Data Science Para leigos

"Big data" é definitivamente um grande jargão, e a maioria das pessoas que se depara com ele percebe que é uma força poderosa em um processo de revolução dos grandes setores. Porém, poucas pessoas conhecem a variedade de ferramentas disponíveis que podem ajudar os negócios de grande e pequeno porte a aproveitar essa revolução. Esta Folha de Cola apresenta uma visão geral dessas ferramentas e mostra como se encaixam no contexto maior do data science.

O QUE VOCÊ PRECISA SABER AO INGRESSAR NO DATA SCIENCE

Tradicionalmente, *big data* é o termo para dados com volume, velocidade e variedade incríveis. As tecnologias tradicionais de banco de dados não conseguem lidar com o big data — por isso a necessidade de inovações de engenharia de dados. Para saber se o seu projeto se qualifica como big data, considere os seguintes critérios:

- **Volume:** Entre 1 terabyte/ano e 10 petabytes/ano
- **Velocidade:** Entre 30 kilobytes/segundo e 30 gigabytes/segundo
- **Variedade:** Combinação de dados não estruturados, semiestruturados e estruturados

DATA SCIENCE E ENGENHARIA DE DADOS NÃO SÃO A MESMA COISA

Gerentes de contratação tendem a confundir os papéis de cientista de dados e engenheiro de dados. Embora seja possível encontrar alguém que desempenhe um pouco das duas funções, ambos os campos são incrivelmente complexos. É improvável encontrar um profissional com habilidades e experiência sólidas nas duas áreas. Por isso, é importante identificar o tipo ideal de especialista para ajudá-lo a atingir seus objetivos específicos. As descrições abaixo irão orientá-lo nesse sentido.

- **Cientistas de dados:** Usam codificação, métodos quantitativos (matemáticos, estatísticos e aprendizagem de máquina) e conhecimento especializado altamente qualificado em sua área de estudo para solucionar complexos problemas corporativos e científicos.
- **Engenheiros de dados:** Usam a ciência da computação e a engenharia de software para projetar sistemas e resolver problemas, e manipular e lidar com grandes conjuntos de dados.

Data Science Para leigos

DATA SCIENCE E INTELIGÊNCIA DE NEGÓCIO TAMBÉM NÃO SÃO IGUAIS

Cientistas de dados centrados no negócio e analistas de negócio que usam a inteligência de negócio são como primos. Os dois tipos de especialistas usam dados para atingir os mesmos objetivos corporativos, mas suas abordagens, tecnologias e funções são distintas. As descrições abaixo esclarecem as diferenças entre as duas funções.

- **Inteligência de negócio (BI):** As soluções de BI geralmente são criadas a partir de conjuntos de dados gerados internamente — em outras palavras, de dentro de uma organização e não de fora. As ferramentas e tecnologias comuns incluem processamento analítico online, extração, transformação, transporte e armazenamento de dados. Embora a BI às vezes envolva métodos avançados como a previsão, esses métodos têm base em inferências matemáticas simples de dados diacrônicos ou correntes.
- **Data science centrada no negócio:** As soluções de data science centrada no negócio são criadas a partir de conjuntos de dados internos e externos a uma organização. As ferramentas, tecnologias e habilidades comuns incluem plataformas de análise baseadas em nuvem, programação estatística e matemática, aprendizagem de máquina, análise de dados com Python e R e visualização de dados avançada. Os cientistas de dados centrados no negócio usam métodos matemáticos ou estatísticos avançados para analisar e gerar previsões a partir de uma grande quantidade de dados de negócio.

O BÁSICO DE ESTATÍSTICA, APRENDIZAGEM DE MÁQUINA E MÉTODOS MATEMÁTICOS EM DATA SCIENCE

Se a estatística foi descrita como uma ciência que deriva informações de dados, qual é a diferença entre um estatístico e um cientista de dados? Boa pergunta! Embora muitas tarefas de data science exijam um pouco mais de prática estatística, o escopo, o alcance da base de conhecimento e a habilidade de um cientista de dados são distintos dos de um estatístico. As principais diferenças são descritas abaixo.

- **Especialização no assunto:** Um dos principais recursos dos cientistas de dados é seu grau sofisticado de especialização na área na qual aplicam seus métodos analíticos. Eles precisam disso para realmente entender as implicações e aplicações das informações geradas a partir dos dados. Um cientista de dados deve ser altamente especializado no assunto para identificar a importância das suas descobertas e decidir, com independência, como prosseguir na análise.

Por outro lado, os estatísticos geralmente têm um conhecimento incrivelmente profundo de estatística, mas pouca especialização nos assuntos em que aplicam seus métodos. Na maioria das vezes, os estatísticos precisam consultar especialistas externos ao assunto para formar uma compreensão sólida do significado das suas descobertas e decidir sobre o melhor modo de avançar em uma análise.

- **Abordagens matemáticas e de aprendizagem de máquina:** Os estatísticos dependem essencialmente de métodos e processos estatísticos para derivar informações dos dados. Por outro lado, os cientistas de dados precisam de uma grande variedade de técnicas para fazê-lo, como métodos estatísticos. Contudo, também necessitam de abordagens não baseadas em estatística — como as encontradas em matemática, agrupamento, classificação e abordagens não estatísticas de aprendizagem de máquina.

A IMPORTÂNCIA DO CONHECIMENTO DE ESTATÍSTICA

Você não precisa ser graduado em estatística para praticar o data science. No entanto, deve, pelo menos, estar familiarizado com alguns dos métodos mais fundamentais usados na análise de dados estatísticos. Isso inclui:

- **Regressão linear:** É útil para modelar as relações entre uma variável dependente e uma ou diversas variáveis independentes. Sua finalidade é descobrir (e quantificar a intensidade de) importantes correlações entre as variáveis dependentes e independentes.
- **Análise da série temporal:** Envolve analisar uma coleção de dados quanto aos valores de atributo no decorrer do tempo, e prever instâncias futuras com base nos dados observacionais do passado.
- **Simulações de Monte Carlo:** O método de Monte Carlo é uma técnica de simulação que você pode usar para testar hipóteses, gerar estimativas de parâmetros, prever os resultados do cenário e validar os modelos. Esse método é poderoso porque pode ser usado para simular muito rapidamente de 1 a 10 mil (ou mais) amostras de simulação para qualquer processo que esteja avaliando.
- **Estatística para dados espaciais:** Uma propriedade fundamental e importante dos dados espaciais é não serem aleatórios. São espacialmente dependentes e autocorrelacionados. Ao modelar dados espaciais, evite métodos estatísticos que supõem que seus dados sejam aleatórios. Krigagem e krige são dois métodos que você pode usar para modelar os dados espaciais. Eles permitem produzir superfícies de previsão para áreas de estudo completas com base em conjuntos de pontos conhecidos no espaço geográfico.

TRABALHANDO COM AGRUPAMENTO, CLASSIFICAÇÃO E MÉTODOS DE APRENDIZAGEM DE MÁQUINA

A aprendizagem de máquina é a aplicação de algoritmos computacionais para aprender (ou deduzir) padrões a partir de conjuntos de dados brutos. O agrupamento (clustering) é um tipo particular de aprendizagem de máquina — *não supervisionada*, especificamente. Ou seja, os algoritmos devem aprender com dados não rotulados e, portanto, usar métodos de inferência para descobrir as correlações.

A *classificação*, por outro lado, é chamada de aprendizagem de máquina supervisionada. Nela, os algoritmos aprendem com os dados rotulados. As seguintes descrições apresentam algumas das abordagens mais básicas de agrupamento e classificação:

- **Agrupamento k-vizinhos próximos:** Geralmente, você utiliza os algoritmos k-vizinhos próximos para subdividir os pontos de dados em um conjunto de clusters (agrupamentos) fundamentados nos valores médios mais próximos. Para determinar a divisão ideal dos seus pontos de dados em agrupamentos, de modo que a distância entre eles seja minimizada, é possível usar o agrupamento k-vizinhos próximos.
- **Algoritmos vizinhos mais próximos:** A finalidade de uma análise vizinha mais próxima é pesquisar e localizar um ponto ou valor numérico mais próximo, dependendo do atributo usado para a base de comparação.
- **Estimativa de densidade do kernel:** Um modo alternativo de identificar os agrupamentos em seus dados é usar uma função de suavização da densidade. A estimativa de densidade do kernel (KDE) funciona colocando uma função de peso kernel útil para quantificar a densidade — em cada ponto de dados no conjunto — e, então, soma os kernels para gerar uma estimativa de densidade do kernel para a região inteira.

MANTENDO MÉTODOS MATEMÁTICOS NA COMBINAÇÃO

Há muita discussão em torno do valor da estatística para a prática do data science, mas os métodos matemáticos aplicados raramente são mencionados. Falando francamente, a matemática é a base de toda a análise quantitativa e, portanto, sua importância não deve ser subestimada. Os dois métodos matemáticos a seguir são particularmente úteis no data science.

- **Tomada de decisão com vários critérios (MCDM):** MCDM é uma abordagem do modelo de decisão matemática que você pode usar quando houver vários critérios ou alternativas a serem avaliados simultaneamente na tomada de uma decisão.

- **Cadeias de Markov:** É um método matemático que encadeia uma série de variáveis geradas aleatoriamente e representa o estado atual para prever como as alterações nas variáveis afetarão os estados futuros.

USANDO TÉCNICAS DE VISUALIZAÇÃO PARA COMUNICAR AS INFORMAÇÕES DO DATA SCIENCE

Todas as informações e critérios no mundo serão inúteis se a comunicação não for possível. Se os cientistas de dados não puderem comunicar claramente suas descobertas para outras pessoas, informações de dados potencialmente valiosas permanecerão inexploradas.


DICA

Adotar as melhores práticas e princípios claros e específicos no design da visualização de dados proporciona o desenvolvimento de visualizações que comunicam seu conteúdo de modo altamente relevante e valioso para os interessados no projeto. A seguir, descrevo um pequeno resumo de algumas das práticas mais importantes no design de visualização de dados.

- **Conheça seu público:** Como as visualizações de dados são planejadas para uma ampla variedade de públicos, finalidades e níveis de habilidades, a primeira etapa para planejar uma ótima visualização de dados é conhecer seu público. Como cada público será composto de uma classe única de clientes, cada uma com necessidades de visualização específicas, é essencial determinar exatamente quem são essas pessoas.
- **Escolha estilos de design adequados:** Depois de determinar o público, escolher o estilo de design mais apropriado também é fundamental. Se o seu objetivo for atrair o público e envolvê-lo de modo mais profundo e analítico na visualização, use um estilo de design que induza uma resposta calculada e exata em seus espectadores. Se quiser que a sua visualização estimule a paixão do seu público, use um estilo emocionalmente convincente.
- **Escolha tipos gráficos de dados inteligentes:** Finalmente, selecione tipos gráficos que indiquem substancialmente as tendências de dados que deseja revelar. É possível exibir a mesma tendência de dados de muitas maneiras, mas alguns métodos transmitem a mensagem visual de modo mais eficiente. Escolha o tipo gráfico que transmita a mensagem de forma mais direta, clara e completa.

VENDO SEU CONJUNTO DE FERRAMENTAS DE CODIFICAÇÃO

D3.js é a linguagem de programação perfeita para criar visualizações da web dinâmicas e interativas. Se você já é programador ou dispõe do tempo necessário para aprender o básico de HTML, CSS e JavaScript, usar o D3.js para planejar visualizações de dados da web interativas é certamente a solução perfeita para muitos dos seus problemas de visualização.

TRABALHANDO COM APLICATIVOS DA WEB

Se você não tem tempo para codificar uma visualização de dados personalizada, não fique triste — existem aplicativos online surpreendentes que podem auxiliá-lo a fazer esse trabalho em pouco tempo. A lista a seguir detalha algumas alternativas excelentes.

- **Watson Analytics:** É a primeira solução completa de análise e data science disponibilizada como oferta 100% em nuvem. O Watson Analytics foi criado com a finalidade de democratizar o poder do data science. É uma plataforma que usuários com todos os níveis de habilidade podem acessar, aprimorar, descobrir, visualizar, relatar e colaborar para as informações baseadas em dados.
- **CartoDB:** Para os que não são programadores ou cartógrafos, o CartoDB é a solução de criação de mapas mais poderosa disponível online. É usado para desenvolver comunicações visuais digitais por pessoas de todos os setores — inclusive serviços de informação, engenharia de software, mídia e entretenimento, e desenvolvimento urbano.
- **Piktochart:** O aplicativo da web Piktochart dispõe de uma interface fácil de usar e em que é possível criar belos infográficos. Ele oferece uma seleção muito ampla de modelos atraentes e profissionalmente projetados. Com o Piktochart, é possível fazer infográficos estáticos ou dinâmicos.

COMBINANDO COM PAINÉIS DE ANÁLISE

Quando a palavra "painel" aparece, muitas pessoas a associam com antigas soluções de inteligência de negócio. Essa associação é falha. Um painel é apenas outro modo de usar métodos de visualização para comunicar informações de dados.

LEMBRE-SE

Embora você realmente possa usar um painel para comunicar as descobertas geradas a partir da inteligência de negócio, também é possível usá-lo para comunicar e transmitir informações valiosas que são derivadas do data science centrado no negócio. Só porque existem há algum tempo, os painéis não devem ser considerados como ferramentas eficientes para comunicar informações de dados valiosas.

APROVEITANDO O SOFTWARE DOS SISTEMAS DE INFORMAÇÕES GEOGRÁFICAS (GIS)

Os sistemas de informações geográficas (GIS) são outro recurso subestimado no data science. Quando o seu objetivo for descobrir e quantificar tendências locais em seu conjunto de dados, o GIS é a solução perfeita para o trabalho. Os mapas são uma forma de visualização de dados espaciais que você pode gerar com o GIS. No entanto, o software GIS também é útil para formas

mais avançadas de análise e visualização. As duas soluções GIS mais populares são detalhadas abaixo.

- **ArcGIS for Desktop:** É o aplicativo de criação de mapas mais usado.
- **QGIS:** Se não tiver dinheiro para investir no ArcGIS for Desktop, poderá usar o `QGIS` de fonte aberta para atingir os mesmos objetivos gratuitamente.

ESCOLHENDO AS MELHORES LINGUAGENS DE PROGRAMAÇÃO PARA O DATA SCIENCE

A codificação é uma das habilidades mais essenciais ao repertório de um cientista de dados. Há aplicativos incrivelmente poderosos que eliminaram a necessidade de codificar em alguns contextos do data science, mas você não poderá usá-los para desenvolver análises e visualizações personalizadas. Em tarefas avançadas, você mesmo terá que codificar usando a linguagem de programação Python ou R.

USANDO O PYTHON PARA O DATA SCIENCE

O Python é uma linguagem de programação legível e fácil de aprender, que você pode usar para fazer a preparação, análise e visualização dos dados. É muito fácil instalá-lo, configurá-lo e aprender a utilizá-lo em comparação com a linguagem de programação R. O Python pode ser executado em Mac, Windows e UNIX

O IPython oferece uma interface de codificação intuitiva para pessoas que não gostam de codificar na linha de comando. Ao baixar e instalar a distribuição **Anaconda Python**, você terá o ambiente IPython/Jupyter, assim como o NumPy, SciPy, MatPlotLib e Pandas e as bibliotecas scikit-learn (entre outras). Você provavelmente precisará desses recursos em seus procedimentos para entender os dados.

O pacote NumPy é o facilitador básico para a ciência da computação no Python. Fornece estruturas de contêineres/array que você pode usar para fazer cálculos com vetores e matrizes (como no R). O SciPy e o Pandas são as bibliotecas Python mais usadas para a computação científica e técnica.

Existem toneladas de algoritmos matemáticos que simplesmente não estão disponíveis para outras bibliotecas Python. As funcionalidades populares incluem álgebra linear, cálculo matricial, funcionalidades de matriz esparsa, estatística e modificação dos dados. O MatPlotLib é a principal biblioteca de visualização de dados do Python.

Por fim, a biblioteca scikit-learn é útil para a aprendizagem de máquina, pré-processamento dos dados e avaliação do modelo.

USANDO R PARA O DATA SCIENCE

O R é outra linguagem de programação popular usada para a computação estatística e científica. O ato de escrever rotinas de análise e visualização em R é conhecido como *script R*. O R foi desenvolvido especificamente para a computação estatística e, como consequência, tem uma oferta mais variada desses pacotes do que o Python.

Além disso, as capacidades de visualização de dados do R são mais sofisticadas do que as do Python e mais fáceis de gerar. Portanto, como linguagem, o Python é mais fácil para os iniciantes.


DICA

O R tem uma comunidade de usuários muito grande e extremamente ativa. Os desenvolvedores estão sempre propondo (e compartilhando) novos pacotes — para mencionar apenas alguns, o pacote `forecast` e os pacotes `ggplot2` e `statnet/igraph`.


DICA

Se quiser fazer uma análise preditiva e previsão em R, o pacote forecast será um bom lugar para começar. Ele oferece ARMA, AR e métodos de suavização exponencial.

Para a visualização dos dados, é possível usar o pacote `ggplot2`, que oferece todos os tipos gráficos de dados padrão e muito mais.

Por fim, os pacotes de análise de rede do R também são muito especiais. Por exemplo, é possível usar o `igraph` e o `StatNet` para fazer análises de rede social, mapeamento genético, planejamento do tráfego e modelagem hidráulica.