White Paper

It is a Mistake to…. Answer Every Inquiry

by John Elder

Founder & CEO

January 2015

ELDER RESEARCH

DATA SCIENCE & PREDICTIVE ANALYTICS

This article is Part 9 (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*.

I'm tempted to start with a kind of query that experience teaches some of us not to answer, like "Does this data make me look fat?" But that actually misleads about the point I'm trying to make. Data Scientists (and their models) *should* answer all queries as truthfully as the evidence allows, regardless of how happy or unhappy that makes the questioner. What I am arguing here is we shouldn't answer when our opinion is unqualified; that is, when there is not enough evidence. I learned this the hard way!

Early in my career, I demonstrated a model built to estimate rocket thrust that used engine temperature, T, as one of the important inputs. A technical gate-keeper for the potential client (who, it turned out, was trying to kill the project to advance his own agenda) slyly suggested we vary some inputs and see what ensued. "Try $T = 98.6$ degrees." (human body temperature, way below the bounds of normal operation.) I argued the test would be senseless, since the input was far outside the model's training bounds, but with much cajoling, I naively complied (See earlier Mistake #7). The model was a nonlinear polynomial network, so when given an input value far outside its training range its output was ridiculous, as expected, but no amount of calm technical explanation around that non-surprising result could erase, in the onlooking decision-makers mind, the negative impact of the breathtakingly bad result that had briefly flashed by. My firm never heard from that company again. Obviously, a model should answer "don't know" for situations in which its training has no standing!

But, how do we know where the model is valid; that is, has enough data close to the query by which to make a useful decision? The simplest approach is to note whether the new point is outside the bounds, on any dimension, of the training data. Yet, especially in high-d, the volume of the populated space is only a small fraction of the volume of the rectangle defined by the univariate bounds. With most real data, inputs are very far from mutually independent, so the occupied fraction of space is very small, even in low-d. (The data often look to me like a folded umbrella packed diagonally in a box.) A second approach, more difficult and rare, is to calculate the convex hull of the sample – essentially, a "shrink wrap" of the data points. Yet even this does not always work well to define the populated space. Figure 1 illustrates a 2-d problem similar to one I encountered in practice (in higher-d) in an aeronautical application. There, practical constraints on joint values of physical variables (e.g., height, velocity, pitch, and yaw) caused the data to be far from i.i.d. (independent and identically distributed.) I noticed then, as in the Figure, that astonishingly, even the sample mean of the data, μ, was outside the true region of populated space!

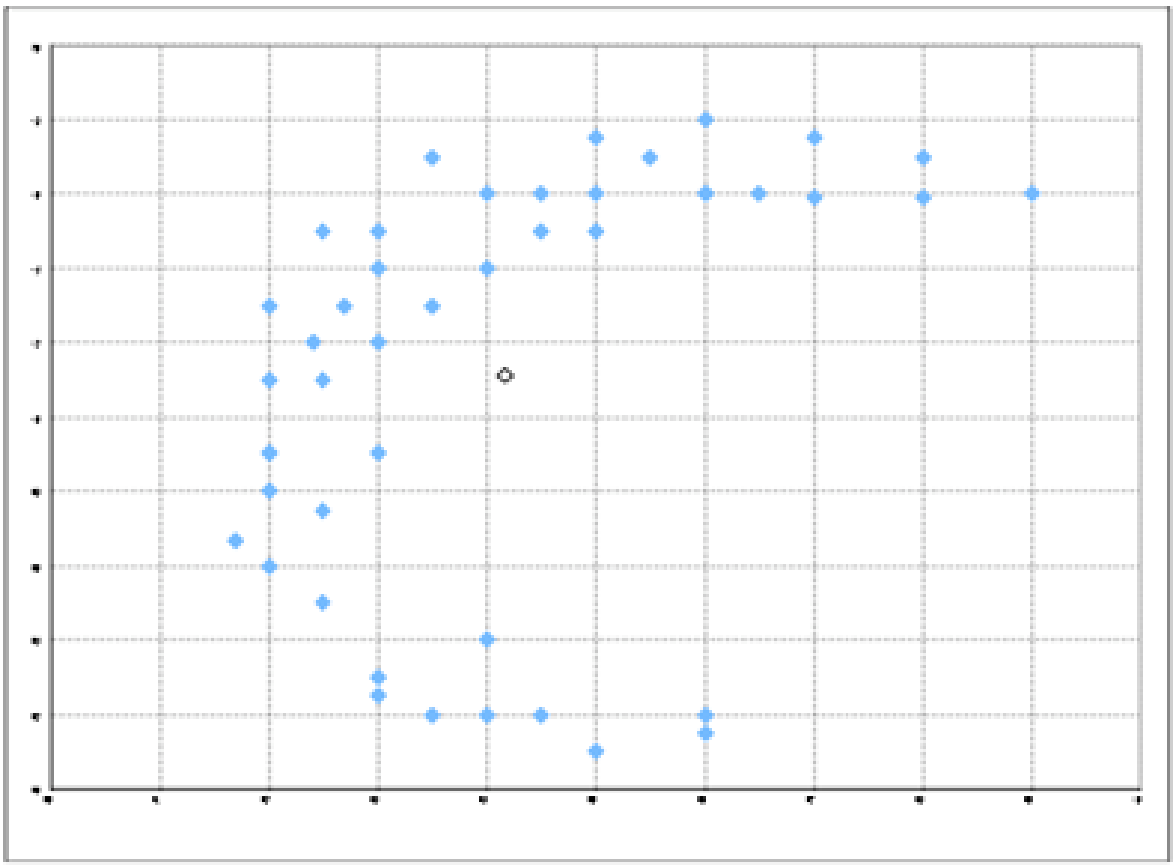

Figure 1: Example 2-dimensional problem for which the data mean (open box symbol) is outside the bounds of the (crescent-shaped) valid space

An approach for flagging some outliers (though perhaps not all) that has helped the few times we've tried it, is to fit a very responsive, nonlinear model to the data, for instance through a polynomial network (Elder & Brown, 1992). High-order polynomials quickly go toward infinity outside the bounds of the training data. If the output estimate resulting from an unbounded, nonlinear (and even overfit) model is well beyond the output bounds, then it is very likely the input point is outside the training data. If a training data point had been near that new input point, it would have better constrained the model's estimate.

Just as it is essential to know where a model has standing – i.e., in what regions of input space its estimates might be valid – it is also useful to know the uncertainty of estimates. Most techniques provide some measure of spread, such as *s*, for the overall accuracy result (e.g., +/- 3% for a political survey), but it is rare indeed to have a conditional standard deviation, $\sigma(x)$, to go with the conditional $\mu(x)$. That is, to have a different uncertainty level for each region of input space, as determined by the data. A valuable area of research, I believe, would be to enhance existing modeling methods to estimate certainty conditioned on where in input space one is inquiring.

I did develop one estimation algorithm, which I call *Delaunay Triangles*, to depend strongly on *σ(x)*; it's goal is to make optimal use of experimental information for global optimization (Elder, 1993). For systems where results are expensive to obtain (e.g., samples from drilling, or other physical experiments), the challenge is to find, as efficiently as possible, the location (input) with the best result (output). If several samples and their results are known, one can model the score surface (relationship between input vector and output score) and rapidly query or traverse the surface of the model to find the best location for the next probe (i.e., experimental settings to employ). If that result isn't yet good enough (and budget remains to keep going), its sample-result information could be used to update the model for use in searching for the new best probe location. The overall estimation surface consists of piecewise planes, as shown in Figure 2, where each region's plane has a quadratic variance "canopy" over it, as shown in Figure 3, revealing how the uncertainty of the estimation grows as one departs from the known points (the corners).[1] This approach worked extremely well, for low (1-12 or so) dimensions, and the resulting multi-modal search algorithm, GROPE (Global R^d Optimization when Probes are Expensive) took the fewest probes of all then-existing algorithms to converge close to the answer on an academic suite of test problems. By having, for every location, *x*, an estimate of the mean, *μ(x),* along with its uncertainty, *σ(x),* the algorithm could, with every new result, refine its estimates and reduce its uncertainty, and thereby zero in on the locations with the greatest potential.

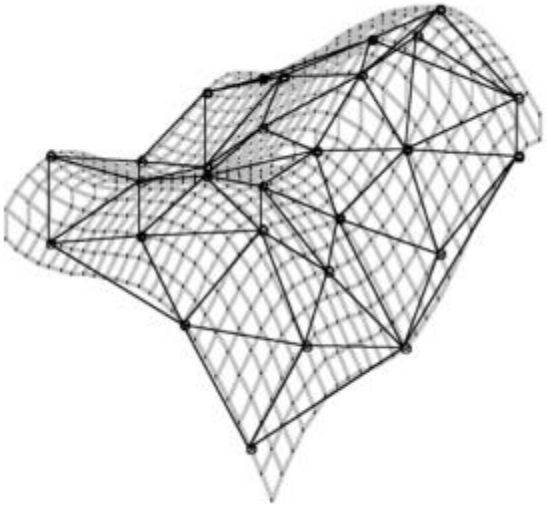

Figure 2: Estimation surface of Delaunay Triangle method (Elder, 1993) is piecewise planar. (The underlying function surface is represented here by a mesh.)

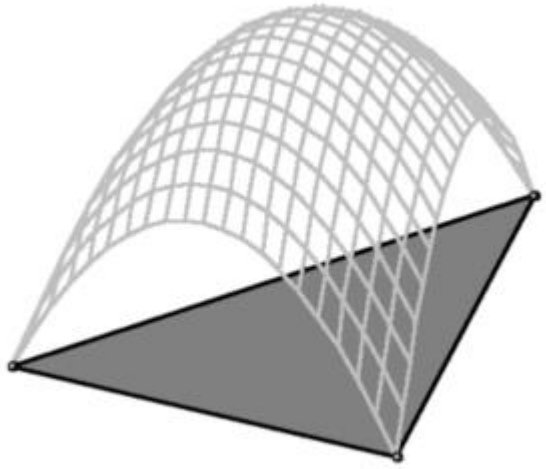

Figure 3: Each simplex (e.g., triangle in 2-dimensions) of the Delaunay method (Elder, 1993) pairs a planar estimation of μ (x) with a quadratic estimation of σ 2(x).

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.

Reference:

Elder, John F. (1993) *Efficient Global Optimization through Response Surface Modeling: A GROPE Algorithm*, Ph.D dissertation, University of Virginia, May.

---

[1] The modeling technique developed for GROPE was driven by the special requirements of optimizing an unknown function – especially, that the response surface model had to agree exactly with the known samples. If one assumes the least about the response surface – that there is Brownian motion (or a random walk) between the known points — then the ideal estimator turns out to be a plane. So, $\mu(x)$ is a piecewise planar collection of simplices (e.g., triangles when there are two input dimensions). The tiling or tessellation of the input space is done in such a way as to create the most uniform simplices (those with the greatest minimum angle), which is performed by Delaunay triangulation (a dual of nearest neighbor mapping). The key though, was to pair this with an estimate of the standard deviation of $\mu(x)$, conditioned on x, $\sigma(x)$. (The Brownian motion assumption drives this to be the square root of a quadratic function of distance from the known corners.) Now, with both parts, $\mu(x)$ and $\sigma(x)$, one can rapidly calculate the location, x, where the probability of exceeding one's result goal is the greatest. So, the model would suggest a probe location, one would perform the experiment, and the result would update the model, with greater clarity on the mean estimates (piecewise planes) and reduced variance (piecewise quadratic "bubbles" over each plane) with each iteration.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673