Hyndsight

A blog by Rob J Hyndman

Why every statistician should know about cross-validation

Published on 4 October 2010

Surprisingly, many statisticians see cross-validation as something data miners do, but not a core statistical technique. I thought it might be helpful to summarize the role of cross-validation in statistics, especially as it is proposed that the Q&A site at stats.stackexchange.com (http://stats.stackexchange.com) should be renamed CrossValidated.com.

Cross-validation is primarily a way of measuring the predictive performance of a statistical model. Every statistician knows that the model fit statistics are not a good guide to how well a model will predict: high R^2 does not necessarily mean a good model. It is easy to over-fit the data by including too many degrees of freedom and so inflate R^2 and other fit statistics. For example, in a simple polynomial regression I can just keep adding higher order terms and so get better and better fits to the data. But the predictions from the model on new data will usually get worse as higher order terms are added.

One way to measure the predictive ability of a model is to test it on a set of data not used in estimation. Data miners call this a "test set" and the data used for estimation is the "training set". For example, the predictive accuracy of a model can be measured by the mean squared error on the test set. This will generally be larger than the MSE on the training set because the test data were not used for estimation.

However, there is often not enough data to allow some of it to be kept back for testing. A more sophisticated version of training/test sets is leave-one-out cross-validation (LOOCV) in which the accuracy measures are obtained as follows. Suppose there are n independent observations, $y_1,\dots,y_n$.

1. Let observation i form the test set, and fit the model using the remaining data. Then compute the error $(e_i^* = y_i - \hat{y}_i)$ for the omitted observation. This is sometimes called a "predicted residual" to distinguish it from an ordinary residual.
2. Repeat step 1 for $i = 1,\dots,n$.
3. Compute the MSE from $e_1^*,\dots,e_n^*$. We shall call this the CV.

This is a much more efficient use of the available data, as you only omit one observation at each step. However, it can be very time consuming to implement (except for linear models — see below).

Other statistics (e.g., the MAE) can be computed similarly. A related measure is the PRESS statistic (predicted residual sum of squares) equal to $n\times$MSE.

Variations on cross-validation include leave-k-out cross-validation (in which k observations are left out at each step) and k-fold cross-validation (where the original sample is randomly partitioned into k subsamples and one is left out in each iteration). Another popular variant is the .632+bootstrap of Efron & Tibshirani (1997) (http://www.jstor.org/stable/2965703) which has better properties but is more complicated to implement.

(http://robjhyndman.com)Rob J Hyndman (http://robjhyndman.com) is Professor of Statistics at Monash University (http://monash.edu), Australia,
and Editor-in-Chief of the International Journal of Forecasting (http://www.sciencedirect.com/science/journal/01692070/).

Twitter: @robjhyndman (http://twitter.com/robjhyndman)
Email: Rob.Hyndman@monash.edu (mailto:Rob.Hyndman@monash.edu)

Tags

beamer computing conferences consulting demography econometrics forecasting fpp graphics humour IJF jobs jokes journals JSS kaggle LaTeX mathematics maxima Monash University obituary organization otexts phd productivity progress publishing R refereeing references reproducible research research team seminars StackExchange statistics supervision tables teaching technology video welfare writing ysc2013

Popular Posts

- A LaTeX template for a CV
- Controlling figure and table placement in LaTeX
- Making a poster in beamer
- Why every statistician should know about cross-validation
- How to fail a PhD
- I'm switching to TeXstudio
- R graph with two y-axes
- Synchronizing WinEdt and pdf files

more complicated to implement.

Minimizing a CV statistic is a useful way to do model selection such as choosing variables in a regression or choosing the degrees of freedom of a nonparametric smoother. It is certainly far better than procedures based on statistical tests and provides a nearly unbiased measure of the true MSE on new observations.

However, as with any variable selection procedure, it can be misused. Beware of looking at statistical tests after selecting variables using cross-validation — the tests do not take account of the variable selection that has taken place and so the p-values can mislead.

It is also important to realise that it doesn't always work. For example, if there are exact duplicate observations (i.e., two or more observations with equal values for all covariates and for the y variable) then leaving one observation out will not be effective.

Another problem is that a small change in the data can cause a large change in the model selected. Many authors have found that k-fold cross-validation works better in this respect.

In a famous paper, Shao (1993) (http://www.jstor.org/stable/2290328) showed that leave-one-out cross validation does not lead to a consistent estimate of the model. That is, if there is a true model, then LOOCV will not always find it, even with very large sample sizes. In contrast, certain kinds of leave-k-out cross-validation, where k increases with n, will be consistent. Frankly, I don't consider this is a very important result as there is never a true model. In reality, every model is wrong, so consistency is not really an interesting property.

Cross-validation for linear models

While cross-validation can be computationally expensive in general, it is very easy and fast to compute LOOCV for linear models. A linear model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and the fitted values can be calculated using

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is known as the "hat-matrix" because it is used to compute $\hat{\mathbf{Y}}$ ("Y-hat").

If the diagonal values of $\mathbf{H}$ are denoted by $h_1, \dots, h_n$, then the cross-validation statistic can be computed using

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}[e_i/(1-h_i)]^2,$$

where e_i is the residual obtained from fitting the model to all n observations. See Christensen's book Plane Answers to Complex Questions (http://www.amazon.com/gp/product/0387953612?ie=UTF8&tag=prorobjhyn-20&linkCode=as2&camp=1789&creative=390957&creativeASIN=0387953612) for a proof. Thus, it is not necessary to actually fit n separate models when computing the CV statistic for linear models. This remarkable result allows cross-validation to be used while only fitting the model once to all available observations.

Relationships with other quantities

Cross-validation statistics and related quantities are widely used in statistics, although it has not always been clear that these are all connected with

Recent comments

- Rob J Hyndman
Thanks Tal. I've submitted a job advert on r-users.com (http://r-users.com)

New jobs in business analytics at Monash (http://robjhyndman.com/hyndsight/monash-business-analytics/) · 2 days ago (http://robjhyndman.com/hyndsight/monash-business-analytics/#comment-1369675348)

- Tal Galili
Hi Rob, Regarding Terry's reaction, it sounds like something my adviser (Yoav Benjamini) would...

New jobs in business analytics at Monash (http://robjhyndman.com/hyndsight/monash-business-analytics/) · 2 days ago (http://robjhyndman.com/hyndsight/monash-business-analytics/#comment-1369531863)

- edyhsgr
Why is MAPE typically used instead of Median Absolute Percent Error? Is MAPE better?

Errors on percentage errors (http://robjhyndman.com/hyndsight/smape/) · 2 days ago (http://robjhyndman.com/hyndsight/smape/#comment-1368799307)

- Rob J Hyndman
Thanks. That's my plan. I was just doing a catch-up of papers that I hadn't mentioned so far....

Great papers to read (http://robjhyndman.com/hyndsight/great-papers/) · 4 days ago (http://robjhyndman.com/hyndsight/great-papers/#comment-1365958117)

- Stephan Kolassa
Nice. How about sharing which papers you discuss on the blog? (Or even better, assign a student...

Great papers to read (http://robjhyndman.com/hyndsight/greatties, although it has not always been clear that these are all connected with cross-validation.

Jackknife

A jackknife estimator is obtained by recomputing an estimate leaving out one observation at a time from the estimation sample. The n estimates allow the bias and variance of the statistic to be calculated.

Akaike's Information Criterion

Akaike's Information Criterion is defined as

$$\text{AIC} = -2\log\mathcal{L} + 2p,$$

where $\mathcal{L}$ is the maximized likelihood using all available data for estimation and p is the number of free parameters in the model. Asymptotically, minimizing the AIC is equivalent to minimizing the CV value. This is true for any model (Stone 1977) (http://www.jstor.org/stable/2984877), not just linear models. It is this property that makes the AIC so useful in model selection when the purpose is prediction.

Schwarz Bayesian Information Criterion

A related measure is Schwarz's Bayesian Information Criterion:

$$\text{BIC} = -2\log\mathcal{L} + p\log(n),$$

where n is the number of observations used for estimation. Because of the heavier penalty, the model chosen by BIC is either the same as that chosen by AIC, or one with fewer terms. Asymptotically, for linear models minimizing BIC is equivalent to leave–v–out cross-validation when $v = n[1 - 1/(\log(n) - 1)]$ (Shao 1997) (http://www3.stat.sinica.edu.tw/statistica/oldpdf/A7n21.pdf).

Many statisticians like to use BIC because it is consistent — if there is a true underlying model, then with enough data the BIC will select that model. However, in reality there is rarely if ever a true underlying model, and even if there was a true underlying model, selecting that model will not necessarily give the best forecasts (because the parameter estimates may not be accurate).

Cross-validation for time series

When the data are not independent cross-validation becomes more difficult as leaving out an observation does not remove all the associated information due to the correlations with other observations. For time series forecasting, a cross-validation statistic is obtained as follows

1. Fit the model to the data $y_1, \dots, y_t$ and let $\hat{y}_{t+1}$ denote the forecast of the next observation. Then compute the error $(e^*_{t+1} = y_{t+1} - \hat{y}_{t+1})$ for

   the forecast observation.
2. Repeat step 1 for $t = m, \dots, n-1$ where m is the minimum number of observations needed for fitting the model.
3. Compute the MSE from $e^*_{m+1}, \dots, e^*_n$.

References

An excellent and comprehensive recent survey of cross-validation results is Arlot and Celisse (2010) (http://dx.doi.org/10.1214/09-SS054)

Related Posts:

- Fast computation of cross-validation in linear models
- Facts and fallacies of the AIC
- Time series cross-validation: an R example
- Fitting models to long time series
- Out-of-sample one-step forecasts

(http://robjhyndman.com/hyndsight/great-papers/) · 4 days ago (http://robjhyndman.com/hyndsight/great-papers/#comment-1365870222)

Recommended books on Amazon

Tags: forecasting, StackExchange, statistics

33 Comments

29 Comments Hyndsight Login

Sort by Oldest Share Favorite

Join the discussion…

Stephan Kolassa · 4 years ago

Very nice article, thanks! The Arlot and Celisse paper is even freely available from Project Euclid... as if I didn't have enough to read already... ;-)

Any thoughts on using cross-validation with mixed linear models, e.g., with repeated measurements on each participant in clinical studies? It seems as if Arlot & Celisse don't explicitly treat this case.

Reply · Share ›

Rob J Hyndman Mod → Stephan Kolassa · 4 years ago

Thanks Stephan. Unfortunately I don't know anything about CV with mixed effects models. However, the following paper may help: http://www.biostat.uzh.ch/rese...

Reply · Share ›

Matt Schneider → Stephan Kolassa · 2 years ago

Hi Stephan,
For linear mixed models, the concept of leverage on the diagonals of the H matrix (above) may not apply since the conditional residuals (given BLUPs: Y-XB_hat-Zu_hat) and the BLUPs (u_hat's) are confounded. Here is some R code and summary documentation from Nobre and Singer (2007, 2011) which may help defining a new form of CV for mixed models if interested. http://www.ime.usp.br/~jmsinge...
http://www.tandfonline.com/doi...

Reply · Share ›

Stephan Kolassa · 4 years ago

Hm, this looks interesting. Thanks!

Reply · Share ›

inwit · 4 years ago

Hi, Rob! This post of yours brought me back one old question about time series and cross-validation. Instead of posting it here, I've sent it to StackExchange. This is a great blog and a good source for inspiration! Keep rocking! :)

Reply · Share ›

Vishal Belsare · 4 years ago

Rob, thanks for a nice post! and for pointing out the paper by Arlot & Celisse.

Reply · Share ›

Abhijit · 4 years ago

Hi Rob,

Very nicely done, indeed.

I'll bite and ask about your comment on consistency. I would agree that any model we create is wrong, but it doesn't follow that there is no underlying true model, however complex, that we're trying to approximate. I would posit that model, however complex, that we're trying to approximate. I would posit that at least locally, and perhaps globally, there is a true regression function E(Y¦X). We can discuss this offline, if you like.

Reply • Share ›

Rob J Hyndman Mod → Abhijit • 4 years ago

I think we are using consistency in two different ways. I agree that consistency of an estimator of E(Y|X) is important when we know X. And we get that with any decent estimator provided the form of Y = f(X,error) is correctly specified. (e.g., if f is linear and error is iid than OLS is consistent.)

But the problem of variable selection is finding E(Y|Z) where Z is unknown and probably unknowable. The BIC provides consistency only when Z is contained within our set of potential predictor variables, but we can never know if that is true. That is why I suggest that consistency is not a useful concept in this context.

Reply • Share ›

Yaroslav Bulatov • 4 years ago

Another issue I have with consistency is that it addresses infinite sample case, but you may never see enough data for the infinite sample properties to matter. For instance, Contrastive Divergence estimator is consistent but for a dense model over n variables it takes in the order of 2^n samples for the estimate to approach true value

Reply • Share ›

Abhijit → Yaroslav Bulatov • 4 years ago

Yes, so a lot of my colleagues look at rates of convergence rather than just consistency. We've also been playing with simulation strategies to see how well a "consistent" method does in the finite sample situations.

Reply • Share ›

Yaroslav Bulatov • 4 years ago

Once you have an estimate of method's performance for finite sample size, why does consistency matter? IE, would you ever have a reason to prefer a consistent estimator over an inconsistent one with better finite-sample properties?

Reply • Share ›

Bob Carpenter • 4 years ago

I think the biggest difference between practitioners of stats and machine learning is what inferences they care about.

Suppose we have data on region of the location you live in, education, sex, age, ethnicity, price of home, and mortgage
on-time payment status, say in a time series over a decade.

With my "machine learning" hat on, I want to predict whether an individual will default in some time frame given the value of the predictors. My imagined "customer" is a bank. I might throw the data through an SVM with a complex kernel if I only care about 0/1 outcome, or through a decision forest, or use K nearest neighbors. All of these are likely to produce reasonable default predictions, and a committee of such even better predictions.

With my "applied statistician" hat on, I might want to estimate the effect of age on mortgage defaults, controlling for the other predictors.

It's just a very different game. Cross-validation makes much more sense in the former game. But as you point out, one needs to be careful. The biggest mistake I see, in practice, is the one you mention -- tuning using cross-validation over all folds then assuming you'll get the same performance on new data.

Both groups tend to forget that neither the population nor effects are stationary (in the statistical sense). That is, the population over which we're predicting isn't the same as the one over which we collected the data. Sure, we can do things like post-stratification to adjust for sampling bias, but there's the underlying change in attitudes, wealth, and so on that are changing in this example.

6 • Reply • Share ›

Chandler Lutz • 3 years ago

Hi Rob,

I really enjoyed the article. Do you have a reference for time series cross-validation technique that you mention at the end?

Thanks, Chandler

• Reply • Share ›

Rob J Hyndman Mod → Chandler Lutz • 3 years ago

That is common practice is forecasting evaluation studies, but I've never seen it in a textbook. I've put it in my new book (incomplete) at http://otexts.com/fpp/2/5/.

• Reply • Share ›

Thomas → Rob J Hyndman • a year ago

Dear Rob,
Many thanks for the article (2 years later...). Do you by any chance have any reference of this technique being used in a published article?
Thanks!
Thomas

• Reply • Share ›

Rob J Hyndman Mod → Thomas • a year ago

Try these: http://goo.gl/GVCSC

• Reply • Share ›

Torsten Seemann • 2 years ago

Mention should be made of the more general information-theoretic approaches such as MML (minimum message length) and MDL (minimum description length) of which AIC and BIC are restricted instances of:

MML: http://en.wikipedia.org/wiki/M...
MDL: http://en.wikipedia.org/wiki/M...

• Reply • Share ›

Jan Galkowski • 2 years ago

Any comments on relative merits of cross-validation and 0.632+ bootstrap, especially for time series?

• Reply • Share ›

Rob J Hyndman Mod → Jan Galkowski • 2 years ago

I don't know much about this. Efron and Tibshirani (http://www.jstor.org/stable/29...) argue for the 0.632 bootstrap over cross-validation but I don't think it has any real theoretical support. I've not thought about how the 0.632 bootstrap would work in the time series context.

• Reply • Share ›

Jan Galkowski • 2 years ago

The bootstrap itself has plenty of theoretical support (*) both in an independent and dependent data contex. (References below.) However, I have not seen much in terms of generalizing the 0.632 (which Hall, at least, argues should really be 0.667). I did read about after posting my question, to see what I could find:

R.M.Kunst, "Cross validation of prediction models for seasonal time series by parametric bootstrapping," Austrian Journal of Statistics, 37(3&4), 2008, 271-284.

D.N.Politis, J.P.Romano, "The stationary bootstrap", JASM, 89(428), 1994, 1303-1313.

(*) S.N.Lahiri, RESAMPLING METHODS FOR DEPENDENT DATA, Springer, 2010.

P.Hall, "On the biases of error estimators in prediction problems", Statistics and Probability Letters 24(3), 15 Aug 1995,

see more

• Reply • Share ›

Rob J Hyndman Mod → Jan Galkowski • 2 years ago

Thanks for the references. I meant that Efron and Tibshirani's 0.632 bootstrap idea was empirically rather than theoretically based. Of course, there are many variations of the bootstrap that have been thoroughly studied from a theoretical perspective.

• Reply • Share ›

Fabio Goncalves • 2 years ago

Hi Rob, thanks for the article! The link to Shao (1995) below actually points to a 1993 paper, which doesn't seem to mention Schwarz's BIC.
"Asymptotically, for linear models minimizing BIC is equivalent to leave—out cross-validation when (Shao 1995)."

Would you be able to confirm this reference?

Thanks!

• Reply • Share ›

Rob J Hyndman Mod → Fabio Goncalves • 2 years ago

Thanks for spotting that error. I've fixed the link to point to Shao (1997).

• Reply • Share ›

Matt Schneider • 2 years ago

To the group: I read various machine learning papers on prediction that select a tuning parameter or number of iterations (let's say for boosting or trees) based on k-fold cross validation. I can see how it makes sense if either of those parameters (tuning, iterations) are chosen on each of k training sets and then we look at the average prediction error (let's say MSE) of the k test sets. However, am I misunderstanding something or is it a misuse when the parameters are chosen based on the aggregate prediction results after doing all the k-folds (arg min (parameters) { total MSE on test sets}) ?

Two thoughts: 1) For inference this may be OK because those "best" tuning parameters accurately model the population well and a typical error of a withheld observation. 2) To call this forecasting, it seems off. Optimal parameters depend on all the data. None of the data was completely withheld. "Prediction error" isn't necessarily "forecast error?"

Agree? Disagree?

• Reply • Share ›

Chong Wu • a year ago

Dear professor Rob J Hyndman

I am Chong Wu from China and I will be finishing a BS degree in Applied Math at the Huazhong University of Science & Technology (Top 10 in China) next year. I like this clear and enlightened article.
I was wondering if you have any plan to recruit new PhD candidate in 2013 fall.

• Reply • Share ›

Rob J Hyndman Mod → Chong Wu • a year ago

We recruit PhD students every year, with applications closing at the end of October. See http://robjhyndman.com/researc... for details.

⌃ | ⌄ • Reply • Share ›

Istvan Hajnal • a year ago

Great overview. Thanks.

⌃ | ⌄ • Reply • Share ›

Econstudent • 10 months ago

Great post, Professor,

I am a little surprised that for time series (or dependent data in general) you did not mention the pertinent reference

P.Burman, E.Chow, D.Nolan, "A cross-validatory method for dependent data", BIOMETRIKA 1994, 81(2), 351-358.

And a more recent contribution is

@article{
Author = {Racine, Jeff},
Title = {Consistent cross-validatory model-selection for dependent data: hv-block cross-validation},
Journal = {Journal of Econometrics},
Volume = {99},
Pages = {39-61},
Year = {2000} }

⌃ | ⌄ • Reply • Share ›

Tim Johnson • 2 months ago

Hi Rob

I think this is a question you can help answer ?

http://stats.stackexchange.com...

⌃ | ⌄ • Reply • Share ›

ALSO ON **HYNDSIGHT** WHAT'S THIS?

Forecasting within limits

1 comment • 3 months ago

Great papers to read

2 comments • 5 days ago

Archives

Select Month ▼