One Statistician's Perspectives on Statistics and "Big Data" Analytics

Some (Ultimately Unsurprising) Lessons Learned

Prof. Stephen Vardeman
IMSE & Statistics Departments

Iowa State University

July 2014

My Background in "Modern Analytics"/"Statistical Learning"

- Teaching PhD and MS Courses in the Area[1]
  - 3 iterations of a 3-credit PhD-level course since Spring 09
  - 1 1-credit summer course beyond 1st 2 versions of the PhD course
  - 1st iteration of a 3-credit MS-level shared course (with Morris and Wu) Spring 14
- Substantial (400-500 hours of) External Work with Corporate Advanced Analytics Groups
- Personal Participation on (Mixed Participant–HS Through PhD) Predictive Analytics Teams at ISU
  - Netflix Challenge
  - Several `kaggle` Contests ("compete as a data scientist for fortune, fame, and fun" — *AND you learn stuff!*)
- Observation of Successful Student Teams

[1] Typed notes and .pdf slides avaialbe at http://www.analyticsiowa.com/course-materials/modern-multivariate-statistical-learning/

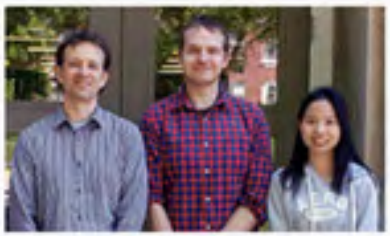
Some (Indirect) Evidence of Real (System) Effectiveness

2013 and 2014 Prudsys AG Data Mining Cup Student Team Performance

5th Place International Finish (99 Entrants) in 2013[2] and **1st Place** International Finish (98 Entrants–28 Countries) in 2014[3][4] by ISU Graduate Student Teams Composed Mostly of Statistics Students

Iowa State's Data-Mining team

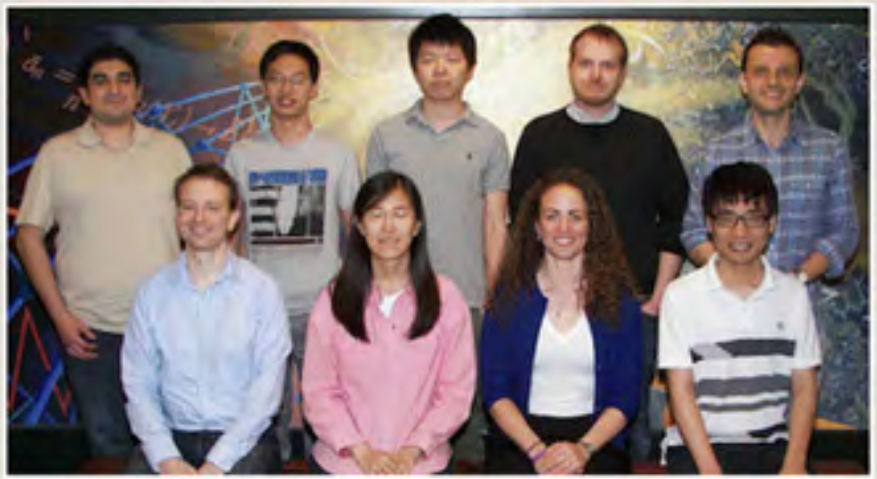
Left to right, Wen Zhou, Jia Liu and Wei Zhang.



Left to right, Cory Lanker, Ian Mouzon and Fangfang Liu.

Iowa State is first U.S. student team to win international data mining competition

Posted Jul 10, 2014 11:46 am



Iowa State data mining team wins international competition.

[2] http://www.news.iastate.edu/news/2013/07/25/datamining
[3] http://www.news.iastate.edu/news/2014/07/10/data-miners
[4] http://www.data-mining-cup.de/dmc-wettbewerb/preistraeger/

Generalities About "Big Data Analytics"

- There is Nothing New Under the Sun ... at a Fundamental Level There Really Are NO Surprises
- Statistical Theory Still Matters
- Linear Models and Linear Algebra Still Matter
- (Like Most Things) Most People Will Need to "Do It" to Learn It
- Learning to "Do It" is Perhaps Harder than Learning "Small Data" Statistics–But Only Because of the Scale Involved
- Practice *Does* Raise Interesting Research Questions (That Can Have Some Emphases Different From Those Statisticians Are Used To)

Some More Specific Topics for Discussion Here

- Up-Front Work
- Data Handling and Handlers
- Analytical Computing Systems and Practices
- The Relevance of Decision Theory and Linear Models
- The (New But Limited) Relevance of Non-Parametric Methods (Smoothing, etc.) and Highly Flexible Parametric Methods
- Increased Attention to Model Bias and Over-Fit in Predictive Analytics
- The Role of "Ensembles"
- The Power of Statistical Theory and First Principles in Practical Data Analyses
- The Necessity of Demystification/Clear Thinking

Up-Front Work

Before

$$\text{Cases} \quad \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \text{Variables} \\ \begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{array} \end{array} \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \\ \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_N \end{array} \end{array}$$

typically comes a *huge* amount of work.

- *Assembling* Data from *Disparate* Sources
- Useful *Definition* of x's and Their *Computation*
  - (just as in small data contexts) this can't be profitably done "synthetically"
  - 80% of corporate project hours and 2/3 of student team project hours
  - serious data handling and record keeping issues must be faced

Data Handlers and Data Handling

- Particularly in Light of the Up-Front Matters, in an Efficient Corporate "Production Mode" a Set of Core Data Analysts
  - is supported by at least an equal number of "data techs" who function more or less like scientific "lab techs"
  - is supported by a good-sized dedicated IT operation (personnel, hardware, and appropriate software)
  - interfaces with a much larger set of less technical business analysts/data scientists placed in business units
- Currently, Much of Our Favorite Software Can Fail to Be Adequate to "Simple" Up-Front Processing
  - R versus MatLab
  - R versus Python
  - need for Perl, SQL

Analytical Computing Systems and Practices

- For Some Kinds of "Big Data" Predictive Modeling, R Can Be Perfectly Adequate
  - it's usually p rather than N that is a problem (if there is one)
  - MatLab and Python can work where R does not
- For Production-Scale (Time Series Type) Corporate Forecasting, SAS FS is Currently the Only Game in Town and Serious IT Support is Essential
- Attention to Standard (CS-type) Code-Sharing and Dataset-Sharing, Versioning and Communication Practice–and Corresponding Resources–are Essential

The Relevance of Decision Theory and Linear Models

- (After Once Passing Quite Out of Vogue During My Professional Lifetime) Decision Theory is Back
  - loss, risk, optimality are the right constructs for guiding analytics ... and modern problem size makes it possible to really use them
  - it tells what is optimal and provides a framework for thinking clearly about what are obstacles ... for example, it promises that in general in predictive analytics

$$\text{Err} = \text{minimum expected loss possible} + \text{modeling penalty} + \text{fitting penalty}$$

  (that among other things puts realistic limits on what is possible)
- Directly or Indirectly Linear Theory Remains at the Center of Methodological Progress and Practice
  - real understanding of everything in modern analytics, from shrinkage to smoothing to ensembles to kernels, turns on linear theory

The New but Limited Relevance of Non-Parametric Methods – And Highly Flexible Parametric Methods

- Splines and Kernel Smoothers, etc. Begin to Look Promising When N is Big
  - big p and the curse of dimensionality dampens enthusiasm
  - some clever piecing together of low-dimensional smoothers (e.g. additive models) must be adopted to progress
- Tree-based Methods Become Effective ... But (*EVEN IN RF Form*) Are No Silver Bullet
- Neural (and Other) "Networks" Are Highly Flexible And Could Be Alternative Smoothers, But Ad Hoc Fitting Makes Efficacy Hard to Assess

Necessary Increased Attention to Model Bias and Over-Fit in Predictive Analytics

- Small Data Statistical Practice Has Typically Focused on "Quality of Fit" For Fixed Predictors–"Fitting Bias and Variance"
- Big Data Statistical Practice Typically Pays Far More Attention to "Model Quality/Class"–"Model Bias"
- Flexibility of a Prediction Method *MUST* Be Balanced Against Propensity for Over-Fit
  - some version of a test set external to a training *must* be used to assess fit
  - cross-validation (CVE) and bootstrapping (OOBE) are the only real options

The Role of "Ensembles"

- For SEL (Unless One Predictor *IS* the Conditional Mean Function) $\exists$ A Linear Combination of Two Predictors That Beats Both
  - this motivates fitting a number of predictors and looking for good linear combinations
  - another version of this is "boosting" which for SEL amounts to successively fitting to residuals and correcting a current predictor by some part of the fit to residuals

BUT NOTICE THAT ONE IS ONLY ATTEMPTING TO BETTER APPROXIMATE THE CONDITIONAL MEAN FUNCTION! (There is NO Magic Here)

- For Classification the Combining of Predictors Might Be Accomplished By Combining "Voting Functions" or "Votes"
  - the former creates a set of classifiers bigger than either produced by two constituent voting functions, and thus again allows for improved classification
  - the latter is less defensible but is often effective

BUT AGAIN, THERE IS NO MAGIC HERE

The Power of Statistical Theory and First Principles in Practical Data Analyses

Basic Statistical Theory Wisely Used Beats Ad Hoc Methodology Every Day

EXAMPLE: A Post-Mortem on the DMC (Classification Problem) Win Strongly Suggests that the Margin of Victory Was Theory-Driven Real Variable Encoding of Categorical Information.

Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ is a set of categorical predictors. What real function(s) of $\mathbf{x}$ can serve as practical numerical predictors? N-P Theory and/or sufficiency considerations point at (log) likelihood ratios–that can (up to some nontrivial number of cells) be estimated as essentially ratios of cell counts

$$\ln \frac{\hat{f}_1(\mathbf{x})}{\hat{f}_0(\mathbf{x})} = \ln \left[\left(\frac{N_0}{N_1} \right) \frac{N_1(\mathbf{x}) + .5}{N_0(\mathbf{x}) + .5} \right]$$

This is MUCH preferable in practice to a set of cell indicators.

The Necessity of Demystification/Clear Thinking

Silliness abounds. Folklore in machine learning is sometimes given more credence than careful thought.

One clear example is the widely held belief that "majority voting by independent classifiers improves error rates." That's just not necessarily true, and obscures the fact that a vector of K classifiers in a 2-class problem takes values in $\{0,1\}^K$ and that (again) the minimum risk function of the vector is a NP test (based on the likelihood ratio).

Another example are persistent claims that some particular methodology (packaged by person or group X) is universally best. Common sense (and "no free lunch" theorems if one really must appeal to them) says that's silly.

In The End

Modern "Big Data Analytics" is fun and a realm where the discipline of Statistics has much to offer.

Thanks

Thanks for your attention!