Choice of K in K-Fold cross validation

I've been using the *K-Fold cross validation* a few times now to evaluate performance of some learning algorithms, but I've always been puzzled as to how I should choose the value of K.

I've often seen and used a value of `K = 10`, but this seems totally arbitrary to me, and I now just use 10 by habit instead of thinking it over. To me it seems that you're getting a better granularity as you improve the value of K, so ideally you should make your K very large, but there is also a risk to be biased.

I'd like to know on what the value of K should depend, and how I should be thinking about this when I evaluate my algorithm. Does it change something if I use the *stratified* version of the cross validation or not?

machine-learning | classification | cross-validation

asked May 4 '12 at 3:52

Charles Menguy
432 1 5 15

2 Answers

The choice of $k = 10$ is somewhat arbitrary. Here's how I decide k:

- first of all, in order to lower the variance of the CV result, you can and should repeat/iterate the CV with new random splits.
  This makes the argument of high k => more computation time largely irrelevant, as you anyways want to calculate many models. I tend to think mainly of the total number of models calculated (in analogy to bootstrapping). So I may decide for 100 x 10-fold CV or 200 x 5-fold CV.
- @ogrisel already explained that usually large k mean less (pessimistic) bias. (Some exceptions are known particularly for $k = n$, i.e. leave-one-out).
- If possible, I use a k that is a divisor of the sample size, or the size of the groups in the sample that should be stratified.
- Too large k mean that only a low number of sample combinations is possible, thus limiting the number of iterations that are different.
  - For leave-one-out: $\binom{1}{n} = n = k$ different model/test sample combinations are possible. Iterations don't make sense at all.
  - E.g. $n = 20$ and $k = 10$: $\binom{2}{n=20} = 190 = 19 \cdot k$ different model/test sample combinations exist. You may consider going through all possible combinations here as 19 iterations of k-fold CV or a total of 190 models is not very much.
- These thoughts have more weight with small sample sizes. With more samples available k doesn't matter very much. The possible number of combinations soon becomes large enough so the (say) 100 iterations of 10-fold CV do not run a great risk of being duplicates. Also, more training samples usually means that you are at a flatter part of the learning curve, so the difference between the surrogate models and the "real" model trained on all n samples becomes negligible.

answered May 4 '12 at 6:04

cbeleites
8,168 12 38

2 (+1) for the elaboration, but (-1) for the repetition counts of the CV. It is true, that the risk of creating **exact** duplicats (looking at the ids of the observations) is small (given enough data etc.), but the risk of creating **pattern/ data structure** duplicates is very high. I would not repeat a CV more than 10 times, no matter what k is ... just to avoid underestimation of the variance. – steffen May 4 '12 at 11:31

1 @steffen, isn't that what ogrisel already pointed out: that the (surrogate) models are not really independent? I completely agree that this is the case. Actually, I try to take this into account by interpreting the results in terms of stability of the (surrogate) models wrt. exchanging "a few" samples (which I didn't want to elaborate here - but see e.g. stats.stackexchange.com/a/26548/4598). And I do **not** calculate standard error but rather report e.g. median and 5^{th} to 95^{th} percentile of the observed errors over the iterations. I'll post a separate question about that. – cbeleites May 4 '12 at 12:35

1 I see. I agree that the approach is valid to estimate the stability of the surrogate. What I had back in mind was the follow-up-statistical test to decide whether one model outperforms another one. Repeating a cv way too often increases the chance of an alpha error unpredictably. So I was confusing the inner with the outer validation (as dikran has put it here). – steffen May 4 '12 at 15:45

@steffen, looking into such model comparisons is next on my list of things that I'd really like to look into. I personally have a lot of doubts about these data-driven optimizations. I've experienced extreme overfitting in the inner CV in some situations - and while I think the outer CV gives a sensible estimate of the resulting model's performance. For the moment I came to the conclusion that I get about as good models by deciding model parameters by my knowledge on the subject and my general experience with the models. – cbeleites May 6 '12 at 13:46

Larger K means less bias towards overestimating the true expected error (as training folds will be closer to the total dataset) but higher variance and higher running time (as you are getting closer to the limit case: Leave-One-Out CV).

If the slope of the learning curve is flat enough at training_size = 90% of total dataset, then the bias can be ignored and K=10 is reasonable.

Also higher K give you more samples to estimate a more accurate confidence interval on you estimate (using either parametric standard error assuming normality of the distribution of the CV test errors or non parametric bootstrap CI that just make the i.i.d assumption which is actually not very true as CV folds are not independent from one another).

Edit: underestimating => overestimating the true expected error

edited May 6 '12 at 2:00

answered May 4 '12 at 4:30

ogrisel
1,939 6 9

can you explain a bit more about the higher variance with large k? As a first approximation I'd have said that the total variance of CV result (= some kind of error calculated from all n samples tested by any of the k surrogate models) = variance due to testing n samples only + variance due to differences between the k models (instability). What am I missing? – cbeleites May 4 '12 at 5:29

Also, did you possibly mean: "less bias towards _over_estimating the ... error" ? – cbeleites May 4 '12 at 5:36

2 By variance I mean variance of the estimated expected test error obtained by taking the median or mean of the CV fold errors w.r.t. the "true distribution", not across CV folds. When k is big your are closer to LOO-CV which is very dependent on the particular training set you have at hand: if the number of samples is small it can be not so representative of the true distribution hence the variance. When k is big, k-fold CV can simulate such arbitrary hard samples of the training set. – ogrisel May 4 '12 at 7:10

1 I fixed the under/over typo (I was probably thinking in terms of scores instead of error...). – ogrisel May 4 '12 at 7:14

6 As an addition: Kohavi studies the bias-variance-tradeoff in validation in chapter 3 of his Phd thesis. I highly recommend it. – steffen May 4 '12 at 15:49
