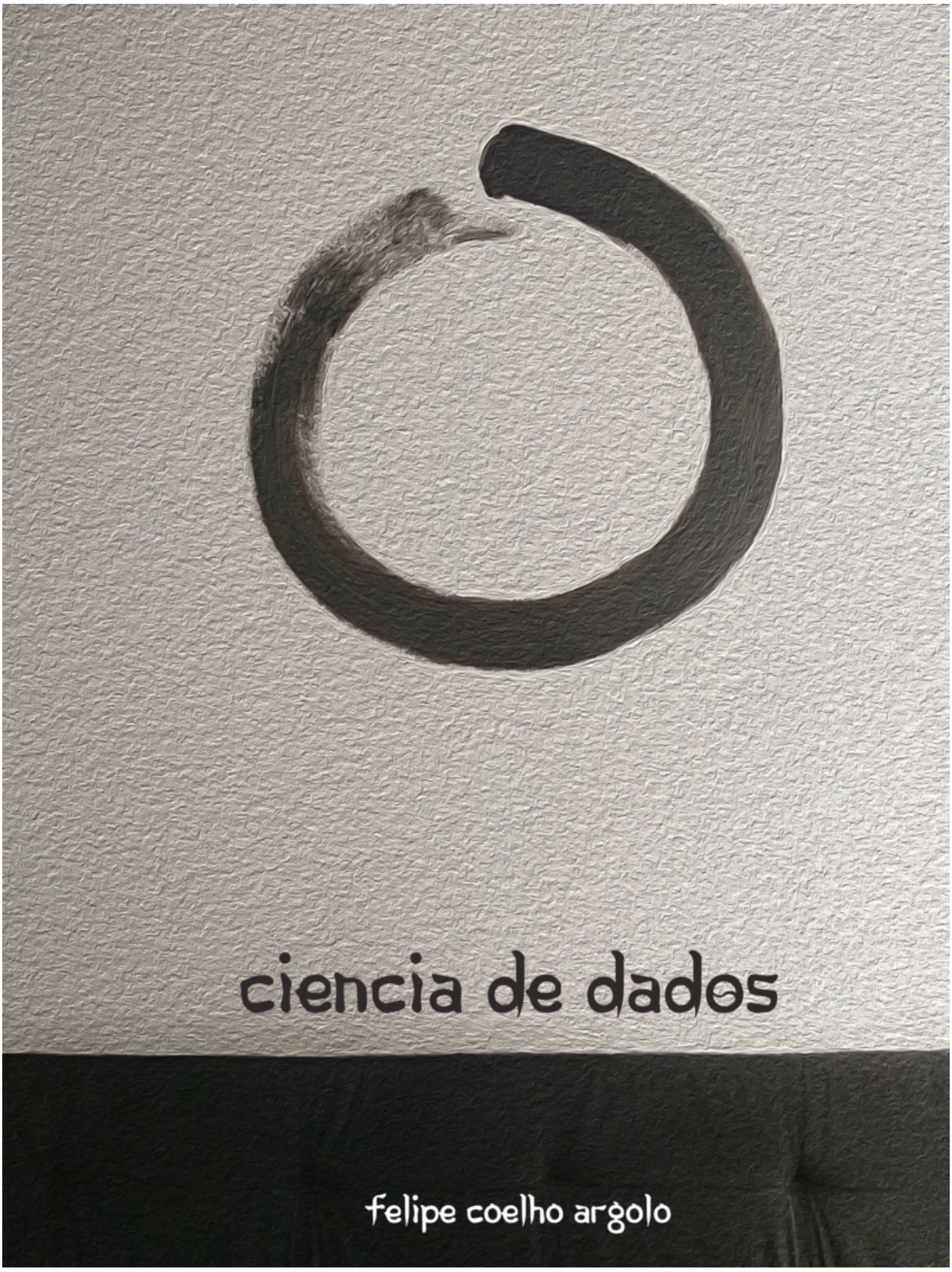
ciencia de dados
felipe coelho argolo

Ciência de dados

Filosofia e aplicações com software

Felipe Coelho Argolo felipe.c.argolo@protonmail.com

São Paulo, 21 de Abril de 2019
Página oficial: https://http://www.leanpub.com/fargolo

Volume 1

Para comentários, críticas, sugestões, ou simplesmente dizer *oi*: felipe.c.argolo@protonmail.com.

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

Nos últimos anos, os termos *inteligência artificial (artificial intelligence)*, *aprendizagem de máquina (machine learning)*, *big data* e *ciência de dados* ganharam forte notoriedade em virtude dos resultados inéditos em problemas de aplicação prática. Avanços em processamento de linguagem natural, visão computacional e algoritmos preditivos foram rapidamentes aplicados em por engenheiros e pesquisadores em finanças, indústria e ciências.
Uma descrição abrangente das técnicas desenvolvidas 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. Inúmeros cursos online e videoaulas são produzidos e disponibilizados por instituições de prestígio (e.g. Curso integral da Oxford em Deep Learning: https://www.youtube.com/watch?v=PlhFWT7vAEw).

No conteúdo disponível, a abordagem costuma ser '*bottom-up*'. Isto é, uma noção do campo é construída através de estudo focado em modelos: cursos específicos para *time series*, *clustering*, *redes neurais* ou ainda ferramentas (e.g. R, Julia, Python, Stan, Matlab. . .). Funciona bem como roteiro natural em cursos de engenharia e ciências exatas.

De cima para baixo
Este texto visita temas num roteiro inverso (*top-down*). Os modelos são contextualizados como ferramentas na exploração de um roteiro com eixo em filosofia das ciências. Assim, as formulações matemática surgem como resposta a exemplos inspirados em fenômenos naturais de biologia (testes estatísticos), psicologia (análise fatorial), saúde pública/economia (correlação, regressão e causalidade) e neurociências (perceptron e redes neurais).

O primeiro capítulo acompanha Darwin estudando tentilhões em Galápagos. Ilustra como o racional hipotético-dedutivo funciona para estudar teorias científicas. O teste t de Student é aplicado para comparação dos bicos das aves. Aborda a relação entre ciências empíricas, o teorema do limite central, a distribuição normal e a distribuição t.

O segundo capítulo destaca o papel descritivo e preditivo de teorias. Além de testar hipóteses, criamos modelos para as relações entre medidas, usando os conceitos de correlações lineares (ρ de Pearson) e tamanho de efeito (*D de Cohen*). Também são introduzidas alternativas não-paramétricas aos procedimentos (ρ de Spearman e teste U de Mann-Whitney). Usamos regressão para fazer predições.

O terceiro capítulo introduz o uso de muitas variáveis (análise multivariada). Grafos são a abstração base para relacionarmos múltiplos conceitos. Estudaremos regressão linear múltipla, colinearidade, mediação e moderação. Conheceremos análise fatorial e sua generalização em equações estruturais: a implementação matemática do abrangente paradigma filosófico para **modelos causais** de Judea Pearl.

O quarto capítulo introduz redes neurais. Começamos da inspiração biológica dos neurônios artificiais. Conheceremos a primeira máquina inteligente da história: o *Mark I Perceptron*. Codificaremos um Mark I virtual do zero (*from scratch*) e observaremos a aprendizagem por *gradient descent*, usando derivativas para minizar erros.

Redes Neurais expandem o poder de um neurônio com múltiplos nodos para a construção de sistemas preditivos complexos. Redes profundas incluem camadas sucessivas, permitindo transformações em sequência para resolver classes mais gerais de problemas. Entenderemos como os neurônios podem propagar erros aos outros, otimizando gradientes em conjunto com o mecanismo de *backpropagation*. Também codificaremos uma rede neural *from scratch*, Mark II.
O quinto capítulo constrasta as duas principais escolas de interpretação da probabilidade: a **frequencista** e a **bayesiana**. O contexto é dado por alternativas ao método hipotético dedutivo: Carnap demonstra a dificuldade de refutações, Feyerabend propõe uma anarquia epistemológica amparada em fatos históricos e W. van Quine pinta um sistema entrelaçado para teorias, hipóteses e observações. Reabordamos alguns

exemplos anteriores usando Stan para inferência bayesiana. Exploramos o poder das simulações através de *Markov Chain Monte Carlo* para obter estimativas difíceis de tratar analiticamente.

Sumário

  - Intuições: prior, likelihood, posterior e probabilidades marginais
  - Comparação de amostras com distribuição normal
  - Correlação linear
- Estimadores e Métodos Markov Chain Monte Carlo
  - Soluções fechadas, Gradient Descent e MCMC

Volume 2 (em construção)
Capítulo 6 - Programação probabilística para contextos gerais

- Processos Gaussianos
- Inferência Bayesiana para cosmologia
- Redes neurais probabilísticas com PyMC

Capítulo 7 - Ambientes desconhecidos

- Aprendizagem não supervisionada
- Redução de dimensões
- Clustering
- Aprendizagem semi-supervisionada
- Reinforcement learning

Capítulo 8 - Tópicos especiais

- Séries temporais
- Sistemas dinâmicos
- Processamento de linguagem natural

Pré-requisitos

Rudimentos em probabilidade, estatística e cálculo são suficientes para compreender quase todos os exemplos. Os programas usam sintaxe semelhante à matemática apresentada no texto então pouca familiaridade com programação não é uma barreira. O capítulo 0 é destinado a isso.

Todos os exemplos podem ser reproduzidos usando software livre.

Leitura recomendada:

Filosofia e divulgação científica

- Surely You're Joking, Mr. Feynman
- O mundo assombrado pelos demônios - Carl Sagan
- A lógica da pesquisa científica - K. Popper
- A estrutura das revoluções científicas - Thomas Kuhn
- Contra o Método - Paul Feyerabend
- Dois dogmas do empiricismo - Willard van Quine
- Stanford Encyclopedia of Philosophy - https://plato.stanford.edu/

Neurociências

- Principles of neural science - Eric Kandel

Matemática/computação

- Coleção '*Fundamentos da matemática elementar*'
- What is mathematics - Courant & Robbins
- Better Explained (https://betterexplained.com/)
- http://material.curso-r.com/
- R Graphics Cookbook
- R Inferno
- Learn you a Haskell for Great Good
- Layered Grammar of Graphics - Hadley Wickham.
- Algorithms unlocked
- Online: Statsexchange, stackoverflow, mathexchange, cross-validated.

Machine Learning

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

Agradecimentos

Minha família, Suzana, Paulo, Isaac e Chris. Amigos Gabriel, Guilherme, Wei.

Aos professores: Carla Daltro, Anibal Neto, Lucas Quarantini, Luis Correia, Rodrigo Bressan, Ary Gadelha.

Aos colegas Fatori, Luccas, Macedo, Walter, Sato, Hiroshi, Lais, Luci, Davi, Oddone, Jones, n3k00n3 (Fernando Pinheiro), userx (victor).