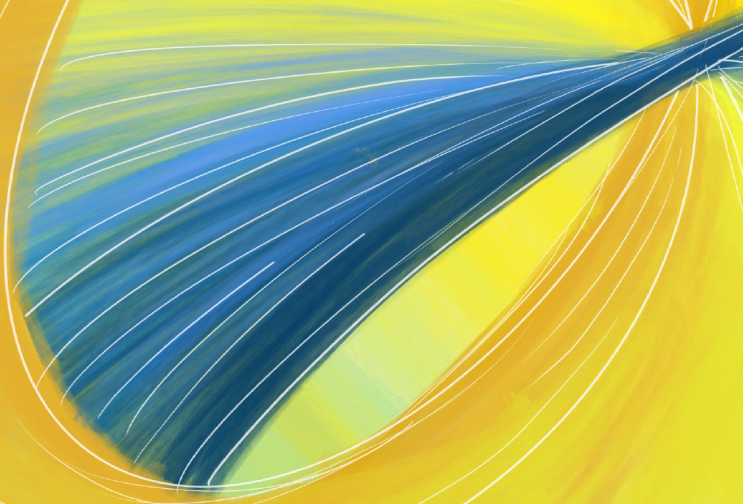
Ciência de dados: algoritmos e aplicações



Luerbio Faria
Fabiano de Souza Oliveira
Paulo Eustáquio Duarte Pinto
Jayme Luiz Szwarcfiter

impa

33º Colóquio Brasileiro de Matemática

Ciência de dados: algoritmos e aplicações

Ciência de dados: algoritmos e aplicações
Primeira impressão, julho de 2021

ISBN 978-65-89124-47-4
MSC (2020) Primary: 68R05, Secondary: 68R99, 05C81, 05C85, 05C99

Coordenação Geral Carolina Araujo

Produção Books in Bytes **Capa** Izabella Freitas & Jack Salvador

Realização da Editora do IMPA
IMPA
Estrada Dona Castorina, 110
Jardim Botânico
22460-320 Rio de Janeiro RJ

www.impa.br
editora@impa.br

Conteúdo

Prefácio

Ciência de Dados é uma área do conhecimento cuja criação é relativamente recente. O interesse por este estudo provém, inicialmente, da quantidade inimaginável de dados que são gerados a cada segundo, por exemplo, na internet e em outros meios de comunicação. Dada a importância dessas informações, e dada a impossibilidade física e lógica de manipular esses dados através dos meios convencionais utilizados até poucos anos atrás, novas técnicas e um novo campo de estudo foram estabelecidos, o que originou a *ciência de dados*. Juntamente com essa nova área foi criada uma nova linha de pesquisa, bastante ampla em sua abrangência. Além disso, abriram-se oportunidades para um novo profissional, denominado *cientista de dados*, cuja procura no mercado de trabalho segue intensa, mesmo no período da presente pandemia.

A ciência de dados poderia ser considerada parte da ciência da computação, mas pela sua imensa aplicação em inúmeras questões do cotidiano e devido ao papel preponderante que a teoria da probabilidade assumiu neste campo de conhecimento, a sua definição como nova área do conhecimento é perfeitamente justificável. Neste aspecto, deve ser ressaltado que ainda não há um consenso geral da abrangência exata da ciência de dados. Assim, dependendo da autoria e da fonte, ciência de dados pode assumir diferentes significados, desde um texto quase puramente matemático, com poucas vinculações diretas relativas a aplicações, até aspectos basicamente humanos de algumas de suas aplicações, sem maiores questões de fundamento. Desta maneira, qualquer texto de ciência de dados reflete a visão de seus autores sobre a matéria. Naturalmente, a presente proposta não poderia deixar de seguir esta linha.

O presente texto se propõe a apresentar um conteúdo de ciência de dados levando em consideração as suas aplicações. Procuramos manter a ênfase algorítmica, ao longo da exposição. Além disso, naturalmente, foi mantida a preocupação pelo rigor matemático necessário para abordar os diversos aspectos da matéria, tanto nas provas matemáticas, quanto na descrição de seu conteúdo.

As partes que compõem este texto de ciência de dados, foram agrupadas em alguns temas, os quais, basicamente compõem os capítulos do presente livro. Os temas englobam conceitos básicos dos seguintes assuntos: *teoria de probabilidades; algoritmos randomizados; análise probabilística de algoritmos; algoritmos de dados massivos; aprendizado de máquina; cadeias de Markov e passeios aleatórios*.

Esses temas são apresentados em uma linguagem acessível a um estudante de graduação, que seria o público alvo desta publicação. Apesar disso, o texto, ou partes do mesmo, podem ser utilizados como material em um curso de pós-graduação.

Gostaríamos de externar os nossas agradecimentos ao Prof. Daniel Ratton, da COPPE/UFRJ, o qual nos sugeriu a área de ciência de dados, através de suas excelentes palestras e conversas particulares.

Agradecemos à Profª Celina Figueiredo, também da COPPE/UFRJ, pela sugestão e incentivo de submeter este texto ao 33º Colóquio Brasileiro de Matemática. Esta submissão certamente veio acelerar a produção do texto.

Não podemos deixar de mencionar a contribuição dos alunos de doutorado Bruno Masquio, Lívia Medeiros, Raquel de Souza, Vitor de Luca, e também daqueles de iniciação científica Antônio de Sousa e Lucas de Carvalho, todos da UERJ, que auxiliaram no processo de revisão e no preparo do curso associado a este texto. Nosso agradecimento a todos eles.

Registramos também o apoio recebido pelo Instituto de Matemática e Estatística da UERJ, instituição de vínculo de seus autores.

Finalmente, agradecemos à Organização do 33º Colóquio Brasileiro de Matemática, bem como à excelência da Comissão Editorial do IMPA, pela cuidadosa revisão e produção do texto.

Rio de Janeiro, Abril 2021
Os Autores

1 Preliminares

1.1 Introdução

Ciência de dados é uma área que se constituiu a relativamente pouco tempo, cujas partes que a compõem ainda não representam um consenso entre seus pesquisadores. Além disso, as partes constituintes ainda se encontram em formação, não estão inteiramente consolidadas.

No presente texto, descrevemos alguns dos principais aspectos da ciência de dados, levando em consideração as suas aplicações, bem como questões de fundamento. É dada ênfase especial aos seus algoritmos, aos métodos gerais dos problemas motivados pelas suas aplicações. Procuramos manter a ênfase algorítmica ao longo do texto, bem como o rigor necessário para descrever os aspectos matemáticos envolvidos.

Os tópicos selecionados foram divididos em capítulos, cujos conteúdos descrevemos a seguir. Uma lista de exercícios e bibliografia comentada encerram cada um dos capítulos.

1.1.1 Introdução à Probabilidade

Nas questões suscitadas em ciência de dados, a probabilidade e a estatística, de um modo geral, assumem um papel preponderante. De fato, não é possível realizar um estudo apropriado desta área sem conhecimentos de probabilidade. Assim sendo, o texto se inicia com um tratamento básico desta matéria, apresentando os principais conceitos e resultados que serão utilizados ao longo do texto. Quando pertinente, as provas dos resultados são incluídas. O texto contém o detalhamento de alguns dos conceitos e propriedades básicos de probabilidade, como probabilidade condicional, independência de eventos, variáveis aleatórias, entre outras. O estudo de alguns tipos de variáveis aleatórias é incluído naquele capítulo com certa ênfase, pois essas variáveis são utilizadas ao longo de todo o livro.

1.1.2 Algoritmos Randomizados

Um dos focos principais do livro é a descrição de algoritmos para resolver os diversos problemas motivados pelo estudo da ciência de dados. Uma grande quantidade desses algoritmos são randomizados e não determinísticos. Assim, é necessário um estudo preliminar desta classe de algoritmos. O estudo se inicia com uma discussão sobre o significado da randomização. Em seguida são abordados os tipos frequentemente utilizados de algoritmos randomizados: os algoritmos de Monte Carlo e de Las Vegas. O estudo é ilustrado por meio da análise de problemas específicos. Para o algoritmo de Monte Carlo, são tratados os problemas da seleção do elemento majoritário de um conjunto; a determinação do corte mínimo de arestas de um grafo; a formulação de um teste de primalidade, entre outros. Em relação ao algoritmo de Las Vegas, são tratados o problema do elemento unitário de um conjunto e o método de Quicksort para ordenação.

1.1.3 Análise Probabilística de Algoritmos

Há um certo número de problemas e algoritmos para os quais a análise de pior caso não é uma abordagem prática ou tão relevante. Um dos exemplos é o método *simplex* para programação linear, cuja complexidade de tempo de pior caso é exponencial, mas que, em termos práticos, obtém resultados quase lineares. Outro exemplo de destaque é o Quicksort, cuja complexidade de pior caso é $O(n^2)$, sendo, na prática, usualmente o algoritmo utilizado, por ser tipicamente de execução muito rápida, superando os algoritmos de complexidade de pior caso $O(n \log n)$. Tais situações podem ser explicadas pelo fato de que entradas "ruins" são de ocorrência

rara. Os bons resultados, quando existentes, são evidenciados pelo estudo da complexidade "típica" do algoritmo, ou seja, a complexidade de caso médio, obtida normalmente por análises probabilísticas. Abordaremos as análises probabilísticas de caso médio para vários algoritmos básicos tais como a busca linear em vetor, busca binária em vetor ordenado e Quicksort.

1.1.4 Algoritmos para Dados Massivos

No estudo clássico da complexidade de algoritmos, um algoritmo é dito eficiente se sua complexidade de tempo é polinomial no tamanho da entrada. Como cada célula de memória alocada por um algoritmo é normalmente inicializada, a complexidade de tempo de um algoritmo é maior ou igual àquela de espaço. Desta maneira, se um algoritmo é eficiente, sua complexidade de espaço também é polinomial.

Tal definição de eficiência, na prática, é insuficiente, pois ela permite polinômios de grau arbitrariamente grande. Um algoritmo com complexidade de tempo expressa por um polinômio de grau 100, por exemplo, é eficiente pela definição, mas não encontra emprego prático. Para uma aplicação real, espera-se que os valores dos graus sejam pequenos, não excedendo algumas poucas unidades. Neste tema, estudaremos algoritmos para os quais o tamanho da entrada é muito grande, além dos limites considerados usuais, de modo que as restrições sobre o grau do polinômio tornam-se ainda mais severas. São fartos os exemplos de aplicações desse tipo particularmente na *web*, onde os serviços *online* servem milhões (ou mesmo, bilhões) de usuários, gerando uma quantidade imensa de dados a serem processados pelas diversas aplicações. Como consequência, a leitura da entrada nestas aplicações deverá ser feita apenas uma única vez, de forma sequencial. Tais algoritmos serão chamados de *algoritmos de dados massivos*. Dada uma sequência de dados massivos, exemplos de problemas que serão abordados incluem: escolher um deles de forma aleatória com distribuição uniforme; estimar a quantidade desses dados que satisfazem certa propriedade, empregando memória muito limitada; determinar a quantidade de dados que são distintos; a pertinência de um elemento arbitrário ao conjunto, entre outros.

1.1.5 Aprendizado de Máquina

Métodos baseados em aprendizado de máquina, para resolver problemas das mais diversas áreas, são cada vez mais utilizados. Em alguns contextos, o aprendizado de máquina é apresentado como a própria ciência de dados, e não como uma parte

da mesma. Esta popularidade se deve, principalmente, ao êxito alcançado por alguns de seus métodos. O sucesso é medido pelos resultados empíricos, às vezes impressionantes. Este livro aborda alguns dos tópicos do aprendizado de máquina. Em especial, o problema de classificação. Uma formulação do algoritmo Perceptron é apresentada em detalhe, incluindo exemplos de aplicação. Descrevemos também a dimensão de Vapnik-Chervonenkis, com uma certa ênfase na utilização dos resultados destes autores em aprendizado de máquina. Finalmente, o texto contém também o método de regressão linear, para o problema de classificação.

1.1.6 Cadeias de Markov

Cadeias de Markov são uma ferramenta poderosa da ciência da computação e da matemática. Suas aplicações incluem a resolução de problemas combinatórios, a previsão do clima, o movimento Browniano, a dispersão de gases, e o espalhamento de doenças.

Vamos ver alguns tópicos relacionados às cadeias de Markov. Mostramos como a matriz potência P^m da matriz de transição P de uma cadeia de Markov é usada para resolver problemas combinatórios. Em especial, o problema da previsão climática. Relacionamos por meio de um algoritmo polinomial, um processo estocástico não Markoviano com um processo Markoviano equivalente. Dentro do estudo das propriedades topológicas, consideramos a classe das cadeias de Markov, cujos estados são unicamente transientes ou absorventes, obtendo os tempos esperados para absorção e a probabilidade de absorção por um determinado estado. E finalmente, mostramos como o método da potência é utilizado para, dada uma cadeia de Markov ergódica, determinar sua distribuição estacionária.

1.1.7 Passeios Aleatórios

Introduzidos em 1905 em um artigo da Nature, os passeios aleatórios são utilizados nas mais diversas áreas de conhecimento. Suas aplicações incluem o estudo do deslocamento de animais, o comportamento de supercordas, e o mercado de ações. Mais recentemente, os passeios aleatórios celebrizaram-se por sua aplicação no algoritmo Pagerank usado como principal ferramenta na máquina de busca da Google.

O capítulo aborda algumas aplicações do tema. Em especial, uma formulação do algoritmo Pagerank é apresentada em detalhe, incluindo um exemplo de aplicação. Na classe dos passeios aleatórios de dimensão 1, estudamos o problema da ruína do jogador e um algoritmo aleatório para o problema 2-SAT.

Títulos Publicados — 33º Colóquio Brasileiro de Matemática

Geometria Lipschitz das singularidades – *Lev Birbrair e Edvalter Sena*

Combinatória – *Fábio Botler, Maurício Collares, Taísa Martins, Walner Mendonça, Rob Morris e Guilherme Mota*

Códigos Geométricos – *Gilberto Brito de Almeida Filho e Saeed Tafazolian*

Topologia e geometria de 3-variedades – *André Salles de Carvalho e Rafał Marian Siejakowski*

Ciência de Dados: Algoritmos e Aplicações – *Luerbio Faria, Fabiano de Souza Oliveira, Paulo Eustáquio Duarte Pinto e Jayme Luiz Szwarcfiter*

Discovering Euclidean Phenomena in Poncelet Families – *Ronaldo A. Garcia e Dan S. Reznik*

Introdução à geometria e topologia dos sistemas dinâmicos em superfícies e além – *Víctor León e Bruno Scárdua*

Equações diferenciais e modelos epidemiológicos – *Marlon M. López-Flores, Dan Marchesin, Vítor Matos e Stephen Schecter*

Differential Equation Models in Epidemiology – *Marlon M. López-Flores, Dan Marchesin, Vítor Matos e Stephen Schecter*

A friendly invitation to Fourier analysis on polytopes – *Sinai Robins*

PI-álgebras: uma introdução à PI-teoria – *Rafael Bezerra dos Santos e Ana Cristina Vieira*

First steps into Model Order Reduction – *Alessandro Alla*

The Einstein Constraint Equations – *Rodrigo Avalos e Jorge H. Lira*

Dynamics of Circle Mappings – *Edson de Faria e Pablo Guarino*

Statistical model selection for stochastic systems – *Antonio Galves, Florencia Leonardi e Guilherme Ost*

Transfer Operators in Hyperbolic Dynamics – *Mark F. Demers, Niloofar Kiamari e Carlangelo Liverani*

A Course in Hodge Theory Periods of Algebraic Cycles – *Hossein Movasati e Roberto Villaflor Loyola*

A dynamical system approach for Lane–Emden type problems – *Liliane Maia, Gabrielle Nornberg e Filomena Pacella*

Visualizing Thurston's Geometries – *Tiago Novello, Vinícius da Silva e Luiz Velho*

Scaling Problems, Algorithms and Applications to Computer Science and Statistics – *Rafael Oliveira e Akshay Ramachandran*

An Introduction to Characteristic Classes – *Jean-Paul Brasselet*

Instituto de
Matemática
Pura e Aplicada

ISBN 978-65-89124-47-4