White Paper

It is a Mistake to Focus on Training Results

John Elder
Founder & CEO

August 2013

ELDER RESEARCH
DATA SCIENCE · AI · MACHINE LEARNING

This article is Part 2 (of 11) of a series by the author on the Top 10 Data Mining Mistakes, drawn from the *Handbook of Statistical Analysis and Data Mining Applications*.

Only out-of-sample results matter; otherwise, a lookup table would always be the best model. Researchers at the MD Anderson medical center in Houston (almost two decades ago) used neural networks to detect cancer. Their out-of-sample results were reasonably good, though worse than training, which is typical. They supposed that longer training of the network would improve it – after all, that's the way it works with doctors – and were astonished to find that running the neural network for a week (rather than a day) led to only slightly better training results and much worse evaluation results. This was a classic case of overfit, where obsession with getting as much as possible out of training cases focuses the model too much on the peculiarities of that data to the detriment of inducing general lessons that will apply to similar, but unseen, data. Early machine learning work often sought, in fact, to continue "learning" (refining and adding to the model) until achieving exact results on known data — which, at the least, insufficiently respects the incompleteness of our knowledge of a situation.

The most important way to avoid overfit is to reserve data. But, since data are precious — especially the cases of greatest interest (like fraud) — one must use resampling tools, such as bootstrap, cross-validation, jackknife, or leave-one-out. Traditional statistical significance tests are a flimsy defense when the model structure is part of the search process, though the strongest penalty-based metrics, such as Bayesian Information Criterion, or Minimum Description Length, can be useful in practice. Further, resampling – as opposed to using a statistical test – doesn't require one to make assumptions about distribution of the data. In practice, this saves a great deal of worry and hesitation!

With resampling, multiple modeling experiments are performed, with different samples of the data, to illuminate the distribution of results. If one were to split the data into training and evaluation subsets a single time, the evaluation accuracy result might largely be due to luck (either good or bad). By splitting it, say, 10 different ways, and training on the 90% sets and evaluating on the out-of-sample 10% sets, one has ten different accuracy estimates. The mean of this distribution of evaluation results tends to be more accurate than a single experiment, and it also provides, in its standard deviation, a confidence measure.

Note that resampling evaluates whatever is held constant throughout its iterations, or "folds". That is, one can set the structure (terms) of a model and search for its parameter values over multiple data subsets; then, the accuracy results would apply to that fixed model structure. Or, one could automate multiple stages of the process – e.g., outlier detection, input selection, interaction discovery – and put that whole process inside a resampling loop. Then, it's the accuracy distribution of that full process that is revealed.

In the end, one has created multiple overlapping models; so which is the model to use? One approach is to choose a single model (perhaps by its beauty rather than its accuracy). Probably the most popular choice, is to re-run the model-building process — training on all the data — and assume that the resulting model inherits the accuracy properties measured by the cross-validation folds of the previous stage. The most sophisticated (and usually most accurate) approach, is to combine the multiple models in an ensemble; for instance, simply averaging all of their outputs. (Ensembles will be described in more detail at the end of this series on mistakes; see also Chapter 13 of the Handbook or our book on Ensembles.)

Bottom Lines:
1) Trying to raise accuracy on the training data extremely high will hurt the models' performance on new data, which is all that really matters.
2) Resampling methods are extremely useful and can experimentally control such overfit in a huge variety of situations

About the Author

Dr. John Elder, Founder and CEO of Elder Research, leads the largest and most experienced data science consulting firm in the U.S. For 20 years, the team has applied advanced analytics to achieve high ROI for investment, commercial and security clients in fields from text mining and stock selection, to credit scoring and fraud detection. John has Engineering degrees from Rice and the University of Virginia, where he's an adjunct professor. He's authored innovative tools, is a popular keynote speaker, and has chaired International Analytics conferences. Dr. Elder served 5 years on a panel appointed by President Bush to guide technology for National Security. He has co-authored three books (on data mining, ensemble modeling, and text mining), two of which won Prose "book of the year" awards.

www.elderresearch.com

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING