Statistical modeling: The two cultures

STK9200

Dennis Christensen

University of Oslo

August 31, 2021

Outline of presentation

1. Brief background
2. Summary of article
3. Discussion
   - Personal beliefs
   - What has happened since 2001?

Background

CP Snow: The Two Cultures (1959) [1]

"A good many times I have been present at gatherings of people who, by the standards of the traditional culture, are thought highly educated and who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the Second Law of Thermodynamics. The response was cold: it was also negative. Yet I was asking something which is the scientific equivalent of: Have you read a work of Shakespeare's? I now believe that if I had asked an even simpler question – such as, What do you mean by mass, or acceleration, which is the scientific equivalent of saying, Can you read? – not more than one in ten of the highly educated would have felt that I was speaking the same language. So the great edifice of modern physics goes up, and the majority of the cleverest people in the western world have about as much insight into it as their neolithic ancestors would have had."

Background

- Leo Breiman
  - 1928 - 2005
  - Statistician at University of California, Berkeley
  - Most known for: Decision tree algorithms; bagging; random forests
- Statistical modeling: The two cultures [2]
  - Published in *Statistical science*, 2001
  - Separates statistics into the *data modeling culture* and the *algorithmic modeling culture*
  - Estimates that 98% of statisticians belong to the former
  - Argues strongly in favour of the latter

The two cultures

Data modeling culture

- Assumes stochastic data model
- Linear regression, logistic regression, Cox model
- Parameters of the model are estimated from the data
- Model validation: yes-no using goodness-of-fit tests

Algorithmic modeling culture

- Assumes *black box* data model
- Estimates black box model by a function f
- Neural networks, decision trees
- Model validation: Predictive accuracy

Projects in consulting

The ozone project

Given over 450 meteorological variables x, construct a function $f(x)$ which predicts ozone levels y accurately 12 hours in advance. Solved using high-dimensional feature set, linear regression and variable selection.

The chlorine project

Given the mass spectrum x of a chemical compound, construct a classifier $f(x)$ which predicts whether the compound contains chlorine. Solved using decision trees.

Takeaways

- Focus on finding a good solution
- Explore data before modeling
- Use predictive accuracy on test sets as criterion for model quality
- Computers are an indispensable partner

Problems with data modeling culture

Breiman argues that the focus in the statistical community on data models has

- Led to irrelevant theory and questionable scientific conclusions
- Kept statisticians from using more suitable algorithmic models
- Prevented statisticians from working on exciting new problems

Return to the university

Annals of Statistics, JASA: "Assume that the data are generated by the following model: ..."

- The conclusions are about the model's mechanisms, and not about nature's mechanism
- If the model is a poor emulation of nature, the conclusions may be wrong

Standard tools for model verification are not sufficiently accurate

- Example: Goodness-of-fit tests will too often accept linear models for high-dimensional data [3]
- Example: Residual plots do not uncover lack of fit in more than four to five dimensions (Cleveland, 1993)

Breiman: The question of how well the model fits the data is of secondary importance compared to the construction of an ingenious stochastic model.

Return to the university

Multiplicity of data models

- Data modeling produces a simple and understandable picture of the relationship between the input variables and responses...
- ...but there are usually several good models
- Which one should we choose?

Limitations of data models

- Parametric models can rarely describe complex generative processes in nature
- "If all a man has is a hammer, then every problem looks like a nail"
- The insistence on the use of parametric models limits the range of problems which can be solved

Lessons from algorithmic modeling

1. Rashomon and multiplicity
   - Multiple models may provide good predictions for a given data set
   - Loosening the focus on explainability allows for the combination of good models, rather than variable deletion
   - Bagging, boosting
2. Occam and simplicity vs. accuracy
   - Usually, accuracy and simplicity are in conflict
   - Decision trees easily explainable, but less accurate
   - Random forests highly accurate, but less explainable

Lessons from algorithmic modeling

3. Bellmann and the curse of dimensionality
   - Data modeling culture: High-dimensional input should be avoided
   - Fewer variables provide explainability
   - Overfitting
   - Algorithmic modeling culture: High-dimensional input is desirable
   - More flexible models
   - Examples: Neural networks, support vector machines
   - Develop techniques to avoid overfitting (implicit/explicit regularisation, drop-out, bagging)

Information from a black box

Example: Variable importance in a survival data set

Survival or nonsurvival of 155 hepatitis patients with 19 covariates

- Logistic regression
  - Predictive error: 17.4%
  - Variable selection: Covariates 7 and 11 most important
  - Logistic regression with only these covariates: error increases to 22.9%
  - Worse than declaring everyone a survivor! Error rate: 20.6%
  - Many different "good" models
- Random forests
  - Predictive error: 12.3%
  - Variable selection performed empirically by permuting each variable randomly
  - Variable selection: Covariates 12 and 17 most important
  - Random forests with either (or both) these covariates: error increases to 14.3%
  - Conclude: Virtually all predictive capability is provided by a single variable, either 12 or 17

Information from a black box

Breiman: The goal is not interpretability, but accurate information.

In this regard, much information is available from an algorithmic model.

Comments from Sir David Cox

- The term “In applications” is to broad - different applications require different methods
- Analysis is guided by a question or a hypothesis rather than data alone
- Some models are useful even though they have little or no predictive power
- Most prediction tasks require some knowledge of the underlying processes generating the data
- Real-data examples are usually included to illustrate the theory, not as the main focus
- Inference tools are to be used as guidelines, but not in a mechanical way

Comments from Brad Efron

- A black box with lots of knobs to twiddle is less preferable than a scientific model
- Breiman's oppenness to new ideas whatever their source is admirable
- New methods always look better than old ones
- Breiman overstates the importance of and lack of interest in prediction
- Algorithmic modeling may be viewed as a third front in the traditional frequentist/Bayesian war

Breiman's response

- To Sir David Cox
  - Algorithmic modeling is not meant to replace data modeling
  - Analysis is guided by *a question about data*
  - Cox ignores major breakthroughs in algorithmic modeling in complex prediction tasks
  - Data modeling is indeed applied in a mechanical way in standard statistics journals
- To Brad Efron
  - Many black box approaches require only a handful of parameters (random forests, support vector machines)
  - Neural networks are admired for their demonstrates usefulness, not their hype
  - Regarding the importance of prediction: Predictive ability is a better tool for evaluating whether a model fits the data

Discussion

Looking back 20 years later: Breiman won

- Deep learning completely dominates the fields of prediction and classification [4, 5, 6, 7]
- A recently educated statistician will almost certainly encounter machine learning
- Machine learning has become a tool for probabilistic inference (deep generative models, contrastive learning, GANs) [8, 5, 9]
- Algorithmic modeling can provide insight into other scientific fields (quantum chemistry, pure mathematics) [10, 11]
- Predictive accuracy has become the main tool for model evaluation, even for data models
- Data models are sometimes used for parameter-tuning in algorithmic models [12, 13]

References I

[1] Charles Percy Snow. *The two cultures*. Cambridge University Press, 1959.

[2] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[3] Peter J Bickel, Ya'acov Ritov, and Thomas M Stoker. Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, 34(2):721–741, 2006.

[4] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

References II

[6] Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–4248. PMLR, 2020.

[9] Samuel Wiqvist, Jes Frellsen, and Umberto Picchini. Sequential neural posterior and likelihood approximation. *arXiv preprint arXiv:2102.06522*, 2021.

[10] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1):1–8, 2017.

References III

[11] Adam Zsolt Wagner. Constructions in combinatorics via neural networks. *arXiv preprint arXiv:2104.14516*, 2021.

[12] Jonas Mockus. *Bayesian approach to global optimization: theory and applications*, volume 37. Springer Science & Business Media, 2012.

[13] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.