Sistemas de Apoio à Decisão

Problemas de Negócios e Soluções de Data Science

Vahid Nikoofard

Universidade do Estado do Rio de Janeiro - Faculdade de Tecnologia
Laboratório Interdisciplinar Virtual de Inteligência Artificial (LIVIA)

Outubro 2022

De Problemas de Negócios a Mineração de Dados

- Cada problema de tomada de decisão baseada em dados é exclusivo.
- Cada problema pode ser dividido em um conjunto de tarefas
- Algumas tarefas são comuns entre problemas diferentes e alguns são exclusivos.
- Uma habilidade crucial para um cientista de dados é decompor um problema em tarefas menores que cada um tenha uma solução mais simples.
- Apesar do grande número de algoritmos específicos de mineração de dados desenvolvidos ao longo dos anos, há apenas um punhado de tipos de tarefas fundamentalmente diferentes tratadas por esses algoritmos.

Algumas Tarefas Fundamentais na Ciência de Dados

Classificação e estimativa de probabilidade de classe tentam prever, para cada indivíduo de uma população, a que (pequeno) conjunto de classes este indivíduo pertence. Geralmente, as classes são mutuamente exclusivas.

- Um exemplo de pergunta de classificação seria: "Entre todos os clientes da MegaTelCo, quais são suscetíveis de responder a determinada oferta?" Neste exemplo, as duas classes poderiam ser chamadas vai responder e não vai responder. Para uma tarefa de classificação, o processo de mineração de dados produz um modelo que, dado um novo indivíduo, determina a que classe o indivíduo pertence.
- Uma tarefa intimamente relacionada é pontuação ou estimativa de probabilidade de classe. O modelo de pontuação aplicado a um indivíduo produz, em vez de uma previsão de classe, uma pontuação que representa a probabilidade (ou outra quantificação de probabilidade) de que o indivíduo pertença a cada classe.

Algumas Tarefas Fundamentais na Ciência de Dados

Regressão ("estimativa de valor") tenta estimar ou prever, para cada indivíduo, o valor numérico de alguma variável.

- Um exemplo de pergunta de regressão seria: "Quanto determinado cliente usará do serviço?" A propriedade (variável) a ser prevista aqui é o uso do serviço, e um modelo poderia ser gerado analisando outros indivíduos semelhantes na população e seus históricos de uso. Um procedimento de regressão produz um modelo que, dado um indivíduo, calcula o valor da variável específica para aquele indivíduo.
- A regressão está relacionada com a classificação, porém, as duas são diferentes. Informalmente, a classificação prevê se alguma coisa vai acontecer, enquanto que a regressão prevê quanto de alguma coisa vai acontecer.

Algumas Tarefas Fundamentais na Ciência de Dados

Combinação por similaridade tenta identificar indivíduos semelhantes com base nos dados conhecidos sobre eles. A combinação de similaridade pode ser usada diretamente para encontrar entidades semelhantes.

- Por exemplo, a IBM está interessada em encontrar empresas semelhantes aos seus melhores clientes comerciais, a fim de concentrar sua força de vendas nas melhores oportunidades. Eles usam a combinação por similaridade com base dos dados "firmográficos", que descrevem as características das empresas.
- A combinação por similaridade é a base de um dos métodos mais populares para se fazer recomendações de produtos (encontrar pessoas semelhantes a você, em termos de produtos que tenham gostado ou comprado). Medidas de similaridade são a base de determinadas soluções ou outras tarefas de mineração de dados, como classificação, regressão e agrupamento.

Algumas Tarefas Fundamentais na Ciência de Dados

Agrupamento tenta reunir indivíduos de uma população por meio de sua similaridade, mas não é motivado por nenhum propósito específico.

- Um exemplo de pergunta de agrupamento seria: "Nossos clientes formam grupos naturais ou segmentos?" O agrupamento é útil na exploração preliminar de domínio para ver quais grupos naturais existem, pois esses grupos, por sua vez, podem sugerir outras tarefas ou abordagens de mineração de dados.
- O agrupamento também é utilizado como entrada para processos de tomada de decisão com foco em questões como: quais produtos devemos oferecer ou desenvolver? Como nossas equipes de atendimento ao cliente (ou equipes de vendas) devem ser estruturadas?

Algumas Tarefas Fundamentais na Ciência de Dados

Agrupamento de coocorrência (mineração de conjunto de itens frequentes) tenta encontrar associações entre entidades com base em transações que as envolvem.

- Um exemplo de pergunta de coocorrência seria: Quais itens são comumente comprados juntos? Por exemplo, analisar os registros de compras de um supermercado pode revelar que carne moída é comprada junto com molho de pimenta com muito mais frequência do que se poderia esperar. Decidir como agir de acordo com essa descoberta pode exigir um pouco de criatividade, mas pode sugerir uma promoção especial, a exibição do produto ou uma oferta combinada.
- Alguns **sistemas de recomendação** também realizam um tipo de agrupamento por afinidade encontrando, por exemplo, pares de livros que são frequentemente comprados pelas mesmas pessoas ("pessoas que compraram X também compraram Y").
- O resultado do agrupamento por coocorrência é uma descrição dos itens que ocorrem juntos. Essas descrições geralmente incluem estatísticas sobre a frequência da coocorrência e uma estimativa do quanto ela é surpreendente.

Algumas Tarefas Fundamentais na Ciência de Dados

Perfilamento (descrição de comportamento) tenta caracterizar o comportamento típico de um indivíduo, grupo ou população.

- Um exemplo de pergunta de perfilamento seria: "Qual é o uso típico de celular nesse segmento de cliente?" O comportamento pode não ter uma descrição simples; traçar o perfil do uso do celular pode exigir uma descrição complexa das médias durante a noite e finais de semana, conteúdos de texto e assim por diante.
- O perfilamento muitas vezes é usado para estabelecer normas de comportamento para aplicações de detecção de anomalias como detecção de fraudes e monitoramento de invasões a sistemas de computador.
- Por exemplo, se sabemos que tipo de compras uma pessoa normalmente faz no cartão de crédito, podemos determinar se uma nova cobrança no cartão se encaixa no perfil ou não. Podemos usar o grau de disparidade como uma pontuação suspeita e emitir um alarme, se for muito elevada.

Algumas Tarefas Fundamentais na Ciência de Dados

Previsão de vínculo tenta prever ligações entre itens de dados, geralmente sugerindo que um vínculo deveria existir e, possivelmente, também estimando a força do vínculo.

- A previsão de vínculo é comum em sistemas de redes sociais: "Como você e Karen compartilham 10 amigos, talvez você gostaria de ser amigo de Karen?" A previsão de vínculo também pode estimar a força de um vínculo.
- Por exemplo, para recomendar filmes para clientes pode-se imaginar um grafo entre os clientes e os filmes que eles já assistiram ou classificaram. No grafo, buscamos vínculos que não existem entre os clientes e os filmes, mas que prevemos que deveriam existir e deveriam ser fortes. Esses vínculos formam a base das recomendações.

Algumas Tarefas Fundamentais na Ciência de Dados

Redução de dados tenta pegar um grande conjunto de dados e substituí-lo por um conjunto menor que contém grande parte das informações importantes do conjunto maior. Pode ser mais fácil de lidar com ou processar um conjunto menor de dados. Além do mais, ele pode revelar melhor as informações.

- Por exemplo, um enorme conjunto de dados sobre preferências de filmes dos consumidores pode ser reduzido a um conjunto de dados muito menor revelando os gostos do consumidor mais evidentes na visualização de dados (por exemplo, preferências de gênero dos espectadores). A redução de dados geralmente envolve perda de informação. O importante é o equilíbrio para uma melhor compreensão.

Algumas Tarefas Fundamentais na Ciência de Dados

Modelagem causal tenta nos ajudar a compreender que acontecimentos ou ações realmente influenciam outras pessoas.

- Por exemplo, considere que usamos modelagem preditiva para direcionar anúncios para consumidores e observamos que, na verdade, os consumidores alvo compram em uma taxa mais elevada após terem sido alvo. Isso aconteceu porque os anúncios influenciaram os consumidores a comprar? Ou os modelos preditivos simplesmente fizeram um bom trabalho ao identificar os consumidores que teriam comprado de qualquer forma?
- Técnicas de modelagem causal incluem aquelas que envolvem um investimento substancial em dados, como experimentos randomizados controlados (por exemplo, os chamados "testes A/B"), bem como métodos sofisticados para obter conclusões causais a partir de dados observacionais.
- Um cuidadoso cientista de dados sempre deve incluir, com uma conclusão causal, os pressupostos exatos que devem ser feitos para que a conclusão causal se mantenha (essas suposições sempre existem — sempre pergunte).

Métodos Supervisionados vs Não-Supervisionados

Duas perguntas semelhantes

- Não-Supervisionado: Nossos clientes naturalmente se encaixam em grupos diferentes?
- Supervisionado:Podemos encontrar grupos de clientes que tenham probabilidades particularmente elevadas de cancelar seus serviços logo após o vencimento de seus contratos

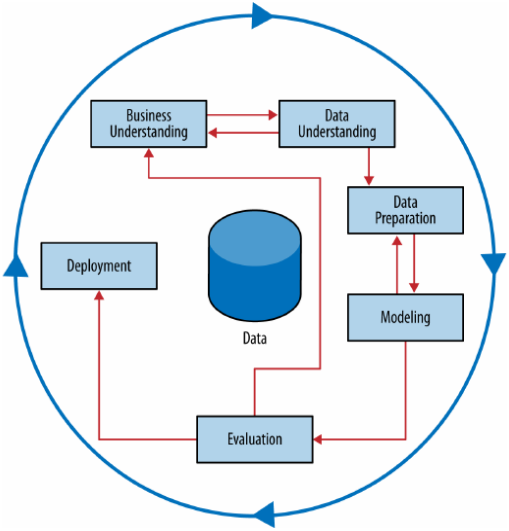
Métodos Supervisionados vs Não-Supervisionados

- Supervisionado: Classificação, regressão e modelagem causal
- Não-supervisionado: Combinação por similaridade, previsão de vínculo, redução dimensional, agrupamento e perfilamento.

O Processo de Mineração de Dados

A mineração de dados é uma arte. Ela envolve a aplicação de uma quantidade substancial de ciência e tecnologia, mas a aplicação adequada ainda envolve arte também. Mas, como acontece com muitas artes maduras, existe um processo bem compreendido que coloca uma estrutura no problema, permitindo **consistência, repetitividade e objetividade razoáveis**. Uma codificação útil do processo de mineração de dados é dada pelo Processo Padrão de Inter-Industrial para Exploração de Dados (CRISP-DM)

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

CRISP-DM: Compreensão do Negócio

- Inicialmente, é vital compreender o problema a ser resolvido.
- A formulação inicial pode não ser completa ou ideal, de modo que diversas repetições podem ser necessárias para que uma formulação de solução aceitável apareça.
- Nesta parte a análise da criatividade desempenha um grande papel.
- Tipicamente, os estágios iniciais do empreendimento envolvem projetar uma solução que tira proveito dos ferramentas discutidos anteriormente.
- O que exatamente queremos fazer? Como faríamos, exatamente? Quais partes deste cenário de uso constituem possíveis modelos de mineração de dados?

CRISP-DM: Compreensão dos Dados

- É importante entender os pontos fortes e as limitações dos dados porque raramente há uma correspondência exata com o problema.
- É comum os custos de dados variarem. Alguns dados estarão disponíveis praticamente de graça, enquanto outros exigirão um esforço para serem obtidos.
- No entendimento de dados, precisamos escavar a superfície e revelar a estrutura do problema de negócios e os dados que estão disponíveis e, em seguida, combiná-los a uma ou mais tarefas de mineração de dados para que possamos ter ciência e tecnologia substanciais para aplicar.

CRISP-DM: Preparação dos Dados

- Exemplos típicos de preparação de dados são sua conversão para o formato tabular, removendo ou inferindo valores ausentes, e convertendo dados para diferentes tipos.
- Os cientistas de dados podem passar um tempo considerável, no início do processo, definindo variáveis a serem utilizadas mais adiante.
- É importante durante a preparação dos dados ter cuidado com "vazamentos"

CRISP-DM: Modelagem

A etapa de modelagem é o principal local onde as técnicas de mineração de dados são aplicadas aos dados. É importante ter alguma compreensão das ideias fundamentais de mineração de dados, incluindo os tipos de técnicas e algoritmos existentes, porque esta é a parte da arte em que a maioria da ciência e da tecnologia podem ser exercidas.

CRISP-DM: Avaliação

- O objetivo da fase de avaliação é estimar os resultados de mineração de dados de forma rigorosa e obter a confiança de que são válidos e confiáveis antes de avançar.
- A fase de avaliação também serve para ajudar a garantir que o modelo satisfaça os objetivos de negócios originais.
- O cientista de dados deve pensar sobre a compreensibilidade do modelo para os investidores (não apenas para os cientistas de dados).
- Avaliação no ambiente real não é simples mas é essencial.

CRISP-DM: Implantação

- Na implantação, os resultados da mineração de dados são colocados em uso real, a fim de se constatar algum retorno sobre o investimento.
- Independentemente da implantação ser bem-sucedida, o processo muitas vezes retorna para a fase de compreensão do negócio.
- Não é necessário falhar na implantação para iniciar o ciclo novamente. A fase de avaliação pode revelar que os resultados não são bons o suficiente para implantação, e precisamos ajustar a definição do problema ou obter dados diferentes.

Implicações na Gestão da Equipe de Data Science

Embora a mineração de dados envolva software, ela também requer habilidades que podem não ser comuns entre os programadores.

- Em engenharia de software, a capacidade de escrever códigos eficientes e de alta qualidade a partir dos requisitos pode ser primordial. Os membros da equipe podem ser avaliados por meio de métricas de software, como a quantidade de código escrito ou o número de entradas de erros resolvidos.
- Em análise, é mais importante para os indivíduos serem capazes de formular bem os problemas, fazer rapidamente protótipos de soluções, fazer suposições razoáveis diante de problemas mal estruturados, projetar experimentos que representem bons investimentos e analisar os resultados. Na construção da equipe de data science, essas qualidades, e não a experiência tradicional em engenharia de software, são habilidades que devem ser buscadas.

Outras Técnicas e Tecnologias Analíticas

A análise de negócios envolve a aplicação de diversas tecnologias para a análise dos dados.

- Estatística: Descritiva e inferencial
- Consulta a Base de Dados: SQL e Processamento Analítico Online(OLAP)
  - SELECT * FROM CLIENTES WHERE IDADE $>$ 45 AND GENERO= 'M' AND DOMICILIO = 'NE'
- Armazenamento de Dados (Data Warehousing)
- Análise de Regressão
- Aprendizado de Máquina e Mineração de Dados

Utilização das Técnicas

Considere um conjunto de perguntas que possam surgir e as tecnologias que seriam adequadas para atendê-las.

- Quem são os clientes mais lucrativos? (Consulta a base de dados)
- Existe mesmo uma diferença entre os clientes lucrativos e o cliente mediano? (Estatística inferencial)
- Mas afinal, quem são esses clientes? Posso caracterizá-los? (Data science)
- Será que algum novo cliente em particular será lucrativo? Quanto rendimento eu devo esperar que esse cliente gere? (Data science)

Observe que as duas últimas perguntas são sutilmente diferentes sobre mineração de dados. A primeira, de classificação, pode ser expressa como uma previsão se um novo cliente será lucrativo (sim/não ou sua probabilidade). A segunda pode ser expressa como uma previsão do valor (numérico) que o cliente trará para a empresa.