Paul Martins

Resampling methods

Notes from ISLR book

Contents

1. Cross-validation

The cross-validation is about splitting your initial data set into a training and hold-out/test data set.

1.1 Validation set

This randomly splits the data set into train and test samples. We build the model with the train sample and assess the (test) MSE with the hold-out set.
Problem 1: The test MSE is highly biased. It depends on the random sampling.
Problem 2: This model is trained on a reduced sample. It does not use all the available observations.

1.2 Leave-One-Out (LOOCV)

Only consider 1 observation in the test sample and assess the test MSE for this observation only (using the model trained on the remaining $n-1$ obs). Finally average all the MSE.

$$MSE_{LOOCV} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \tag{1}$$

Problem 1: It can be computationally expensive if n is big (have to retrain the model n times). In some cases, like least square regression, we have shortcuts such as

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

Problem 2: Despite having lower bias than the validation set approach (and k-fold), this has high variance. Indeed, all of the individual MSE_i are highly correlated since they are obtained from models trained on very similar training sets. The mean of many highly correlated quantities has high variance.

1.3 k-fold cross-validation

Split the data set in k samples of equal sizes. Assess the test MSE on each of this k sample, while training the model on the $k-1$ remaining samples (k=5 or 10, like in Figure 1)

$$MSE_{k-fold} = \frac{1}{k}\sum_{k=1}^{K} MSE_k \tag{3}$$

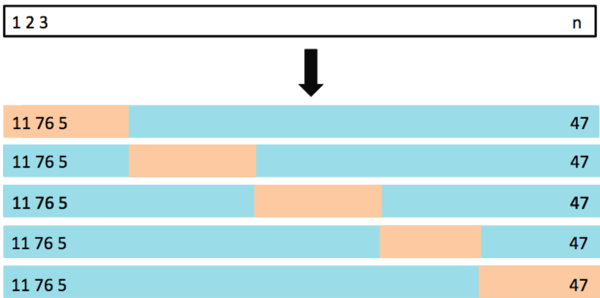

Figure 1. Illustration of 5-fold cross-validation.

This has the advantages of being much faster than LOOCV (only train the model k times) and it suffers much less of the high variance problem.

For classification, we can replace the MSE with the number of misclassified observations.

2. Bootstrap

This technique is used to quantify the uncertainty of a given estimator or learning method. It consists in taking a random sample with replacement of your data set and calculate the bootstrap value of the parameter $\hat{\alpha}^{*1}$ using that first sample. Repeat this B times and then the standard deviation of $\hat{\alpha}$ is

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2} \tag{4}$$

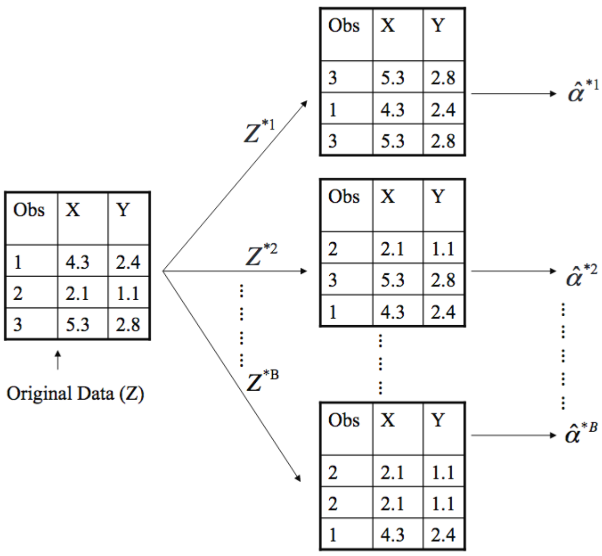

Figure 2. Example of bootstrap.