Steampunk Machine Learning

VICTORIAN CONTRIVANCES FOR MODERN DATA SCIENCE

TERENCE KELLY

BOREDOM PUNCTUATED BY ERROR

Like all engineers, programmers and datacenter operators employ models to describe and predict the behavior of complex production systems. For example, a performance expert might define a parameterized equation that maps a computer system's workload to its application-level performance under ordinary conditions. Fitting such a model to data—setting the free parameters so that the equation agrees closely with measurements of the running system—is straightforward in principle.

In practice, however, production systems impose challenging complications. Modern datacenters suffer sudden failures, unforeseen overload, and myriad mishaps—rare, qualitatively anomalous episodes that beget wild extremes and outright absurdities in the data to which models are fitted. Such anomalous data points are noise because they reflect system *derailment* rather than normal system operation; noise can mislead unwary model-fitting techniques into spouting nonsense.

Scrubbing out bogus data points before fitting a model would be difficult even for a human expert working at a leisurely pace, because high-dimensional measurement data defy visualization and resist analysis; furthermore, armchair analysis isn't an option for datacenter automation. To be practical, a methodology

must fit models to routine data while effectively ignoring interspersed anomalous data, without human assistance. As a bonus side effect, a successful approach will flag anomalies as the odd data points that flout an otherwise-satisfactory fitted model.

This episode of Drill Bits first reviews a popular and versatile model-fitting technique that can work well for well-behaved data. It fails spectacularly, however, on data contaminated with anomalies typical of production computing systems. We then meet an obscure yet equally versatile alternative, invented centuries ahead of its time, that does a far better job of fitting models to real-world data. We explain the superior method's uncanny knack for separating wheat from chaff, and describe its application to performance-anomaly detection and performance prediction in real datacenters. A sidebar disentangles terms and concepts from machine learning and statistics; example code fits models to example data using the popular R analytics platform.

SENDING WIGGLES INTO FITS

The synthetic data of figure 1 (open circles) idealize three features of workload time-series data from production systems: a long-term upward trend, diurnal cycles, and random short-term fluctuations. Outliers are conspicuously absent. Such well-mannered data make it easy to fit an explicable and parsimonious linear model, but first let's dispel a widespread misconception.

The *linear* aspect of linear modeling isn't nearly as restrictive as newcomers might think. It does *not* mean that "models must be straight lines" in the x-y plane or

FIGURE 1: FRIENDLY DATA AND FITTED RAMP + OSCILLATION MODEL

hyperplanes in higher dimensions. For example, given observations of a response variable z and predictor variables x and y , the model equation $z = \alpha_0 + \alpha_1 e^x + \alpha_2 y^2$ is fair game: The right-hand side is not linear in the predictor variables but *is* linear in the α parameters, which is what matters. Linear modeling is thus very versatile—a fact that deserves more prominence given the confusing terminology. One lengthy text¹⁰ clarifies the matter on page 224, but it really belongs on the cover.

For the data of figure 1, a commonsense model equation is $y = \alpha_0 + \alpha_1 x + \alpha_2 \sin(t)$, where response variable y is hourly request rate (vertical axis), predictor x is time in hours (horizontal axis), and time-of-day t is derived from x and expressed as an angle in radians: $t = 2\pi((x \bmod 24)/24)$. The α parameters represent y -intercept, slope, and the amplitude of diurnal cycles.

Parameter values can be derived from (x, y)

[XKCD 2048](#)

CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

observations in many different ways.

The most popular is least-squares regression (LSQ), which minimizes the sum of squared *residuals*—discrepancies between the data points and the fitted model. In the example code, LSQ regression is a breeze: Transforming a `.csv` file into α values takes a few keystrokes. The resulting fitted model, $y = 3009.3 + 78.036x + 1953.5\sin(t)$, is shown in red in figure 1. Eyeballing suggests a good fit: The residuals appear to be small.

Appearances can deceive and automated modeling can't use eyeballs, so diligent practitioners formally validate fitted models. Validation methodology is beyond the scope of this column but is explained elsewhere in the context of R modeling,¹² computer performance modeling,⁸ and linear modeling in general.¹⁰

FIT FOR PURPOSE

Figure 2 adds a pair of anomalous incidents to the data: demand overload on the first day and a complete outage on the last. These incidents yield *clusters* of outliers far from the data points corresponding to ordinary operating conditions, a feature common in data from production computing systems. Shown in red is a least-squares fit of the same model equation fitted in figure 1. The clusters of outliers work as a team, wrenching the LSQ fit so far

XKCD 2048

clockwise that the slope goes negative and dragging the red curve far away from nearly all data points.

Is the model equation that nicely fit the data of figure 1 inapplicable when the outliers of figure 2 are present? Or does the problem lie not with the model *equation* but rather with the *fitting method*?

In the lingo of statistics, a regression procedure that resists influence by outliers is termed *robust*. Least squares is not robust, but many alternatives are.¹⁶ If rogue data points are not excessively numerous, they won't drag a robust fit away from the mundane data points.

The oldest and conceptually simplest robust procedure is LAR [least absolute residuals] regression. LAR minimizes the sum of *absolute* residuals—a more obvious and more intuitive objective than the LSQ criterion. First proposed in the mid-18th century,¹⁴ LAR predates LSQ by a generation and modern statistics by a century.

Mathematical-statistical niceties^{4,9,10} sometimes

recommend least squares over LAR or vice versa. Computational differences, however, largely explain the popularity of LSQ and the obscurity of LAR: Whereas matrix operations suffice for LSQ, general LAR regression is a linear programming problem,⁴ which is much more computationally demanding. LAR became practical only with the advent of high-speed electronic computers and efficient algorithms,^{1,2} two centuries after it was first proposed, whereupon its advantages quickly became apparent.⁹

The blue curve in figure 2 is an LAR fit of the same model equation: $y = 4523.2 + 62.301x + 1754.0\sin(t)$. Slight clockwise torque is evident, but overall the LAR-fitted model agrees closely with the routine data points while ignoring the anomalous ones almost entirely.

SORE THUMBS IN THE TRENCHES

Figure 3 shows the cumulative distributions of residuals from the least-squares and LAR fits of figure 2. The distribution

FIGURE 3: CUMULATIVE DISTRIBUTIONS OF RESIDUALS

of LSQ residuals is unimodal and spans a wide range. In contrast, the LAR distribution is trimodal: The majority of residuals are small and span a narrow range in the center; these correspond to routine operating conditions. Residuals corresponding to anomalous conditions are large and stand out prominently, far to the sides.

Fitting a model so that anomalies manifest via residuals is overkill for this simple two-dimensional example dataset, whose anomalies are obvious by design. High-dimensional data from the real world, however, are not so easy to visualize; anomalies don't announce themselves loudly. Often the best way to detect when a real production system has gone haywire is the approach used here: Encapsulate expert knowledge of the system in a model equation that describes its behavior well under ordinary conditions, fit the model to production measurements using a robust procedure, and direct suspicion toward large residuals.

My colleagues and I applied this methodology to identify performance anomalies in a real production system.¹³ Our response variable was application-level performance, our predictors were workload characteristics, and we used queuing-theoretic model equations that map workload to performance. Under routine operating conditions, workload explained performance well. When the system went off the rails, workload did *not* explain performance, resulting in glaring residuals. Our performance models could also predict the consequences of server consolidation and never-before-seen workload changes.

LAR yielded consistently better results than least-squares regression in our work. We also tried several other

**Prediction
is difficult,
especially about
the future**

— attributed to Niels
Bohr, Yogi Berra,
and others

robust regression procedures, some of which require fussy tuning. We prefer LAR.

ONE-DIMENSIONAL ROBUSTNESS

LAR's robustness stems from its optimization criterion. To understand the connection, work through a simple special case: Compute a “representative” value for a given set of numbers by treating them as observations of a response variable with no corresponding predictor variables; fit an intercept-only model via regression.

Figure 4 shows how LAR and LSQ optimize their respective fitting criteria for the set of numbers $\{1, 2, 3, 4, 99\}$. The smooth red LSQ curve on the right plots the sum of squared residuals—the quadratic spelled out on the right-hand axis. To find the minimum, take the derivative, set it equal to zero, and solve for n , which yields the formula for the *arithmetic mean* of the given numbers (try it). Least-squares regression generalizes the mean and

FIGURE 4: LAR AND LSQ OPTIMIZATION CRITERIA

is sensitive to outliers for the same reasons.

The chisel-pointed blue LAR curve on the left-hand side of figure 4 plots the sum-of-absolute-residuals function on the left-hand axis, which attains a minimum at the *median* of the given numbers. LAR regression generalizes the median, which is insensitive to outliers. If the outlier in our set of numbers increases from 99 to a higher value, the mean can increase arbitrarily but the median won't budge.

Of course, the pros and cons of robustness depend on context. Care is required even for the seemingly simple choice among mean, median, and other central-tendency measures for one-dimensional data (see the flowchart on page 185 of Jain⁸). Knowing that LSQ and LAR generalize the mean and median, respectively, doesn't recommend one over the other; it merely explains why they respond differently to outliers.

DRILLING DEEPER

Both experience and formal study are essential to the art of modeling. Textbook study provides a necessary mathematical foundation; practical experience exposes difficulties often overlooked in the classroom. As you learn techniques, consider their uses in four broad categories: exploratory data analysis,¹⁵ descriptive statistics, statistical inference, and prediction.

Bits

Example data and code are available at https://queue.acm.org/downloads/2021/Drill_Bits_O6_example_code.tar.gz. The code uses R's "L1pack" package, which may be installed from the R command prompt.

Drills

1. Shift the data of figure 1 horizontally so that the diurnal cycle is out of phase with a sine wave. Adapt the model-fitting procedure to this change.
2. LSQ yields a unique model fit, but LAR may not. Is this a problem?
3. Compile and run the FORTRAN LAR code in Barrodale and Roberts.²
4. LAR regression is a special case of linear programming (LP), so a general LP solver can compute LAR parameters. What are the advantages of using an LP solver? (Hint: Think about constraints on parameters.)
5. When do LAR and LSQ regression provide maximum-likelihood parameters? When is this important? See Neter, et al.¹⁰ for LSQ and general background, and Bloomfield and Steiger⁴ for LAR.
6. Outliers can occur in both response variables and predictors. How does each arise? Which hurts more in practice? See “leverage points” on pages 417 and 452 of Wilcox.¹⁶
7. Count the many names of LAR regression: “least absolute deviations,” “ L_1 ,” etc. (see list on pages 33–34 of Bloomfield and Steiger⁴). For bonus points, invent a new name and use it in a publication.

Acknowledgments

Charlotte Zhuang, data scientist Alex Zhang, and *Programming Pearls* author Jon Bentley reviewed drafts and provided helpful feedback. Weiwei Gu read a draft and tested the example code.

Machine Learning, Statistics, And Damned Lies

Machine learning (ML) and statistics share tools and methods but differ in terms of their emphases, objectives, and historical development. Fitting models to data via regression was originally unrelated to either field, having been proposed much earlier: least squares in the early 1800s and LAR circa 1757.

Modern statistics emerged in the late 19th and early 20th centuries. Statistics emphasizes inferring properties of large unknown populations from small known samples; its mathematical style resembles probability theory turned backwards. When undertaken in the spirit of statistics, regression modeling often employs expert-crafted model equations that are easy to justify in terms of the problem domain. In the best case, a model equation simply re-states settled science; if most data points can be fitted closely to such an equation, remaining data points with large residuals may correspond to qualitatively anomalous phenomena.

In statistics, validating fitted models emphasizes testing hypotheses about the parameter values. For example, if regression yields a nonzero parameter value, validation tests the null hypothesis that the “true value” is zero (the contrary regression result being a chance artifact of sampling). The reasoning behind such tests often depends on detailed statistical assumptions about the origins of the sample data—assumptions that are sometimes implausible or difficult to check. Practitioners may restrict themselves to purely *descriptive* statistics if inference is unnecessary⁷ or if its underlying assumptions don’t hold.

ML appeared in the mid-20th century. It often strives to develop predictive models by discovering generalizable patterns in data.⁵ Where predictive accuracy is the paramount goal, parsimony, clarity, and scientific justifiability may suffer; ML models can be unwieldy black boxes

disconnected from domain knowledge. When used in the spirit of ML, regression might occupy an inner loop within an outer loop that explores a space of candidate model equations. Validating “trained” (fitted) models emphasizes testing whether predictive accuracy generalizes beyond training data.

All models are wrong and some are harmful. Historically, the development of modern statistics was deeply intertwined with the eugenics movement and other misanthropic pseudoscience.^{3,6} Given its pervasive use, today’s data science has comparable potential for harm.¹¹

References

1. Barrodale, I., Roberts, F.D.K. 1973. An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis* 10(5), 839–848.
2. Barrodale, I., Roberts, F.D.K. 1974. Algorithm 478: solution of an overdetermined system of equations in the l_1 norm. *Communications of the ACM*, 17(6), 319–320; <https://dl.acm.org/doi/10.1145/355616.361024>.
3. Black, E. 2003. *War Against the Weak*. Four Walls Eight Windows.
4. Bloomfield, P., Steiger, W.L. 1983. *Least Absolute Deviations: Theory, Applications, and Algorithms*. Birkhäuser.
5. Bzdok, D., Altman, N., Krzywinski, M. 2018. Statistics versus machine learning. *Nature Methods* 15(4), 233–234; <https://www.nature.com/articles/nmeth.4642>.
6. Gould, S.J. 1996. *The Mismeasure of Man*. W.W. Norton and Company.
7. Hartmann, H. 2016. Statistics for engineers. *Queue*, 14(1), 23–52; <https://dl.acm.org/doi/10.1145/2857274.2903468>.

8. Jain, R. 1991. *The Art of Computer Systems Performance Analysis*. John Wiley and Sons.
9. Narula, S.C. 1987. The minimum sum of absolute errors regression. *Journal of Quality Technology* 19(1), 37–45; <https://www.tandfonline.com/doi/abs/10.1080/00224065.1987.11979031>.
10. Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. 1996. *Applied Linear Statistical Models*, fourth edition. McGraw-Hill.
11. O'Neil, C. 2016. *Weapons of Math Destruction*. Crown Books.
12. Schmuller, J. 2017. *Statistical Analysis with R*. John Wiley and Sons.
13. Stewart, C., Kelly, T., Zhang, A. 2007. Exploiting nonstationarity for performance prediction. In *Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems*, 31–44; <https://dl.acm.org/doi/10.1145/1272996.1273002>.
14. Stigler, S.M. 1984. Boscovich, Simpson and a 1760 manuscript note on fitting a linear relation. *Biometrika* 71(3), 615–20; <https://academic.oup.com/biomet/article-abstract/71/3/615/258808>.
15. Tukey, J.W. 1977. *Exploratory Data Analysis*. Pearson.
16. Wilcox, R.R. 2005. *Introduction to Robust Estimation and Hypothesis Testing*, second edition. Elsevier.

Terence Kelly (tpkelly@acm.org) is easily distracted by outliers.

Copyright © 2021 held by owner/author. Publication rights licensed to ACM.