

# **$k$ -fold cross-validation ( $k$ -fold CV)**

## **Author and editors**

Alexander Jung\*, Department of Computer Science, Aalto University, Espoo, Finland

Editors: Konstantina Olioumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

\*Corresponding author: alex.jung@aalto.fi

## **Abstract**

$k$ -fold cross-validation ( $k$ -fold CV) estimates the risk of an empirical risk minimization (ERM)-based machine learning (ML) method by dividing a dataset into  $k$  folds of equal size and using each fold once as the validation set while the other  $k - 1$  folds form the training set. The  $k$  resulting validation errors are averaged into a single estimate. Every data point serves both for training and for validation, unlike a single split into training set and validation set. The choice of  $k$  trades bias against variance: small  $k$  leaves each training set a small fraction of the dataset and overestimates the risk, while large  $k$  makes the  $k$  training sets overlap in almost every data point and their validation errors highly correlated. Leave-one-out cross-validation (LOO-CV) is the extreme case with one data point per fold, and stratified  $k$ -fold cross-validation preserves the class proportions of the full dataset in every fold.

## **Definition**

Weather recordings from fifteen days relate a day's maximum daytime temperature  $y$  to its morning minimum  $x$ . Splitting them once into a training set  $\mathcal{D}^{(\text{train})}$  of twelve days and a validation set  $\mathcal{D}^{(\text{val})}$  of three computes the validation error from three data points only. Which days are held back also changes the verdict: Fig. 1 shows two such splits, and the hypothesis that empirical risk minimization (ERM) on a linear model delivers has a different slope in each.  $k$ -fold CV avoids the choice. The fifteen days are divided into five folds of three, each fold serves once as  $\mathcal{D}^{(\text{val})}$  while the other four form  $\mathcal{D}^{(\text{train})}$ , and the validation error of each iteration enters an average. Those five validation errors are 0.77, 1.03, 0.90, 1.05, and 1.17, and their average 0.99 is the  $k$ -fold CV estimate of the expected squared error loss of ERM on the linear model.

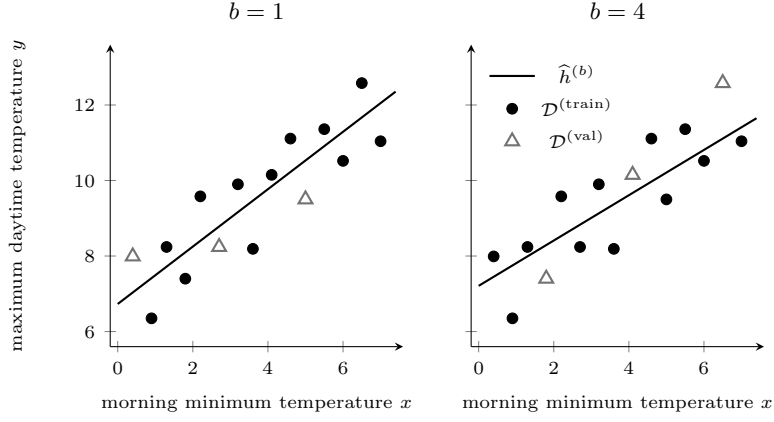


Figure 1: Two of the five iterations of  $k$ -fold CV on the same fifteen days, with  $k = 5$ . In the iteration  $b$ , the three days of fold  $b$  are the validation set (open triangles) and the remaining twelve are the training set (filled circles); the line is the hypothesis  $\hat{h}^{(b)}$  that ERM on a linear model delivers from that training set. Its slope is 0.76 for  $b = 1$  and 0.60 for  $b = 4$ , and the validation error is 0.77 against 1.05: both the learned hypothesis and its assessment move with the split

$k$ -fold CV estimates the expected loss (or risk) of an ERM-based machine learning (ML) method (Hastie et al., 2009, Sect. 7.10; Stone, 1974). It divides a dataset  $\mathcal{D}$  evenly into  $k$  subsets, the folds  $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(k)}$  (Fig. 2). For each fold  $b = 1, \dots, k$ , the model is trained on the union of all folds except  $\mathcal{D}^{(b)}$  and the resulting hypothesis is evaluated on the held-out fold  $\mathcal{D}^{(b)}$ . The  $k$  per-fold validation errors are then averaged to obtain the CV estimate of the expected loss.

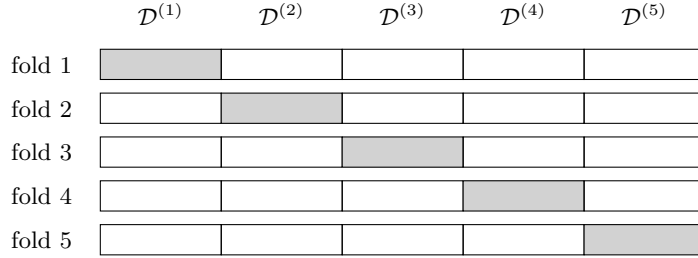


Figure 2: An instance of  $k$ -fold CV with  $k = 5$ : the available dataset  $\mathcal{D}$  is evenly divided into five folds  $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(5)}$ . Each row corresponds to one iteration; the shaded cell marks the fold used as validation set, while the remaining  $k - 1$  folds form the training set

Compared to a single split into training set and validation set,  $k$ -fold CV uses

each data point for both training and validation. In Fig. 2, for example, each data point has been used four times for training and once for validation after all five iterations. For a typical ERM-based method, the  $k$ -fold CV estimate is more accurate than one from a single validation set holding a fraction  $1/k$  of the entire dataset (Blum et al., 1999).

The choice of  $k$  trades off bias against variance of the  $k$ -fold CV estimate (Hastie et al., 2009, Sect. 7.10.1). Small  $k$  (e.g.,  $k = 2$ ) gives higher bias because each training set contains only a fraction  $(k - 1)/k$  of the full dataset; if the learning curve is steep at this size,  $k$ -fold CV overestimates the true expected loss. Choosing large  $k$  tends to reduce bias but can increase variance because the  $k$  training sets differ in only a few data points, making the per-fold estimates highly correlated (Bengio and Grandvalet, 2004).

Several variants of  $k$ -fold CV adapt it to specific dataset structures:

- *Stratified  $k$ -fold cross-validation* preserves the class proportions of the full dataset in every fold, important for classification with imbalanced data (Kohavi, 1995).
- *Leave-one-out cross-validation (LOO-CV)* is the extreme case  $k = m$ , where each fold contains a single data point (Stone, 1974).
- *Repeated  $k$ -fold CV* averages the results of several runs with different random fold assignments, which reduces the variance contributed by the partitioning (Kim, 2009).

## Cross References

validation, validation error, generalization gap, learning curve, LOO-CV, stratified  $k$ -fold CV, model selection

## References

1. Bengio and Grandvalet (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research*, 5, pp. 1089–1105.
2. Blum et al. (1999). Beating the Hold-Out: Bounds for K-fold and Progressive Cross-Validation. In: *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT)*, pp. 203–208.
3. Hastie et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer Science+Business Media, New York, NY, USA.
4. Kim (2009). Estimating Classification Error Rate: Repeated Cross-Validation, Repeated Hold-Out and Bootstrap. *Computational Statistics & Data Analysis*, 53(11), pp. 3735–3745.

5. Kohavi (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1143.
6. Stone (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), pp. 111–147.