Aprendizagem estatística para ciências.

Aplicações com software.

Felipe Coelho Argolo

Página intencionalmente deixada em branco.

Versão 0.2: Introdução; Capítulo 0; Capítulo 1. 26 de Dezembro de 2018

Prefácio

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful
George Box & Norman R. Draper, Empirical Model-Building and Response Surfaces

Uma antiga aplicação da matemática é fazer inferências com base em observações de cenários parecidos. Civilizações antigas, como os babilônios, usavam interpolação linear para estimar informações. Fazendo o censo populacional com intervalo de anos, estimavam o valor dos anos não medidos, supondo que eles eram medidas centrais daquelas que ao seu redor. Métodos iterativos também foram usados para aproximar a raiz quadrada de números naturais ($\sqrt{2}$) e números irracionais π.

Essas técnicas deram fruto a abstrações mais gerais, aos campos da estatística e dos métodos numéricos. Em particular, o último século contou com a invenção do computador universal e dos processadores eletrônicos, impulsionando o poder de cálculos vertiginosamente.

O aperfeiçoamento teórico e instrumental trouxe ferramentas mais adequadas para cientistas e também algoritmos mais potentes para aplicações práticas.

Nos últimos anos, o campo ganhou forte notoriedade social e acadêmica em virtude dos resultados inéditos em problemas de predição com aplicação prática. Avanços em processamento de linguagem natural, visão computacional e algoritmos preditivos foram rapidamentes aplicados pela indústria e por pesquisadores.

Uma descrição abrangente pode facilmente alcançar 1,000 páginas de texto sucinto, como o clássico 'Deep Learning (Adaptive Computation and Machine Learning)' de Goodfellow, Bengio and Courville. Outra obra de escopo e tamanho semelhante é a "Neural networks and learning machines", de Simon Haykin.

Objetivos

Este texto oferece uma introdução intuitiva ao campo, contextualizando-o epistemologicamente. O campo de aprendizagem estatística tem definição pouco estabelecida. Abrange aspectos de matemática pura e aplicada. Com uma perspectiva mais geral, a matemática pura desenvolve abstrações básicas, descrevendo o comportamento de números, probabilidades, funções e outras entidades. Veremos que progressos fundamentais foram feitos por nomes como De Moivre, Euler e Gauss.

Em matemática aplicada, especialistas estudam a relação dessas abstrações com fenômenos observáveis. Estas pessoas empregam métodos quantitativos a contextos restritos: por exemplo, James Clerk Maxwell deduziu (1860) a distribuição estatística e velocidade de partículas em um gás ideal, conhecida como distribuição de Maxwell–Boltzmann. Em estatística, veremos a descoberta da distribuição t para as estimativas de uma média por Wlliam Gosset.

São exemplos de campos que fazem uso extenso das ferramentas descritas: neurociências (modelos lineares em fmri), psicometria (análise fatorial), ecologia, biologia molecular (testes estatísticos), ciências clínicas (meta-análises e inferência causal), economia, marketing, algotrading.

Este texto introduz e fornece um guia para aplicações práticas destas ferramentas a fenômenos observáveis. É destinados aos profissionais e pesquisadores trabalhando na fronteira entre matemática aplicada e ciências naturais.

O primeiro capítulo ilustra como o racional hipotético-dedutivo funciona para estudar teorias científicas. Aborda a relação entre ciências empíricas e três abstrações matemáticas: a distribuição normal, a distribuição t e o teorema do limite central. O segundo capítulo aborda correlações e modelos preditivos lineares. Um framework frequencista e linguagem R são usados para demonstrações de exemplos e exercícios.

O terceiro capítulo apresenta um racional diferente para os procedimentos. Usando o conceito de holismo epistemológico (van Quine), reproduzidos as análises anteriores usando inferência bayesiana. Fazemos perguntas diferentes para obter informações de nossos dados. No capítulo quatro, o foco está em modelos classificatórios e na função logística. Usamos R, Stan e um framework bayesiano para modelos simples e hierárquicos. Exploramos o poder das simulações através de Markov Chain Monte Carlo para obter estimativas

difíceis de tratar analiticamente.
O quarto capítulo ilustra o uso de grafos/redes para a construção de modelos preditivos. Os exemplos são de Support Vector Machine e Redes Neurais. Modelos são construídos do zero (from scratch) para ilustrar dois mecanismos importantes de otimização (gradient descent e back propagation).

Sumário

- Reinforcement learning

Pré-requisitos

Para uma leitura fluida do texto, recomenda-se a compreensão de rudimentos em probabilidade, estatística e cálculo (análise real). Os exemplos com ferramentas computacionais (exceto gráficos) usam sintaxe semelhante à matemática apresentada no texto. Assim, baixa familiaridade com linguagens de programação não é uma barreira.

Todos os exemplos podem ser reproduzidos usando software livre.

Leitura recomendada:

Neurociências

- Principles of neural science - Eric Kandel

Matemática pura e programação

- Better Explained (https://betterexplained.com/)
- What is mathematics - Courant & Robbins
- Fundamentos da matemática elementar - Iezzi (Vol. 5)
- MOOCs sobre estatística básica usando R (e.g.: https://www.coursera.org/specializations/statistics)
- Cálculo Diferencial e Integral - Piskounov =)
- http://material.curso-r.com/
- R Graphics Cookbook
- R Inferno
- Learn you a Haskell for Great Good
- Layered Grammar of Graphic - Hadley Wickham.
- The art of computer programming
- Algorithms unlocked
- Portais: statsexchange, stackoverflow, mathexchange, cross-validated.

Machine Learning

- An Introduction to Statistical Learning: with Applications in R
- Neural Networks and Learning Machines - Simon Haykin
- Stanford course on computer vision: http://cs231n.stanford.edu/
- Deep learning at Oxford 2015: (https://www.youtube.com/watch?v=dV80NAlEins&list=PLE6Wd9FR--EfW8dtjAuPoTuPcqmOV53Fu)