MACHINE LEARNING

CROSS-VALIDATION

Cross-validation

- OK, we train our regression models on particular datasets
- Is it going to work fine on independent datasets as well?
- We want to estimate how accurately a predictive model will perform in practise
- One round of cross-validation involves partitioning a sample of data into complementary subsets
- Training set / test set / validation set
- **PROBLEM**: conventional partitioning (70% - 30%)is not a good estimator of performance because the data is not enough or the partition is not appropriate
- Cross-validation is better !!!

Techniques

1.) exhaustive cross-validation: learn and test on all possible ways to divide the original sample into a training and a validation set

2.) non-exhaustive cross-validation: these methods do not compute all ways of splitting the original sample

~ **k-fold cross validation** !!!

k-fold cross validation

- The original dataset is randomly partitioned into **k** equal sized subsamples
- Of the **k** subsamples → a single subsample is retained as the validation data for testing the model and the remaining **k-1** subsamples are used as training data
- This cross-validation process is then repeated **k** times („folds") with each of the subsamples used exactly once as the validation data
- The **k** results from the folds can then be averaged and combined to produce a single estimation
- **ADVANTAGE**: all observations are used for both training and validation + each observations are used for validation exactly once