Big Data for Public Policy

Statistical Learning [Part 2]

Malka Guillot

ETH Zürich | 860-0033-00L

Turn on recording

Table of contents

Prologue

Coming back on the homework

- Most challenging homework
- Great spirit on the forum!
- Organizing an intro to python session (voluntary participation)

By now you should have:

- Installed Anaconda, with Jupyter-notebook and Spyder
- (Installed Git)
- Created a GitHub account
- Joined the Moodle class
- Registered for a presentation

Last week

- What is statistical learning?
- Statistics in social science – causality.
- Statistics in machine learning – prediction.
- Accuracy v. interpretability.

Today

- Model accuracy
- The bias-variance tradeoff.
- Classification

Reference: JWHT, chap 2.2 & 5.1

Model Accuracy

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- **Regression setting**: the **mean squared error** is a metric of how well a model fits the data.
- But it's **in-sample**.
- What we are really interested in is the **out-of-sample** fit!

Measuring fit (1)

- We would like $(y_0 - \hat{f}(x_0))^2$ to be small for some (y_0, x_0), not in our training sample $(x_i, y_i)_{i=1}^n$.
- Assume we had a large set of observations (y_0, x_0) (a test sample),
- then we would like a low

$$Ave(y_0 - \hat{f}(x_0))^2$$

- i.e a low average squared prediction error (test MSE)

Measuring fit (2)

To estimate model fit we need to partition the data:

1. **Training set**: data used to **fit** the model
   - **Training MSE**: how well our model fits the training data.
2. **Test set**: data used to **test the fit**
   - **Test MSE**: how well our model fits new data

We are most concerned in **minimizing test MSE**

Training MSE, test MSE and model flexibility

Overfitting

- As model flexibility increases, training MSE will decrease, but the test MSE may not.
- When a given method yields a small training MSE but a large test MSE, we are said to be **overfitting** the data.
- (We almost always expect the training MSE to be smaller than the test MSE)
- Estimating test MSE is important, but requires training data...

The Bias-Variance Trade-Off

Decomposing the expected (test) MSE

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

3 components:

1. $Var(\hat{f}(x_0)) =$ Variance of the predictions
   - how much would $\hat{f}$ change if we applied it to a different data set
2. $[Bias(\hat{f}(x_0))]^2 =$ Bias of the predictions
   - how well does the model fit the data?
3. $Var(\epsilon) =$ variance of the error term

The bias-variance tradeoff

Accuracy in Classifications

$$(\text{training})\ \text{error rate} = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \neq \hat{y}_i)$$

$$(\text{test})\ \text{error rate} = Ave(1(y_0 \neq \hat{y}_0))$$

- MSE in the context of regression (continuous predictor).
- Modifications in the setting in which we're interested in prediction classes
- We are essentially interested in what % of classifications are correct.
- For cross-validation we could also use the estimated test error rate

How to choose training and test set?

Resamling methods

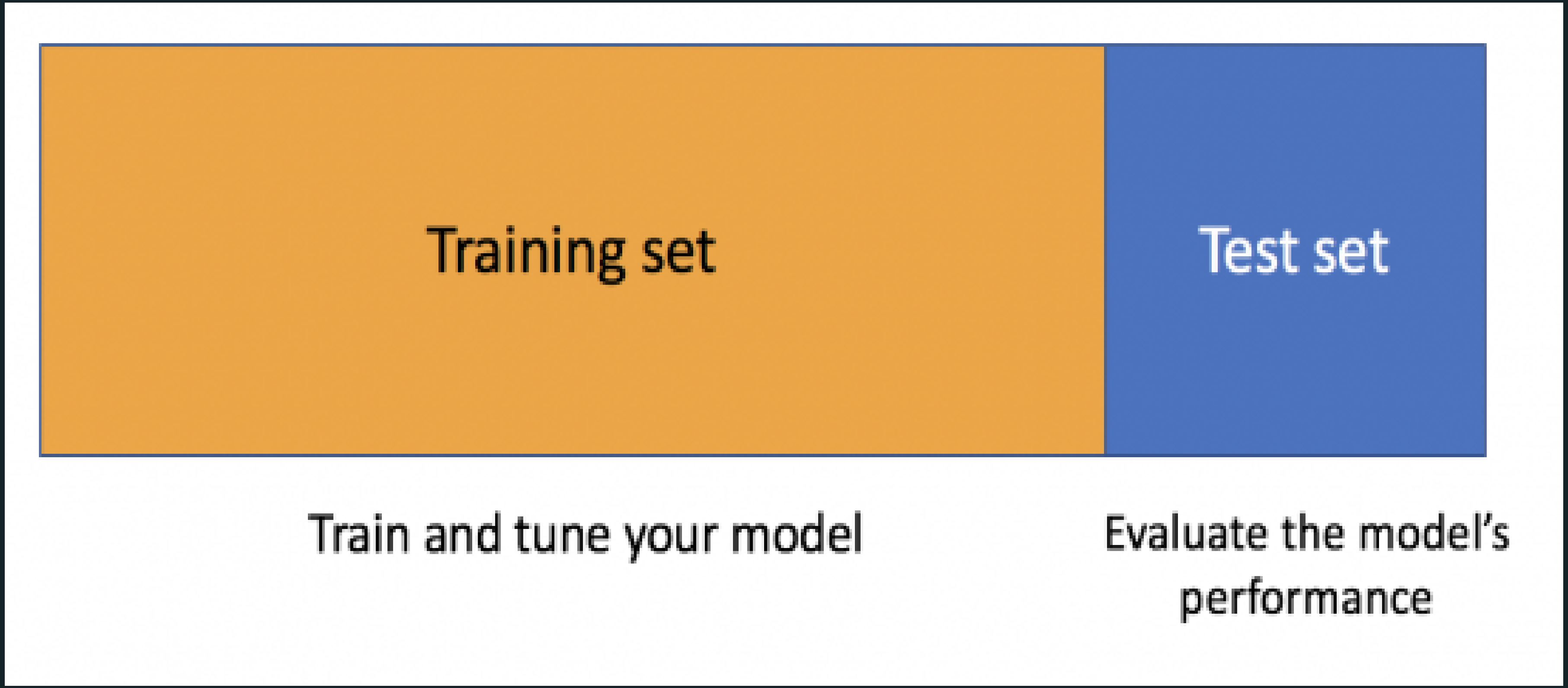
Estimate the test error rate by

holding out a subset of the training observations from the fitting process,

$+$ then **applying** the statistical learning method to those held out observations

Validation set approach

- Randomly divide labeled data **randomly** into two parts: training and test (validation) sets.



Two concerns

- Arbitrariness of split
- Only use parts of the data for estimation

  $\rightarrow$ we tend to overestimate test MSE because our estimate of $f(x)$ is less precise

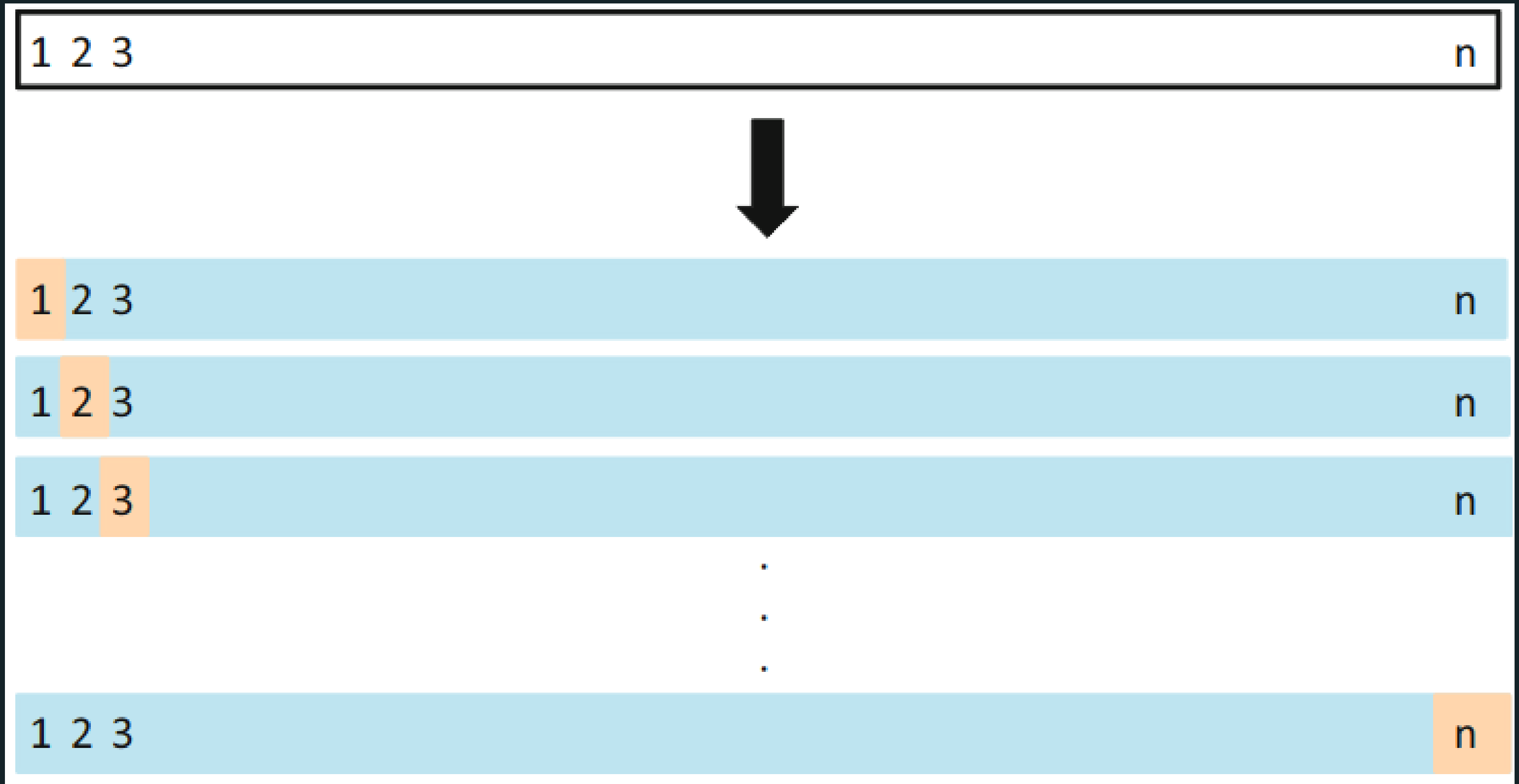
Leave-One-Out Cross-Validation (LOOC)

- Fit on $n-1$ training observations, and a prediction the Last
- Iterate n times
- Assess the average model fit across each test set.

Estimate for the test MSE:

$$CV_n = \sum_{i=1}^{n} MSE_i$$

Leave-One-Out Cross-Validation (LOOC)



- less bias than the validation set approach
- always yield the same results
- potentially too expensive to implement

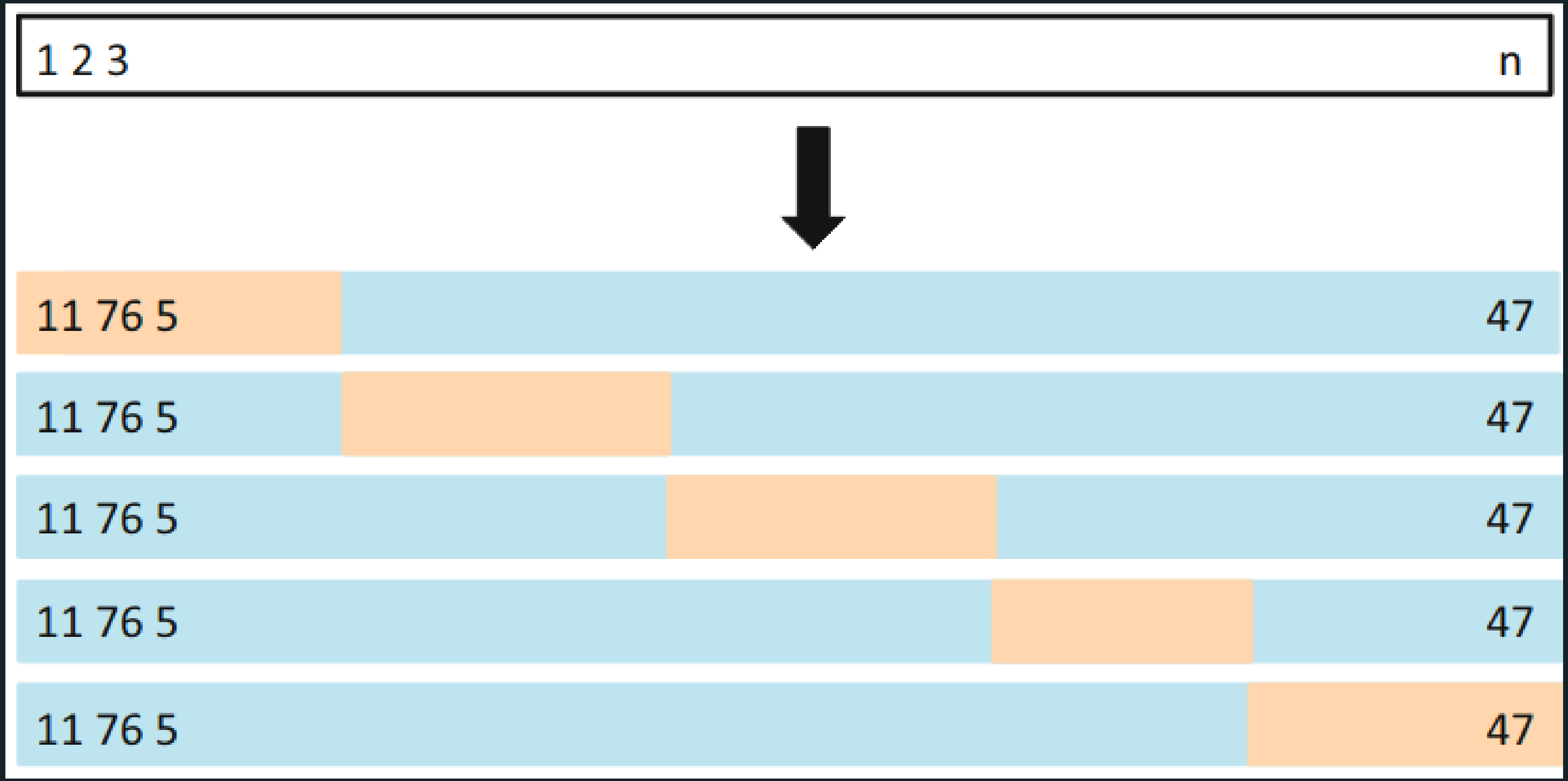
k-fold Cross-validation

- Leave-One-Out Cross-Validation with $k = 1$
- Randomly dividing the data into the set of observations into k groups
- 1st fold is treated as a validation set, and the method is fit on the remaining $k - 1$ folds
- Iterate k times

Estimate for the test MSE:

k

k-fold Cross-validation



$\Rightarrow$ Arguably the contribution to econom(etr)ics: Cross-validation (to estimate test MSE)!

Bias-Variance Trade-Off f-Fold Cross-Validation

Bias

- **validation set approach** can lead to overestimates of the test error rate
- **1-fold validation**: almost unbiased estimates of the test error
- **k-fold validation** is in between

Variance

- **1-fold validation**: higher variance
- **k-fold validation**: lower variance

$k = 5$ or $k = 10$ is a good benchmark

Turn off recording