**MÓDULO 1**: Exploratory Analytics, Predictive Analytics & Machine Learning

DISCENTE:

**Science and data science** (p. XX a XX)

David M. Bleia, and Padhraic Smythd. PNAS | August 15, 2017 | vol. 114 | no. 33 | 8689–8692.

**Fichamento de Conteúdo**

|  |
| --- |
| Resumo  A ciência de dados atraiu muita atenção, prometendo transformar grandes quantidades de dados em previsões e insights úteis. Neste artigo, pergunta-se por que os cientistas deveriam se preocupar com a ciência de dados. Para responder, discute-se a ciência de dados a partir de três perspectivas: *estatística, computacional e humana*. Argumenta-se *que a combinação eficaz de todos os três componentes é a essência do que trata a ciência de dados*. |

Aqui, discute-se a ciência de dados sob a perspectiva da pesquisa científica. O que é ciência de dados? Por que os cientistas se preocupam com isso? A perspectiva dos autores é que a ciência de dados é surge da estatística e da ciência da computação.

Segundo os autores, há problemas, pois os métodos existentes de estatística e computação não foram configurados para resolver os tipos de problemas que os cientistas modernos enfrentam.

Alguns problemas são computacionais, como trabalhar com conjuntos de dados enormes e complexos. Algumas questões são estatísticas, como as ricas interações de muitas variáveis ​​relacionadas e as dificuldades teóricas e práticas em torno das estatísticas de alta dimensão.

Finalmente, algumas questões são mais confusas e filosóficas, como modelos necessariamente mal especificados do mundo, dificuldades em identificar causalidade a partir de dados empíricos e desafios para atender objetivos disciplinares em torno da exploração e compreensão de dados.

O pensamento estatístico é um componente essencial. A estatística fornece as técnicas básicas para analisar e raciocinar sobre os dados.

Os métodos computacionais também são essenciais, especialmente quando os cientistas enfrentam dados grandes e complexos e têm restrições computacionais, como tempo e memória.

Finalmente, há o ângulo humano, a realidade de que a ciência de dados não pode ser totalmente automatizada. A aplicação de ferramentas estatísticas e computacionais modernas a questões científicas modernas requer julgamento humano significativo e conhecimento disciplinar profundo.

**Perspectiva Estatística**

As discussões sobre ciência de dados geralmente se concentram nos aspectos de grande escala de dados e computação. Essas questões são importantes, mas esse enfoque ignora que **os objetivos fundamentais da ciência de dados dependem do pensamento estatístico**. Desde o seu início, a estatística tem servido à ciência para orientar a coleta e análise de dados. Embora muitos aspectos da relação entre ciência e dados tenham mudado - os domínios em que usamos a análise de dados, a escala dos dados e a natureza das questões científicas - os princípios básicos são os mesmos.

Em termos gerais, a estatística trata do desenvolvimento de métodos para dar sentido aos dados. À medida que o campo evoluiu, esses métodos foram amplamente lançados nas linguagens da matemática e da probabilidade.

A estatística usa uma variedade de suposições funcionais e distribucionais para modelar relacionamentos entre variáveis ​​e entidades no mundo real e usa dados observados para fazer inferências e fazer previsões sobre tais relacionamentos.

Todos os conjuntos de dados envolvem incerteza. Pode haver incerteza sobre como eles foram coletados, como foram medidos ou o processo que os criou. A modelagem estatística ajuda a quantificar e raciocinar sobre as incertezas de uma forma sistemática. Ele fornece ferramentas e teoria que orientam as inferências e previsões para problemas específicos e dados reais. A estatística se relaciona com a ciência de dados por meio de vários subcampos estatísticos. Aqui, discutimos três: dados complexos e estruturados, alta dimensionalidade e causalidade.

Os conjuntos de dados modernos são complexos. Por exemplo, considere um problema de pesquisa envolvendo dados climáticos. Pode haver diferentes tipos de dependências nos dados: dependências ao longo do tempo, dependências em várias escalas espaciais e dependências entre diferentes variáveis, como precipitação, pressão e temperatura. As estatísticas fornecem uma linguagem rica para modelar parcimoniosamente essas dependências.

Essa linguagem ajuda a codificar o conhecimento do mundo em distribuições de probabilidade formais, compartilhar força estatística entre componentes relacionados de um problema e capturar regularidades sequenciais e espaciais entre as variáveis.

Muitos desses benefícios são encontrados nas estatísticas Bayesianas, uma estrutura que ajuda a articular suposições sobre os dados em um modelo formal e, em seguida, prescreve os métodos correspondentes para analisar dados para fazer inferências sobre o mundo.

Métodos bayesianos e técnicas relacionadas para modelagem de probabilidade expressiva têm o potencial de fornecer as ferramentas necessárias para combinar o conhecimento do domínio científico com a inferência estatística de dados.

Um subcampo relacionado de estatísticas diz respeito a dados de alta dimensão, onde medimos milhares ou mesmo milhões de variáveis ​​por ponto de dados. À medida que a medição científica se tornou cada vez mais sofisticada, a inferência estatística de dados de alta dimensão tornou-se mais importante para muitas disciplinas científicas. Para lidar com esses dados, estatísticos e cientistas da computação desenvolveram métodos poderosos que envolvem robustez, regularização e estabilidade. Além disso, dados de alta dimensão frequentemente surgem em problemas de reconhecimento de padrões, onde fazemos uma previsão sobre uma variável desconhecida com base em um grande conjunto de variáveis ​​ou parâmetros relacionados. Técnicas de aprendizado de máquina, como o aprendizado profundo, têm se mostrado particularmente eficazes nesse contexto.

Eles fornecem maneiras flexíveis de como a variável de destino pode depender dos preditores, e agora eles podem ser dimensionados para conjuntos de dados muito grandes.

A promessa implícita de conjuntos de dados ricos é que eles podem ajudar a aprofundar nossa compreensão de como o mundo funciona, e usar dados para obter tal compreensão é o objetivo elevado da inferência causal.

O pensamento estatístico sobre causalidade remonta ao final de 1800, com o desenvolvimento de ideias influentes em torno da diferença entre correlação e causalidade e como projetar experimentos significativos. Hoje, a causalidade cresceu e se tornou um campo rico, com contribuições significativas da ciência da computação, ciências sociais e estatísticas. O desenvolvimento de novos métodos de inferência causal - como escalar para grandes conjuntos de dados, como desenvolver inferências a partir de dados observacionais, como desenvolver inferências a partir de dados de interação (como em uma rede social) e como projetar experimentos na era do computador - é um caminho perfeito para contribuições estatísticas para a ciência de dados.

**Perspectiva Computacional**

O pensamento estatístico fornece métodos para responder a questões científicas com dados. O pensamento computacional se concentra na implementação algorítmica desses métodos e fornece uma maneira de entender e comparar suas pegadas computacionais. O pensamento computacional é particularmente importante na análise de dados moderna, onde frequentemente enfrentamos uma troca entre precisão estatística e recursos computacionais, como tempo e memória.

Um exemplo bem conhecido de pensamento computacional gira em torno da otimização. Muitos métodos de ciência de dados envolvem a maximização de uma função dos dados. (Um exemplo principal disso é quando tentamos maximizar a probabilidade dos dados em relação aos parâmetros de um modelo de probabilidade.) A maneira mais comum de maximizar uma função é escalá-la, calculando iterativamente a direção para viajar e movendo-a livremente parâmetros ao longo dessa direção.

No contexto da otimização, o pensamento computacional envolve a compreensão de como calcular melhor a direção, quando as direções aproximadas são suficientes, a distância percorrida em cada iteração,

e quanta precisão sacrificamos quando paramos de escalar cedo para economizar computação.

Outro exemplo de pensamento computacional são os métodos de amostragem. Os métodos de amostragem ajudam a calcular soluções aproximadas de problemas de análise de dados onde às soluções exatas são muito complexas para a análise matemática direta. Por exemplo, o bootstrap (18) é uma forma de calcular intervalos de confiança em situações muito complexas. Ele faz uma amostragem repetida dos dados para aproximar os tipos de grandezas que seriam impossíveis (ou quase impossíveis) de derivar analiticamente. O bootstrap, em sua simplicidade, teve um grande impacto na prática da estatística na ciência moderna. Outra aplicação de amostragem amplamente utilizada é na análise de dados Bayesianos, onde um dos métodos computacionais mais prevalentes é a cadeia de Markov Monte Carlo (MCMC).

Os algoritmos MCMC amostram os parâmetros de um modelo estatístico para produzir distribuições posteriores aproximadas, distribuições de grandezas ocultas condicionadas aos dados. Como o bootstrap, o MCMC transforma cálculos matemáticos difíceis em procedimentos baseados em amostragem. Desde a década de 1990, essa transformação abriu a porta para modelos, métodos e aplicações inimagináveis ​​para as estatísticas bayesianas.

Um exemplo final de pensamento computacional está na análise de escala de dados com computação distribuída. Agora podemos distribuir grandes conjuntos de dados em vários processadores (para velocidade) e vários dispositivos de armazenamento (para memória), e há uma variedade de software para suportar computação distribuída. Os avanços na computação distribuída baseiam-se na pesquisa da década de 1970 em computação científica em grande escala, bem como nas inovações mais recentes desenvolvidas na indústria de tecnologia. As mesmas ideias que permitiram que as empresas de tecnologia escalonassem seus métodos para a Internet crescente podem permitir que os cientistas escalem seus conjuntos de dados crescentes. Esses exemplos são apenas algumas das maneiras como o pensamento computacional desempenha um papel na ciência de dados. De forma mais ampla, o pensamento computacional ajuda a orientar como contabilizamos os recursos ao analisar os dados. Enquanto o pensamento estatístico oferece um conjunto de métodos para a compreensão de dados, o pensamento computacional fornece as considerações cruciais de como equilibrar a precisão estatística com recursos computacionais limitados.

**Perspectiva Humana**

Descrevemos o pensamento estatístico e o pensamento computacional, dois componentes essenciais da ciência de dados que fornecem ferramentas gerais para a análise de dados. A arte da ciência de dados é entender como aplicar essas ferramentas no contexto de um conjunto de dados específico e para responder a questões científicas específicas.

A ciência de dados combina pensamento estatístico e computacional, mas muda seu foco e prioriza os objetivos tradicionais de cada um. Ele conecta modelos estatísticos e métodos computacionais para resolver problemas específicos de disciplinas. Em particular, coloca um rosto humano no processo de análise de dados: compreender um domínio de problema, decidir quais dados adquirir e como processá-los, explorar e visualizar os dados, selecionar modelos estatísticos e métodos computacionais apropriados e comunicar os resultados das análises. Essas habilidades geralmente não são ensinadas nas aulas de estatística ou ciência da computação tradicional, mas, em vez disso, são adquiridas por meio da experiência e da colaboração com outras pessoas. Essa perspectiva da ciência de dados é holística e concreta. Para cada problema científico, o cientista de dados desenvolve uma compreensão de seu contexto: como os dados foram coletados, teorias existentes e conhecimento de domínio e os objetivos gerais da disciplina. Crucialmente, o cientista de dados resolve o problema de forma iterativa e colaborativa com o especialista no domínio. (Observamos que eles não precisam ser duas pessoas diferentes; o cientista de dados e o especialista no domínio podem ser simplesmente dois "chapéus" para a mesma pessoa.) Juntos, eles desenvolvem dados computacionais e estatísticos ferramentas para explorar dados, questões e métodos a serviço do objetivos da disciplina.

Como exemplo, considere um neurocientista computacional. Nova tecnologia de imagem permite que seus neurônios camundongos imagem enquanto eles atuam em um labirinto com outros ratos. Amplo financiamento e equipamento permitiram que ela rodasse centenas de ratos, resultando em terabytes de dados de vídeo e dados de imagens cerebrais. Com um cientista de dados, ela pode desenvolver métodos que testam as teorias existentes do comportamento do mouse, produzir hipóteses sobre como o comportamento é controlado pelo cérebro e lidar com a alta resolução e complexidade dos dados de vídeo e do cérebro por meio de algoritmos. Além disso, o cientista de dados ajuda a desenvolver métodos que abordam as limitações da nova tecnologia, especialmente como diferentes execuções do experimento podem exibir diferentes (irrelevantes) condições que confundem os resultados da análise. O projeto bem-sucedido resulta em novos resultados da neurociência e no desenvolvimento de novos métodos de ciência de dados.

A perspectiva humana revela como aspectos do processo de análise de dados, como metadados, proveniência de dados, fluxos de trabalho de análise de dados e reprodutibilidade científica, são críticos para a pesquisa científica moderna. Precisamos de boas ferramentas de software e infraestrutura que possam registrar, replicar e facilitar a forma como os pesquisadores interagem com seus dados (26, 27). De forma mais ampla, a prática da ciência de dados não é apenas uma única etapa de análise de um conjunto de dados. Em vez disso, ele circula entre o pré-processamento, exploração, seleção, transformação, análise, interpretação e comunicação dos dados. Uma das principais prioridades da ciência de dados é desenvolver ferramentas e métodos que facilitem esse ciclo.

**Resumo**

Apresentamos uma visão holística da ciência de dados, uma visão que tem implicações para a prática, pesquisa e educação. Ele sugere o potencial de integração de pesquisas que cruzam as fronteiras estatísticas, computacionais e humanas. Além disso, coloca em foco que, para resolver problemas do mundo real, um cientista de dados precisará realizar tarefas que vão além de seu treinamento tradicional. A ciência de dados é mais do que a combinação de estatística e ciência da computação - requer treinamento em como entrelaçar técnicas estatísticas e computacionais em uma estrutura maior, problema por problema, e como abordar questões específicas de disciplinas. A ciência de dados holística requer que entendamos o contexto dos dados, avaliemos as responsabilidades envolvidas no uso de dados privados e públicos e comuniquemos claramente o que um conjunto de dados pode ou não nos dizer sobre o mundo.