Cross - validation

Machine Learning

Outline

- Motivation
- Cross - validation
- Cross - validation in R

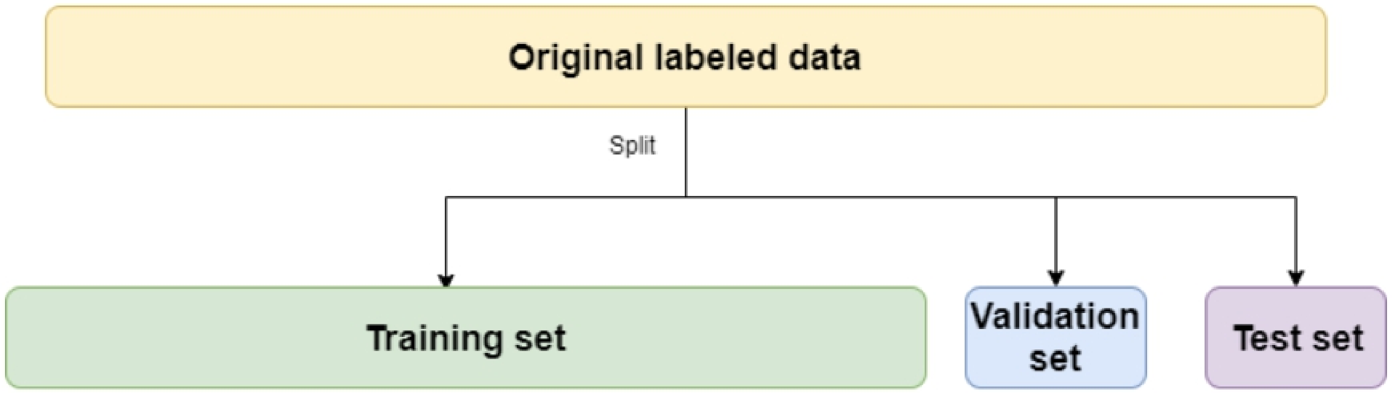
Cross - validation and Bootstrap

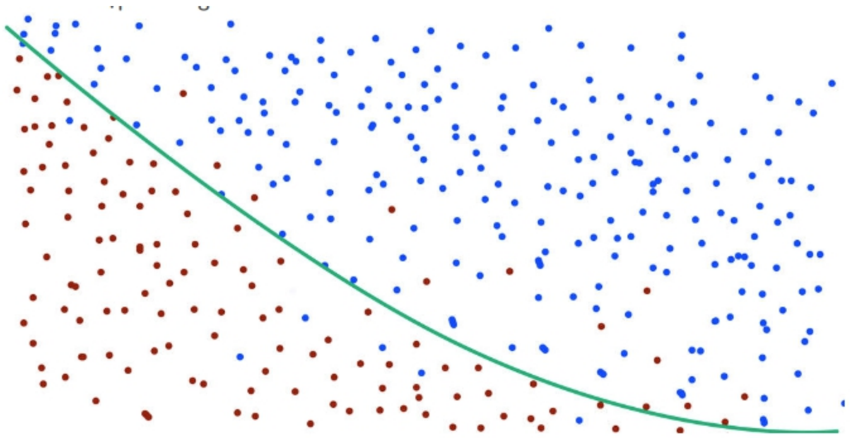
In the section we discuss two resampling methods: cross-validation and the bootstrap.

- These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
- For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates

Cross - validatiion: Motivation

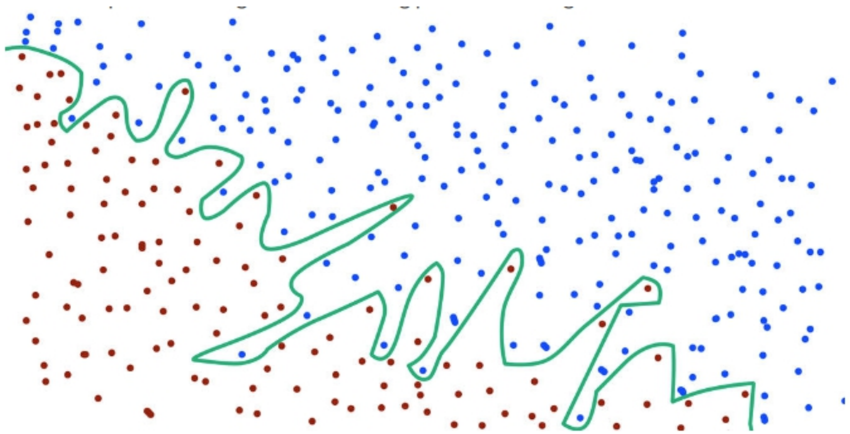
- A cross-validation is a technique to evaluate the model with different subsets of training data. It helps to improve model accuracy and to avoid overfitting in an estimation





Train data

Overfitting ?



Test data

Cross - validatiion: Background

- Cross validation is a widely used strategy:

  - Estimating the predictive accuracy of a model

  - Performing model selection e.g:

    + Choose among variables in a regression or the degree of freedom of a nonparametric model (selection for identification)

    + Parameter estimation and tuning (selection for estimation)

- Main features:

  - Main idea: test the model on data not used in estimation

  - Split data once or several times

  - Part of data is used for training each model (the training sample), and the remaining part is used for estimating the prediction error of the model (the validation sample)

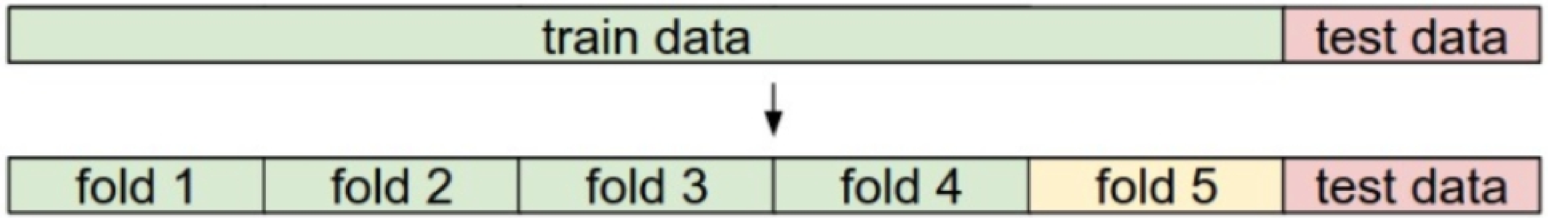
Cross - validation: How it work?

Example:

DATA set (train)
50000 pictures

Spilt: 5 fold

each fold
10000 pictures



Train				Validation
fold 1	fold 2	fold 3	fold 4	fold 5
fold 1	fold 2	fold 3	fold 5	fold 4
fold 1	fold 2	fold 4	fold 5	fold 3
fold 1	fold 3	fold 4	fold 5	fold 2
fold 2	fold 3	fold 4	fold 5	fold 1

Validation-set approach

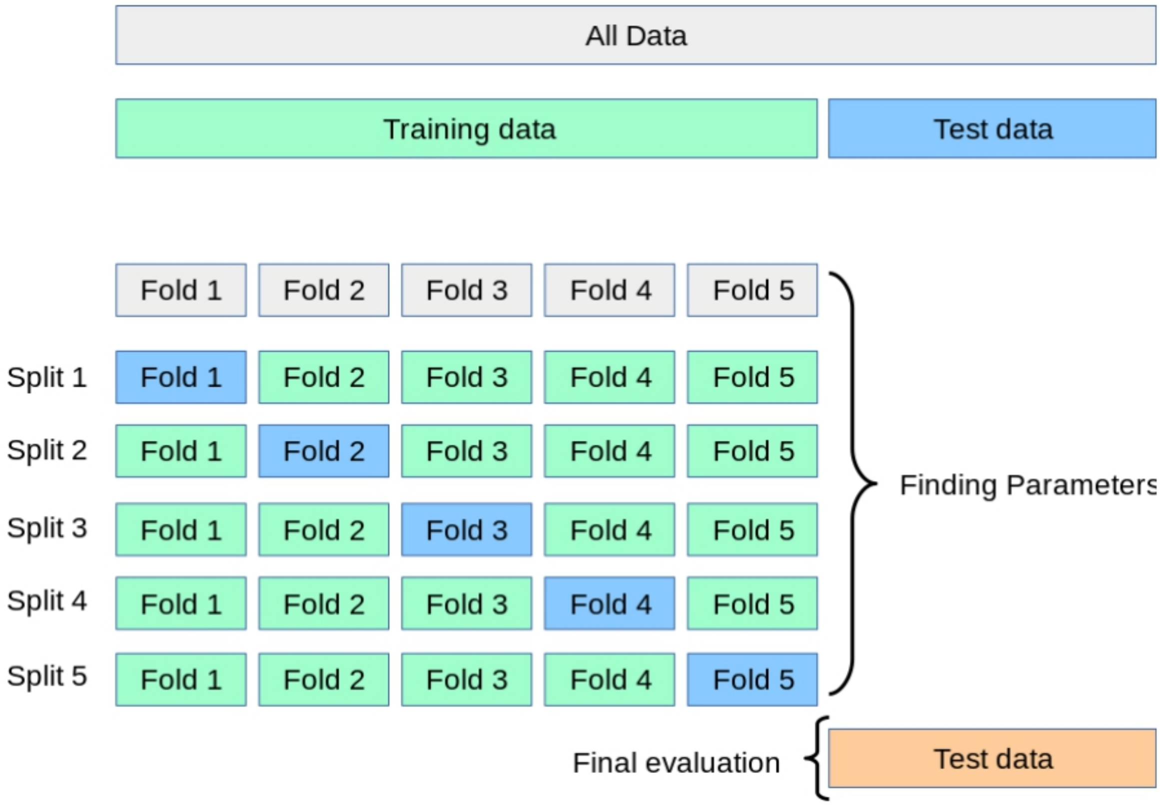
- Here we randomly divide the available set of samples into two parts: a training set and a validation or hold-out set
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. This is typically assessed using MSE

$$MSE = \frac{1}{n} \Sigma \underbrace{\left(y - \widehat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$

Mean squared error formula

K -fold Cross-validation

- Widely used approach for estimating test error
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other K − 1 parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part k = 1, 2, . . . K, and then the results are combined.



How many folds are needed (K =?)

- large: small bias, large variance as well as computational time
- small: computation time reduced, small variance, large bias
- A common choice for K-Fold Cross Validation is K=5

The details

- Let the K parts be C1, C2, . . . CK, where Ck denotes the indices of the observations in part k. There are nk observations in part k: if N is a multiple of K, then nk = n/K.
- Compute

$$\mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k$$

where $\mathrm{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.

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

A nice special case!

- With least-squares linear or polynomial regression, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit! The following formula holds:

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

where $\hat{y}_i$ is the ith fitted value from the original least squares fit, and h_i is the leverage (diagonal of the "hat" matrix; see book for details.) This is like the ordinary MSE, except the ith residual is divided by $1 - h_i$.

- LOOCV sometimes useful, but typically doesn't shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance
- A better choice is K = 5 or 10.

Cross-validation: right and wrong

Consider a simple classifier applied to some two-class data:

1. Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.

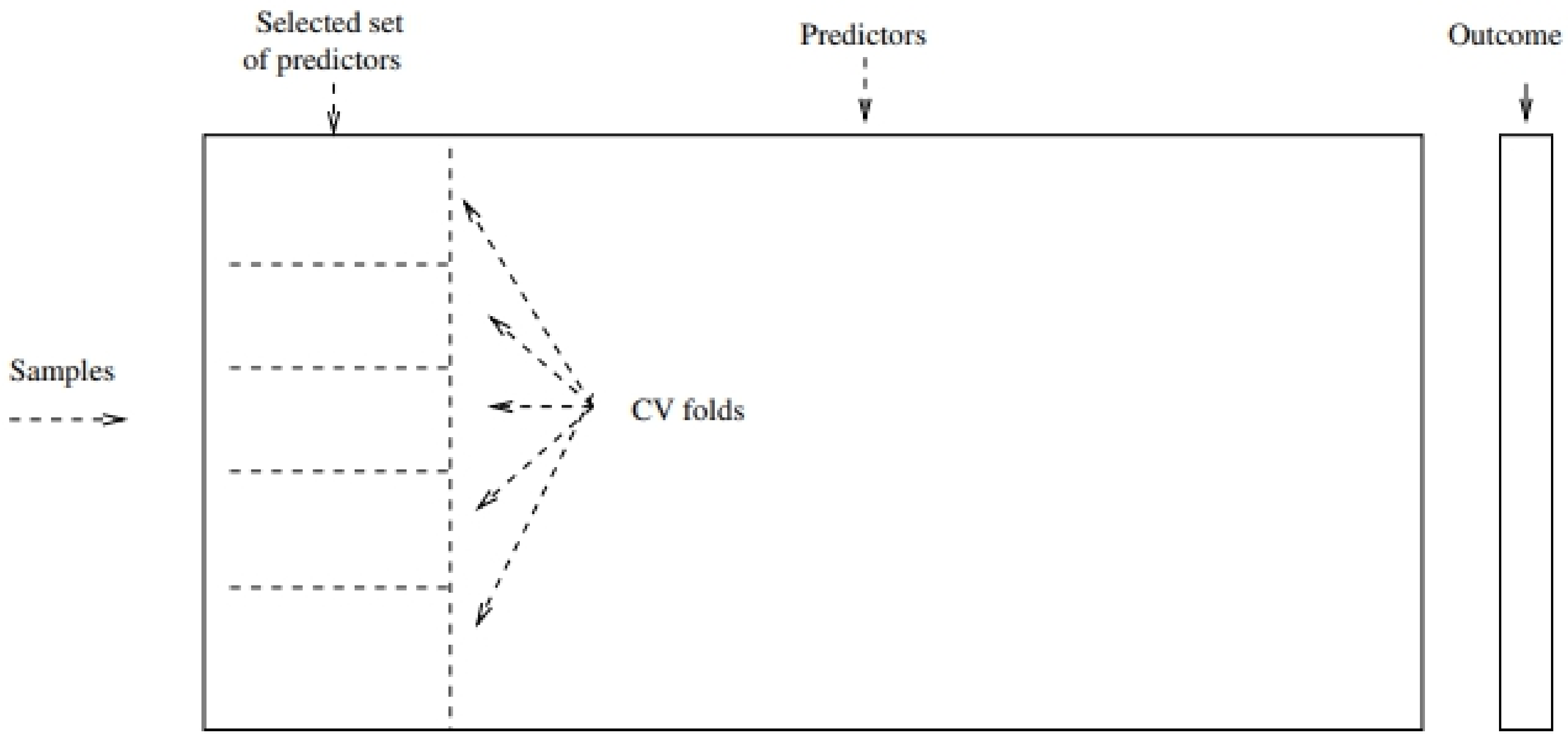
2. We then apply a classifier such as logistic regression, using only these 100 predictors.

How do we estimate the test set performance of this classifier?

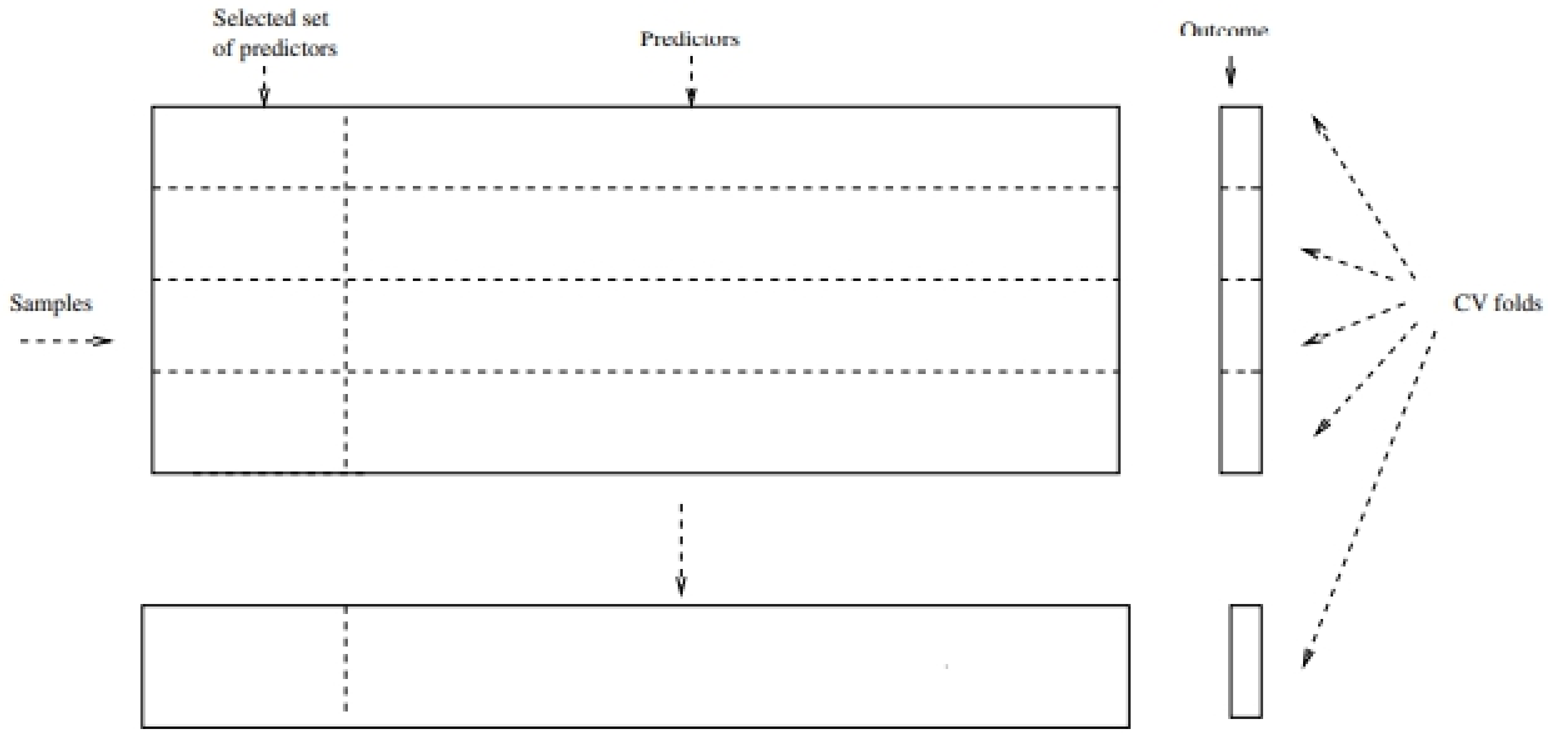
Can we apply cross-validation in step 2, forgetting about step 1?

- **Wrong: Apply cross-validation in step 2.**
- **Right: Apply cross-validation to steps 1 and 2.**

Wrong Way


Selected set
of predictors
Predictors
Outcome
Samples
CV folds

Right Way



Selected set
of predictors
Predictors
Outcome
Samples
CV folds