White Paper

It is a Mistake to Ask the Wrong Questions

John Elder
Founder

November 2013

ELDER RESEARCH
DATA SCIENCE · AI · MACHINE LEARNING

This article is Part 4 (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*.

It is very important to have the right project goal; that is, to aim at the right target. This was exemplified (in a positive way) by a project at Shannon Labs, led by Daryl Pregibon, to detect fraud in international calls. Rather than use a conventional approach, which would have tried to build a model to distinguish (rare, but expensive) fraud from (vast examples of) non-fraud, for any given call, the researchers characterized normal calling patterns for each account (customer) separately. When a call departed from what was the normal pattern for that account, an extra level of security, such as an operator becoming involved, was initiated. For instance, if one typically called a few particular countries each week, briefly, during weekdays, a long call to a different region of the world on the weekend would bear scrutiny. Efficiently reducing historical billing information to its key features, creating a mechanism for the proper level of adaptation over time, and implementing the models in real time for vast streams of data, provided interesting research challenges. Still, the key to success was asking the right question of the data. The ongoing "account signature" research won technical awards (at a KDD conference, for example) but, more importantly, four researchers, part time in a year, were able to save their company enough money to pay the costs of the entire Shannon Labs (of 400 people) for the next year[1] — an impressive example of Data Mining return on investment (ROI).

Even with the right project goal it is essential to also have an appropriate model goal. You want the computer to "feel" about the problem like you do – to share your multi-factor score function, just as stock grants or options are supposed to give key employees a similar stake as owners in the fortunes of a company.[2] But, analysts and tool vendors almost always use squared error as the criterion, rather than one tailored to the problem. Lured by the incredible speed and ease of using squared error in algorithms, we are like drunks looking for lost keys under the lamppost – where the light is better – rather than at the bar where they were likely dropped.

For instance, imagine that we're trying to decide whether to invest in our company's stock as a pension option, and we build a model using squared error. Say it forecasts that the price will rise from $10 to $11 in the next quarter, and it goes on to actually rise to $14. We've enjoyed a positive surprise; we expected a 10% gain, but instead got 40%.[3] However, when entering in that data for the next go-round of analysis, the computer has a different response; it sees an error of $3, between the truth and the estimate, and squares that to a penalty of 9. It would have more than twice "preferred" it if the price had dropped -$1 to $9; then, its squared error would only have been 4. A criterion which instead punishes negative errors much more than positive errors would better reflect our preferences.

Though conventional squared error can often put a model into a serviceable region of performance, the function being optimized has a thorough effect on the suitability of the final model. “Inspect what you expect”, a retired IBM friend often says about managing projects. For instance, you won’t produce the best-spelling students if your grading has focused on penmanship. When performance is critical, have the computer do not what’s easiest for it (and thereby, us) but what’s most useful. To best handle custom metrics, analysts need a strong multi-dimensional (and preferably, multi-modal) optimization algorithm.[4] Still, using even a simple random search with a custom score function is better than not customizing our criteria of merit.

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.

---

[1] They were rewarded, as we techno-nerds like, with bigger toys. The group got a “Data Wall” – a 10’x20′ computer screen, complete with couch, with which to visualize data. As it was often commandeered by management for demonstrations, the research group was eventually provided a second one to actually use.

[2] Unfortunately, the holder of an option has a different score function from the owner of the stock. The option is very valuable if the company thrives, but only worthless if it doesn’t. Yet, the owner can be seriously hurt by a downturn. Thus, a manager’s rational response to having options is to take on increased risk — to “shoot the moon” for the potential up-side reward. To better align owner and management interests, it is better to grant stock outright, rather than (cheaper, but more inflammatory) options.

[3] Of course, our joy is short-lived, as we kick ourselves for not mortgaging the house and betting even more! Fear and greed are always at war when dealing with the markets.

[4] In my 1993 PhD dissertation, I introduced a global search algorithm for multi-modal surfaces which updates a piecewise planar model of the score surface as information is gathered. It is very efficient, in terms of function evaluations, but its required overhead restricts it to a dozen or so dimensions (simultaneous factors) in practice.

www.elderresearch.com

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING