

k -fold Cross Validation

k -fold cross validation is a common technique for estimating the performance of a classifier. Given a set of m training examples, a single run of k -fold cross validation proceeds as follows:

1. Arrange the training examples in a random order.
2. Divide the training examples into k folds. (k chunks of approximately m/k examples each.)
3. For $i = 1, \dots, k$:
 - Train the classifier using all the examples that do not belong to Fold i .
 - Test the classifier on all the examples in Fold i .
 - Compute n_i , the number of examples in Fold i that were wrongly classified.
4. Return the following estimate to the classifier error:

$$E = \frac{\sum_{i=1}^k n_i}{m}$$

To obtain an accurate estimate to the accuracy of a classifier, k -fold cross validation is run several times, each with a different random arrangement in Step 1. Let E_1, \dots, E_t be the accuracy estimates obtained in t runs. Define:

$$e = \frac{\sum_{j=1}^t E_j}{t}, \quad V = \frac{\sum_{j=1}^t (E_j - e)^2}{t - 1}, \quad \sigma = \sqrt{V}$$

The estimate for the algorithm performance is an error of e with standard-deviation of σ .