

MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
<http://www.ime.usp.br/~pam>

Aula 1

13 de março de 2023

Sumário

- 1 As Origens
- 2 Inferência Bayesiana
- 3 Inferência Frequentista
- 4 Era Moderna
- 5 Estatística

Paradigma

1. Modelo, padrão a ser seguido.
2. Um pressuposto filosófico, uma teoria, um conhecimento, que origina o estudo de um campo científico.
3. Aquilo que os membros de uma comunidade científica partilham.

Exemplos:

- Inferência Frequentista, Inferência Bayesiana
- Data Mining, Neural Networks, Data Science
- Statistical Learning, Machine Learning

Aprendizado com Estatística

- **Aprendizado com Estatística (AE)/Statistical Learning (SL)**: nomenclatura nova, mas a maioria dos conceitos foram desenvolvidos desde o Século 19. Métodos estatísticos para previsão, classificação, análise de agrupamentos etc. Inferência é o objetivo e interpretação é importante.
- **Aprendizado com (ou de) Máquina (AM)/Machine Learning (ML)**: métodos para “aprender” padrões ocultos em dados. Usados para previsão, classificação, reconhecimento de padrões, análise de agrupamentos etc. Pouca atenção à inferência (do ponto de vista computacional) e à interpretabilidade.
- When ML methods are **statistically sound** they are called **Statistical Learning (SL)** methods.
- Nosso foco: Métodos de AE.

Probabilidade

1. Início em 1654 com Fermat (1601-1665), Pascal (1623-1662): jogos de dados
2. Huygens (1629-1695): primeiro livro de probabilidade em 1657.
3. Bayes (1702-1761): primeira versão do Teorema de Bayes, publicado em 1763.

Gauss e Legendre

1. Gauss (1777-1856) inventou o método de mínimos quadrados (MQ) na última década do século 18 (1795) e o usou regularmente depois de 1801 em cálculos astronômicos.
2. Legendre (1752-1833): publicou no apêndice de "Nouvelles Methodes pour la Détermination des Orbites des Còmetes". Nenhuma justificação.
3. Gauss (1809): deu justificativa probabilística do método. Em "The Theory of the Motion of Heavenly Bodies".
4. Implementaram o que é hoje chamado de **regressão linear**.

Bayes ou Laplace?

1. Laplace (1749-1761): desenvolveu o Teorema de Bayes independentemente, publicado em 1774.
2. 1812: *Théorie Analytique des Probabilités*: aplicações científicas e práticas.
3. 1814: *Essais Philosophiques sur les Probabilités*: interpretação Bayesiana das probabilidades
4. Inferência Bayesiana \leftrightarrow Inferência Laplaciana. Usada a partir de 1800.
5. Fisher e Neyman: início do século 20.
6. Jeffrey (1939). *Theory of Probability*. Considerado como o re-início da Inferência Bayesiana
7. de Finetti, Savage, Lindley etc.

Fisher e Neyman

1. Inferência Frequentista (testes de hipóteses, estimação, planejamento de experimentos e amostragem) foi iniciada por R. Fisher(1890-1962) e J. Neyman (1894-1981).
2. Fisher (1925): Statistical Methods for Research Workers. (14^a Edição: 1970)
3. Fisher (1935): The Design of Experiments (8^a Edição: 1966)
4. Fisher (1936): propõe a **análise discriminante linear**.

Gosset/Student

1. W. Gosset (1876-1937): Em 1908 publicou sob o pseudônimo de Student um artigo que iniciou um novo paradigma em "Pequenas Amostras". Resultado provado por Fisher em 1912-1915.
2. Student (1908). The probable error of a mean. *Biometrika*, 6, 1-25.
3. Fisher (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Royal Society, A*, 222, 308-368.
4. Stigler: "The most influential article on Theoretical Statistics in the 20th Century"
5. Hald: "For the first time in the history of Statistics a framework for frequency-based general theory of parametric statistical inference was clearly formulated."

Neyman e Pearson

1. Neyman and Pearson (1933a) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Royal Society*, A, 231, 289-331.
2. Neyman and Pearson (1933b). On the testing of statistical hypothesis in relation to probabilities a priori. *Proc. Cambridge Philos. Society*, 24, 429-510.
3. Livros de E. Lehmann sobre estimação e testes de hipóteses. 1967, 1983.
4. Era do "Small Data".

1940 - 2000

1. 1940: propostas de abordagens alternativas: regressão logística.
2. 1970: Nelder e Wedderburn: generalized linear models para uma classe de métodos que incluem regressão linear e logística como casos especiais.
3. 1970: Efrom: bootstrap; Hoerl e Kennard: ridge regression.
4. até o final de 1970: métodos lineares.
5. 1980: tecnologia computacional possibilita a aplicação de métodos não lineares.
6. 1984: Friedman et al. introduzem CART (Classification and regression trees) e propõem uma implementação prática de método para seleção de modelos, incluindo CV (cross validation)
7. 1986: Hastie e Tibshirani: estendem os MLG pra os modelos GAM (generalized additive models).
8. 1996: Tibshirani: introduz o LASSO; extensões para outros métodos de regularização.

O que é Estatística?

- [1] Coleta de dados: amostras, planejamento de experimentos, estudos observacionais.
- [2] Modelagem e análise de dados.
- [3] Tomada de decisões.

Eras da Estatística

A história da Estatística pode ser dividida em três eras:

(1) A era de Quetelet (astrônomo, matemático, estatístico belga, 1796-1874) e seus sucessores, na qual o objetivo era obter grandes conjuntos de dados (censos) em ciências sociais.

(2) O período clássico de Pearson (1857-1936), Fisher (1890-1962), Neyman (1894-1981), Hotelling (1895-1973) e seus sucessores, que desenvolveram a teoria de inferência ótima; métodos apropriados para *small data sets*.

(3) A era da produção de dados em massa (*big data sets*), com novas tecnologias como *microarrays*, dados de alta frequência em finanças, dados astronômicos etc.

Algoritmo e Inferência

Análise Estatística:

- (a) algoritmica
- (b) inferencial

Exemplo: Considere estimar a média $\mu = E(X)$ de uma v.a. X , definida sobre uma população.

Para uma AAS X_1, \dots, X_n , considere o estimador

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Este é o **algoritmo**.

Quão acurado e preciso é \bar{X} . Esta é a parte da **inferência**.

Algoritmo e Inferência

Algoritmos: é o que os estatísticos fazem.

Inferência: porque os estatísticos usam os algoritmos. (Efrom, 2016).

Conjuntos de dados enormes (*Big Data*) requerem novas metodologias. Esta demanda está sendo atendida por algoritmos estatísticos baseados em computação intensiva.

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C++, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Estatística e Computação

- Avanços em Estatística diretamente relacionados com avanços na área computacional.
- → 1960: máquinas de calcular manuais, elétricas, eletrônicas.
- 1960→ 1980: "grandes computadores": IBM 1620, CDC 360, VAX etc; cartões e discos magnéticos; FORTRAN.
- 1980→: computadores pessoais; supercomputadores; computação paralela; "clouds"; C, C+, S.
- Pacotes estatísticos: S-Plus, SPSS, Minitab etc. Repositório R.
- Era do "Big Data" e da "Data Science".

Referências

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Berlin: Springer.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*. Second Edition. Springer

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P.A. and Singer, J.M. (202). *Estatística e Ciência de Dados*. Rio de Janeiro:LTC.