NOTES 5

Resampling Methods

Acknowledgement: some of the contents are borrowed with or without modification from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R.Tibshirani.

Model Validation

- Training error is different from testing error
- To validate the model performance, hold out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations
- Divide the available set of samples into two parts: a training set and a validation or hold-out set.
- The resulting validation-set error provides an estimate of the test error.

Overfitting

- When a given method yields a small training MSE but a large test MSE, we are said to be *overfitting* the data.
- Mean Square Error (MSE):a measure of the overall model fit

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

Holdout Procedures

- Common case: data set large but limited
- Problems:
  - Want both sets to be representative
  - Want both sets as large as possible

100

	A	B
	80	20

eg:

		A		B	
Train:	80	~~80~~	64	~~0~~	16
Test:	20	~~0~~	16	~~20~~	4

- Simple check: Are the proportions of classes about the same in each data set?
- Stratified holdout
  - Guarantee that classes are (approximately) proportionally represented
- Repeated holdout
  - Randomly select holdout set several times and average the error rate estimates

Want both sets as large as possible…

- In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.

- The limited size of the training data can also affect the estimation of testing error

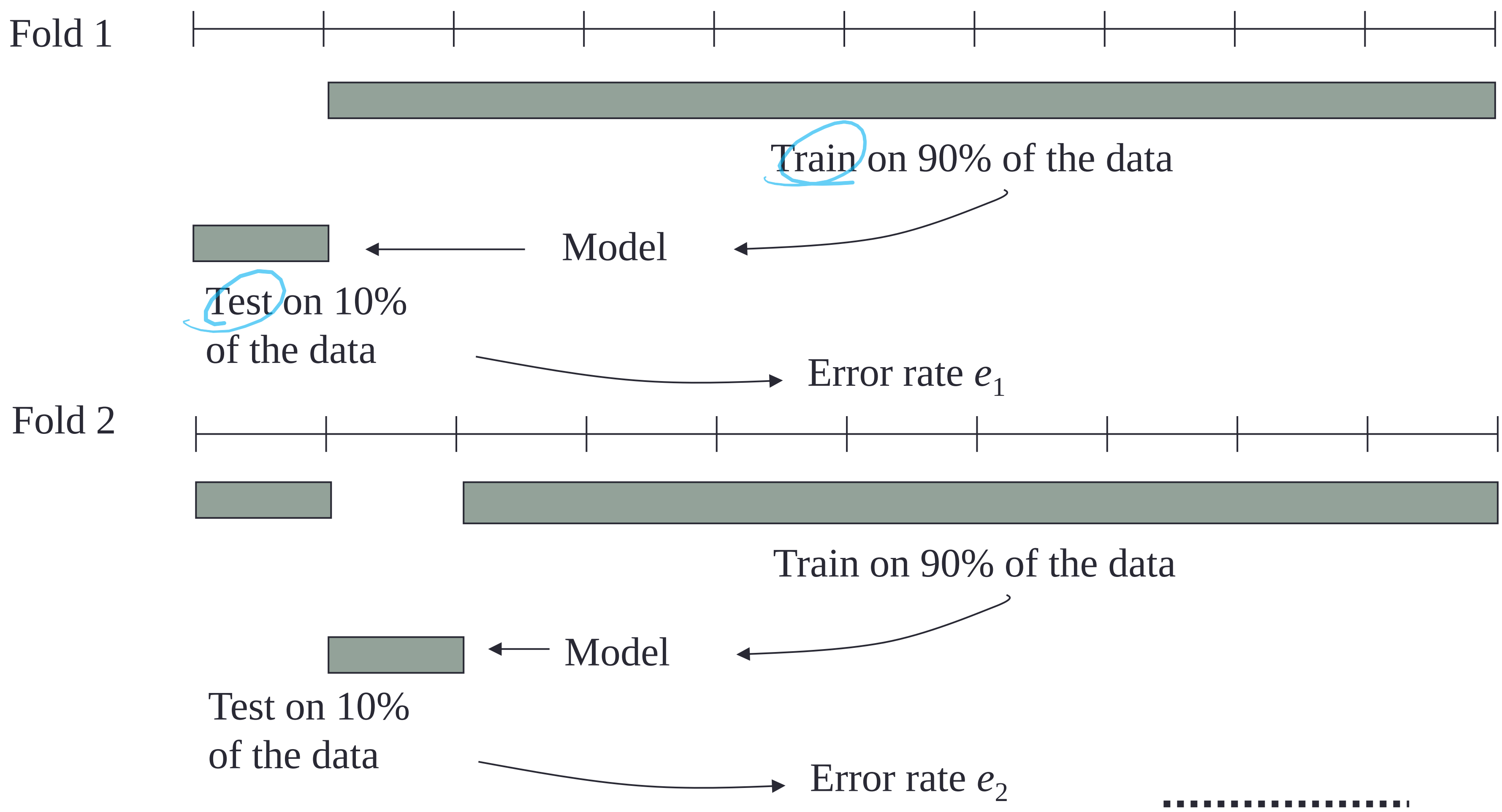
- Use K-fold Cross Validation

Cross-Validation

divide all data into [] folds
use 1 of the [] for testing

- Cross-validation
  - Fixed number of partitions of the data (*folds*), e.g. 5 and 10
  - In turn: each partition used for testing and remaining instances for training
  - May use stratification and randomization
  - Standard practice:
    - Stratified tenfold cross-validation
    - Instances divided randomly into the ten partitions

Cross Validation



Cross-Validation

- Final estimate of error

$$\bar{\delta} = \frac{1}{k}\sum_{i=1}^{k}\delta_k$$

- MSE is used in the case of a quantitative response and misclassification rate in the case of a qualitative response.

Leave-One-Out Holdout

- Setting K = n yields n-fold or leave-one out cross-validation (LOOCV). (n instance set)

num. of observation

- Use all but one instance for training
  - Maximum use of the data
  - High computational cost
  - Non-stratified sample

Lmz testing data always belongs to 1 category

Bootstrap

when we don't have enough observation for even cross-validation (~ 100 to 200)

- If the size of the original data set is n, sample it **with** replacement n times

  Creat a new dataset with the same size of the original dataset

  - Use as training data
  - Use instances not in training data for testing
- How many test instances are there? $\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$

size

likelihood of an observation not selected = select for testing

$(1-\frac{1}{n})^n \approx e^{-1} \approx 0.368$

0.632 Bootstrap

- On the average $e^{-1}n = 0.368\ n$ different instances will be in the test set, where n is the size of the original data set
- Thus, on average we have 63.2% of instance in training set
- Estimate error rate

$$e = 0.632\ e_{test} + 0.368\ e_{train}$$

On average, there'll be 63.2% be selected in testing dataset

Sample With Replacement

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original Data (Z)

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set.

What to Do?

- “Large” amounts of data
  - Hold-out 1/3 of data for testing
  - Train a model on 2/3 of data
  - Estimate error (or success) rate and calculate CI
- “Moderate” amounts of data
  - Estimate error rate:
    - Use 10-fold *cross-validation* with stratification,
    - or use *bootstrap*.
  - Train model on the entire data set

usually used

accurate