Cross-Validation (CV)

Generalizes the case of splitting data into training, validation, and test data.

Useful when data is scarce

Provides an estimate of the expected generalization error $\mathrm{Err} = \mathbf{E}[L(Y, \hat{f}(X))]$

We consider K - fold cross validation

K-Fold Cross Validation

Idea:

1. split data into K parts of (more or less) the same size;
2. Repeat K times:
   1. train on $K-1$ parts, leaving the *i*-th part out
   2. estimate prediction error on *i*-th part
3. Combine the K estimates of prediction error

K-Fold Cross Validation

Example $K = 5$

Leave first part out:

1	2	3	4	5
Validation	Train	Train	Train	Train

Leave second part out...

Leave third part out:

1	2	3	4	5
Train	Train	Validation	Train	Train

Leave fourth part out...

Leave fifth part out...

In General

Let $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$ be *indexing function*

- partitions data points, assigning each to one part

Let models be $h(x, \alpha)$ (α = tuning parameter)

Let $\hat{h}^{-k}(x, \alpha)$ fit for the model with α and k-th part of data removed.

Cross-validation estimate of prediction error:

$$\mathrm{CV}(\hat{h}, \alpha) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{h}^{-\kappa(i)}(x_i, \alpha))$$

Choose model minimizing $\mathrm{CV}(\hat{h}, \alpha)$ and fit it using all data

How to choose *K* ?

Typical choice: K=5 or K=10

Leave-one-out cross-validation: $K = N$

- nice theoretical properties: approximately unbiased estimator for expected prediction error
- high variance! [WHY?]
- computationally expensive

K=5 or K=10 may give a biased estimate of expected prediction error, but smaller variance: good compromise in practice.