Jorge Buescu – Professor na Faculdade de Ciências da Universidade de Lisboa

O CAMPEONATO DOS CÉREBROS

Como fazer da ciência

É um lugar-comum dizer que vivemos numa Sociedade da Informação. Hoje em dia, há mais dados instantaneamente acessíveis sobre qualquer assunto do que alguma vez na história da Humanidade. O que gera um problema radicalmente novo: como é possível sobreviver a este tsunami de informação sem que ele nos afogue? Como é possível extrair conhecimento útil a partir do oceano de dados em que estamos mergulhados? Como é possível transformar a Sociedade da Informação numa verdadeira Sociedade do Conhecimento?

Embora possam parecer vagas, estas perguntas são tudo menos ociosas. Suponha por exemplo o leitor que pertence ao Departamento de Análise de uma grande cadeia de supermercados. Tem ao seu dispor dezenas de milhões de dados sobre os padrões de consumo dos seus clientes. O seu objectivo é prever o comportamento futuro de cada cliente, tentando, a partir da imensidão dos dados, deduzir que o cliente A está fidelizado ou que o cliente B precisa de algum incentivo para voltar, talvez cupões de desconto.

Este tipo de fenómeno ocorre hoje em dia com maior frequência em diferentes contextos. Podem ser supermercados, empresas de cartões de crédito, bancos, seguradoras, problemas científicos ou mesmo companhias de apostas: há cada vez mais situações em que um dilúvio de dados torna muito difícil a construção de um modelo de previsão a médio prazo. Por outro lado, a existência de modelos sofisticados de evolução a médio prazo é extraordinariamente útil: permite a um banco, por exemplo, recusar um empréstimo a um cliente se o seu risco de incumprimento for demasiado grande, ou a uma seguradora aumentar o prémio de seguro por risco acrescido. Mas mesmo estas decisões têm de ser tomadas de forma muito selectiva e tendo em conta os dados: nem os bancos nem as seguradoras querem perder os seus clientes levando-os a bater à porta da concorrência!

Eis o problema, então: como é que, perante uma gigantesca massa de dezenas ou centenas de milhões de dados, que descrevem a evolução de um sistema – financeiro, industrial, comercial, de gestão – conseguimos construir o melhor modelo matemático para previsão da evolução futura do sistema, de forma a antecipar as tendências?

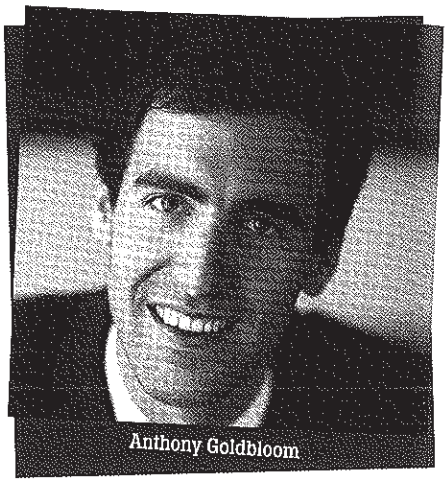
Classicamente, as grandes empresas para quem estes resultados eram importantes provavelmente fariam a sua modelação a partir dos seus departamentos de análise. Com a actual – e crescente – inundação de informação, muitas destas empresas voltam-se para o exterior, encomendando soluções a companhias externas, especializadas neste tipo de modelação matemática preditiva.

No entanto, este facto gera um problema bastante grave: há inúmeras abordagens diferentes a problemas de previsão. Talvez uma destas companhias propusesse uma solução baseada em máquinas de apoio vectorial; outra talvez em redes neuronais... como é que o comprador de uma destas soluções poderia saber *a priori* qual a metodologia mais apropriada para o seu caso? Sobretudo se cada proposta custasse da ordem de um milhão de dólares?

E aqui entra uma ideia tão luminosa quanto prometedora: o Kaggle.

Em 2010 o australiano Anthony Goldbloom transformou aquilo que era um problema numa oportunidade. Se modelos matemáticos diferentes dão respostas diferentes com base nos mesmos dados, a solução é óbvia: *organizar uma competição entre eles para ver qual é o melhor*.

O Kaggle, uma companhia *start-up* criada por Goldbloom (www.kaggle.com), é, nem mais nem menos, uma plataforma *online* que promove competições que seleccionam, para cada problema, o melhor modelo possível. Um verdadeiro campeonato de cérebros.



Anthony Goldbloom

Eis como funciona: as empresas e investigadores disponibilizam os seus dados no Kaggle, abrindo uma competição com um prazo bem definido e um prémio financeiro para as melhores soluções. O concurso é aberto à participação de todos os analistas, cientistas e estatísticos que pretendam registar-se, individualmente ou em equipa, e ao longo da competição submetem ao Kaggle as suas propostas.

Durante a competição os resultados são afixados em tempo real, para que cada equipa possa saber a sua classificação durante o con-

a dos dados um desporto!

curso. É permitido a uma equipa submeter novas versões de um modelo, de forma que, assim que uma equipa entra no concurso, tem sempre o incentivo de melhorar o seu modelo (além de saber instantaneamente quando e por quem foi ultrapassado). No final do concurso ganha a equipa que tiver atingido o primeiro lugar. O prémio (em geral, financeiro) será dela; o modelo produzido passará para a posse da empresa que patrocinou a competição.

O próprio sistema de classificação na competição utiliza dados reais. Por exemplo, se uma empresa possui dados correspondentes a 24 meses de evolução do problema em questão, disponibiliza apenas os primeiros 18 meses, comparando a "previsão" de cada modelo para os seis meses seguintes com o que se passou na realidade. Quanto mais próxima a previsão estiver da realidade melhor será o modelo.

Eis um exemplo do que já aconteceu no Kaggle. Em 2010, um académico da Universidade de Drexel, na Filadélfia, propôs uma competição para a criação de um modelo para a progressão do VIH. Durante os três meses da competição, 109 equipas de cientistas usaram os registos de mil doentes para prever a evolução da doença de acordo com a constituição genética. No final, o modelo vencedor revelou ter uma eficiência de 77%, em comparação com 70% dos modelos convencionais.

Entre os clientes do Kaggle encontram-se fontes tão diversas como a Deloitte (consultadoria) e a NASA, que abriu uma competição com o objectivo de modelar as possibilidades de detecção de matéria escura: o vencedor foi um glaciologista que de outra forma nunca se teria sequer dedicado ao problema. Entre outras competições contam-se, por exemplo, uma companhia de automóveis usados que pretende saber a probabilidade de um carro adquirido num leilão ser uma má compra e de um banco que pretende modelar a probabilidade de um cliente pedir um empréstimo nos próximos dois anos.

A maior das competições até hoje, que permanecerá em aberto até 3 de Abril de 2013 e tem um prémio final de 3 milhões de dólares, tem o nome de Heritage Provider Network Health Prize. A descrição da competição é a seguinte: "Mais de 71 milhões de indivíduos dão entrada em hospitais nos EUA em cada ano. Vários estudos concluíram que em 2006 foram gastos mais de 30 mil milhões de dólares em internamentos hospitalares desnecessários. Existirá uma melhor forma de lidar com a situação? Poderemos identificar os pacientes mais em risco e dar-lhes o tratamento de que precisam antes de surgir a necessidade de internamento? A Heritage Provider Network acredita que sim". Existem, na altura da escrita, 1.344 equipas e 9.255 entradas concorrentes.

O Kaggle tem características únicas e fascinantes, entre as quais está o facto de todos ganharem com este processo.

Em primeiro lugar, ganham as equipas concorrentes – não apenas as que ganham a competição, mas também todas as outras, porque têm uma oportunidade única de treinar e desenvolver as suas competências de análise de dados e construção de modelos perante dados do mundo real, enriquecendo assim a sua experiência profissional. Mesmo uma menção honrosa no Kaggle é já hoje algo que se deve colocar no *Curriculum Vitae*. Isto, é claro, para além da adrenalina da competição intelectual. Como diz a página de entrada do Kaggle, "estamos a fazer da ciência dos dados um desporto".

Em segundo lugar, ganham as empresas que propõem as competições. Propondo um desafio no Kaggle, têm acesso a um universo de cerca de 17 mil cientistas e analistas de dados, espalhados, literalmente, por todo o Mundo, muitos deles de grande nível. Nunca até hoje um desafio lançado no Kaggle deixou de melhorar os melhores modelos disponíveis, o que só em si é significativo.

E, finalmente, ganha o próprio Kaggle – por enquanto uma pequena empresa com meia dúzia de funcionários, que já teve de mudar da Austrália para Silicon Valley e atraiu em Novembro de 2011 um financiamento de 11 milhões de dólares.

A plataforma Kaggle é aberta, pelo que o leitor é livre de participar no Kaggle de ambas as formas: ou formando uma equipa de modelação para entrar numa, ou em várias competições em curso; ou, se dispuser de um problema com uma quantidade gigantesca de dados que quisesse ver tratado, propondo uma nova competição (há, naturalmente, vários graus de confidencialidade que podem ser negociados pelos clientes do Kaggle).

O Kaggle introduziu um novo paradigma na ciência da análise de dados. Os investidores reconhecem-lhe claramente grande potencial. Conseguirá ele transformar esta excitação num crescimento sustentado e significativo?

Ninguém sabe. Talvez as equipas de modeladores do Kaggle possam responder!