What is an Easy Guide to K-Fold Cross-Validation?

Authored by

stats writer

April 22, 2024

An Easy Guide to K-Fold Cross-Validation is a method used in machine learning and statistics to evaluate the performance of a predictive model. It involves dividing a dataset into k equal subsets, where one subset is used as the testing set and the remaining k-1 subsets are used as the training set. This process is repeated k times, with each subset being used as the testing set once. The results from each iteration are then averaged to obtain a more accurate estimate of the model's performance. K-Fold Cross-Validation is a popular and reliable technique for assessing the generalizability of a model and is commonly used to select the best performing model for a given dataset.

An Easy Guide to K-Fold Cross-Validation

To evaluate the performance of some model on a dataset, we need to measure how well the predictions made by the model match the observed data.

The most common way to measure this is by using the mean squared error (MSE), which is calculated as:

MSE = (1/n)*Σ(yi - f(xi))2

where:

n: Total number of observationsyi: The response value of the ith observationf(xi): The predicted response value of the ith observation

The closer the model predictions are to the observations, the smaller the MSE will be.

In practice, we use the following process to calculate the MSE of a given model:

1. Split a dataset into a training set and a testing set.

2. Build the model using only data from the training set.

3. Use the model to make predictions on the testing set and measure the test MSE.

The test MSE gives us an idea of how well a model will perform on data it hasn't previously seen. However, the drawback of using only one testing set is that the test MSE can vary greatly depending on which observations were used in the training and testing sets.

One way to avoid this problem is to fit a model several times using a different training and testing set each time, then calculating the test MSE to be the average of all of the test MSE's.

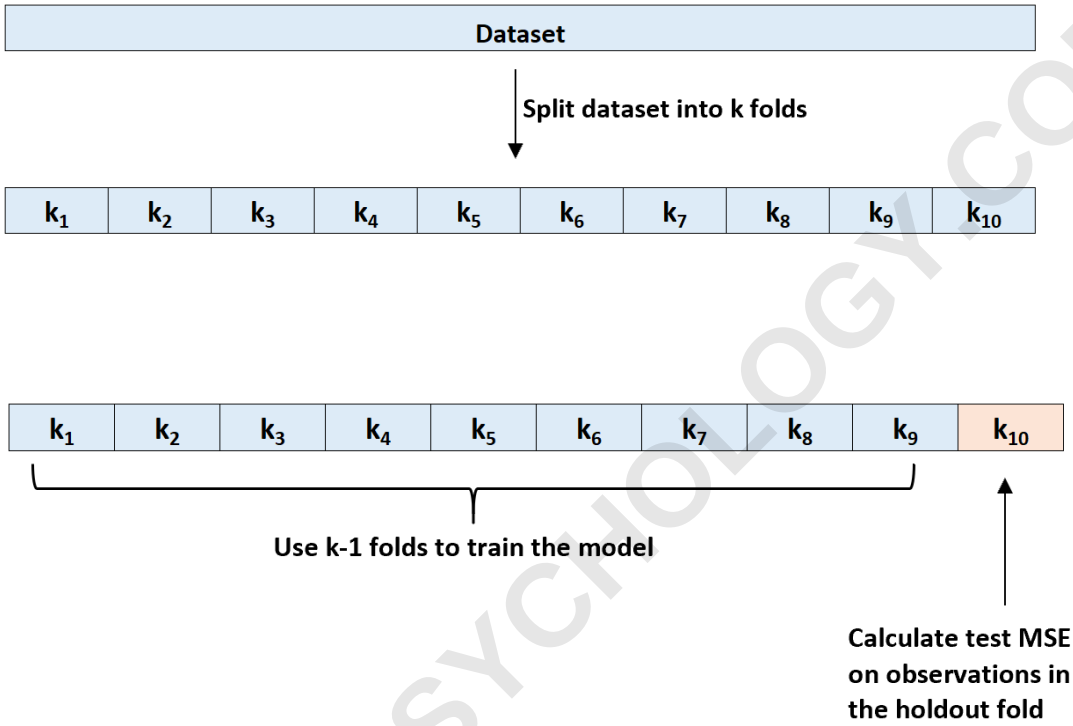
This general method is known as cross-validation and a specific form of it is known as k-fold cross-validation.

K-Fold Cross-Validation

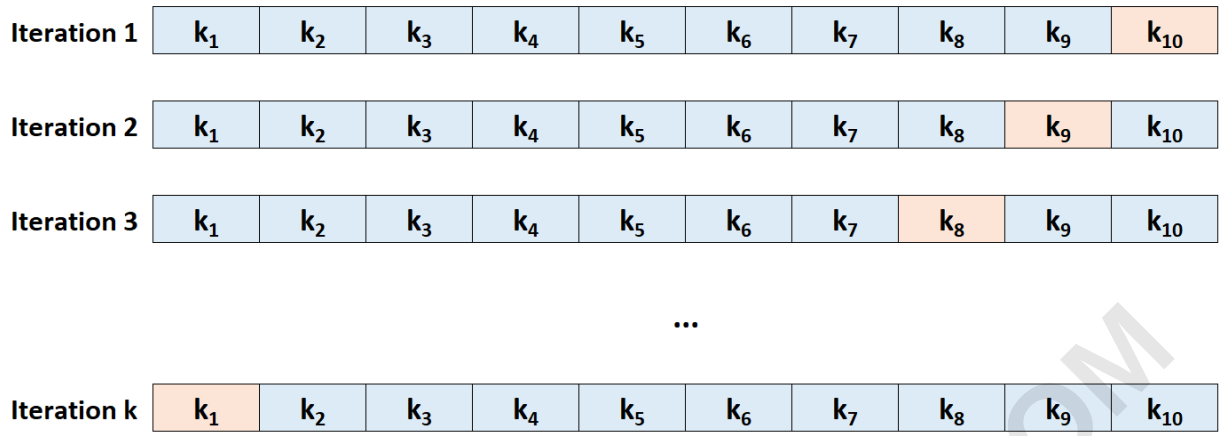
K-fold cross-validation uses the following approach to

evaluate a model:

Step 1: Randomly divide a dataset into *k* groups, or "folds", of roughly equal size.



Step 3: Repeat this process *k* times, using a different set each time as the holdout set.



Step 4: Calculate the overall test MSE to be the average of the *k* test MSE's.

Test MSE = (1/k)*ΣMSEi

where:

k: Number of foldsMSEi: Test MSE on the ith iteration

How to Choose K

In general, the more folds we use in k-fold cross-validation the lower the bias of the test MSE but the higher the variance. Conversely, the fewer folds we use the higher the bias but the lower the variance. This is a classic example of the bias-variance tradeoff in machine learning.

In practice, we typically choose to use between 5 and 10 folds. As noted in *An Introduction to Statistical Learning*, this number of folds has been shown to offer an optimal balance between bias and variance and thus provide reliable estimates of test MSE:

To summarize, there is a bias-variance trade-off associated with the choice of k in k-fold cross-validation.

Typically, given these considerations, one performs k-fold cross-validation using k = 5 or k = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

-Page 184, *An Introduction to Statistical Learning*

Advantages of K-Fold Cross-Validation

When we split a dataset into just one training set and one testing set, the test MSE calculated on the observations in the testing set can vary greatly depending on which observations were used in the training and testing sets.

By using k-fold cross-validation, we're able to use calculate the test MSE using several different variations of training and testing sets. This makes it much more likely for us to obtain an unbiased estimate of the test MSE.

K-fold cross-validation also offers a computational advantage over leave-one-out cross-validation (LOOCV) because it only has to fit a model *k* times as opposed to *n* times.

For models that take a long time to fit, k-fold cross-validation can compute the test MSE much quicker than LOOCV and in many cases the test MSE calculated by each approach will be quite similar if you use a sufficient number of folds.

Extensions of K-Fold Cross-Validation

There are several extensions of k-fold cross-validation, including:

Repeated K-fold Cross-Validation: This is where k-fold cross-validation is simply repeated *n* times. Each time the training and testing sets are shuffled, so this further reduces the bias in the estimate of test MSE although

this takes longer to perform than ordinary k-fold cross-validation.

Leave-One-Out Cross-Validation: This is a special case of k-fold cross-validation in which *k*=*n*. You can read more about this method **here**.

Stratified K-Fold Cross-Validation: This is a version of k-fold cross-validation in which the dataset is rearranged in such a way that each fold is representative of the whole. As noted by **Kohavi**, this method tends to offer a better tradeoff between bias and variance compared to ordinary k-fold cross-validation.

Nested Cross-Validation: This is where k-fold cross validation is performed within each fold of cross-validation. This is often used to perform hyperparameter tuning during model evaluation.