Cross-Validation

Probably the simplest and most widely used method for estimating prediction error is *cross-validation*. This method directly estimates the expected extra-sample error

$$\mathrm{Err} = \mathbb{E}\left[L\left(Y, \hat{f}(X)\right)\right], \tag{1}$$

the average generalization error when the method $\hat{f}(X)$ is applied to an independent test sample form the joint distribution of X and Y- As mentioned earlier, we might hope that cross-validation estimates the conditional error, with the training set $\mathcal{T}$ held fixed. But cross-validation typically estimates well only the expected prediction error.

K-Fold Cross-Validation

Ideally, if we had enough data, we would set aside a validation set and use it to assess the performance of our prediction model. Since data are often scarce, this is usually not possible. To finesse the problem, K-fold cross-validation uses part of the available data to fit the model, and a different part to test it. We split the data into K roughly equal-sized parts; for example, when $K = 5$, the scenario looks like this:

For the kth part (third above), we fit the model to the other $K - 1$ parts of the data, and calculate the prediction error of the fitted model when predicting the kth part of the data. We do this for $k = 1, 2, \ldots, K$ and combine the K estimates of prediction error.
Here are more details. Let $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$ be an indexing function that indicates the partition to which observation i is allocated by the randomization. Denote by $\hat{f}^{-k}(x)$ the fitted function, computed with the kth part of the data removed. Then the cross-validation estimate of prediction error is

$$\mathrm{CV}\left(\hat{f}\right) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(x_i)\right). \tag{2}$$