CIÊNCIA DOS DADOS E A ANÁLISE PREDITIVA

Extrair conhecimento dos dados para tomar melhores decisões

ABSTRATO

Num mundo onde tudo acontece muito depressa e a mudança é constante, o que podemos fazer?

Porque a qualquer momento podem surgir novos concorrentes, novos produtos, novos clientes, a importância de manter uma vantagem competitiva é fundamental.

A ciência dos dados e a análise preditiva têm vindo a ser adotadas por cada vez mais organizações, sendo um requisito necessário para manter a vantagem competitiva.

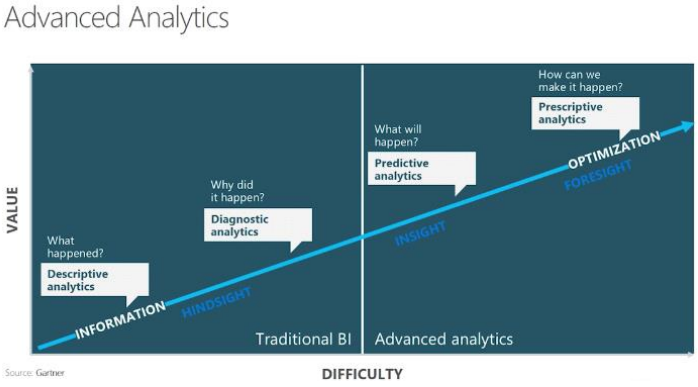
Mas quais os problemas que a análise preditiva pode ajudar a resolver? E como?

ANTÓNIO CRUZ

Ciência dos dados e análise preditiva

A ciência dos dados é o processo de extrair e examinar conjuntos de dados de forma a extrair conhecimento e tirar conclusões acerca da informação neles contida. A ciência dos dados e as suas técnicas são utilizadas nas organizações de forma a permitir a tomada de decisões informadas ou baseadas em factos.

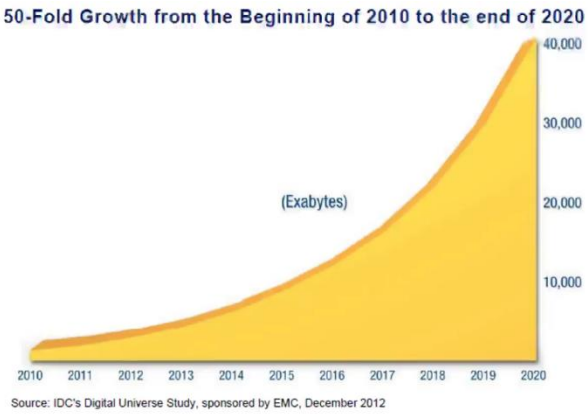
A ciência dos dados utiliza técnicas e teorias de diversos campos do conhecimento como a matemática, estatística, ciência da computação, ciências sociais, etc. E a análise preditiva é uma parte importante da ciência dos dados.



A Gartner® define a analítica avançada como a capacidade de prever o que vai acontecer e de que forma podemos influenciar acontecimentos futuros.

Apesar da atual elevada capacidade computacional e o seu baixo custo, a mesma não tem acompanhado o crescimento dos dados.

Estima-se que em 2005 o mundo produziu 130 *exabytes* de dados, em 2010 foram 1200 *exabytes* e em 2015 foram 7900. A EMC® prevê um crescimento exponencial nos próximos anos e que em 2020 sejam um valor próximo dos 40000 *exabytes*.



Não existem muitas coisas que se podem comparar a um *exabyte,* mas para termos uma ideia da enormidade do número, diz-se que 5 *exabytes* seriam equivalentes a todas as palavras faladas pela humanidade.

Não é possível, utilizando os métodos tradicionais, retirar todo o conhecimento útil desta quantidade de dados. E é aqui que entra a inteligência artificial e o *machine learning*. Só com métodos avançados e com capacidade de "aprenderem" sem ajuda humana, será possível vencer esse desafio.

Podemos definir análise preditiva como a tecnologia que aprende com base na experiência para prever o comportamento futuro.

Considera-se aqui que experiência é uma lista de dados históricos ou passados.

Machine learning é o motor por trás da análise preditiva. Os métodos (algoritmos) de *machine learning* processam dados e produzem modelos que permitem responder a diversas questões.

A definição mais simples de *machine learning* é que é o ramo da inteligência artificial que explora maneiras dos computadores melhorarem o seu desempenho com base na experiência, i.e., que dão aos computadores a capacidade de aprenderem sem serem explicitamente programados para isso.

Existem 3 grandes áreas onde a analítica preditiva e o *machine learning* podem fazer a diferença: aumentos das receitas, diminuição dos custos e na minimização dos riscos.

Mas quais os tipos de questões que a análise preditiva pode ajudar a responder e como?

Os algoritmos de *machine learning* podem ser agrupados em famílias em função do tipo de questão que pretendem responder.

Aprendizagem supervisionada

A aprendizagem supervisionada consiste na construção de uma função por parte do algoritmo de *machine learning* a partir de um conjunto de exemplos onde é conhecido o resultado que se pretende analisar.

Por exemplo, se eu tiver informação sobre as caraterísticas e comportamento de empresas que pediram dinheiro ao banco, e se tiver também a informação se no passado essas empresas deixaram, ou não, de cumprir com os planos de pagamento, os algoritmos de *machine learning* permitem construir um modelo, como base nesses exemplos, que podem prever quais as empresas que no futuro irão deixar de cumprir os seus planos de pagamento.

Isto é A ou B?

Esta família é formalmente conhecida como classificação de duas classes. É útil para responder a qualquer questão do tipo sim ou não, é ou não é, compra ou não compra. Muitas das questões no âmbito da ciência de dados podem ser respondidas desta forma. Alguns exemplos:

- Vai este cliente renovar o seu contrato?
- Vai o equipamento falhar nos próximos 30 dias?
- Se eu baixar o preço de venda em 5%, vou aumentar as minhas vendas?
- O email é spam?

Isto é A, B, C ou D?

Esta família é conhecida como classificação multi-classes. É útil para responder a uma pergunta cuja resposta pode ter mais de duas possibilidades: que sabor, que pessoa, que parte, que empresa, que candidato? Esta família é maioritariamente uma extensão da classificação de duas classes. Alguns exemplos de questão são:

- Que carater é este (OCR - *optical character recognition*)?

- Qual a musica que está a tocar?
- Quem é que está a falar?
- Qual o tópico do artigo?

Que quantidade ou que valor?

Quando a pergunta que fazemos é respondida por um numero, e não uma categoria ou conjunto como nos exemplos anteriores, os algoritmos pretendem responder a questões do tipo:

- Quantas refeições vou vender na próxima semana?
- Qual vai ser a temperatura na próxima sexta-feira?
- Quais vão ser as vendas no quarto trimestre?
- Quantos novos seguidores do Facebook irei ter na próxima semana?
- Quantos trabalhadores vão faltar no próximo mês?

Aprendizagem não supervisionada

Nesta família de algoritmos os modelos não são treinados com exemplos, i.e., eles à partida não sabem o que estamos à procura ou quais as categorias que constituem o nosso contexto. O que diferencia este conjunto de técnicas da aprendizagem supervisionada é o facto de não existir previamente um numero ou uma classe que nos indique a que grupo aquela observação pertence, o que representa esse grupo ou mesmo quantos grupos devem existir.

Isto é esquisito?

Esta família de algoritmos é conhecida como deteção de anormalidades. Têm como objetivo a identificação de dados que não são normais.

O papel do algoritmo é detetar quais os comportamentos anómalos existentes no conjunto dos nossos dados. A utilidade deste tipo de algoritmos é grande e alguns exemplos da sua possível utilização são:

- Existe um tráfego anómalo na rede (pode significar vírus ou ataques)?
- As leituras de pressão de um determinado equipamento são anómalas?
- A compra deste cliente é anómala?

Como estão os dados organizados?

Estamos a falar de aprendizagem não supervisionada. Existem um conjunto de técnicas que pretendem agrupar dados, sem a existência de exemplos anteriores, com base na distância entre as observações.

Um dos exemplos mais comuns é o cluster, que também podemos chamar de segmentação.

Estas técnicas pretendem separar as observações em grupos naturais, para permitir tratar esses grupos de forma mais homogénea.

Alguns exemplos de questão são:

- Quais os compradores com gostos semelhantes?
- Quais os espetadores que gostam dos mesmos filmes?
- Quais as impressoras que apresentam o mesmo tipo de problemas?
- Quais os dias ou horas da semana que apresentam padrões de consumo semelhantes?
- Em quantos tópicos devemos separar um conjunto de documentos de texto?

Outro conjunto de técnicas de aprendizagem não supervisionada é a redução de dimensionalidade. Mais uma vez pretendemos simplificar os dados de forma a tornar mais fácil a sua análise, a explicar comportamentos, comunicar conclusões, ocupar menos espaço de armazenamento e aumentar a velocidade de processamento.

Basicamente estamos a falar em reduzir o número de variáveis que explicam padrões ou comportamentos. Por exemplo, é relativamente normal termos centenas ou milhares de variáveis de análise na indústria automatizada, na área financeira, na análise de redes sociais, etc.

O grande número de variáveis independentes (explicativas) a multiplicar pelo número de observações tornam as computações (dependentes de intensos cálculos matriciais) muito demoradas. Ao reduzir o número de variáveis, das centenas para poucas dezenas ou menos, sem perder muito poder explicativo, as vantagens são tremendas.

É também impossível ao ser humano conseguir entender dezenas ou centenas de variáveis, ou explicá-las a terceiros. Reduzir essas variáveis para 2 ou 3 torna tudo mais fácil.

Alguns exemplos:

- Quais as variáveis que explicam comportamentos nas redes sociais tendem a mudar da mesma forma?
- Quais os padrões mais comuns nas variáveis que explicam alterações nos preços de produtos?
- Quais o conjunto de palavras que tendem a ocorrer conjuntamente nos documentos?

O que devo fazer?

Um outro conjunto de técnicas chamada de reforço de aprendizagem, são um pouco diferentes da aprendizagem supervisionada e não supervisionada. Podemos prever com uma regressão ou previsão de séries temporais qual vai ser a temperatura de amanhã, mas isso não nos diz o que fazer. As técnicas de reforço de aprendizagem vão um passo mais longe e propõem ou escolhem uma ação, dentro de um espaço pré-determinado de ações.

Este conjunto de técnicas foi inicialmente desenvolvido baseada em como os cérebros dos ratos e humanos reagem ao prémio ou punição. O algoritmo escolhe a técnica que maximiza o prémio, pelo que ele deve saber quais as possíveis ações antes de decidir e

precisam de ter um feedback da ação que tomaram, e ela foi asneira, espetacular ou assim-assim.

Estas técnicas são ótimas para situações onde é preciso tomar rapidamente um conjunto muito grande de pequenas ações, que tornam a intervenção humana difícil ou inviável.

Elevadores, controlo de luminosidade, temperatura, etc., são ótimos candidatos.

Exemplos:

- Em que sítio da minha página web devo colocar a publicidade de forma a maximizar os cliques nela?
- Quantas ações desta empresa devo comprar agora?
- Acelero, travo ou mantenho a velocidade perante o sinal amarelo?

Esta área tem vindo recentemente a ter mais atenção e requer normalmente um maior esforço para ser implementada porque implica uma interação grande com variados sistemas.

O interessante é que estes sistemas podem começar a funcionar sem dados e conforme vão recolhendo e aprendendo, vão evoluindo.

Conclusões

A informação é a principal matéria-prima para qualquer organização seja ela pública ou privada, industrial, comercial ou de serviços.

O conhecimento que resulta da informação é uma vantagem competitiva. Um diferenciador estratégico. Para qualquer organização.

Utilizar a informação para gerar conhecimento e decidir baseado em factos já não é uma opção, é um requisito.

Faça sentido dos dados. Use a analítica preditiva para tomar boas decisões.

Este artigo teve como inspiração um *post* no blog de *machine learning* da Microsoft® (Brandon Rohrer, 2015).

Referências

Brandon Rohrer, S. D. (27 de August de 2015). *What Types of Questions Can Data Science Answer?* Obtido de Cortana Intelligence and Machine Learning Blog: https://blogs.technet.microsoft.com/machinelearning/2015/08/27/what-types-of-questions-can-data-science-answer/