IBM Analytics
White Paper

Metodologia de Base para Ciência de Dados

No domínio da ciência de dados, solucionar problemas e responder perguntas por meio da análise de dados é a prática padrão. Geralmente, cientistas de dados constroem um modelo para prever resultados ou descobrir padrões subjacentes, com o objetivo de obter insights. As organizações podem, então, usar esses insights para executar ações que melhorem realmente os resultados futuros.

Há muitas tecnologias em rápida evolução para analisar dados e construir modelos. Em muito pouco tempo, elas progrediram de desktops a warehouses paralelos massivos com enormes volumes de dados e funcionalidade analítica em bancos de dados relacionais e Apache Hadoop. A análise de texto em dados não estruturados ou semiestruturados está se tornando cada vez mais importante como uma forma de incorporar sentimento e outras informações úteis do texto nos modelos preditivos, levando muitas vezes a melhorias significativas na qualidade e na precisão do modelo.

As abordagens de análise emergentes buscam automatizar diversas das etapas na construção e aplicação de modelos, tornando a tecnologia de learning machine mais acessível aos que não possuem muita aptidão quantitativa. Além disso, ao contrário de usar uma abordagem “top-down”, onde primeiro define-se o problema de negócios e, em seguida, os dados são analisados para encontrar uma solução, alguns cientistas de dados podem usar uma abordagem “bottom-up”. Nesse caso, o cientista de dados analisa grandes volumes de dados para ver qual meta de negócio pode ser sugerida pelos dados e, então, resolve esse problema. Como a maioria dos problemas é tratada de uma maneira "top-down", a metodologia neste material reflete essa visão.

Uma metodologia de ciência de dados de 10 estágios que abrange tecnologias e abordagens

À medida que os recursos de análise de dados se tornam mais acessíveis e dominantes, os cientistas de dados precisam de uma metodologia de base capaz de fornecer uma estratégia orientadora, independentemente das tecnologias, dos volumes de dados ou das abordagens envolvidas (veja a Figura 1). Essa metodologia possui algumas similaridades com metodologias 1 a 5 reconhecidas para mineração de dados, mas enfatiza diversas das novas práticas em ciência de dados, como o uso de volumes de dados muito grandes, a incorporação de análise de texto na modelagem preditiva e a automação de alguns processos.

A metodologia consiste em 10 estágios que formam um processo interativo para usar dados para descobrir insights. Cada estágio desempenha uma função vital no contexto da metodologia geral.

O que é uma metodologia?

Uma metodologia é uma estratégia geral que orienta os processos e atividades dentro de um determinado domínio. A metodologia não depende de tecnologias ou ferramentas específicas, nem é um conjunto de técnicas ou receitas. Em vez disso, uma metodologia fornece aos cientistas de dados uma estrutura para proceder de acordo com os métodos, processos e heurísticas que serão usados para obter respostas ou resultados.

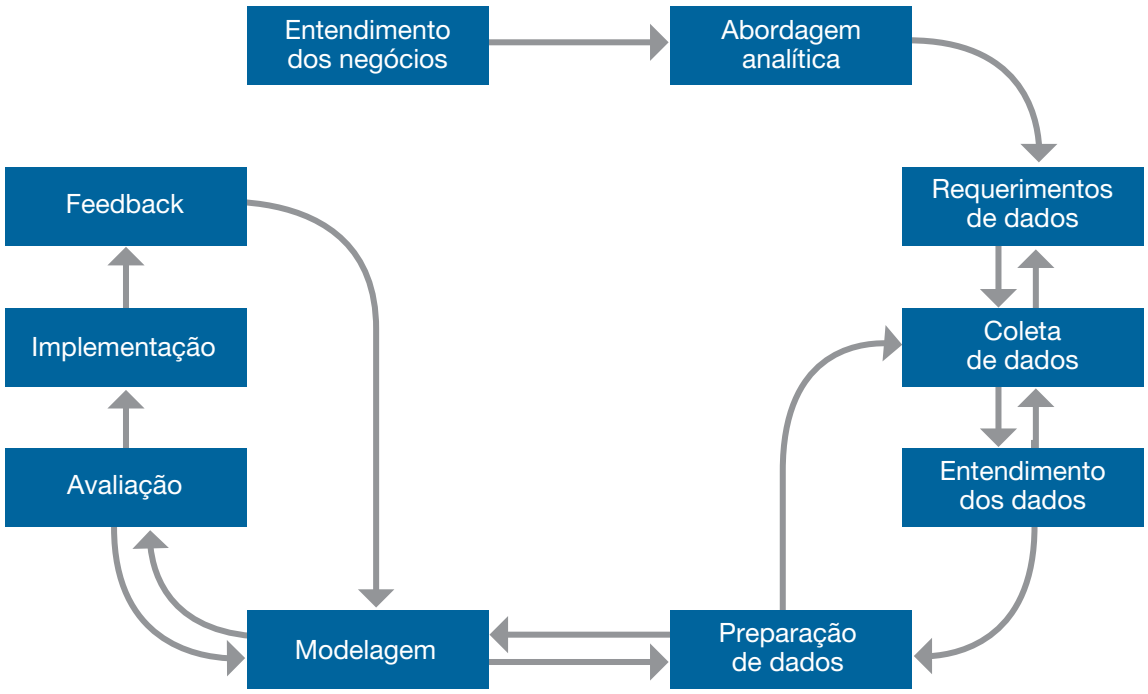

Figure 1. Metodologia de Base para Ciência de Dados.

Estágio 1: Entendimento de negócios

Cada projeto inicia com o entendimento de negócios. Os patrocinadores de negócios que precisam da solução de análise desempenham a função mais crítica nesse estágio, definindo o problema, os objetivos do projeto e os requisitos da solução a partir de uma perspectiva do negócio. Esse primeiro estágio forma a base para uma resolução bem-sucedida do problema de negócio. Para ajudar a garantir o sucesso do projeto, os patrocinadores devem estar envolvidos em todo o projeto, para fornecer conhecimento de domínio, revisar provas intermediárias e assegurar que o trabalho permaneça sob controle para gerar a solução desejada.

Estágio 2: Abordagem analítica

Assim que o problema de negócio tiver sido claramente identificado, o cientista de dados poderá definir a abordagem analítica para resolver o problema. Esse estágio implica em expressar o problema no contexto das técnicas de estatística e aprendizado de máquina, de modo que a organização possa identificar aquelas mais adequadas para obter o resultado desejado. Por exemplo, se o objetivo fosse prever uma resposta como “sim” ou “não”, a abordagem analítica poderia ser definida como a construção, o teste e a implementação de um modelo de classificação.

Estágio 3: Requisitos de dados

A abordagem analítica escolhida determina os requisitos de dados. Especificamente, os métodos analíticos a serem usados requerem determinados conteúdos, formatos e representações de dados, orientados pelo conhecimento de domínio.

Estágio 4: Coleta de dados

No estágio da coleta de dados iniciais, os cientistas de dados identificam e reúnem os recursos de dados disponíveis, estruturados, não estruturados e semiestruturados, relevantes para o domínio do problema. Geralmente, eles devem escolher se devem fazer investimentos adicionais para obter elementos de dados menos acessíveis. Pode ser melhor adiar a decisão de investimento até que se tenha mais conhecimento sobre os dados e o modelo. Se houver lacunas na coleta de dados, o cientista de dados poderá precisar revisar os requisitos de dados adequadamente e coletar dados novos e/ou adicionais.

Embora a amostragem e a subconfiguração de dados ainda sejam importantes, as plataformas de alta performance e a funcionalidade de análise dentro do banco de dados atuais permitem que os cientistas de dados usem conjuntos de dados muito maiores contendo muitos ou, até mesmo, todos os dados disponíveis. Ao incorporar mais dados, os modelos preditivos podem representar melhor eventos raros como a incidência de danos ou a falha do sistema.

Estágio 5: Entendimento dos dados

Após a coleta de dados original, os cientistas de dados geralmente usam estatísticas descritivas e técnicas de visualização para entender o conteúdo dos dados, avaliar a qualidade dos dados e descobrir insights iniciais sobre os dados. Pode ser necessário coletar dados adicionais para preencher lacunas.

Estágio 6: Preparação de dados

Esse estágio abrange todas as atividades para construir o conjunto de dados que será usado no estágio de modelagem subsequente. As atividades de preparação de dados incluem limpeza de dados (lidar com valores ausentes ou inválidos, eliminar duplicatas, formatar adequadamente), combinar dados de diversas fontes (arquivos, tabelas, plataformas) e transformar dados em variáveis mais úteis.

Em um processo chamado engenharia de recurso, os cientistas de dados podem criar variáveis explanatórias adicionais, também referidas como preditores ou recursos, por meio de uma combinação de conhecimento de domínio e variáveis estruturadas existentes. Quando dados de texto estão disponíveis, como logs da central de atendimento do cliente ou notas de médicos em formato não estruturado ou semiestruturado, a análise de texto é útil na derivação de novas variáveis estruturadas para enriquecer o conjunto de preditores e melhorar a precisão do modelo.

A preparação de dados geralmente é a etapa mais demorada em um projeto de ciência de dados. Em muitos domínios, algumas etapas de preparação de dados são comuns entre diferentes problemas. Automatizar determinadas etapas de preparação de dados com antecedência pode acelerar o processo, minimizando o tempo de preparação ad hoc. Com os sistemas paralelos massivos de alta performance atuais e a funcionalidade analítica residindo onde os dados estão armazenados, os cientistas de dados podem preparar dados de maneira mais fácil e rápida, usando conjuntos de dados muito grandes.

Estágio 7: Modelagem

Iniciando com a primeira versão do conjunto de dados preparado, o estágio de modelagem foca no desenvolvimento de modelos preditivos ou descritivos de acordo com a abordagem analítica definida anteriormente.

Com os modelos preditivos, os cientistas de dados usam um conjunto de treinamento (dados históricos nos quais o resultado de interesse é conhecido) para construir o modelo. O processo de modelagem em geral é altamenteiterativo, pois as organizações ganham insights intermediários, que levam a refinamentos na preparação de dados e na especificação de modelo. Para uma determinada técnica, os cientistas de dados podem experimentar diversos algoritmos com seus respectivos parâmetros para localizar o melhor modelo para as variáveis disponíveis.

Estágio 8: Avaliação

DDurante o desenvolvimento do modelo e antes da implementação, o cientista de dados avalia o modelo para entender sua qualidade e assegurar que ele trate de maneira completa e adequada o problema de negócio. A avaliação de modelo implica na computação de diversas medidas de diagnóstico e outros resultados, como tabelas e gráficos, permitindo que o cientista de dados interprete a qualidade do modelo e sua eficácia na resolução do problema. Para um modelo preditivo, os cientistas de dados usam um conjunto de teste, que é independente do conjunto de treinamento, mas segue a mesma distribuição de probabilidade e possui um resultado conhecido. O conjunto de testes é usado para avaliar o modelo para que ele possa ser refinado conforme necessário. Às vezes, o modelo final também é aplicado em um conjunto de validação para uma avaliação final.

Além disso, os cientistas de dados podem designar testes de significância estatística para o modelo como prova adicional de sua qualidade. Essa prova adicional pode ser instrumental, justificando a implementação do modelo, ou a execução de ações quando as apostas são altas — como um protocolo médico suplementar caro ou um sistema de voo crítico.

Estágio 9: Implementação

Quando um modelo satisfatório tiver sido desenvolvido e aprovado pelos patrocinadores de negócios, ele será implementado no ambiente de produção ou em um ambiente de teste comparável. Geralmente, ele é implementado de uma maneira limitada, até que sua performance tenha sido completamente avaliada.

A implementação pode ser tão simples quanto gerar um relatório com recomendações ou tão complexo quanto integrar o modelo em um fluxo de trabalho complexo e pontuar o processo gerenciado por um aplicativo customizado. A implementação de um modelo em um processo de negócios operacional geralmente envolve grupos, aptidões e tecnologias adicionais de dentro da empresa. Por exemplo, um grupo de vendas pode implementar um modelo de propensão de resposta por meio de um processo de gerenciamento de campanha criado por uma equipe de desenvolvimento e administrado por um grupo de marketing.

Estágio 10: Feedback

Ao coletar resultados do modelo implementado, a organização obtém feedback sobre a performance do modelo e seu impacto no ambiente no qual ele foi implementado. Por exemplo, o feedback poderia ter o formato de taxas de resposta para uma campanha promocional que visa a um grupo de clientes identificados pelo modelo como respondentes de alto potencial. A análise desse feedback permite que os cientistas de dados refinem o modelo para melhorar sua precisão e utilidade. Eles podem automatizar algumas ou todas as etapas de reunião de feedback e avaliação, refinamento e reimplementação do modelo para acelerar o processo de atualização do modelo para obter melhores resultados.

Fornecendo valor contínuo para a organização

O fluxo da metodologia ilustra a natureza iterativa do processo de resolução de problemas. À medida que os cientistas de dados sabem mais sobre os dados e a modelagem, eles retornam frequentemente a um estágio anterior para fazer ajustes. Os modelos não são criados uma vez, implementados e deixados no lugar como estão; em vez disso, por meio do feedback, do refinamento e da reimplementação, os modelos são melhorados e adaptados continuamente às condições evolutivas. Dessa maneira, o modelo e o trabalho por trás dele podem fornecer valor contínuo à organização pelo período em que a solução for necessária.

Para mais informações

Um novo curso sobre a Metodologia de Base para Ciência de Dados está disponível por meio da Big Data University. O curso on-line gratuito está disponível em: http://bigdatauniversity.com/bdu-wp/bdu-course/data-science-methodology

Para obter exemplos de trabalho de como essa metodologia foi implementada em casos de uso reais, visite:

- http://ibm.co/1SUhxFm
- http://ibm.co/1IazTvG

Confirmações

Obrigado a Michael Haide, Michael Wurst, Ph.D., Brandon MacKenzie e Gregory Rodd por seus comentários úteis e a Jo A. Ramos por seu papel no desenvolvimento dessa metodologia ao longo de seus anos de colaboração

Sobre o autor

John B. Rollins, Ph.D., é um cientista de dados na organização IBM Analytics. Seu conhecimento é em engenharia, mineração de dados e econometria em diversos segmentos de mercado. Ele possui sete patentes e é autor de um livro de engenharia campeão de vendas e de diversos documentos técnicos. Ele possui doutorado em engenharia petrolífera e economia pela Texas A&M University.

IBM Analytics
Route 100
Somers, NY 10589

Produzido nos Estados Unidos da América em junho de 2015

[1] Brachman, R. & Anand, T., "The process of knowledge discovery in databases," in Fayyad, U. et al., eds., Advances in knowledge discovery and data mining, AAAI Press, 1996 (pp. 37-57)

[2] SAS Institute, http://en.wikipedia.org/wiki/SEMMA, www.sas.com/en_us/software/analytics/enterprise-miner.html, www.sas.com/en_gb/software/small-midsize-business/desktop-data-mining.html

[3] Wikipedia, "Cross Industry Standard Process for Data Mining," http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining, http://the-modeling-agency.com/crisp-dm.pdf

[4] Ballard, C., Rollins, J., Ramos, J., Perkins, A., Hale, R., Dorneich, A., Milner, E., and Chodagam, J.: Dynamic Warehousing: Data Mining Made Easy, IBM Redbook SG24-7418-00 (Sep. 2007), pp. 9-26.

[5] Gregory Piatetsky, CRISP-DM, still the top methodology for analytics, data mining, or data science projects, Oct. 28, 2014, www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Recicle

Fale com especialista

IMW14828-BRPT-01