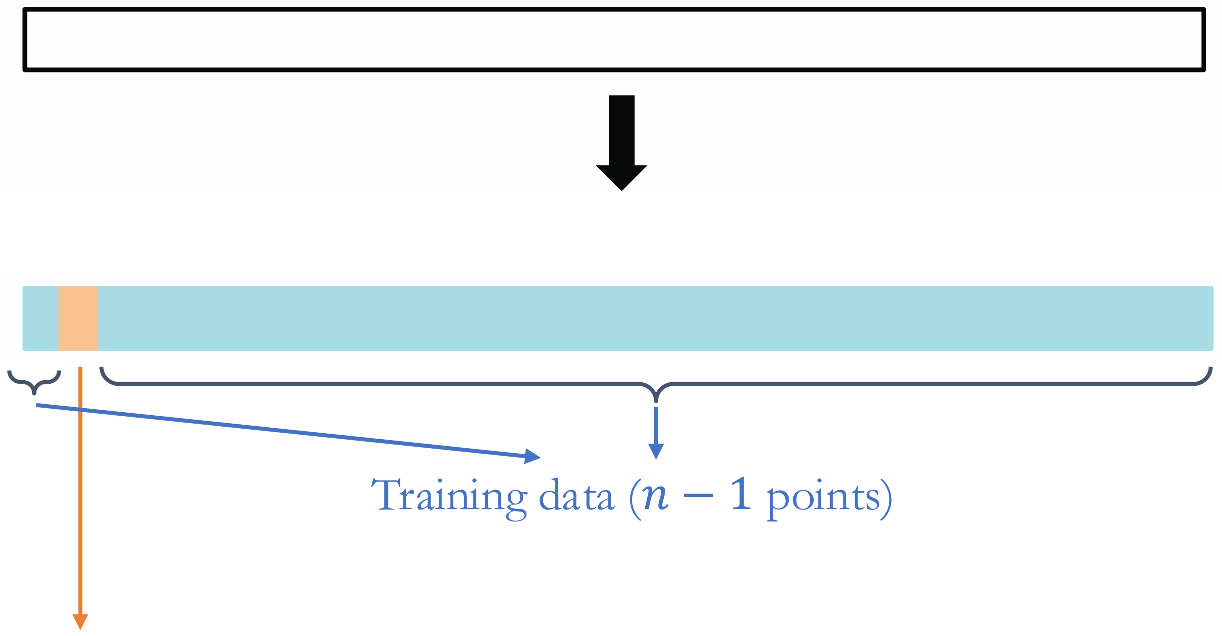


Validation data

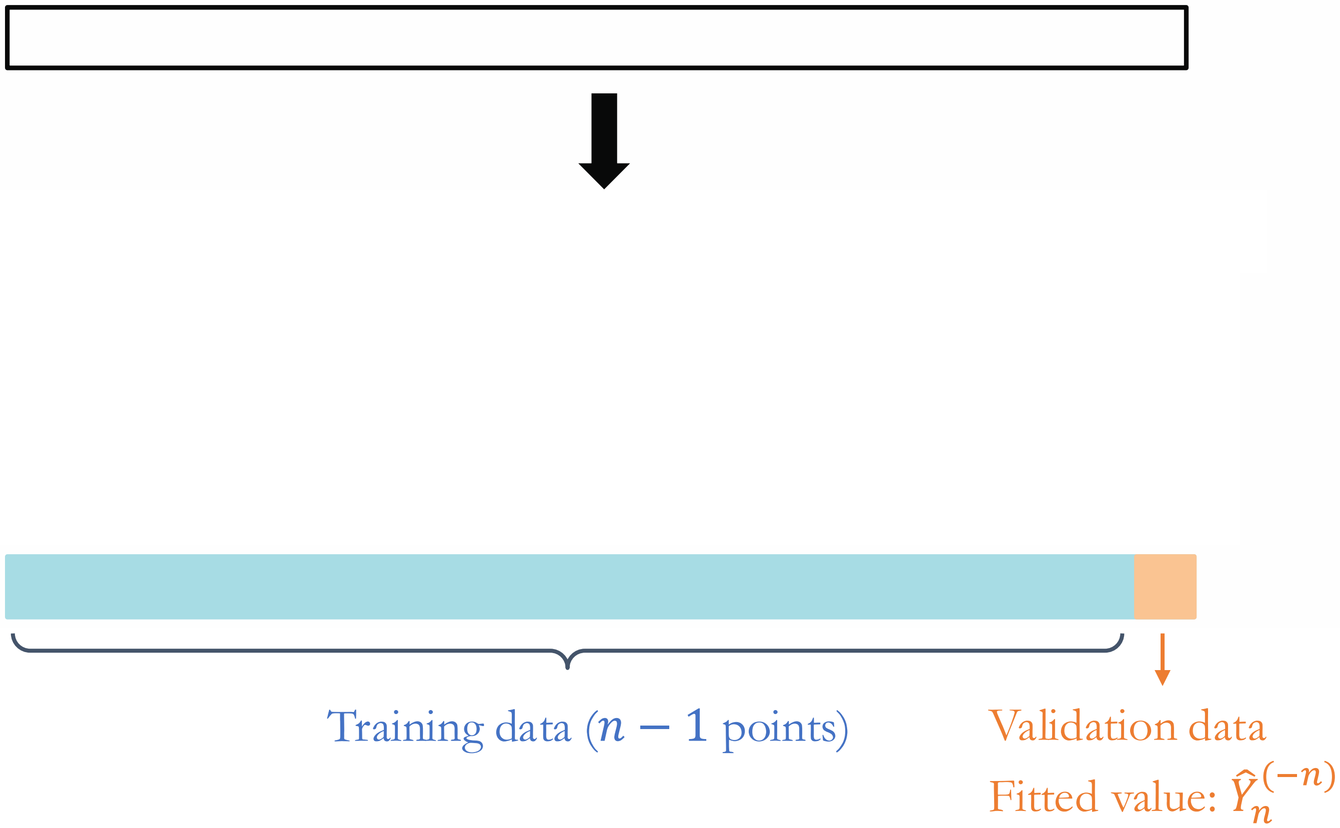
Fitted value: $\hat{Y}_1^{(-1)}$

Training data ($n - 1$ points)

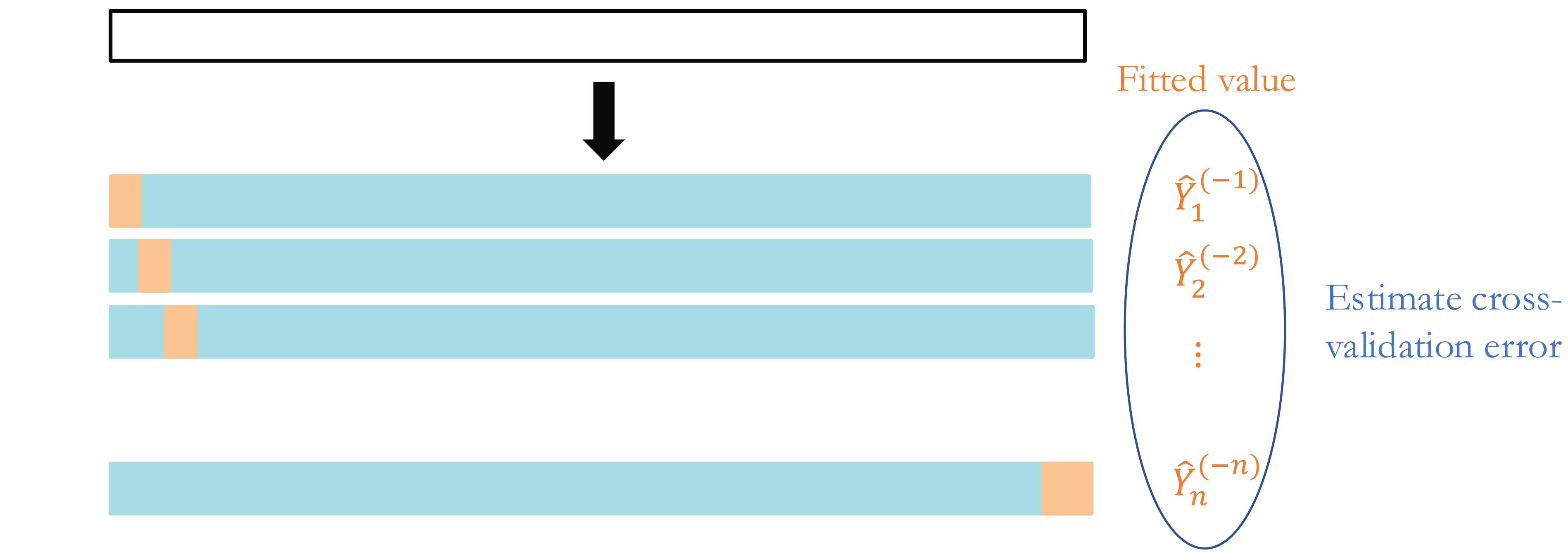
Leave-one-out cross-validation



Leave-one-out cross-validation



Leave-one-out cross-validation



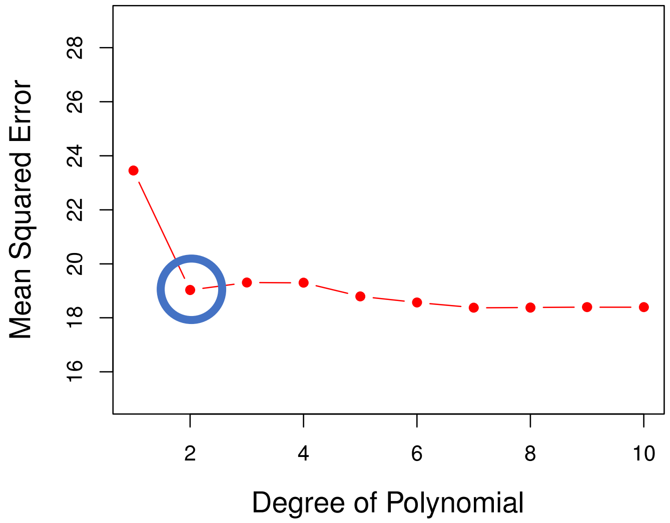
Leave one out cross-validation

- **Regression** with mean squared loss
  - $\hat{y}_i^{(-i)}$: Prediction for the ith sample without using the ith sample
  - $CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2$

- **Classification** with zero-one loss
  - $\hat{y}_i^{(-i)}$: Prediction for the ith sample without using the ith sample
  - $CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} 1\left[y_i \neq \hat{y}_i^{(-i)}\right]$

Example

- Estimate miles per gallon (mpg) from engine horsepower
- The LOOCV error curve



LOOCV has low bias and no randomness

- Each training set in LOOCV has $n - 1$ observations, almost as many as are in the entire data set
  - LOOCV tends not to overestimate the test error rate by too much (low bias)
  - There is no randomness in the training/validation set splits

Computational concerns

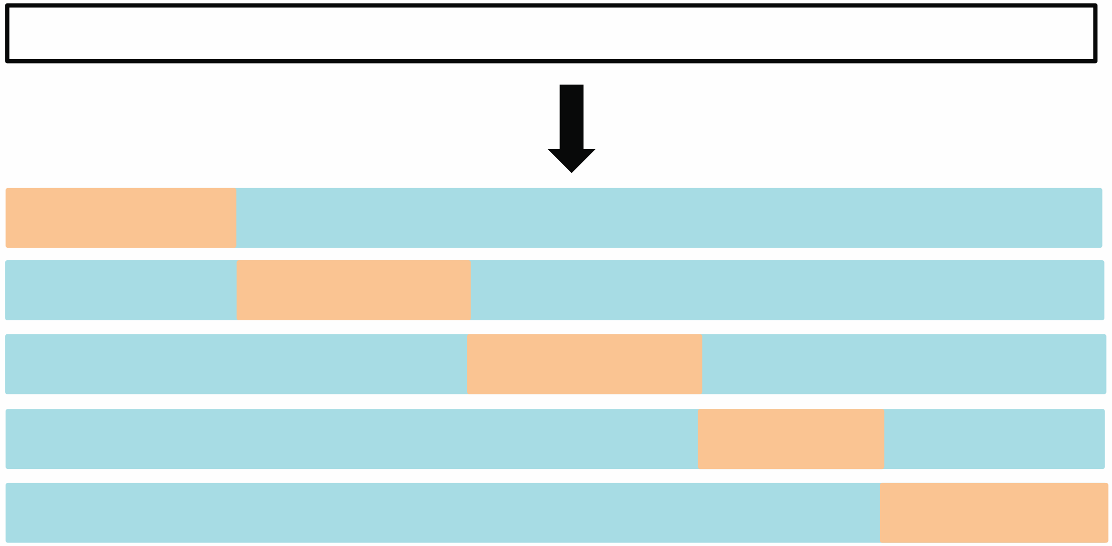
- Computing $CV_{(n)}$ can be computationally expensive, since it involves fitting the model n times

- What if we use a model other than linear or polynomial regression?

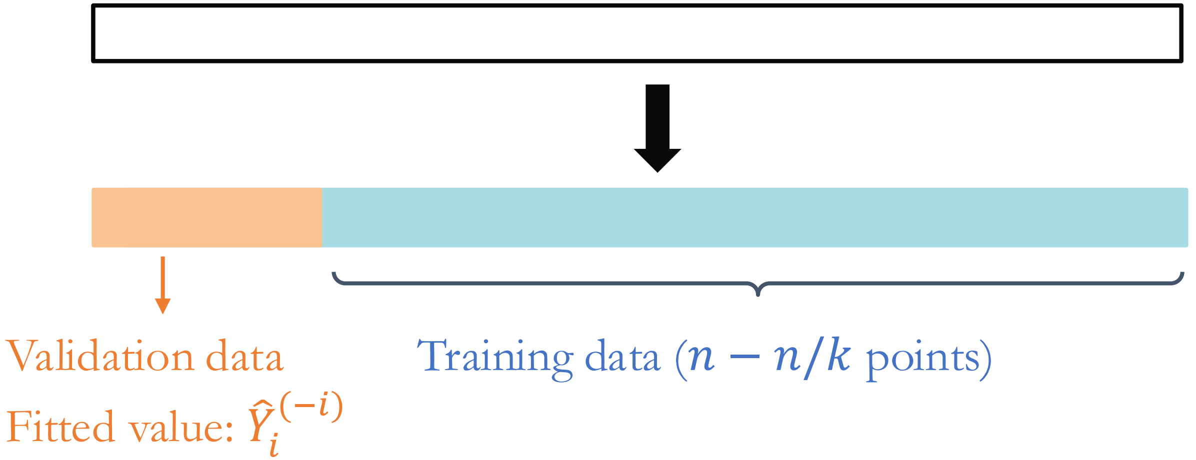
- k-fold cross-validation: Split the data into k equal sized subsets
    - Only requires fitting the model k times
    - $\frac{n}{k}$ times speed up over leave one out cross-validation

k-fold cross-validation

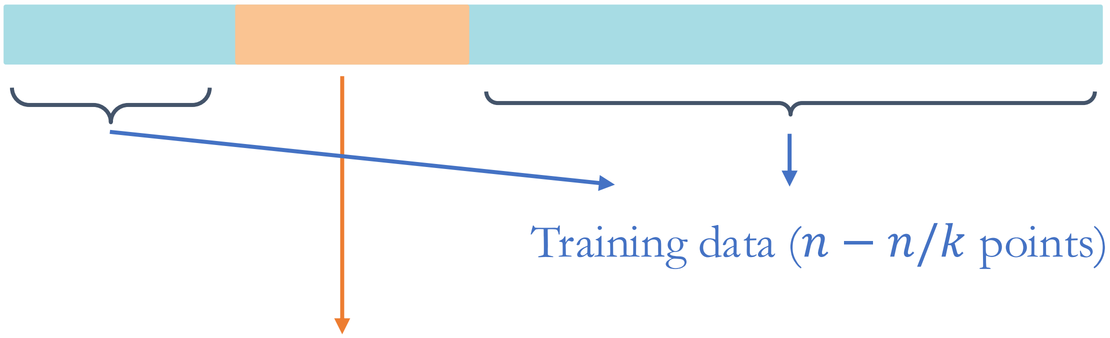
- Split the data into k subsets or *folds*
- For every $i = 1, \cdots, k$:
  - Train the model on every fold except the ith fold
  - Compute the test error on the ith fold
  - Average the test errors



k-fold cross-validation



k-fold cross-validation

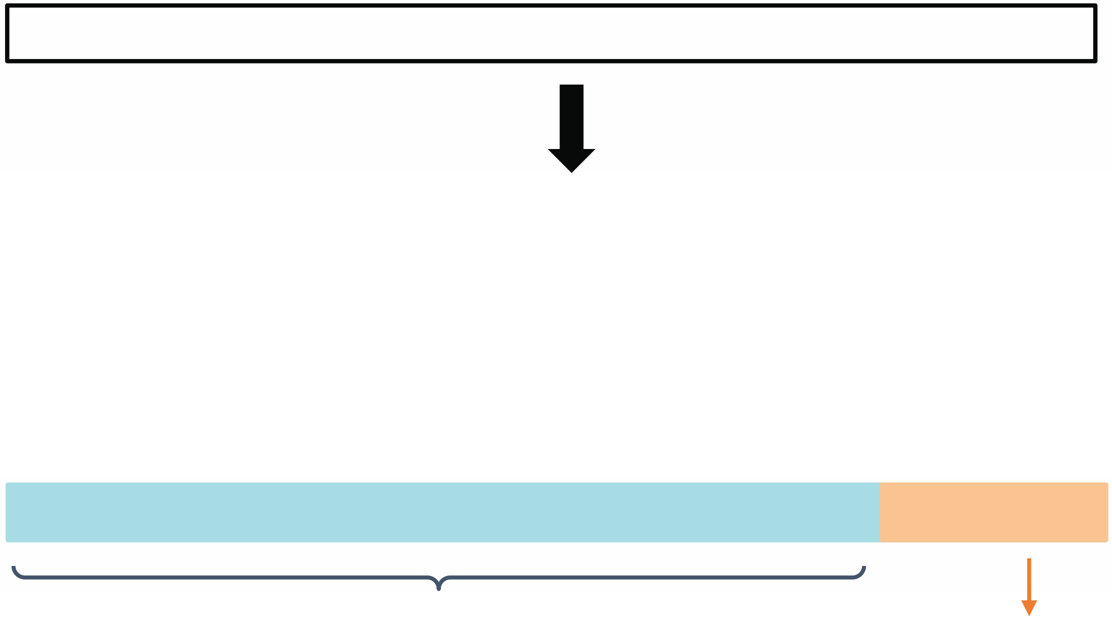


Training data ($n - n/k$ points)

Validation data

Fitted value: $\hat{Y}_i^{(-i)}$

k-fold cross-validation

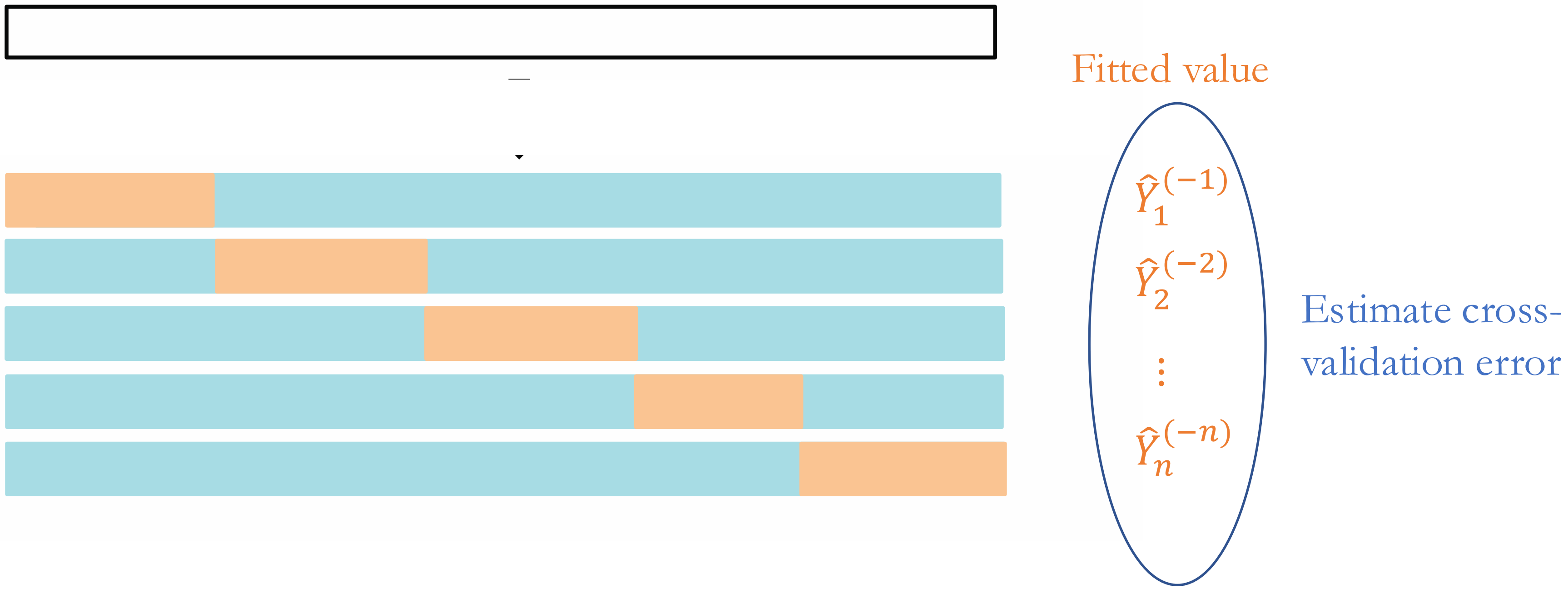


Training data ($n - n/k$ points)

Validation data

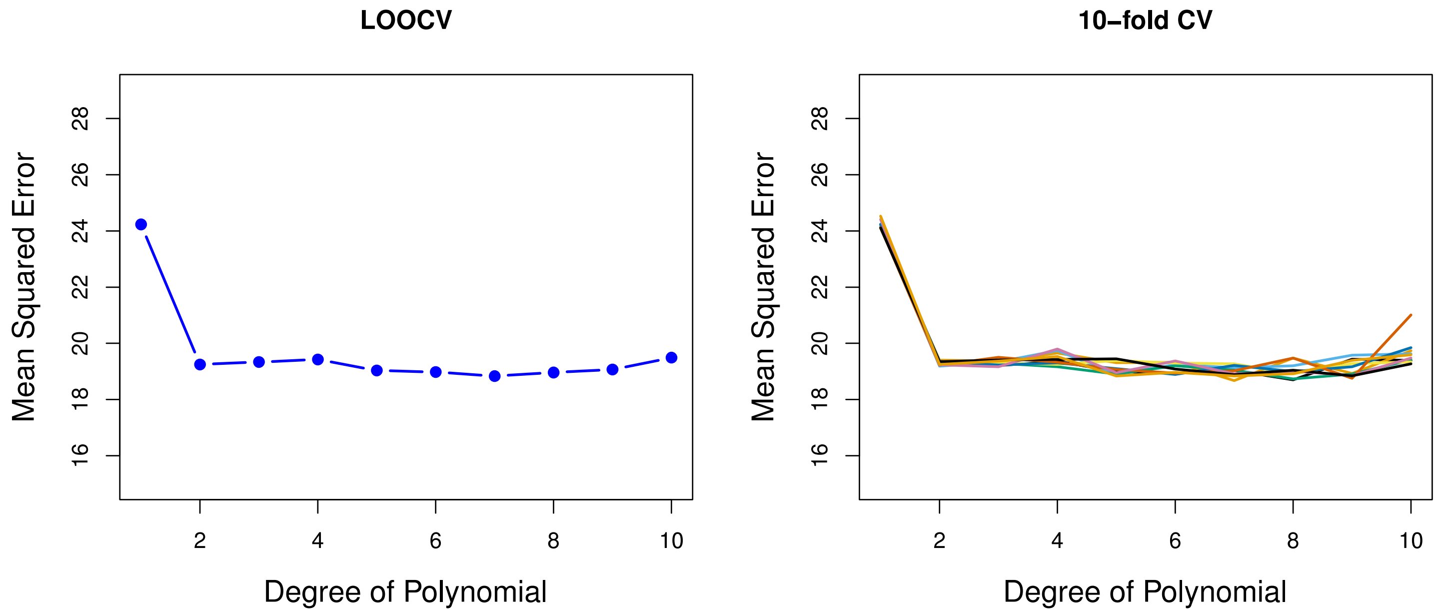
Fitted value: $\hat{Y}_i^{(-i)}$

k-fold cross-validation



LOOCV vs. k-fold CV

- Estimate miles per gallon (mpg) from engine horsepower
- The LOOCV error curve vs. 10-fold error curve



LOOCV vs. k-fold CV: Bias-variance tradeoff

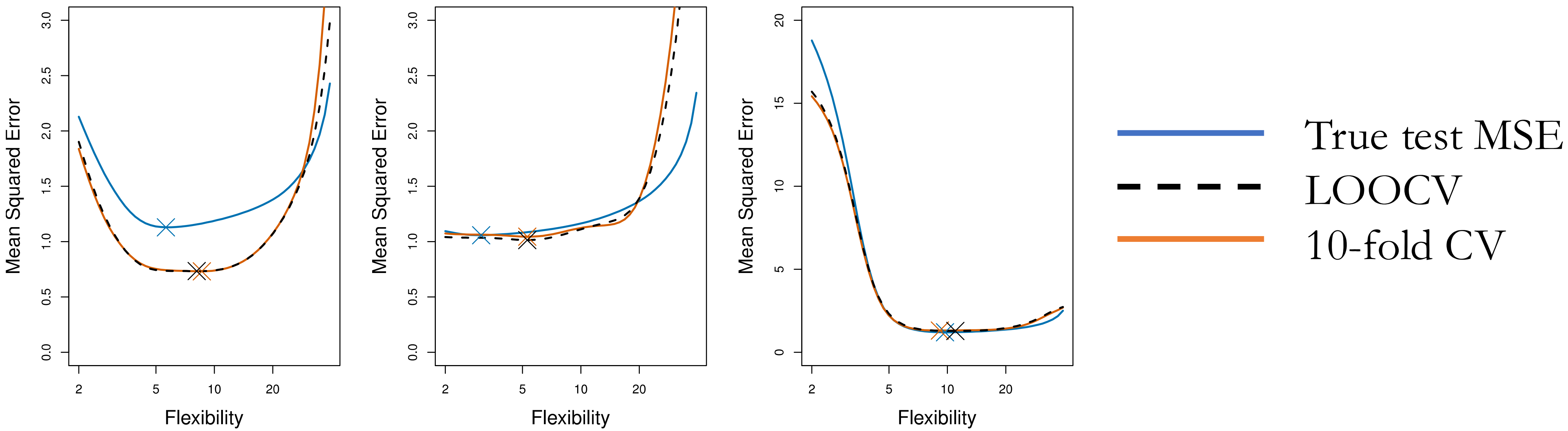
- Leave one out cross-validation
  - **Low bias**: LOOCV gives approximately unbiased estimates of the test error, as each training set contains $n - 1$ observations
  - **High variance**: LOOCV is an average of n fitted models, each of which is trained on an almost identical set of observations

- k-fold cross-validation
  - **Intermediate bias**: k-fold CV leads to an intermediate bias, as each training set contains $n - n/k$ observations
  - **Intermediate variance**: k-fold CV is an average of k fitted models that are less correlated with each other (overlapping training observations are $n - 2 \cdot n/k$)

- **Rule of thumb**: Use $k = 5$ or $k = 10$

Choosing an optimal model

- In some cases, we are only interested in the location of the minimum point in the tested test MSE curve
- **Rule of thumb**: The model with the minimum CV error often has the lowest test error
- **Example**: Regression with simulated data



Choosing an optimal model

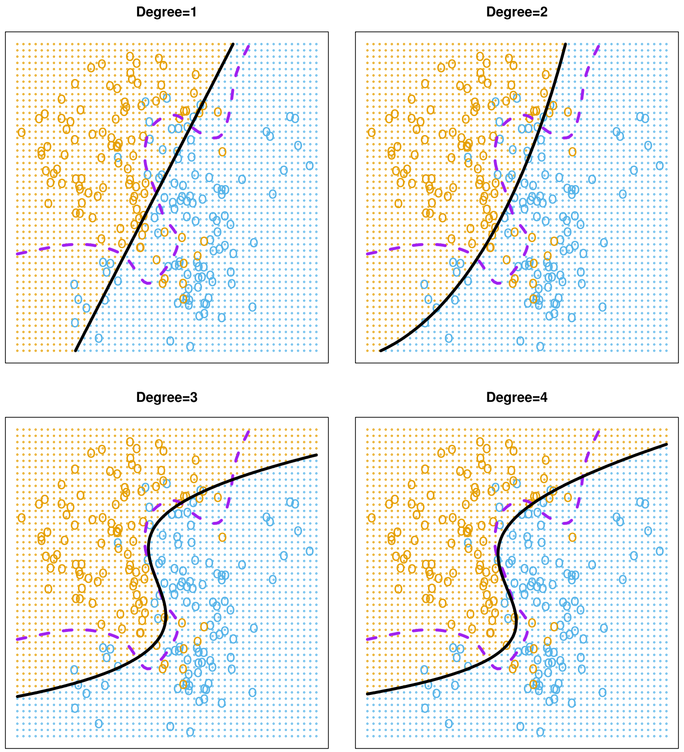
- **Example**: Classification with simulated data
  - Logistic regression with polynomial features
  - $\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_{1,1}X_1 + \cdots + \beta_{1,q}X_1^q + \beta_{2,1}X_2 + \cdots + \beta_{2,q}X_2^q$

Test error rate: 0.201

Test error rate: 0.197

Test error rate: 0.160

Test error rate: 0.162

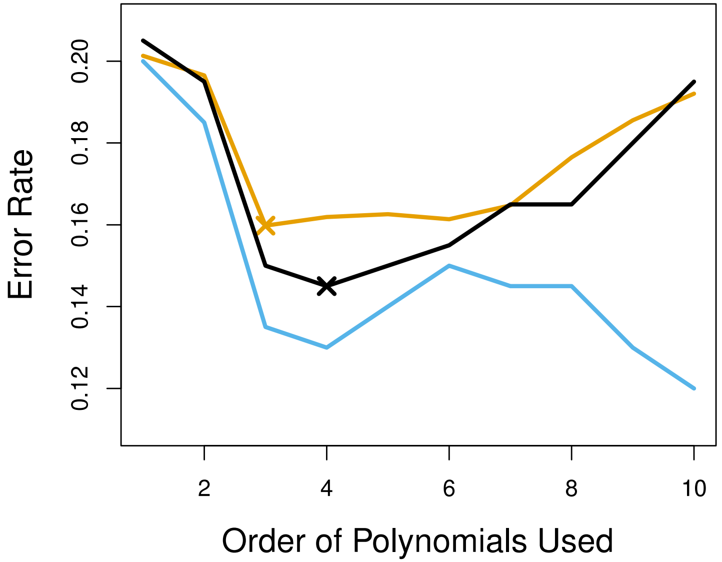


True decision boundary

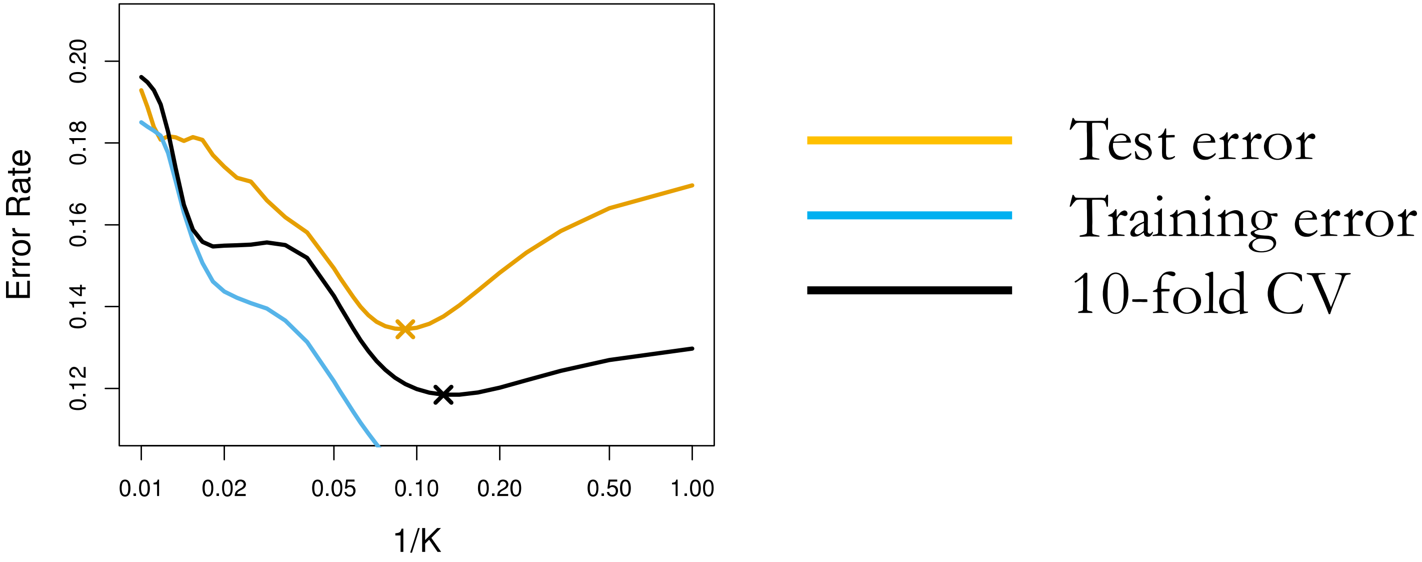
Logistic regression

Choosing an optimal model

- **Example:** Classification with simulated data
  - Logistic regression with polynomial features
  - $\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_{1,1}X_1 + \cdots + \beta_{1,q}X_1^q + \beta_{2,1}X_2 + \cdots + \beta_{2,q}X_2^q$



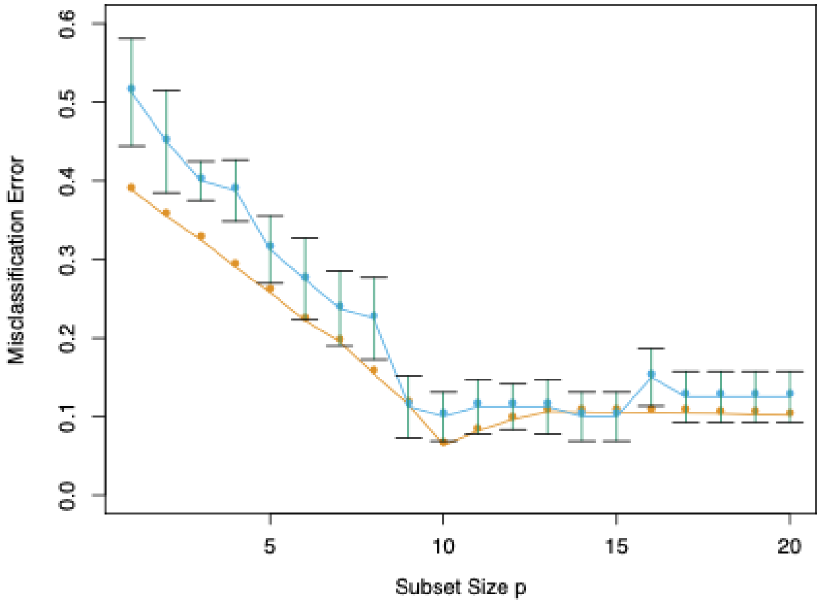
Logistic regression



KNN

Choosing an optimal model

- Example
  - A few models with have the same CV error
  - The vertical bars represent one standard error in the test error from the 10 folds



Blue: 10-fold cross validation
Yellow: True test error

- **Rule of thumb**: Choose the simplest model whose CV error is less than one standard error above the model with the lowest CV error