K-fold Crossvalidation & Bootstrapping

Johnny & Quentin

January 5, 2016

Outline

- Introduction
- Resampling methods:
  - Validation/training set approach
  - Leave one out crossvalidation (LOOCV)
  - K-fold crossvalidation
- Bootstrapping
- Questions + discussion

Introduction

- Application:
  - Evaluate model performance
  - Model selection
- Through estimating the error rate of the model in supervised learning
  - Categorization
  - Regression
- Predict external validity of the model

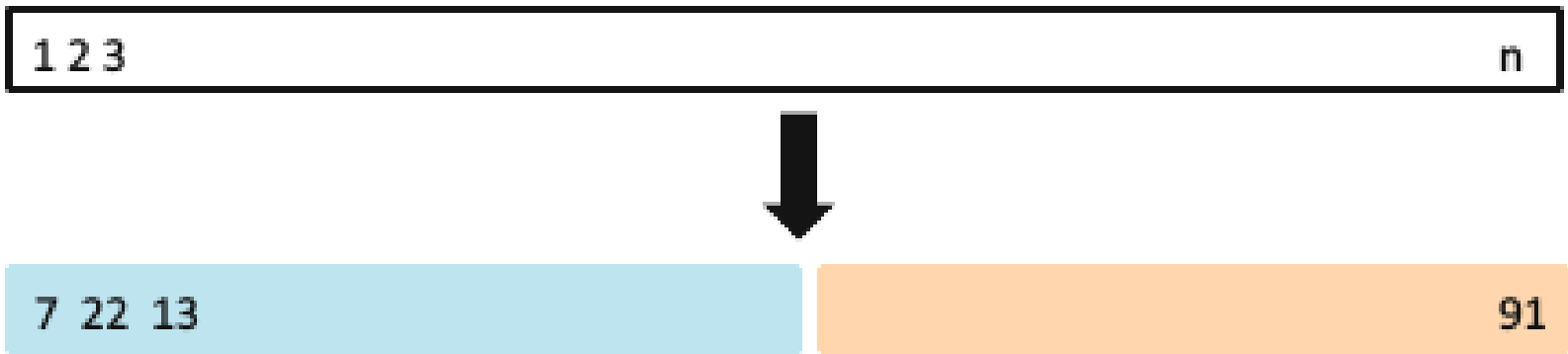
Assessing Fit

- Mean squared error (MSE) consists of:
  - Bias: failure to capture trends in the data - underfitting
    - $\mathrm{E}[\hat{f}(x)] - f(x)$
  - Variance: sensitivity to small fluctuations in the training set - overfitting
    - $\mathrm{E}[(\hat{f}(x) - \mathrm{E}[\hat{f}(x)])^2]$
  - Irreducible error
    - σ^2
- There is a trade-off between bias and variance

Validation Set Approach

- Training set & validation set
- Training error: residuals in training data
- Validation error: residuals in validation data
  - used to estimate test error rate (model performance)

Validation Set Approach



- Illustration of the validation set approach

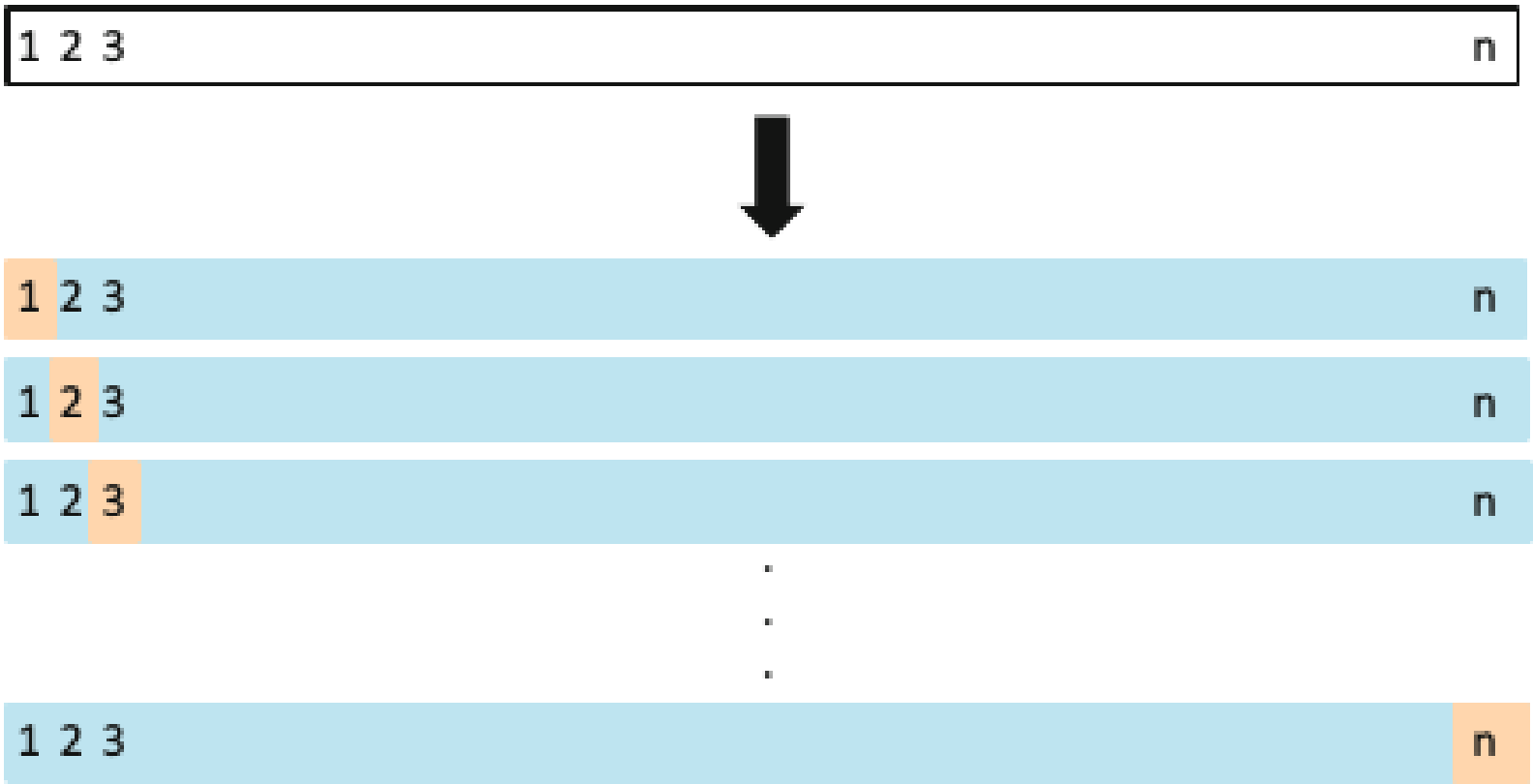
Validation Set Approach: Drawbacks

- High variability depending on the random split
- Overestimation of test error
  - Model not based on full data: generally less accurate
- How to improve this high bias and variance?

Leave One Out Cross-Validation (LOOCV)

- Use $n-1$ observations as training set
- Test the model on the remaining observation
  - Calculate $\text{MSE}_1 = (y_1 - \hat{y}_1)^2$
- Repeat n times:
  - test error estimate: $CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{MSE}_i$

Leave One Out Cross-Validation (LOOCV)



- Illustration of the LOOCV algorithm

LOOCV: Advantages

- Less bias: no overestimation of test error
  - Use almost full data set
- No randomness in the splits

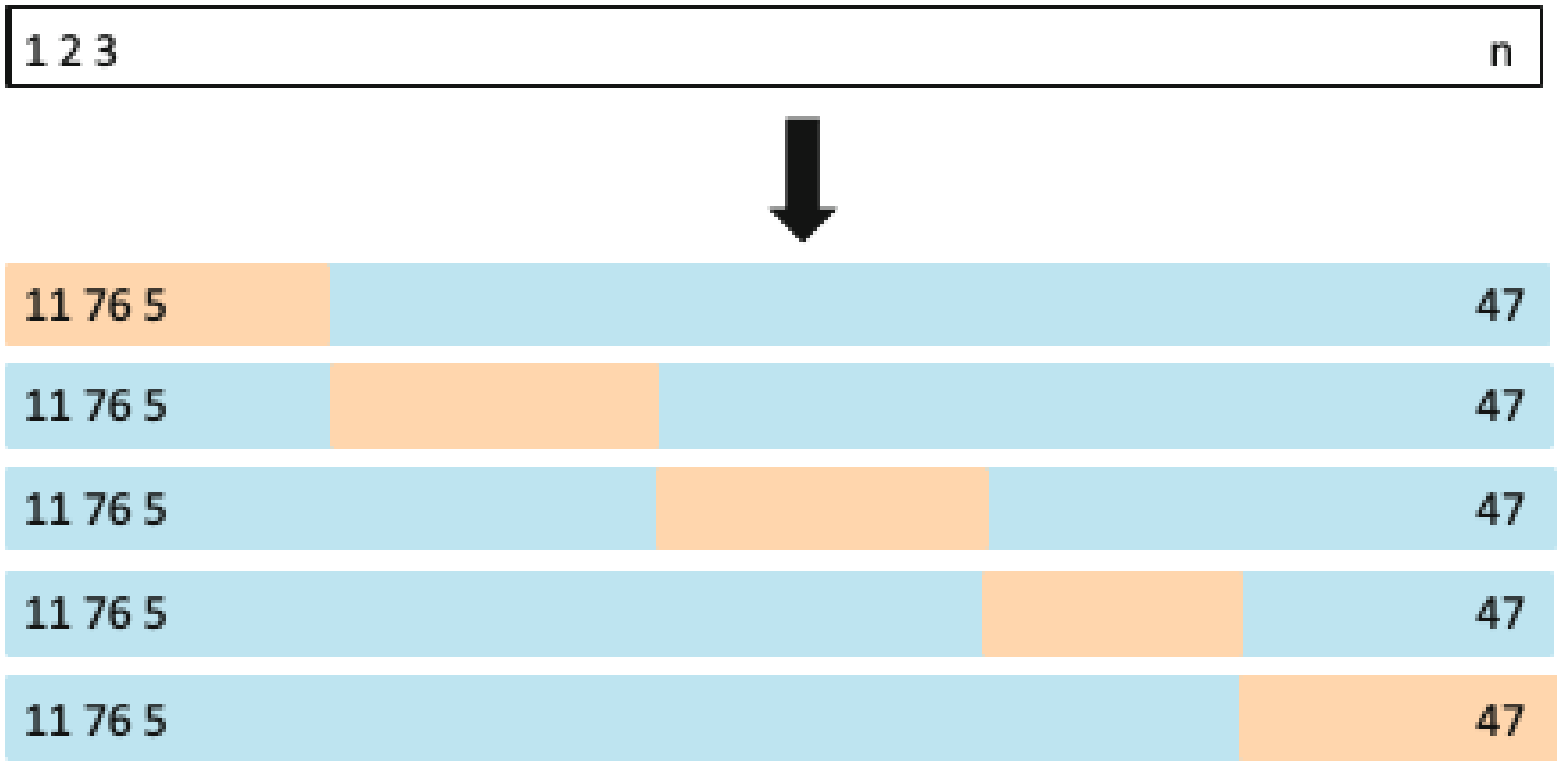
LOOCV: Drawbacks

- Computationally expensive: model has to be fit n times
- High variance of the individual fits that are averaged:
  - Validation errors are based only on 1 data point

K-Fold Cross-Validation: the Generalization

- Divide the data in k groups (=folds) of equal size
- Use $k-1$ groups as training set
- Test the model on the remaining group
- Repeat k times:
  - test error estimate: $(CV_{(k)}) = \frac{1}{k}\sum_{i=1}^{k} \text{MSE}_i$

K-Fold Cross-Validation



- Illustration of the K-fold cross-validation algorithm

K-Fold vs. LOOCV

- LOOCV is KFCV with $k = n$
- LOOCV fits the model n times, KFCV fits the model k times
- Trade-off between bias and variance
  - LOOCV: low bias in test error estimation, high variance
    - Highly correlated estimates, a lot of overlap between folds
  - KFCV: higher bias, lower variance
    - High number of permutations: estimates are all over the place

Extending to Classification

- Instead of MSE, classification error is used
  - $\text{Err}_i = I(y_i \neq \hat{y}_i)$
- The algorithm remains the same

Illustrating bias

```
## Shiny App
```

How to deal with the trade-off

- The higher k, the more risk of overfitting
- The lower k, the more uncertain the estimate of model fit (MSE)
- Suggested to use $k = 5$ or $k = 10$
- Resampling as a powerful alternative to training/validation set approach

However...

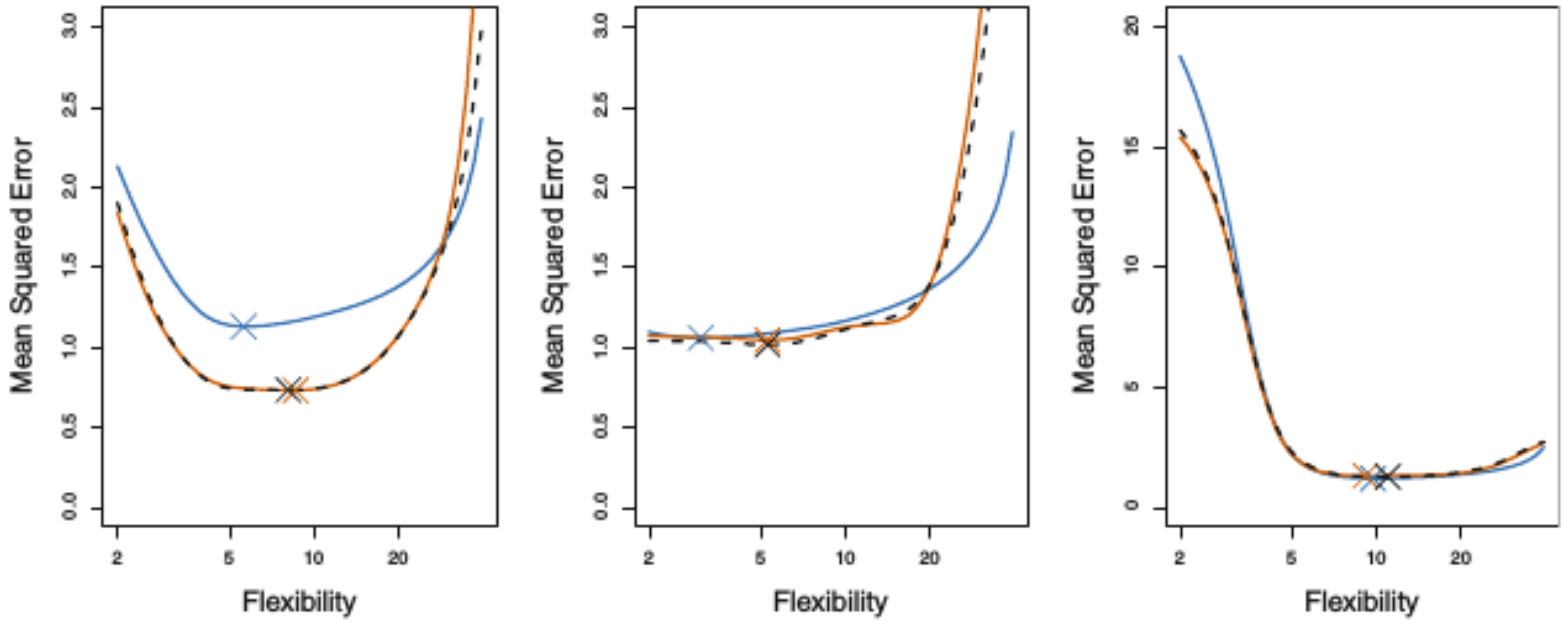


FIGURE 5.6. *True and estimated test MSE for the simulated data sets in Figures 2.9 (left), 2.10 (center), and 2.11 (right). The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the 10-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.*

- Depending on whether goal is model assessment or model selection, the estimated value of MSE may not matter, only the shape of the curve

Quentin Time!!