Recap 6

Generalization

ISLR 5, ESL 7,8

Jilles Vreeken
Krikamol Muandet

UNIVERSITÄT DES SAARLANDES

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

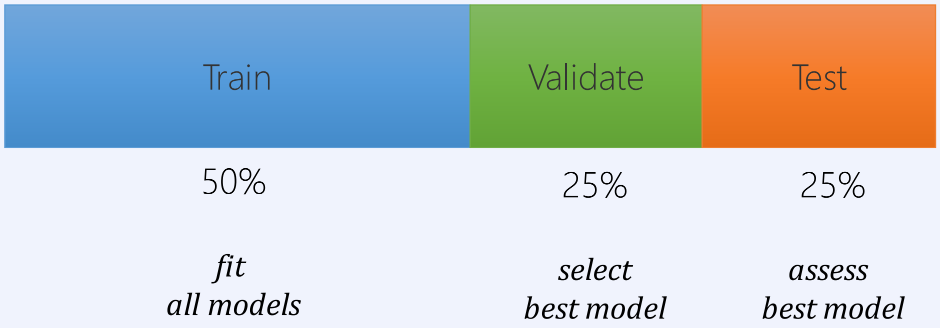
Lecture Recap 1

- Validation
  - To estimate how well a model generalizes, we should test on different data than we trained on.
  - Therefore one can divide the data into 2 parts, train and validation data. However, a more conservative pipeline is to have a 3 way division of train, validation and test.

Train-Validation-Test Paradigm

More conservatively, we can divide the data **three-way**

1. **training set** for fitting models
2. **validation set** for comparatively assessing model performance in order to select a model
3. **test set** in order to assess the performance of the selected model



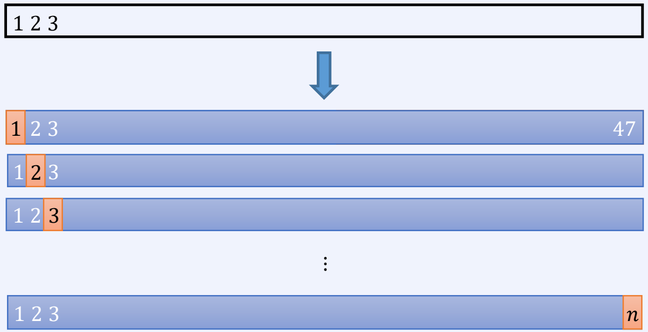
Lecture Recap 1

- Leave-one-out Cross Validation (LOOCV)
  - Idea is to leave one point from data aside to test. Large train data means little bias in training but test data is small so high variance.

Leave-one-out Cross Validation (LOOCV)

Key idea: set only **one data point aside** for testing

- training set is now as large as can be, so little bias
- but, only one point to test on, so high variance



Repeating for every data point averages out variance

$$MSE_i = (y_i - \hat{y}_i)^2$$

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

- process is **deterministic**, repeating always gives same result
- for least-squares linear or polynomial regression we have

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

Lecture Recap 1

- K-Fold Cross Validation
  - Divide the training data into random $\boldsymbol{k}$ folds, train on $\boldsymbol{k}$-1 folds and validate on held out data.
  - Larger relative size of training data reduces bias but increases the variance due to smaller val data.

k-fold Cross Validation

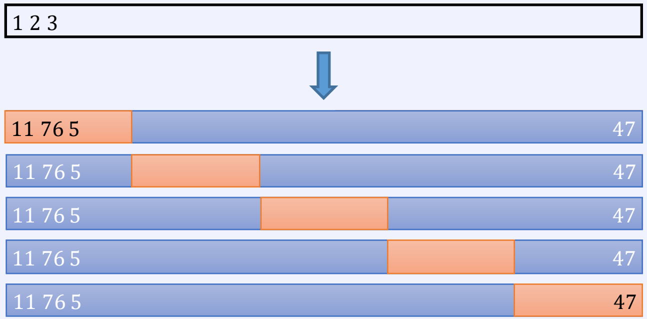
Randomly divide the data into k folds

- train on $k-1$ folds, test on the remaining 1 fold
- repeat such that all folds have been tested on
- gives k estimates of the test error, the final estimate is

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

- in practice, we use $k = 5$ or 10
- LOOCV is k-fold CV with $k = n-1$
- k-fold CV is more efficient but has higher bias than LOOCV

- In general due to bias variance tradeoff, k-fold CV often gives more accurate error estimates than LOOCV! Since k-fold CV has less overlap in training data, and hence less correlated estimates



Lecture Recap 1

- Bootstrap
  - Bootstrap is used to quantify the uncertainty of a given estimator
  - Is applicable to all kinds of methods for which no theory exists

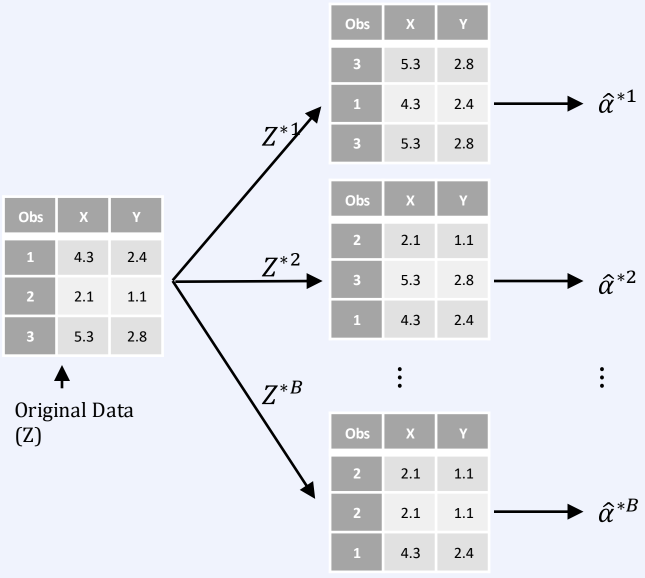
Bootstrap

Key idea: sample subset of data for training:

- training set is sampled from original set with replacement.
- Calculate the statistic of interest. Example: Train the model and compute error.
- Repeat the above two steps a large number of times.

Bootstrap samples are highly correlated, which increases the variance of the error estimate.

However, re/sub-sampling methods like bootstrap allow to learn about the variability of the fitted models, as the training set changes.



Bootstrap on a data set of 3 rows