Estatística, Ciência de Dados & Aprendizagem de Máquina: Transformando dados em informação*

Hedibert Lopes & Paulo Marques

28 de Junho de 2017

Leo Braiman e as duas culturas. Há quase duas décadas, Leo Breiman iniciava os comentários finais do seu artigo *Modelagem estatística: as duas culturas*[1] com o seguinte parágrafo[2]:

> *A meta em Estatística é utilizar dados para predizer e extrair informação sobre o mecanismo de geração de tais dados. Em nenhum lugar está escrito qual tipo de modelo deveria ser usado para problemas envolvendo dados. Para tornar minha posição mais clara, não sou contra modelos para dados propriamente ditos. Em algumas situações eles são a maneira mais apropriada para resolver o problema. Entretanto, a ênfase precisa estar no problema e nos dados.*

Igualmente revelador é o último parágrafo:

> *As raízes da Estatística, assim como da Ciência, residem no trabalho com dados e na checagem da teoria contra os dados. Espero que nesse século nosso campo retorne às suas raízes. Existem sinais de que essa esperança não seja ilusória. Nos últimos 10 anos houve um movimento na direção de trabalhos estatísticos aplicados a problemas do mundo real e também da busca, por parte dos estatísticos, de esforços colaborativos com outras disciplinas. Eu acredito que essa tendência continuará e, de fato, tem que continuar se desejamos sobreviver como um campo energético e criativo.*

Escolhemos começar com esses dois parágrafos para enfatizar dois pontos muito importantes na estatística moderna: i) Devemos focar no problema a ser resolvido/abordado e nos dados coletados durante o processo ou devemos priorizar a modelagem estatística de um suposto processo gerador dos dados? ii) a importância da multi-disciplinaridade na Estatística do século XXI.

*Estas notas refletem nossas experiências e opiniões atuais a respeito dos tema. Portanto, sente-se confortavelmente, reflita e use estas informações por sua conta e risco.

[1] Breiman (2001) Statistical Modeling: The Two Cultures. *Statistical Science*, **16**, 3, 199-231.

[2] Todas as traduções são de nossa inteira responsabilidade.

O papel da Estatística na Ciência de Dados. As dificuldades para abordar a primeira questão estão relacionadas à necessidade do envolvimento de diversos agentes, de diferentes disciplinas, no processo de solução do problema. A Ciência de Dados é a tentativa de abarcar todas as etapas desse processo. Veja-se, por exemplo, a iniciativa da Associação Americana de Estatística (ASA em Inglês) de elaborar um documento apontando *o papel da Estatística na Ciência de Dados*[3]:

> *Embora ainda não se tenha um consenso a respeito do que precisamente se constitue* Ciência de Dados*, três comunidades profissionais, todas dentro de Ciência da Computação e Estatística, estão se estabelecendo como fundamentais: (i)* Gerenciamento de bases de dados *trata da transformação, conglomeração e organização coerente de fontes de dados; (ii)* Estatística e Aprendizagem de Máquina *convertem dados em conhecimento; (iii)* Sistemas distribuídos e paralelos *cuidam da infraestrutura computacional para a execução eficiente da análise de dados.*
>
> *Em nível mais fundamental, vemos Ciência de Dados como uma colaboração mutuamente benéfica entre essas três comunidades profissionais, complementada com interações significativas entre diversas disciplinas relacionadas. Para que a Ciência de Dados atinja completamente seu potencial requer-se colaboração máxima e multifacetada entre esses grupos.*

Nessa mesma linha, David Donoho apresentou em 2015, durante um workshop para celebrar o centenário de John Tukey, o artigo entitulado *50 anos de ciência dos dados*[4]. Ele começa seu manuscrito com o seguinte texto que cita o aclamado artigo de Tukey[5]:

> *Há mais de 50 anos, John Tukey conclamava pela reforma da Estatística na academia. No artigo "O futuro da análise de dados", ele aponta para a existência de uma ciência ainda desconhecida, cujo objeto de interesse era da aprendizagem através de dados, ou 'análise de dados'. Há quase vinte anos, John Chambers, Bill Cleveland e Leo Breiman, independentemente, mais uma vez pediram urgência da Estatística acadêmica na expansão de seus limites para além do domínio clássico da Estatística teórica; Chambers pedia mais ênfase na preparação e apresentação dos dados ao invés da modelagem estatística; enquanto Breiman pedia mais ênfase em predição ao invés de inferência. Cleveland sugeriu o nome "Ciência de Dados" para o campo que ele envisionava.*

Ciência Estatística e Ciência de Dados: para onde vamos a partir daqui? A Professora da Universidade de Toronto, Nancy Reid, uma das mais importantes estatísticas do mundo, ministrou a XXXV Fisher Memorial Lecture[6] na Royal Statistical Society (RSS) em 2016 com o título *Ciência Estatística e Ciência de Dados: Para onde vamos a partir daqui?* Um resumo de sua palestra apareceu na revista StatsLife da RSS[7]

[3] van Dyk, Fuentes, Jordan, Newton, Ray, Lang and Wickham (2015) The role of statistics in data science. `http://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science`

[4] `http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf`

[5] John W. Tukey (1962) The future of data analysis. *The Annals of Mathematical Statistics*, **33**(1), 1-67.

[6] `https://www.statslife.org.uk/events/eventdetail/699/11/data-science-the-view-from-the-mathematical-sciences`

[7] `https://www.statslife.org.uk/features/3072-event-report-the-35th-fisher-memorial-lecture-by-nancy-reid`

> *A pesquisa em Ciências de Dados poderia cobrir coleta e qualidade de dados, base de dados com grande n e pequeno p, tanto quanto bases de dados com pequeno n e grande p. Poderia ainda examinar novos tipos de dados como* networks*, grafos, textos digitais e imagens. Poderia incluir assuntos como limpeza de dados, gerenciamento de dados (isto é, conversão de dados crus em dados "analisáveis"), programação de software, colaboração e gerenciamento de projeto.*
>
> *Nancy espera que a área de Ciência de Dados venha a descobrir que o "núcleo velho" é importante, e que cientistas estatísticos estão geralmente tentando resolver uma vasta gama de problemas além do simples reconhecimento de padrões. Estatísticos são frequentemente criticados por serem muito cautelosos, diz ela. Entretanto, muitas promessas têm sido feitas em torno de* Big Data*. Retornando à sua "*hype curve*", Nancy conclui que a próxima grande novidade a aparecer será "*Smart Data*".*

Estatística e aprendizagem de máquina. Um dos mais prolíficos estatísticos da atualidade, Michael I. Jordan é professor do departamento de Engenharia Elétrica e de Ciência da Computação e do departamento de Estatística da Universidade da Califórnia em Berkeley. Com mestrado em Matemática e doutorado em Ciências Cognitivas, Jordan é membro da Academia Nacional de Ciências dos EUA, da Academia Nacional de Engenharia e da Academia Americana de Artes e Ciências. Além disso, ele é *fellow* de inúmeras associações científicas, entre elas AAAI, ACM, ASA, CSS, IEEE, IMS, ISBA, SIAM e a Associação Americana para o Avanço da Ciência.

Jordan falou recentemente para o blog *reddit Machine Learning*[8] sobre Estatística e Aprendizagem de Máquina enfatizando mais as similaridades das duas áreas do que suas diferenças:

> *Durante os anos 80 e 90, é surpreendente o número de vezes que pessoas trabalhando dentro da comunidade ML (Machine Learning) perceberam que suas idéias já tinham existências pré-históricas dentro Estatística . . . árvores de decisão, mais próximo vizinho, regressão logística, kernels, análise de componentes principais, correlação canônica, modelos gráficos, k-médias e análise discriminante vem à memória, e também muitos princípios metodológicos gerais (por exemplo, método dos momentos, métodos de inferência Bayesiana, bootstrap, validação cruzada, EM, ROC, e gradiente descendente estocástico), e muitas outras ferramentas teóricas (grandes desvios, desigualdades de concentração, processos empíricos, Bernstein-von Mises, estatística U, etc).*

Ele continua com uma provocação:

> *Quando Leo Breiman desenvolveu florestas aleatórias, ele era um estatístico ou um* machine learner*? Quando eu e meus colegas desenvolvemos a* latent Dirichlet allocation*, éramos estatísticos ou* machine learners*? São Máquina de Vetor de Suporte (SVM) e* boosting *aprendizagem de máquina, enquanto regressão logística é estatística, embora essencialmente resolvam os mesmos problemas de otimização?*

Finalmente, ele sugere que ambas comunidades, Estatística e de Aprendizagem de Máquina, se beneficiam e se beneficiarão crescentemente através de suas interações.

[8] `https://www.reddit.com/r/MachineLearning/comments/2fxi6v/ama_michael_i_jordan`

> *Acredito que a comunidade de ML tem sido extremamente criativa ao combinar idéias existentes de vários campos e misturá-las para resolver problemas em domínios emergentes, e também acredito que a comunidade se destacou fazendo uso criativo de novas arquiteturas de computação. Vejo tudo isso como a emergência de uma contra-partida, por parte da Engenharia, das investigações puramente teóricas que aconteceram classicamente dentro da Estatística e da Otimização. Entretanto, não se deve equacionar Estatística e Otimização à Aprendizagem de Máquina e suas aplicações. A "comunidade estatística" tem sido também bastante aplicada, mas, por razões históricas, suas colaborações tendem a se focar mais na Ciência, Medicina e decisão do que na Engenharia. O surgimento da "comunidade de ML" ajudou a expandir o escopo da "Inferência Estatística aplicada". Começou-se a quebra de algumas barreiras entre pensamento de engenharia (engineering thinking) e o pensamento inferencial (inferential thinking).*

Outras discussões acerca das interações entre Estatística, Ciência de Dados, Aprendizagem de Máquina, Mineração de Dados, *Big Data* e temas afins são listadas abaixo (textos em vermelho refletem *hyperlinks*):

1. John Tukey (1962) The future of data analysis
2. David Hand (2013) Data mining: statistics and more?
3. Marie Davidian (2013) Aren't we data science?
4. Hal Varian (2014) Big data: new tricks for econometrics
5. Einav and Levin (2014) Economics in the age of big data
6. Athey and Imbens (2015) Lectures on machine learning
7. David Donoho (2015) 50 years of data science
8. Peter Diggle (2015) Statistics: a data science for the 21st century
9. van Dyk *et al.* (2015) Role of statistics in data science
10. Francis Diebold (2016) Machine learning versus econometrics
11. Uchicago (2016) Machine learning: what's in it for economics?
12. Coveney, Dougherty, Highfield (2016) Big data need big theory too
13. Franke *et al.* (2016) Statistical inference, learning and models in big data

Veja também discussões que aparecem na AmStat News da ASA:

1. Davidian (1 jul 2013) Aren't we data science?
2. Bartlett (1 oct 2013) We are data science
3. Matloff (1 nov 2014) Statistics losing ground to computer science
4. van Dyk *et al.* (1 oct 2015) Role of statistics in data science
5. Jones (1 nov 2015) The identity of statistics in data science
6. Priestley (1 jan 2016) Data science: the evolution or the extinction of statistics?
7. See also Press (28 may 2013) A very short history of data science

Estatística Bayesiana moderna. Terminamos listando algumas publicações científicas do primeiro autor dessa nota[9], que se encaixam nessa interface entre modelagem Estatística, séria análise de dados e computação científica.

1. Efficient Bayesian inference for multivariate factor SV models
2. Particle learning for Bayesian non-parametric MSSV model
3. Scalable semiparametric inference for the means of heavy-tailed distributions
4. Rational Sunspots
5. Put option implied risk-premia in general equilibrium under recursive preferences
6. On the long run volatility of stocks: time-varying predictive systems
7. Parsimonious Bayesian factor analysis when the number of factors is unknown
8. Parsimony inducing priors for large scale state-space models
9. Bayesian factor model shrinkage for linear IV regression with many instruments
10. Cholesky realized stochastic volatility model
11. Sequential Bayesian learning for stochastic volatility with variance-gamma jumps in returns
12. Particle learning for fat-tailed distributions
13. Online Bayesian learning in dynamic models: An illustrative introduction to particle methods
14. Evaluation and analysis of sequential parameter learning methods in Markov switching SV models
15. Sequential parameter learning and filtering in structured AR models
16. Analysis of exchange rates via multivariate Bayesian factor stochastic volatility models
17. Tracking epidemics with Google Flu Trends data and a state-space SEIR model
18. Measuring vulnerability via spatially hierarchical factor models
19. A semiparametric Bayesian approach to extreme value estimation
20. Bayesian Statistics with a Smile: a Resampling-Sampling Perspective

[9]Todos os artigos podem ser obtidos na página `http://hedibert.org/scientific-papers`