Validation

Normally, we use our validation set to estimate the out-of-sample error for a single hypothesis (ie. fitted model).

Sometimes, we don't have enough data, and instead, we want to use all the data we have for validation.

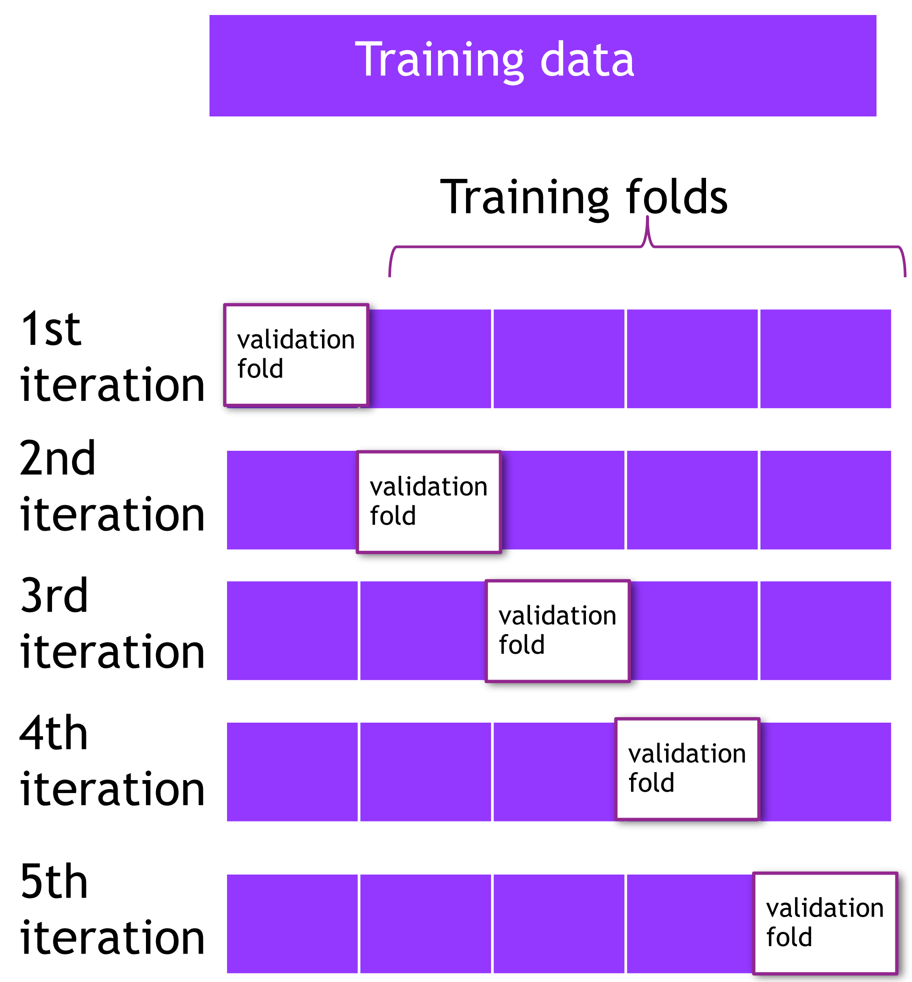
K-Fold Cross Validation

So, we will give up getting an estimate for a single hypothesis and instead get an estimate for this one approach.

In this case, we use K-fold cross-validation to estimate the out-of-sample error for this approach (approach is hypothesis class, hyper-parameters, optimizer).
Note: it is not an estimate for the out-of-sample error for a single hypothesis.

Edited by Haoran Chen

Estimating the out of sample error for $\mathcal{H}_i$ using K-fold cross-validation for N examples



$e_1 = E_{val}(g_1^-)$

$e_2 = E_{val}(g_2^-)$

$e_3 = E_{val}(g_3^-)$

$e_4 = E_{val}(g_4^-)$

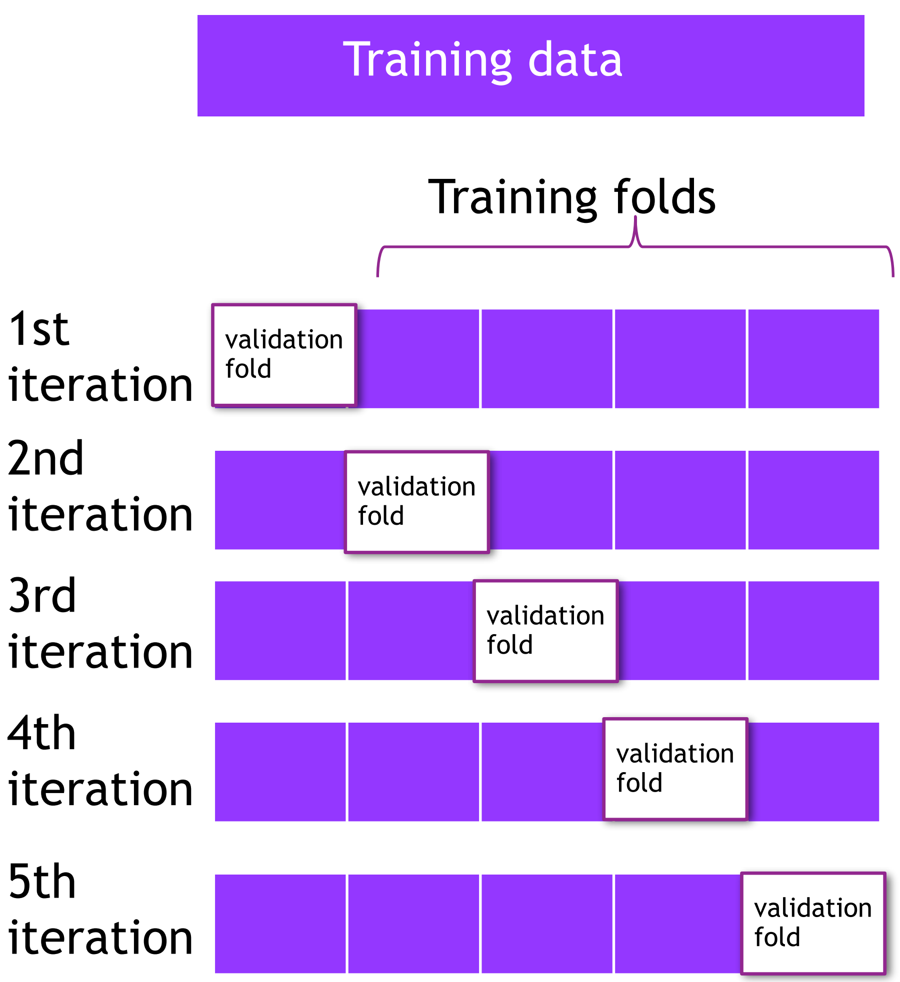
$e_5 = E_{val}(g_5^-)$

not independent error measures

cross validation error

Cross-Validate(D,k)
Divide training data (*randomly*) into K disjoint sets: $D_1, D_2, \ldots, D_K$
For i = 1 to K
    Train on all examples except those in the ith set $D - D_i$
    Let g_i^- be the fitted model
    Validation error $e_i = E_{val}(g_i^-)$
Return $E_{cv} = \frac{1}{K}\sum_{i=1}^{K} e_i$

Estimating the out of sample error for $\mathcal{H}i$ using K-fold cross-validation for N examples



$e_1 = E_{val}(g_1^-)$

$e_2 = E_{val}(g_2^-)$

$e_3 = E_{val}(g_3^-)$

$e_4 =$

$e_5 =$

not independent error measures

E_{cv} is expected error for all data sets of size N-(N/K)

$|D_i| = N/K$

Cross-Validate(D,k)
Divide training data (*randomly*) into K disjoint sets: $D_1, D_2, \ldots, D_K$
For i = 1 to K
  Train on all examples except those in the ith set $D - D_i$
  Let g_i^- be the fitted model
  Validation error $e_i = E_{val}(g_i^-)$
Return $E_{cv} = \frac{1}{K}\sum_{i=1}^{K} e_i$

Which of the M models should we choose

Cross validation error:

$$E_{cv} = \frac{1}{K}\sum_{i=1}^{K} e_i$$

$$\underset{E_{cv,1}}{\mathcal{H}_1}, \underset{E_{cv,2}}{\mathcal{H}_2}, \dots, \underset{E_{cv,M}}{\mathcal{H}_M}$$

We select the model $\mathcal{H}_{m}*$ with the lowest cross-validation error.

$$m^* = \arg \min_{i \in \{1,\dots,M\}} \{E_{cv,1}, E_{cv,2}, \dots, E_{cv,M}\}$$

During this process, for one model class, you computed K different hypotheses. If you wish to use this model class to predict in the future, run the algorithm again on **all** the data. Or average the result of your K hypotheses.