# Optimal number of folds in $K$-fold cross-validation: is leave-one-out CV always the best choice?

Asked 6 years, 1 month ago Active 1 year ago Viewed 8k times

47

30

Computing power considerations aside, are there any reasons to believe that **increasing the number of folds** in cross-validation leads to better model selection/validation (i.e. that the higher the number of folds the better)?

Taking the argument to the extreme, does leave-one-out cross-validation necessarily lead to better models than $K$-fold cross-validation?

Some background on this question: I am working on a problem with very few instances (e.g. 10 positives and 10 negatives), and am afraid that my models may

edited Jul 17 '18 at 21:55
amoeba
**65k** 17 218 273

asked Jun 12 '13 at 13:24
Amelio Vazquez-Reina
**6,961** 14 62 97

1 An older related thread: Choice of K in K-fold cross-validation. – amoeba Oct 19 '15 at 23:15

This question is not a duplicate because it restricts to small datasets and "Computing power considerations aside". This is a severe limitation, making the question inapplicable to those with large datasets and the training algorithm with computational complexity at least linear in the number of instances (or prediction in at least the square root of the number of instances). – Serge Rogatch Sep 4 '16 at 8:29

## 2 Answers

46

✓

Leave-one-out cross-validation does not generally lead to better performance than K-fold, and is more likely to be *worse*, as it has a relatively high variance (i.e. its value changes more for different samples of data than the value for k-fold cross-validation). This is bad in a model selection criterion as it means the model selection criterion can be optimised in ways that merely exploit the random variation in the particular sample of data, rather than making genuine improvements in performance, i.e. you are more likely to over-fit the model selection criterion. The reason leave-one-out cross-validation is used in practice is that for many models it can be evaluated very cheaply as a by-product of fitting the model.

If computational expense is not primarily an issue, a better approach is to perform repeated k-fold cross-validation, where the k-fold cross-validation procedure is repeated with different random partitions into k disjoint subsets each time. This reduces the variance.

If you have only 20 patterns, it is very likely that you will experience over-fitting the model selection criterion, which is a much neglected pitfall in statistics and machine learning (shameless plug: see my paper on the topic). You may be better off choosing a relatively simple model and try not to optimise it very aggressively, or adopt a Bayesian approach and average over all model choices, weighted by their plausibility. IMHO optimisation is the root of all evil in statistics, Note also if you are going to perform model selection, you need to use something like nested cross-validation if you also need a performance estimate (i.e. you need to consider model selection as an integral part of the model fitting procedure and cross-validate that as well).

answered Jun 12 '13 at 13:58

Dikran Marsupial
**37.1k** 4 106 151

8 +1. I like your "optimisation is the root of all evil in statistics" message... – Stephan Kolassa Jun 13 '13 at 8:10

5 Thanks @DikranMarsupial . I don't quite follow. Why would models learned with *leave-one-out* have higher variance than with regular *k-fold* cross validation? My intuition tells me that, since across folds we are only shifting one data point, the training sets across folds overlap heavily, so I would expect to see little variance between models. Or going in the other direction, in K-fold, if K is low, the training sets for each fold would be quite different and the resulting models are more likely to be different. Am I wrong? – Amelio Vazquez-Reina Jun 14 '13 at 13:18

That is a very good question in its own right, so I suggest you ask it as a new question, and I will have a think about how to answer it! – Dikran Marsupial Jun 14 '13 at 13:22

Thank you @DikranMarsupial I followed your advice and started a separate question here. – Amelio Vazquez-Reina Jun 14 '13 at 20:15

1 @DikranMarsupial I thought I would mention here that I have started one more thread inspired by your "optimization in statistics" comment in this answer. Your comment made me look at overfitting from a broader perspective that I am used to. – Amelio Vazquez-Reina Jul 11 '13 at 1:05

8

## Choosing the number K folds by considering the learning curve

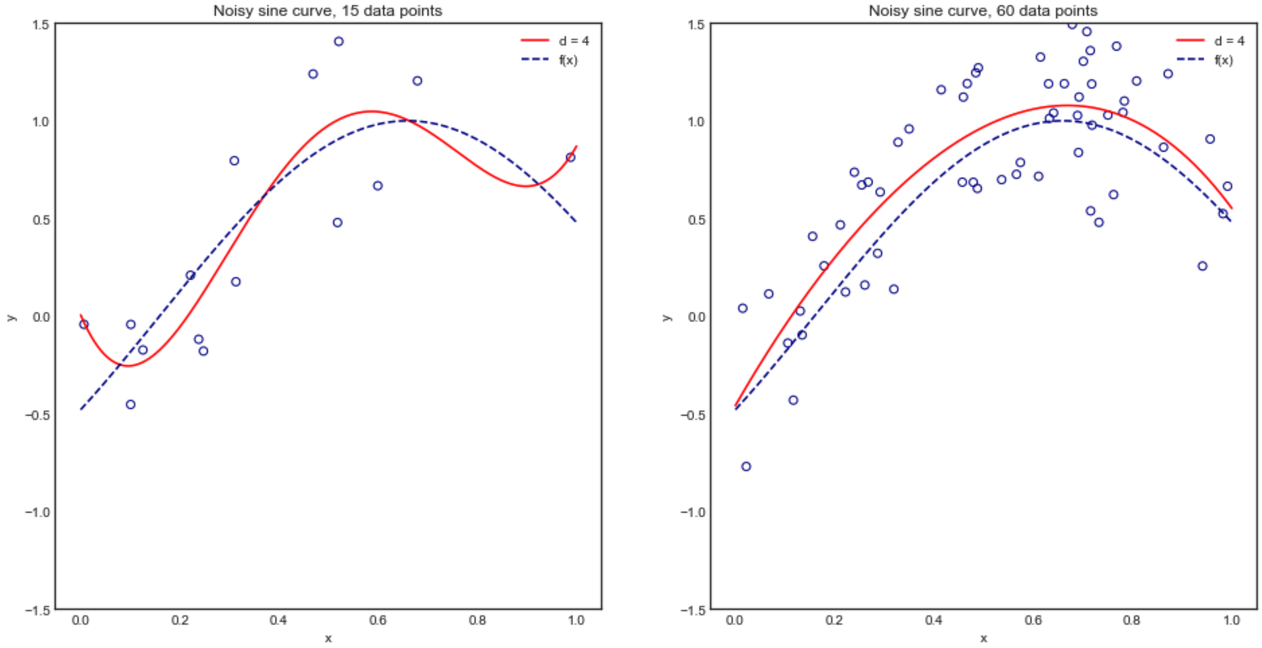
I would like to argue that choosing the appropriate number of $K$ folds depends a lot on the shape and position of the learning curve, mostly due to its impact on the **bias**. This argument, which extends to leave-one-out CV, is largely taken from the book "Elements of Statistical Learning" chapter 7.10, page 243.

For discussions on the impact of $K$ on the **variance** see here

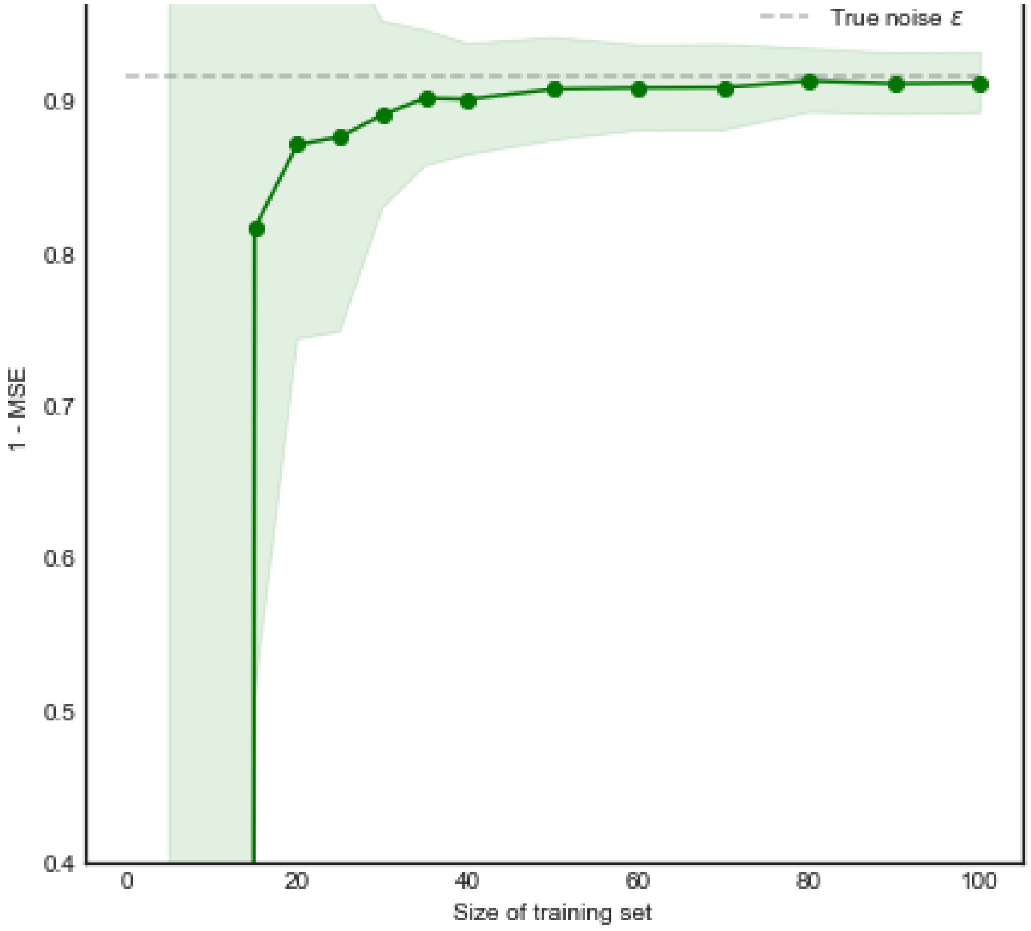
> prediction error. Whether this bias is a drawback in practice depends on the objective. On the other hand, leave-one-out cross-validation has low bias but can have high variance.

## An intuitive visualization using a toy example

To understand this argument visually, consider the following toy example where we are fitting a degree 4 polynomial to a noisy sine curve:



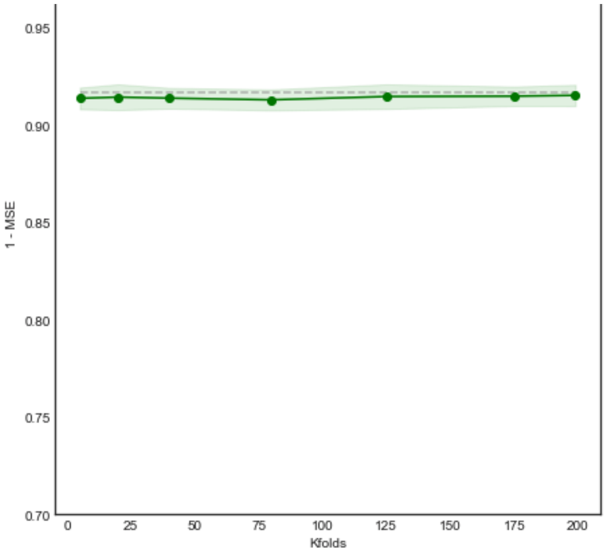
Intuitively and visually, we expect this model to fare poorly for small datasets due to overfitting. This behaviour is reflected in the learning curve where we plot $1-$ Mean Square Error vs Training size together with ± 1 standard deviation. *Note that I chose to plot 1 - MSE here to reproduce the illustration used in ESL page 243*

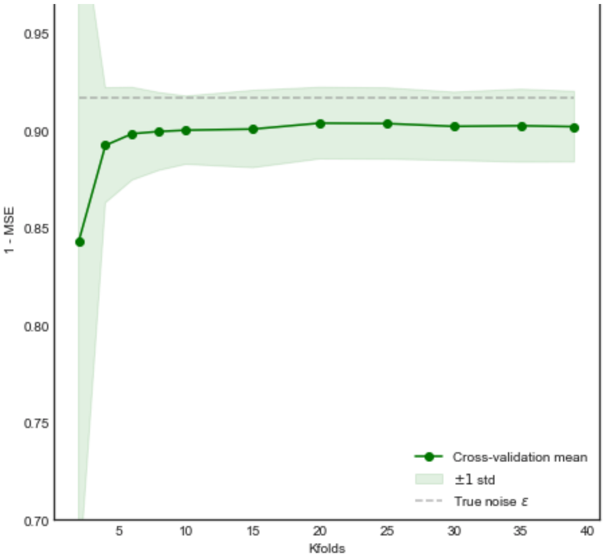


## Discussing the argument

The performance of the model improves significantly as the training size increases to 50 observations. Increasing the number further to 200 for example brings only small benefits. Consider the following two cases:

1. If our training set had 200 observations, $5$ fold cross validation would estimate the performance over a training size of 160 which is virtually the same as the performance for training set size 200. Thus cross-validation would not suffer from much bias and increasing $K$ to larger values will not bring much benefit (*left hand plot*)
2. However if the training set had $50$ observations, $5$ fold cross-validation would estimate the performance of the model over training sets of size 40, and from the learning curve this would lead to a biased result. Hence increasing $K$ in this case will tend to reduce the bias. (*right hand plot*).





## [Update] - Comments on the methodology

You can find the code for this simulation here. The approach was the following:

1. Generate 50,000 points from the distribution $sin(x) + \epsilon$ where the true variance of $\epsilon$ is known
2. Iterate $i$ times (e.g. 100 or 200 times). At each iteration, change the dataset by resampling $N$ points from the original distribution
3. For each data set $i$:
    - Perform K-fold cross validation for one value of $K$
    - Store the average Mean Square Error (MSE) across the K-folds
4. Once the loop over $i$ is complete, calculate the mean and standard deviation of the MSE across the $i$ datasets for the same value of $K$
5. Repeat the above steps for all $K$ in range $\{5, \ldots, N\}$ all the way to LOOCV

An alternative approach is to *not resample* a new data set at each iteration and instead reshuffle the same dataset each time. This seems to give similar results.

edited Jul 18 '18 at 11:43

answered Jul 17 '18 at 12:28

Xavier Bourret Sicotte

**4,036** 2 19 48

Let us continue this discussion in chat. – Xavier Bourret Sicotte Jul 17 '18 at 14:10

KFold (Shuffle = True) :) – meTchaikovsky Sep 17 '18 at 7:54

@me_Tchaikovsky Recall that the MSE of the predictor can be decomposed as $MSE = Var + Bias^2$ and assuming no bias when the model matches the true underlying function then we are left with the variance of the error term $\epsilon \sim U(-.5, .5)$. Variance of uniform r.v. is $1/12(b-a)^2$ so $1/12$ in this case – Xavier Bourret Sicotte Sep 17 '18 at 15:45