10 Modern Statistical Concepts Discovered by Data Scientists

Posted by Dr. Vincent Granville on February 19, 2015 at 7:00pm View Blog

You sometimes hear from some old-fashioned statisticians that data scientists know nothing about statistics, and that they - the statisticians - know everything. Here we prove that actually it is the exact opposite: data science has its own core of statistical science research, in addition to data plumbing, statistical API's, and business / competitive intelligence research. Here we highlight 11 major data science contributions to statistical science. I am not aware of any statistical science contribution to data science, but if you know one, you are welcome to share.

"Data don't make any sense,
we will have to resort to statistics."

Here's the list:

1. Clustering using tagging or indexation methods (see section 3 after clicking on the link), allowing you to cluster text (articles, websites) much faster than any traditional statistical technique, with a scalable algorithm very easy to implement
2. Bucketization - the science and art of identifying the right homogeneous data buckets (millions of buckets among billions of observations), to provide highly localized (or segment-targeted) predictions, or to smooth regression parameters across similar buckets, with strong statistical significance. It is equivalent to joint (not sequential) binning in multiple dimensions, which is a combinatorial optimization problem. While decision trees also produce some bucketization, the data science approach is more robust, simple, scalable and model-free. It does not directly produce decision trees, and lead to easy

interpretation (each data bucket corresponding to a specific type of fraud, in a fraud detection problem). A related problem is bucket clustering, via standard hierarchical clustering techniques.

3. Random number generation, a 3,000 year old problem, benefited from data science advances: for instance, using the digits of irrational numbers such as Pi or SQRT(2), produced with very fast algorithms, to simulate randomness.
4. Model-free confidence intervals, getting rid of *p*-value, hypothesis testing, asymptotic analysis, errors due to poor model-fitting or outliers, and of a bunch of obscure statistical old-fashioned concepts
5. Variable / feature selection and data reduction, without using L2-based, model-based techniques such as PCA, potentially numerically unstable, which are sensitive to outliers, and lead to difficult interpretation
6. Hidden decision trees, an hybrid technique combining some sort of averaged decision trees and Jackknife regression, more accurate, and far easier to code, implement, and interpret than either logistic regression or traditional decision trees. Not subject to over-fitting, unlike its ancestor statistical techniques.
7. Jackknife regression, a universal, simplified regression technique, easy to code and to integrate in black-box analytical products. Traditional statistical science offers hundreds of regression techniques, nobody but statisticians know which one to use, and when, obviously a nightmare in production environments.
8. Predictive power and other synthetic metrics designed for robustness rather than for mathematical elegance
9. Identification of true signal in data subject to the curse of big data (spurious correlations)
10. New data visualization techniques - in particular using data video to display insights
11. Better goodness-of-fit and yield metrics, based on robust L1 rather than outlier-sensitive L2 metrics.

All this research is available for free.

Additional Reading

- Data Scientist Reveals his Growth Hacking Techniques
- Top data science keywords on DSC
- 4 easy steps to becoming a data scientist
- 13 New Trends in Big Data and Data Science
- 22 tips for better data science
- Data Science Compared to 16 Analytic Disciplines
- How to detect spurious correlations, and how to find the real ones
- 17 short tutorials all data scientists should read (and practice)
- 10 types of data scientists
- 66 job interview questions for data scientists
- High versus low-level data science

Follow us on Twitter: @DataScienceCtrl | @AnalyticBridge

Source: www.datasciencecentral.com/profiles/blogs/10-modern-statistical-concepts-discovered-by-data-scientists