QTM 347 Machine Learning

Lecture 7: Cross-Validation

Ruoxuan Xiong

Suggested reading: ISL Chapter 5

EMORY

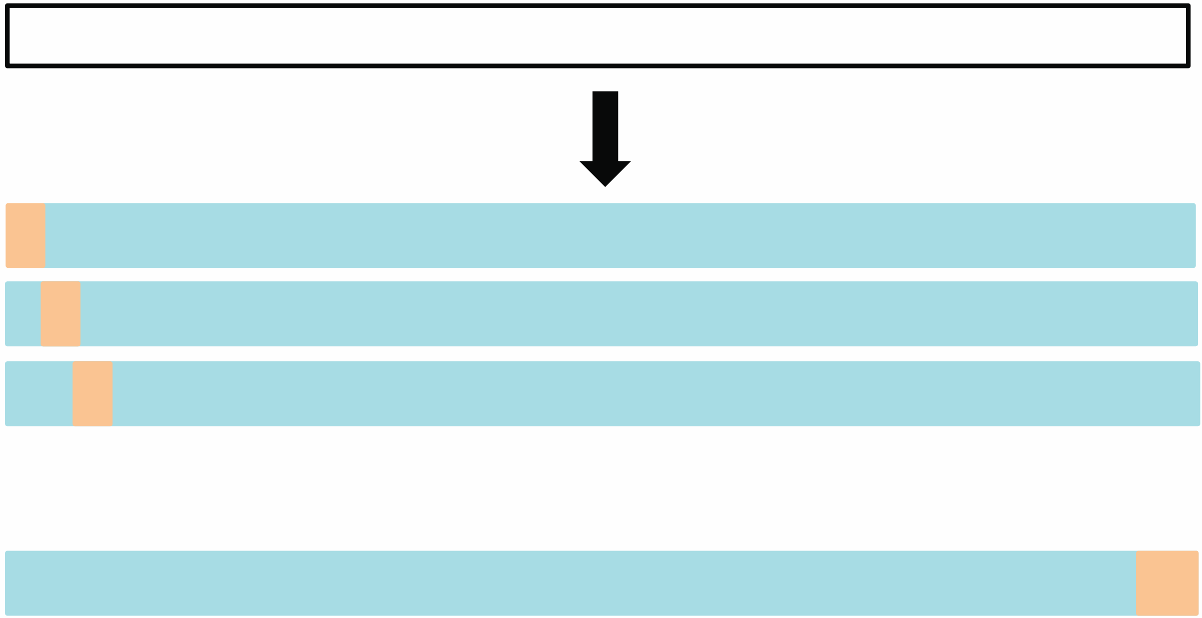
Lecture plan

- Cross validation

- Bootstrap

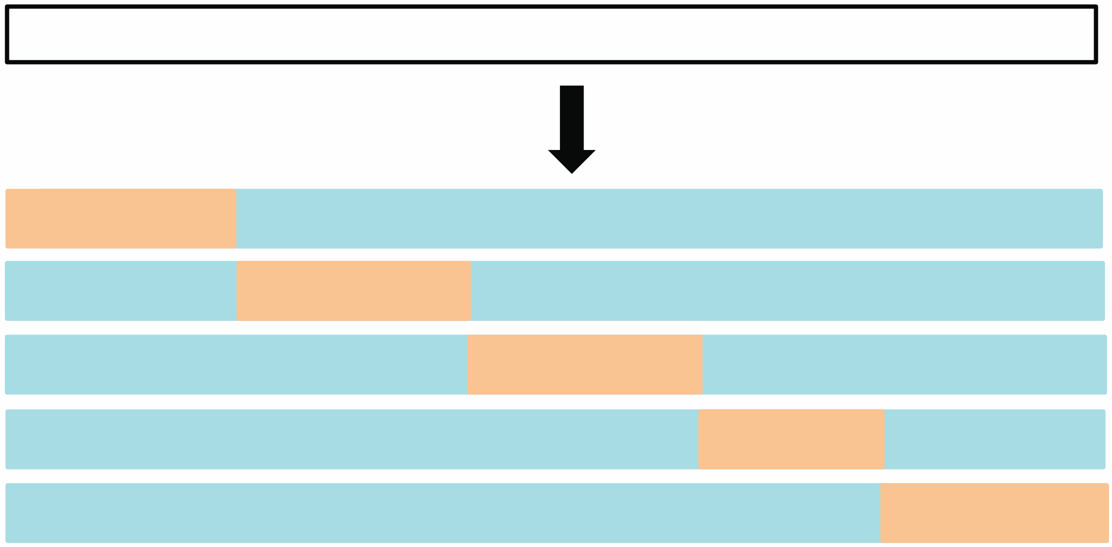
Leave one out cross-validation

- 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

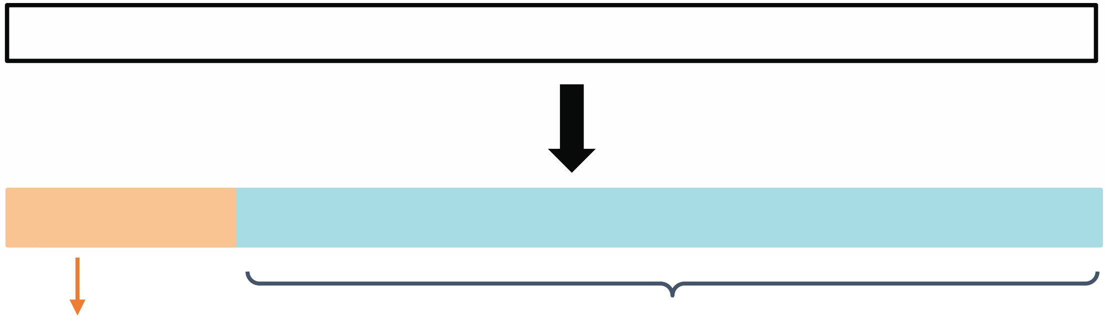


k-fold cross-validation

- 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

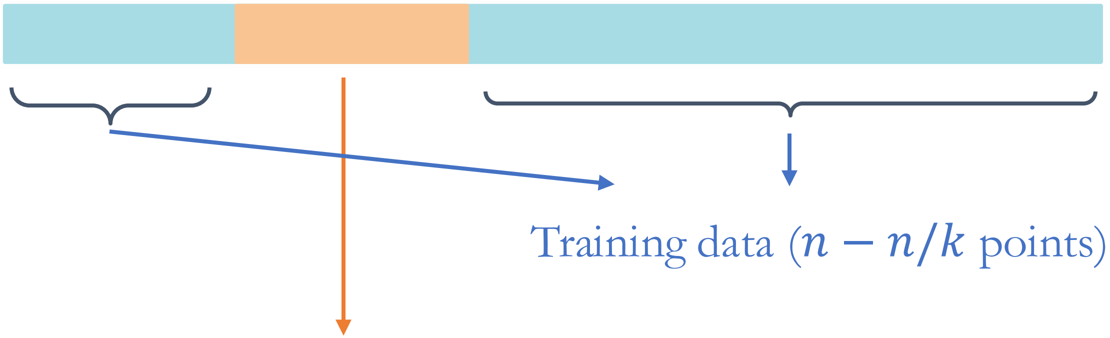


Validation data

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

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

k-fold cross-validation

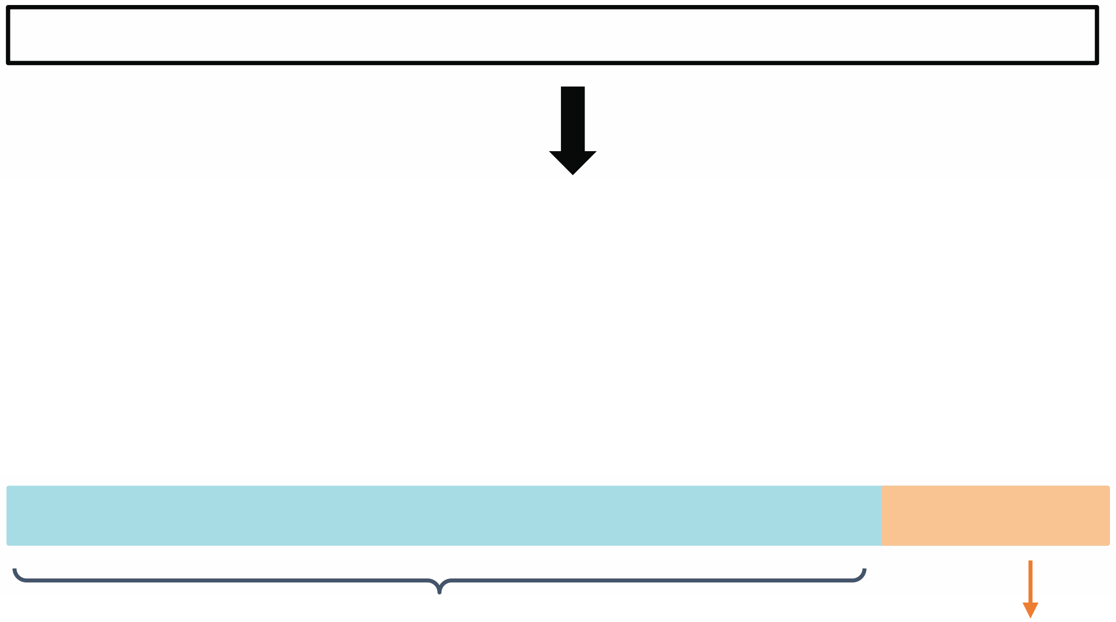


Training data ($n - n/k$ points)

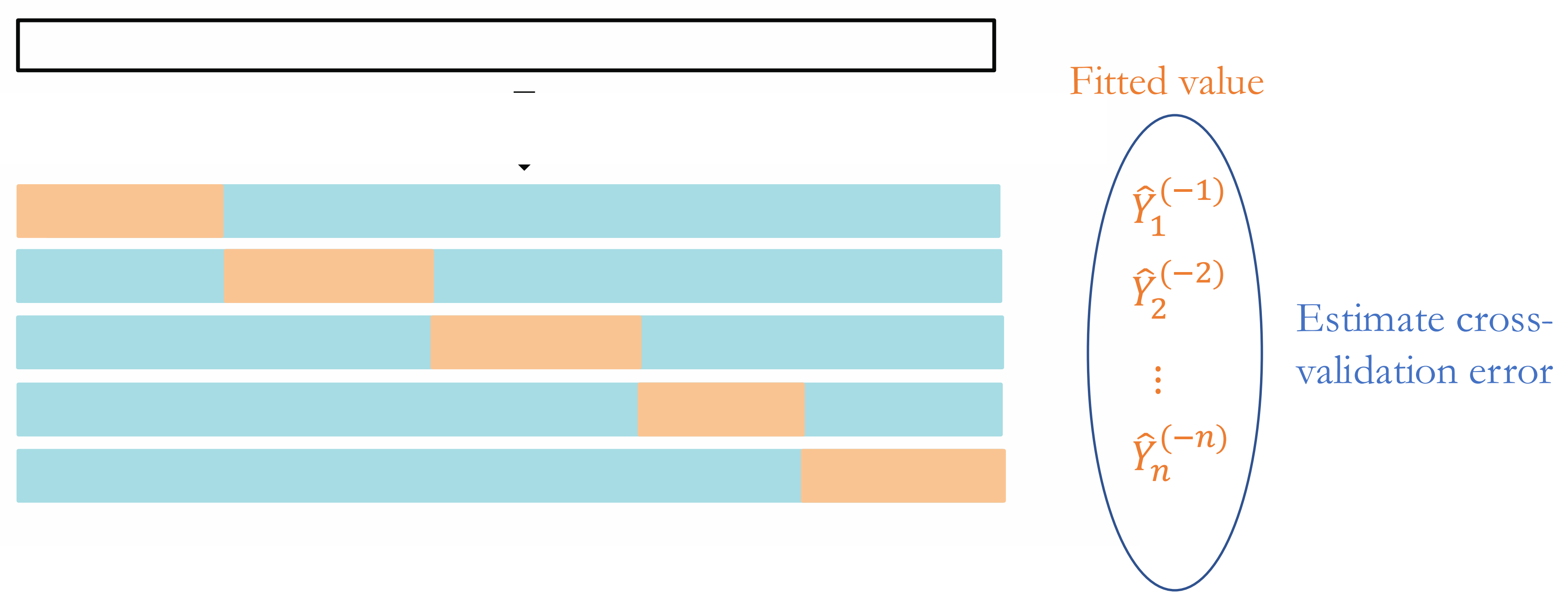
Validation data

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

k-fold cross-validation



k-fold cross-validation



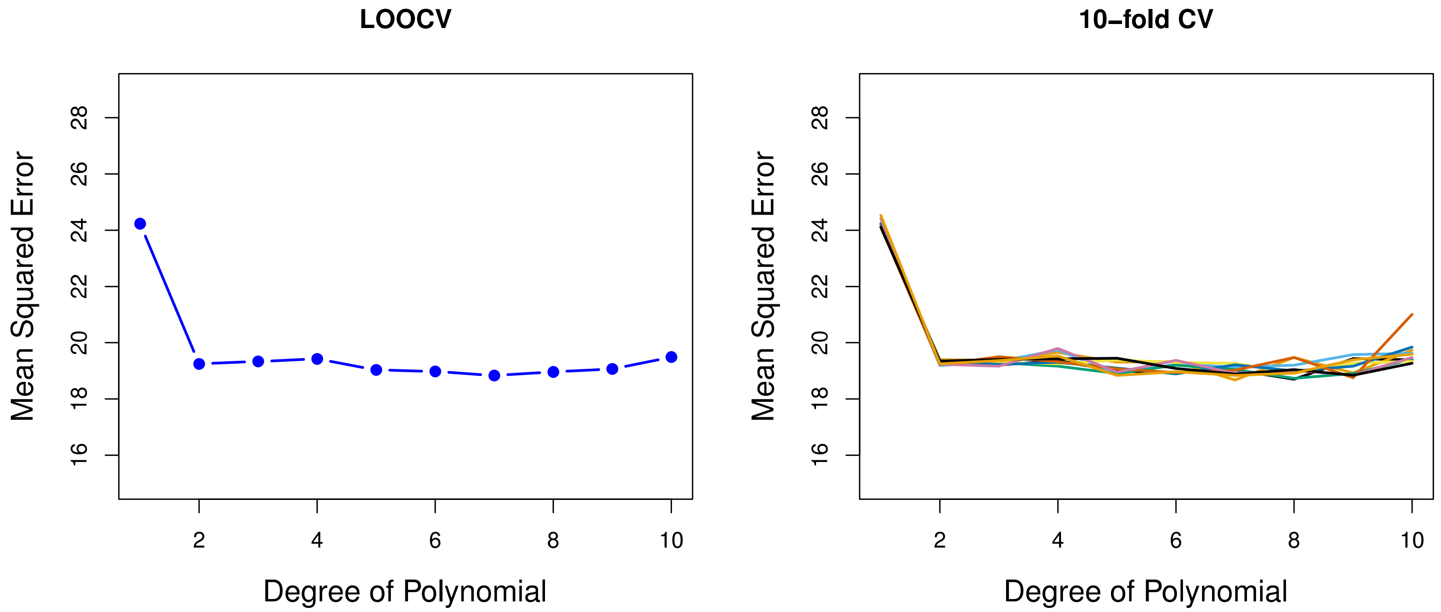
Cross-validation error

- **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
  - $CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} 1\left[Y_i \neq \hat{Y}_i^{(-i)}\right]$

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

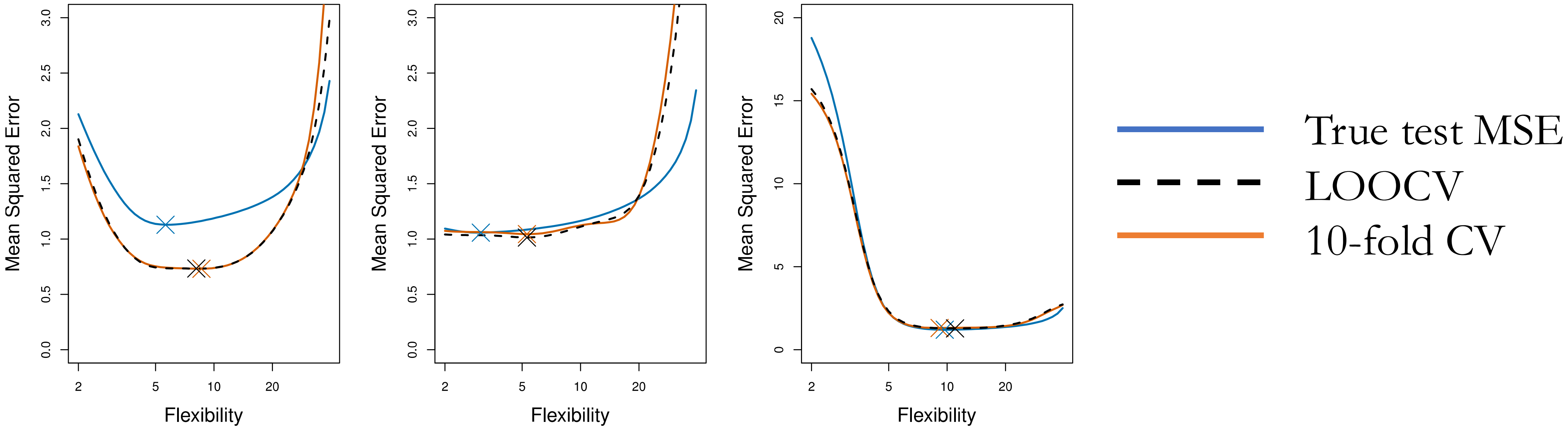
- Leave one out cross-validation
  - **Low bias**: LOOCV gives approximately unbiased estimates of the test error, as each training set contains $n-1$ observations
  - **High variance**: LOOCV is an average of n fitted models, each of which is trained on an almost identical set of observations

- 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

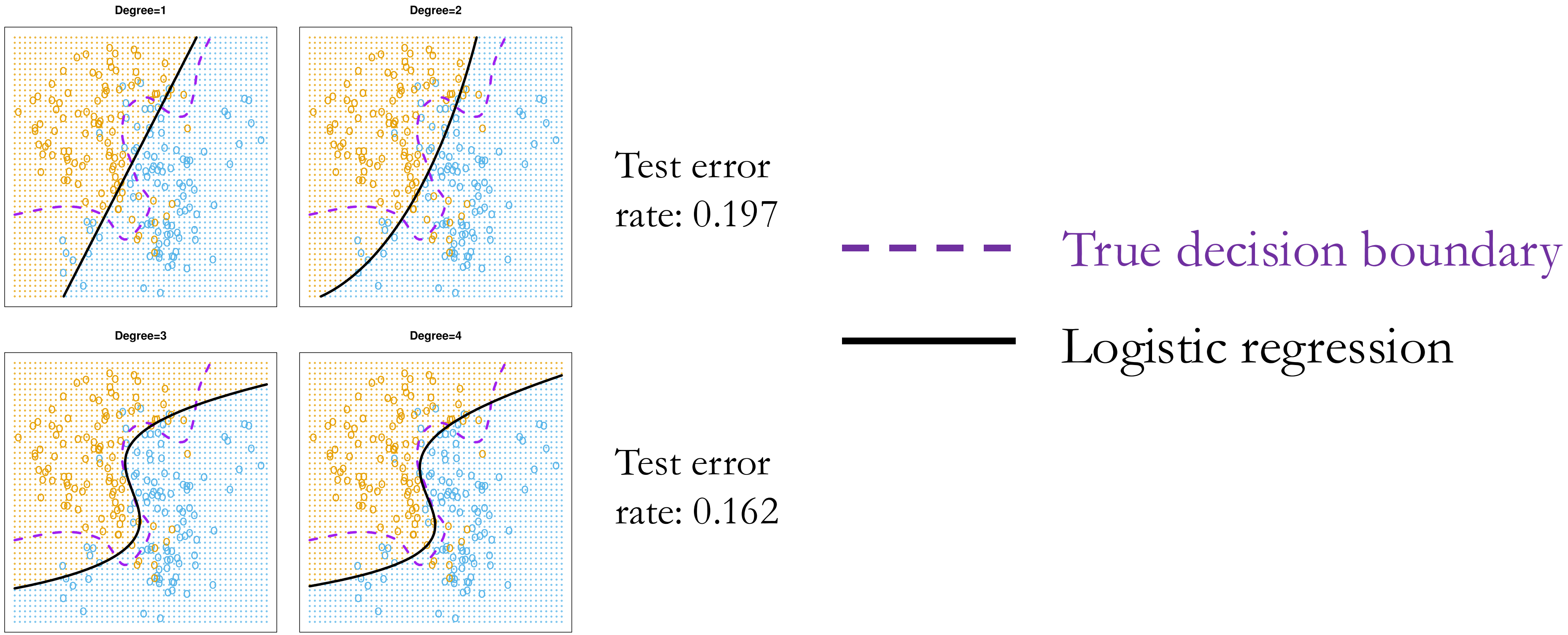
- In some cases, we are only interested in the location of the minimum point in the tested test MSE curve
- **Rule of thumb**: The model with the minimum CV error often has the lowest test error
- **Example**: Regression with simulated data



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



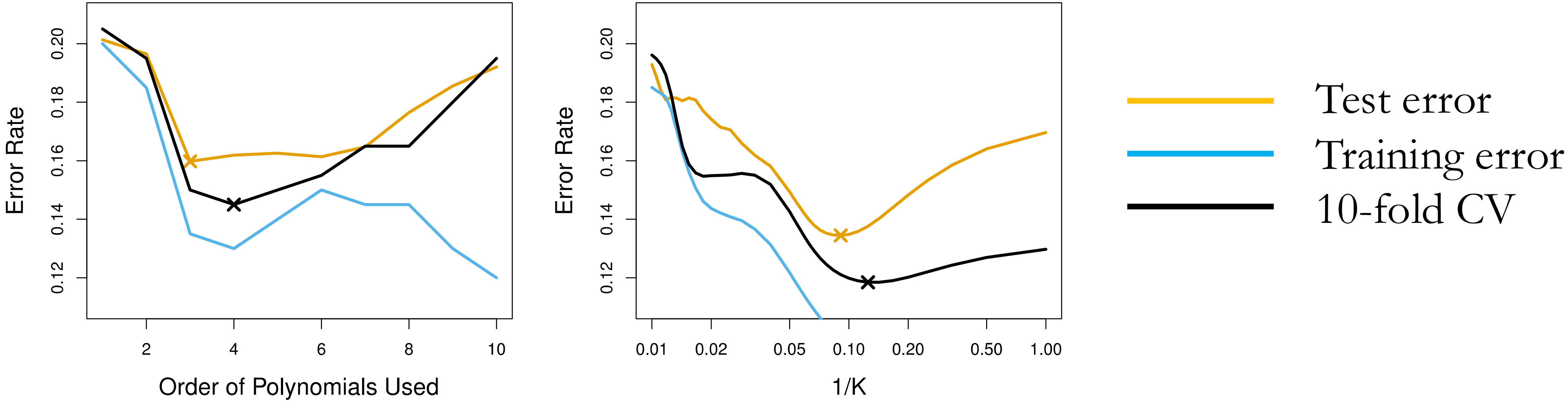
Test error rate: 0.197

Test error rate: 0.160

Test error rate: 0.162

Choosing an optimal model

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

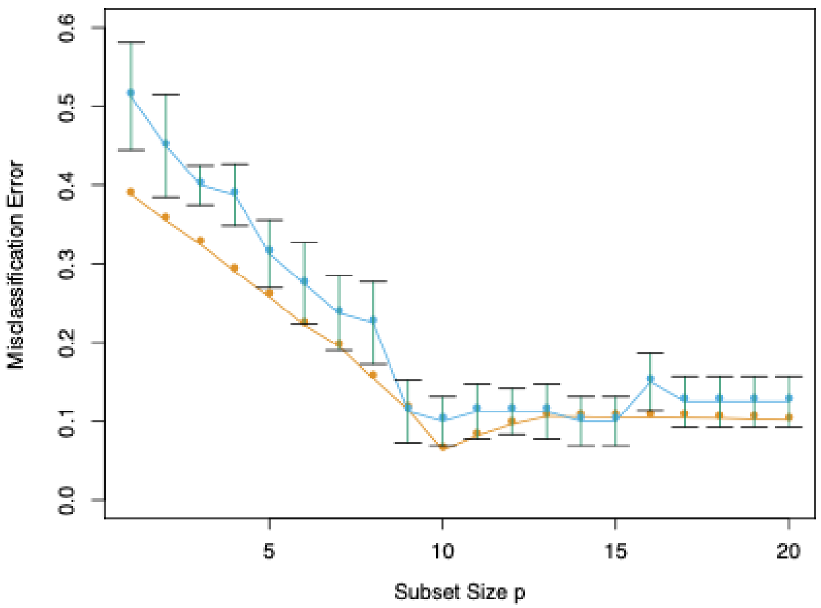


Logistic regression

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