Introdução à Inteligência Artificial

Uma abordagem não técnica

Tom Taulli

novatec

Apress®

Introdução à Inteligência Artificial

Uma abordagem não técnica

Tom Taulli

Apress®

Novatec

São Paulo | 2020

First published in english under the title Artificial Intelligence Basics; A Non-Technical Introduction by Tom Taulli, edition: 1

Cover designed by eStudioCalamar.

Publicação original em inglês intitulada Artificial Intelligence Basics; A Non-Technical Introduction por Tom Taulli, edição: 1

Capa desenvolvida por eStudioCalamar

Editor: Rubens Prates

Tradução: Luciana do Amaral Teixeira Revisão gramatical: Tássia Carvalho Editoração eletrônica: Carolina Kuwabata

ISBN: 978-85-7522-XXX

Histórico de edições impressas:

Janeiro/2020 Primeira edição

Novatec Editora Ltda.

Rua Luís Antônio dos Santos 110
02460-000 – São Paulo, SP – Brasil Tel.: +55 11 2959-6529
Email: novatec@novatec.com.br
Site: www.novatec.com.br
Twitter: twitter.com/novateceditora
Facebook: facebook.com/novatec
LinkedIn: linkedin.com/in/novatec

Sumário

Sobre o autor

Tom Taulli desenvolve software desde os anos 1980. Na faculdade, fundou sua primeira empresa, voltada ao desenvolvimento de sistemas de e-learning. Criou também outras empresas, incluindo a Hypermart.net, vendida para a InfoSpace em 1996. Ao longo de sua jornada, Tom escreveu colunas para publicações online como businessweek.com, techweb.com e Bloomerang.com. Atualmente, escreve sobre inteligência artificial na Forbes.com e atua como consultor de várias empresas na mesma área. É possível contactá-lo pelo Twitter (@ttaulli) ou em seu website (`www.taulli.com`).

Prefácio

Como este livro demonstra, a adoção da inteligência artificial (IA) será um importante ponto de inflexão na história da humanidade. Assim como outras tecnologias igualmente inovadoras, a forma como é administrada e aqueles com acesso a ela vão moldar a sociedade pelas próximas gerações. Contudo, a IA se destaca em relação a outras tecnologias inovadoras dos séculos 19 e 20 – pense em máquina a vapor, rede elétrica, genômica, computadores e Internet – porque não depende apenas de uma infraestrutura física criticamente cara para ser adotada; afinal, muitos de seus benefícios podem ser disponibilizados por meio de hardware já existente e que carregamos por aí nos bolsos. O principal fator limitante no tocante à adoção em massa da tecnologia da IA, entretanto, é nossa infraestrutura compartilhada: educação, compreensão e visão.

Essa é uma diferença crucial, pois, se corretamente administrada, a IA pode atuar como uma força democratizante arrebatadora. Ela vem eliminando de nossas vidas o trabalho árduo do passado e vai liberar uma quantidade enorme de energia e capital humanos. Esse "se", entretanto, está longe de ser certo. Implementada de maneira irresponsável, a IA tem o poder de desestabilizar grandes partes da economia mundial, causando, como muitas pessoas temem, um encolhimento na força de trabalho, uma diminuição no poder aquisitivo da classe média e uma economia sem base ampla e estável e que é alimentada por uma espiral de dívida infinita.

No entanto, antes de sucumbir ao pessimismo relacionado à IA, é preciso dar uma olhada no passado. Embora a capacidade de transformação da IA possa ser histórica – e ela é –, essas mesmas questões têm sido discutidas no cenário econômico há décadas ou mesmo séculos. A IA é, afinal, uma extensão de uma tendência para a automação que está em cena desde Henry Ford. Na verdade, a própria Zoho nasceu da tensão entre a automação e os princípios econômicos igualitários. De volta ao início dos anos 2000, chegamos a uma percepção que moldou nossa abordagem à tecnologia: pessoas comuns – proprietários de pequenas empresas, aqui e no exterior – devem ter acesso às mesmas automações avançadas de negócios que as empresas da Fortune 500 têm; caso contrário, uma enorme faixa da população será excluída da economia.

Naquela época, o poderoso software era quase unanimemente fechado por trás de contratos rígidos, com preços exorbitantes e implementações complicadas. Grandes empresas poderiam assumir o fardo de tais sistemas, enquanto os operadores

menores eram excluídos; o que os deixava em tremenda desvantagem. Procuramos interromper essa situação com a promessa de tecnologia para um público cada vez mais amplo. Nas últimas duas décadas, nos esforçamos para aumentar o valor de nossos produtos sem aumentar o preço, explorando a escalabilidade da tecnologia da nuvem. Nosso objetivo é capacitar pessoas em todos os níveis da sociedade empurrando para baixo o preço do software de negócios enquanto ampliamos o poder das ferramentas. O acesso ao capital não deve limitar o sucesso; as empresas devem crescer ou diminuir com base na força de sua visão para o futuro.

Vista dessa forma, a IA representa o cumprimento da promessa da tecnologia, liberando as pessoas das restrições de tempo e lhes permitindo uma liberação do tedioso ou desagradável trabalho de rotina. Ela ajuda a identificar padrões em escalas microscópicas e macroscópicas às quais os seres humanos não estão naturalmente adaptados para perceber. Ela pode prever problemas e corrigir erros; além de economizar dinheiro, tempo e até mesmo vidas.

Buscando democratizar esses benefícios, assim como foi feito para o software de negócios em geral, a Zoho tem incluído IA em todo seu conjunto de aplicativos. Passamos os últimos seis anos desenvolvendo silenciosamente nossa própria tecnologia interna de IA, construída sobre a base de nossos próprios princípios. O resultado é Zia, um assistente de IA que é inteligente, mas não esperto. Essa é uma distinção crucial. Um sistema inteligente dispõe de informações e funcionalidades que empoderam a visão e intuição únicas de um operador ativo. Um sistema esperto esconde o funcionamento interno do processo, reduzindo o ser humano a um usuário passivo que simplesmente consome as informações dadas pela máquina. A IA deve ser uma ferramenta a ser utilizada, não uma lente por meio da qual vemos o mundo. Para lidar com uma ferramenta tão poderosa, devemos estar equipados com o conhecimento para compreendê-la e operá-la sem que se comprometa a qualidade humana dos sistemas humanos.

A necessidade de se manter atualizado sobre essa tecnologia é exatamente a razão para que um livro com noções básicas de inteligência artificial seja tão importante no mundo de hoje. É a infraestrutura intelectual que permitirá às pessoas – pessoas normais – explorar o poder da IA. Sem esse tipo de iniciativa, a IA abala o equilíbrio de poder em favor de grandes empresas com grandes orçamentos. É crucial que a população em geral se equipe com as habilidades para entender os sistemas de IA, porque eles definirão cada vez mais como interagimos e navegamos pelo mundo. Em breve, a informação contida neste livro não será meramente um tema de interesse, mas um pré-requisito para a participação na economia moderna.

É assim que uma pessoa comum pode apreciar os frutos da revolução da IA. Nos próximos anos, a forma como definimos o trabalho e as atividades que carregam valor econômico mudará. Temos de admitir o fato de que o futuro do trabalho pode

ser tão estranho para nós quanto seria o trabalho administrativo para nossos antepassados distantes. Contudo, precisamos – e devemos – ter fé na capacidade humana de inovar nas formas de trabalho, mesmo que esse trabalho não se pareça com aquele com o qual estamos familiarizados. O primeiro passo, antes de tudo, é aprender mais sobre essa nova, excitante e fundamentalmente democrática tecnologia.

– Sridhar Vembu, cofundador e CEO da Zoho

Introdução

Para um usuário, o aplicativo Uber é simples. Com apenas alguns cliques, é possível chamar um motorista em poucos minutos.

Nos bastidores, entretanto, existe uma avançada plataforma tecnológica que depende muito da inteligência artificial (IA). A seguir estão listados apenas alguns dos recursos:

- Sistema de NLP (Natural Language Processing – Processamento de Linguagem Natural) capaz de compreender conversas, permitindo uma experiência simplificada.
- Software de visão computacional que verifica milhões de imagens e documentos como licenças de motoristas e cardápios de restaurante.
- Algoritmos de processamento de sensores que ajudam a melhorar a precisão em áreas urbanas densas, incluindo detecção automática de acidentes por meio da identificação de movimentos inesperados do telefone de um motorista ou passageiro.
- Algoritmos sofisticados de machine learning (aprendizado de máquina) que preveem disponibilidade de motoristas, demanda de corridas e previsão de chegada.

Essas tecnologias não são somente surpreendentes, mas também necessárias. Não havia como a Uber experimentar seu crescimento – que envolveu o gerenciamento de mais de 10 bilhões de viagens – sem a IA. Diante disso, não deve ser surpresa que a empresa gaste centenas de milhões nessa tecnologia e tenha um grande grupo de especialistas em inteligência artificial na equipe.[1]

A IA, entretanto, não é apenas para startups como a Uber. A tecnologia também está se mostrando uma prioridade crítica para empresas tradicionais. Basta olhar para o McDonald's. Em 2019, a empresa investiu $300 milhões de dólares para adquirir uma startup de tecnologia, a Dynamic Yield. Foi o maior negócio da empresa desde que comprou o Boston Market em 1999[2].

A Dynamic Yield, fundada em 2011, é pioneira em alavancar a IA para criar interações personalizadas com o cliente na web, aplicativos e email. Alguns de seus clientes incluem Hallmark Channel, IKEA e Sephora.

Voltando ao McDonald's, a empresa tem se submetido a uma transformação digital – e a IA é parte fundamental da estratégia. Com a Dynamic Yield, a rede planeja usar a tecnologia para redesenhar seu drive-thru, que responde pela maior parte de

sua receita. Analisando dados como tempo, tráfego e hora do dia, os menus digitais serão alterados dinamicamente para maximizar as oportunidades de receita. Parece ainda que o McDonald's usará geocerca (geofencing)[3] e até mesmo reconhecimento da imagem das placas dos veículos para aprimorar a personalização.

Mas isso é apenas o começo. A rede de fast-food espera usar a IA para quiosques e sinalização em lojas, bem como para cadeia de suprimentos.

A empresa percebe que o futuro é tanto promissor quanto perigoso. Se as organizações não forem proativas com relação às novas tecnologias, podem acabar fracassando. Basta olhar para a forma como a Kodak foi lenta na adaptação às câmeras digitais ou para como a indústria de táxi não mudou quando confrontada com a chegada da Uber e da Lyft.

Por outro lado, as novas tecnologias podem ser quase um elixir para uma empresa. É preciso, contudo, que exista estratégia sólida, boa compreensão do que é possível e vontade de assumir riscos. Então, neste livro, vou fornecer ferramentas para ajudar com tudo isso.

OK, então, quão grande vai ser a IA? De acordo com um estudo da PWC, ela vai adicionar impressionantes $15,7 trilhões de dólares ao PIB global até 2030, o que é mais do que a produção da China e da Índia juntas. Os autores do relatório ressaltam: "A IA relaciona-se a quase todos os aspectos de nossas vidas e ela está só começando"[4].

É verdade, quando se trata de prever tendências, pode haver uma boa dose de exagero. No entanto, a IA pode ser diferente porque tem o potencial de se transformar em uma tecnologia de uso geral. Um paralelo a isso é o que aconteceu no século 19 com o surgimento da eletricidade, que teve um impacto transformador em todo o mundo.

Como sinal da importância estratégica da IA, empresas de tecnologia como Google, Microsoft, Amazon.com, Apple e Facebook fizeram investimentos substanciais nesse setor. Por exemplo, o Google se autodenomina uma empresa "AI-First" (IA primeiro) e gastou bilhões comprando empresas do ramo, bem como contratando milhares de cientistas de dados.

Em outras palavras, mais e mais empregos exigirão conhecimentos em IA. Contudo, isso não significa que você precisará aprender linguagens de programação ou entender estatísticas avançadas. Será fundamental, entretanto, ter uma base sólida dos fundamentos.

Neste livro, o objetivo é fornecer conselhos práticos que podem fazer grande diferença em sua empresa e carreira. Você não encontrará explicações profundamente técnicas, trechos de código ou equações. Em vez disso, este livro traz respostas para as principais perguntas que os gerentes têm: onde IA faz sentido?

Quais são as pegadinhas? Como você avalia a tecnologia? Que tal começar um plano piloto de IA?

Este livro também apresenta uma visão de mundo real da tecnologia. Uma grande vantagem de ser escritor da Forbes.com e consultor no mundo da tecnologia é que tenho a oportunidade de falar com muitas pessoas talentosas no campo da IA – e isso me ajuda a identificar o que é realmente importante no setor. Também tenho acesso a estudos de caso e exemplos do que funciona.

O livro está organizado de forma a abranger os principais tópicos da IA – e não é necessário ler os capítulos na ordem. Considere esta publicação como um manual.

A seguir estão breves descrições dos capítulos:

- *Capítulo 1 – Fundamentos da IA*: esta é uma visão geral da rica história da IA, que remonta à década de 1950. Você aprenderá sobre pesquisadores brilhantes e cientistas da computação como Alan Turing, John McCarthy, Marvin Minsky e Geoffrey Hinton. Também serão abordados conceitos-chave, como o teste de Turing, que verifica se uma máquina atingiu a verdadeira IA.
- *Capítulo 2 – Dados*: os dados são a força vital da IA. Com eles, os algoritmos conseguem encontrar padrões e correlações para fornecer informações. Entretanto, há minas terrestres nos dados, como qualidade e tendência. Esse capítulo oferece um framework para o trabalho com dados em um projeto de IA.
- *Capítulo 3 – Machine learning*: este é um subconjunto da IA que envolve técnicas estatísticas tradicionais, como regressões. Nesse capítulo, contudo, também abordaremos algoritmos avançados, como o k-NN (k-Nearest Neighbor – k-ésimo vizinho mais próximo) e o classificador Naive Bayes. Além disso, daremos uma olhada em como montar um modelo de machine learning (aprendizado de máquina).
- *Capítulo 4 – Deep learning*: trata-se de outro subconjunto da IA e é claramente o que tem visto grande parte da inovação durante a última década. Deep learning (aprendizado profundo) aborda o uso de redes neurais para encontrar padrões que imitam o cérebro. Nesse capítulo, daremos uma olhada nos principais algoritmos, como redes neurais recorrentes (RNNs – Recurrent Neural Networks), redes neurais convolucionais (CNNs – Convolutional Neural Networks) e redes adversárias generativas (GANs – Generative Adversarial Networks). Também discutiremos conceitos-chave, como retropropagação (backpropagation).
- *Capítulo 5 – Automação robótica de processos*: aqui discutimos como são utilizados sistemas para automatização de procedimentos repetitivos, como a inserção de dados em um sistema de CRM (Customer Relationship Management). A automação robótica de processos (RPA – Robotic Process

Automation) tem experimentado um enorme crescimento durante os últimos anos por causa do alto ROI (Return On Investment – retorno sobre investimento). Essa tecnologia também tem sido uma forma introdutória para que as empresas implementem a IA.

- *Capítulo 6 – Natural Language Processing (NLP)*: essa forma de IA, que envolve a compreensão de conversas, é a mais onipresente e pode ser encontrada em recursos como Siri, Cortana e Alexa. Entretanto, os sistemas de NLP (processamento de linguagem natural), como os chatbots, também se tornaram críticos no mundo corporativo. Esse capítulo mostra maneiras de usar essa tecnologia de maneira eficaz e discute como evitar questões complicadas.
- *Capítulo 7 – Robôs físicos*: a IA está começando a ter um grande impacto sobre essa indústria. Com deep learning, está ficando mais fácil para os robôs compreenderem seus ambientes. Nesse capítulo, vamos dar uma olhada tanto nos robôs domésticos quanto nos industriais, bem como em uma infinidade de estudos de caso.
- *Capítulo 8 – Implementação da IA*: vamos abordar passo a passo a criação de um projeto de IA, desde o conceito inicial até sua implantação. Esse capítulo também abordará diversas ferramentas – como Python, TensorFlow e PyTorch.
- *Capítulo 9 – O futuro da IA*: esse capítulo discutirá algumas das maiores tendências da IA, como carro autônomo, militarização da IA, desemprego causado pelas mudanças tecnológicas, desenvolvimento e aprovação de novos medicamentos.

Ao final do livro, você encontrará um apêndice de recursos para estudos futuros e um glossário de termos comuns relacionados à IA.

Material de acompanhamento

Todas as atualizações serão fornecidas no meu site em `www.Taulli.com`.

---

1 `www.sec.gov/Archives/edgar/data/1543151/000119312519120759/d647752ds1a.htm#toc647752_11`

2 `https://news.mcdonalds.com/news-releases/news-release-details/dynamic-yieldacquisition-release`

3 N.T.: Geocerca é um serviço de aplicativo que usa GPS ou RFID para definir uma área geográfica (perímetro virtual) que acionará uma resposta quando um dispositivo entrar ou sair dessa área, como enviar notificações push, disparar mensagens de texto e alertas.

4 `www.pwc.com/gx/en/issues/data-and-analytics/publications/artificialintelligence-study.html`

CAPÍTULO 1

Fundamentos da IA

Lições de história

Inteligência artificial seria a versão final do Google. A máquina de busca definitiva, que compreenderia tudo na web. Entenderia exatamente o que você deseja e lhe daria a coisa certa. Não estamos nem perto de fazer isso agora. No entanto, podemos ficar cada vez mais perto disso; o que é basicamente no que estamos trabalhando.

– Larry Page, cofundador da Google Inc. e CEO da Alphabet[1]

No conto "Resposta", escrito por Fredric Brown em 1954, todos os computadores dos 96 bilhões de planetas no universo estão conectados a uma supermáquina. Perguntaram, então, "Existe um Deus?", e ele respondeu: "Sim, agora existe um Deus".

Não há dúvidas de que a história de Brown foi certamente inteligente – bem como um pouco cômica e arrepiante! A ficção científica tem sido uma forma de compreender as implicações das novas tecnologias, e a inteligência artificial (IA) tem sido um tema importante. Alguns dos personagens mais memoráveis na ficção científica envolvem androides ou computadores que se tornam autoconscientes, como nos filmes *Exterminador, Blade Runner, 2001: uma odisseia no espaço* e até mesmo *Frankenstein*.

No entanto, com o ritmo implacável de novas tecnologias e inovação hoje em dia, a ficção científica está começando a se tornar real. Agora, podemos conversar com nossos smartphones e obter respostas, nossas contas de mídia social nos mostram o conteúdo pelo qual estamos interessados, nossos aplicativos bancários nos fornecem lembretes e muito mais. Essa criação de conteúdo personalizado quase parece mágica, mas está rapidamente se tornando normal em nossa vida cotidiana.

Para compreender a IA, é importante conhecer o básico de sua rica história. Você vai ver como o desenvolvimento dessa indústria tem sido cheio de avanços e contratempos. Há também um elenco de pesquisadores e acadêmicos brilhantes que ampliou os limites da tecnologia, como Alan Turing, John McCarthy, Marvin Minsky e Geoffrey Hinton. Por meio de tudo isso, houve progresso constante.

Vamos começar.

Alan Turing e o teste de Turing

Alan Turing é uma figura imponente em ciência da computação e IA; tanto que muitas vezes é chamado de "pai da IA".

Em 1936, escreveu um artigo chamado "On Computable Numbers" ("Sobre números computáveis"), no qual estabeleceu os conceitos fundamentais de um computador e que se tornou conhecido como a máquina de Turing. Lembre-se de que os verdadeiros computadores não seriam desenvolvidos até mais de uma década mais tarde.

No entanto, foi seu artigo "Computing Machinery and Intelligence" ("Máquinas computacionais e inteligência") que se tornou histórico para a IA. Ele se concentrou no conceito de uma máquina que era inteligente. Para fazer isso, entretanto, era necessário existir uma maneira de avaliá-la. O que é inteligência – pelo menos para uma máquina?

Foi aqui que ele teve a ideia do famoso "teste de Turing". Trata-se essencialmente de um jogo com três participantes: dois humanos e um computador. O avaliador, um humano, faz perguntas abertas aos outros dois (um humano, um computador) com o objetivo de determinar qual deles é o humano. Se o avaliador não puder fazer distinção, presume-se que o computador é inteligente. A Figura 1.1 mostra o fluxo básico de funcionamento do teste de Turing.

O que é genial nesse conceito é que não há necessidade de verificar se a máquina realmente sabe algo, é autoconsciente ou mesmo se está correta. Em vez disso, o teste de Turing indica que uma máquina pode processar grandes quantidades de informações, interpretar a fala e comunicar-se com seres humanos.

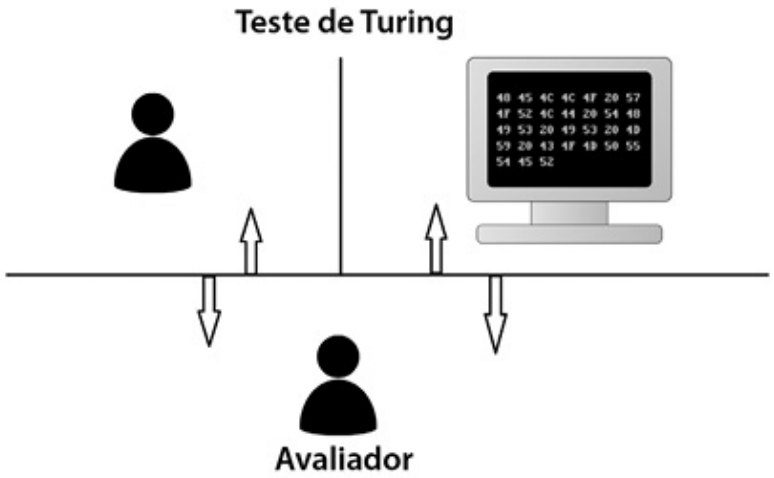

Figura 1.1 – Fluxo básico de funcionamento do teste de Turing.

Turing acreditava que somente perto da virada do século uma máquina passaria em seu teste. Sim, essa foi uma das muitas previsões da IA que aconteceram antes do imaginado.

Então, como o teste de Turing foi realizado ao longo dos anos? Bem, o teste se provou difícil de falhar. Lembre-se de que há concursos, como o Loebner Prize e a Turing Test Competition para incentivar as pessoas a criarem sistemas de software inteligentes.

Em 2014, houve um caso no qual uma máquina superou o teste de Turing. Um computador que disse ter 13 anos de idade[2] enganou avaliadores humanos, provavelmente porque algumas respostas continham erros.

Tempos depois, em maio de 2018, na Google's I/O Conference, o CEO Sundar Pichai deu uma demonstração de destaque do Google Assistant[3]. Diante de um público ao vivo, ele usou o dispositivo para telefonar para um cabeleireiro local e agendar um horário. A pessoa do outro lado da linha agiu como se estivesse conversando com uma pessoa!

Incrível, não é? Definitivamente. No entanto, o assistente provavelmente ainda não passou no teste de Turing. A razão é que a conversa foi focada em um tópico – não em perguntas abertas.

Como não deve ser surpresa, tem havido controvérsia com relação ao teste de Turing, com algumas pessoas sugerindo que ele pode ser manipulado. Em 1980, o filósofo John Searle escreveu um artigo famoso, intitulado "Minds, Brains and Programs" ("Mentes, cérebros e programas"), no qual descreveu seu próprio experimento de pensamento, chamado de "argumento do quarto chinês", para destacar as falhas.

Funcionava da seguinte maneira: suponha que John está em uma sala e não entende o idioma chinês. No entanto, ele tem manuais que fornecem regras fáceis de usar para traduzi-lo. Do lado de fora da sala está Jan, que entende a língua e envia caracteres chineses para John. Depois de algum tempo, ela vai obter uma tradução precisa de John. Como tal, é razoável supor que Jan acredita que John pode falar chinês.

Conclusão de Searle:

> *O argumento é: se o homem na sala não entende chinês com base na implementação do programa apropriado para compreender chinês, então nenhum outro computador digital compreenderá o idioma, porque nenhum deles tem algo que o homem não tem.*[4]

Foi um argumento muito bom – e tem sido tema de debates acirrados em círculos de IA desde então.

Searle também acreditava que havia duas formas de IA:

- *IA forte*: é quando uma máquina realmente entende o que está acontecendo. Podem existir emoções e criatividade inclusive. Na maior parte, é o que vemos em filmes de ficção científica. Esse tipo de IA também é conhecido como

inteligência artificial geral (AGI – Artificial General Intelligence). Observe que apenas poucas empresas se concentram nessa categoria, como a DeepMind, do Google.

- *IA fraca*: aqui, uma máquina realiza a correspondência entre padrões (pattern matching) e costuma estar focada em tarefas específicas. Exemplos incluem Siri, da Apple, e Alexa, da Amazon.

A realidade é que a inteligência artificial se encontra nas fases iniciais da IA fraca. Alcançar o ponto de IA forte pode facilmente levar décadas. Alguns pesquisadores acham que isso pode nem chegar a acontecer.

Dadas as limitações do teste de Turing, algumas alternativas surgiram, entre elas:

- *Teste de Kurzweil-Kapor*: criado pelo futurologista Ray Kurzweil e pelo empresário tecnológico Mitch Kapor, esse teste requer que um computador mantenha uma conversa por duas horas e que dois de três juízes acreditem se tratar de uma conversa humana. Kapor não acredita que isso seja alcançado antes de 2029.
- *Teste do café*: criado pelo cofundador da Apple, Steve Wozniak, esse teste sugere que um robô deve ser capaz de entrar na casa de um estranho, localizar a cozinha e fazer uma xícara de café.

O cérebro é... uma máquina?

Em 1943, Warren McCulloch e Walter Pitts se conheceram na Universidade de Chicago. Eles se tornaram amigos rapidamente apesar de suas origens serem totalmente diferentes, assim como suas idades (McCulloch tinha 42 anos e Pitts tinha 18). McCulloch cresceu em uma família rica e frequentou escolas de prestígio. Pitts, por outro lado, cresceu em um bairro de baixa renda e chegou a ser um sem-teto quando adolescente.

Apesar de tudo isso, a parceria se transformaria em uma das mais significativas para o desenvolvimento da IA. McCulloch e Pitts desenvolveram novas teorias para explicar o cérebro, as quais muitas vezes fora contra a sabedoria convencional da psicologia freudiana. Ambos acreditavam que a lógica poderia explicar o poder do cérebro e investigaram as reflexões de Alan Turing. A partir daí, escreveram juntos, em 1943, um artigo chamado "A Logical Calculus of the Ideas Immanent in Nervous Activity" ("Um cálculo lógico das ideias inerentes à atividade nervosa"), publicado pelo *Bulletin of Mathematical Biophysics*. A tese era que as funções principais do cérebro, como neurônios e sinapses, poderiam ser explicadas por meio de lógica e matemática, com operadores lógicos como E, Ou e Não. Com isso, seria possível construir uma rede complexa capaz de processar informações, aprender e pensar.

Ironicamente, o artigo não conseguiu despertar o interesse dos neurologistas.

Contudo, a publicação chamou a atenção daqueles que trabalhavam com computadores e IA.

Cibernética

Embora Norbert Wiener tenha criado várias teorias, a mais famosa era sobre cibernética e se concentrava na compreensão do controle e das comunicações com animais, pessoas e máquinas – mostrando a importância dos loops de feedback.

Em 1948, Wiener publicou *Cybernetics: Or Control and Communication in the Animal and the Machine ("Cibernética: ou controle e comunicação no animal e na máquina")*. Apesar de ser um trabalho acadêmico – repleto de equações complexas –, o livro se tornou um bestseller, entrando na lista de mais vendidos do *New York Times*.

Os temas abordados eram bastante variados. Alguns dos tópicos incluíam mecânica newtoniana, meteorologia, estatística, astronomia e termodinâmica. Esse livro anteciparia o desenvolvimento da teoria do caos, das comunicações digitais e até mesmo da memória do computador.

O livro também foi significativo para a IA. Como McCulloch e Pitts, Wiener comparou o cérebro humano ao computador. Além disso, especulou que um computador seria capaz de jogar xadrez e acabar vencendo grandes mestres. A principal razão para tal é que ele acreditava que uma máquina era capaz de aprender enquanto jogava. Wiener chegou até mesmo a pensar que os computadores seriam capazes de se replicar.

Mas *a cibernética* também não era utópica. Wiener foi prudente na compreensão das desvantagens dos computadores, tais como o potencial de desumanização, e chegou a pensar que as máquinas tornariam as pessoas desnecessárias.

Tratava-se, definitivamente, de uma mensagem mista. As ideias de Wiener, no entanto, eram poderosas e estimulariam o desenvolvimento da IA.

História da origem

O interesse de John McCarthy pelos computadores surgiu em 1948, quando ele participou de um seminário chamado "Cerebral Mechanisms in Behavior" ("Mecanismos cerebrais no comportamento"), que abordou o tema de como as máquinas acabariam sendo capazes de pensar. Alguns dos participantes incluíam os principais pioneiros no campo, como John von Neumann, Alan Turing e Claude Shannon.

McCarthy continuou a mergulhar na indústria emergente dos computadores – o que incluiu uma passagem pelo Bell Labs – e, em 1956, organizou um projeto de pesquisa de dez semanas na Universidade de Dartmouth. Ele o chamou de "um

estudo da inteligência artificial". Foi a primeira vez que o termo foi usado.

Entre os participantes estavam acadêmicos como Marvin Minsky, Nathaniel Rochester, Allen Newell, O. G. Selfridge, Raymond Solomonoff e Claude Shannon. Todos eles se tornariam grandes nomes da IA.

Os objetivos do estudo eram definitivamente ambiciosos:

> *O estudo visa proceder com base na conjectura de que todos os aspectos da aprendizagem ou qualquer outra característica da inteligência pode, em princípio, ser tão precisamente descrito que uma máquina pode ser construída para simulá-los. Será feita uma tentativa para descobrir como fazer com que as máquinas usem linguagem, formulem abstrações e conceitos, resolvam problemas reservados aos seres humanos e melhorem a elas mesmas. Pensamos que um avanço significativo pode ser feito em um ou mais desses problemas se um grupo cuidadosamente selecionado de cientistas trabalhar em conjunto ao longo de um verão.*[5]

Na conferência, Allen Newell, Cliff Shaw e Herbert Simon apresentaram um programa de computador chamado Logic Theorist (Teórico da Lógica), desenvolvido na Research and Development (RAND) Corporation. A inspiração principal veio de Simon (que receberia o prêmio Nobel de Economia em 1978). Quando viu como os computadores imprimiam palavras em um mapa para sistemas de defesa aérea, percebeu que essas máquinas poderiam ser usadas para além do processamento de números. Seria capaz de ajudar também com imagens, caracteres e símbolos – todos os quais poderiam levar a uma máquina de pensar.

Em relação ao Logic Theorist, o foco estava na resolução de vários teoremas matemáticos do *Principia Mathematica*. Uma das soluções do software acabou sendo mais elegante – e o coautor do livro, Bertrand Russell, ficou encantado.

Criar o Logic Theorist não foi uma tarefa fácil. Newell, Shaw e Simon usaram um IBM 701, que usava linguagem de máquina. Então, criaram uma linguagem de alto nível, chamada IPL (Information Processing Language – Linguagem de Processamento de Informações), que acelerou a programação. Por vários anos, essa foi a linguagem escolhida para a IA.

O IBM 701 também não tinha memória suficiente para o Logic Theorist; o que levou a outra inovação: processamento de listas. A técnica permitiu alocar e desalocar dinamicamente a memória à medida que o programa era executado.

Conclusão: o Logic Theorist é considerado o primeiro programa de IA já desenvolvido.

Apesar disso, ele não despertou muito interesse! A conferência de Dartmouth foi uma enorme decepção. Até mesmo o termo "inteligência artificial" foi criticado.

Pesquisadores tentaram pensar em alternativas, como "processamento de informações complexas", mas as sugestões não eram tão atraentes quanto IA – e o

termo foi mantido.

Quanto a McCarthy, ele continuou em sua missão de alavancar a inovação em IA. Considere os seguintes eventos:

- Durante o final da década de 1950, ele desenvolveu a linguagem de programação Lisp, frequentemente usada para projetos de IA devido à facilidade de manipular dados não numéricos. O pesquisador também criou conceitos de programação como recursão, tipagem dinâmica e coleta de lixo. Lisp continua a ser usada hoje em dia, em especial em robótica e aplicações de negócios. Enquanto McCarthy estava desenvolvendo a linguagem, também ajudou a fundar o Laboratório de IA do MIT.
- Em 1961, o pesquisador formulou o conceito de time-sharing de computadores; o que teve um impacto transformador na indústria e levou ao desenvolvimento da Internet e da computação em nuvem.
- Alguns anos mais tarde, fundou o Laboratório de Inteligência Artificial de Stanford.
- Em 1969, escreveu um artigo chamado "Computer-Controlled Cars" ("Carros controlados por computador"), no qual descreveu como uma pessoa poderia informar direções com um teclado e uma câmera de televisão navegaria o veículo.
- Em 1971, recebeu o prêmio Turing, considerado o prêmio Nobel da Ciência da Computação.

Em um discurso em 2006, McCarthy comentou que estava muito otimista em relação ao progresso da IA forte. Segundo ele, "nós, humanos, não somos muito bons em identificar a heurística que nós mesmos usamos"[6].

Era de ouro da IA

De 1956 a 1974, o campo da IA foi um dos mais movimentados no mundo tecnológico. Um grande catalisador foi o rápido desenvolvimento na tecnologia dos computadores. Eles passaram de sistemas maciços – baseados em tubos de vácuo – para sistemas menores que funcionavam com circuitos integrados muito mais rápidos e dispunham de maior capacidade de armazenamento.

O governo federal também estava investindo bastante em novas tecnologias. Em parte, devido aos ambiciosos objetivos do programa espacial Apollo e às difíceis demandas da Guerra Fria.

Quanto à IA, a principal fonte de financiamento foi a Advanced Research Projects Agency (ARPA – Agência de Projetos de Pesquisa Avançada), lançada no final da década de 1950 após o choque do Sputnik da Rússia. Os gastos com projetos em geral vinham com poucos requisitos e o objetivo era inspirar avanços inovadores.

Um dos líderes de ARPA, J. C. R. Licklider, tinha o lema de "financiar pessoas, não projetos". Em grande parte, quase todos os financiamentos vinham de Stanford, MIT, Lincoln Laboratories e Universidade Carnegie Mellon.

Com exceção da IBM, o setor privado teve pouco envolvimento no desenvolvimento da IA. Lembre-se de que – em meados da década de 1950 – a IBM recuaria e se concentraria na comercialização de seus computadores. Havia realmente o medo por parte dos clientes de que essa tecnologia conduziria a uma perda significativa de empregos. Então, a IBM não queria ser responsabilizada.

Em outras palavras, grande parte da inovação em IA aconteceu no círculo acadêmico. Em 1959, por exemplo, Newell, Shaw e Simon continuaram a expandir os limites do campo com o desenvolvimento de um programa chamado "General Problem Solver" ("Solucionador de problemas gerais"). Como o nome sugeria, tratava-se de um recurso para resolver problemas de matemática, como a Torre de Hanói.

Havia, no entanto, muitos outros programas que tentavam alcançar algum nível de IA forte. Entre os exemplos, tem-se:

- *SAINT (Symbolic Automatic INTegrator – Integrador Automático Simbólico) (1961)*: esse programa, criado pelo pesquisador do MIT James Slagle, ajudou a resolver problemas de cálculo de calouros. Seria incorporado a outros programas, chamados SIN e MACSYMA, capazes de lidar com matemática mais avançada. SAINT era realmente o primeiro exemplo de um sistema especialista, uma categoria da IA sobre a qual falaremos mais adiante neste capítulo.
- *ANALOGY (1963)*: esse programa foi criado pelo professor Thomas Evans, do MIT. Ele demonstrou que um computador poderia resolver problemas de analogia de um teste de QI.
- *STUDENT (1964)*: sob a supervisão de Minsky no MIT, Daniel Bobrow criou essa aplicação de IA para sua tese de doutorado. O sistema utilizava Natural Language Processing (NLP) para resolver problemas de álgebra para estudantes do ensino médio.
- *ELIZA (1965)*: o professor Joseph Weizenbaum, do MIT, criou esse programa que se transformou em um grande sucesso e chegou a despertar o interesse da imprensa. Foi nomeado Eliza em homenagem ao livro *Pygmalion*, de George Bernard Shaw, e atuava como um psicanalista. Um usuário podia digitar perguntas e Eliza forneceria conselhos (esse foi o primeiro exemplo de um chatbot). Algumas pessoas que o utilizaram acreditaram que o programa era uma pessoa real; o que preocupou profundamente Weizenbaum, já que a tecnologia era razoavelmente básica. É possível encontrar exemplos de Eliza na Web, como em `http://Psych.Fullerton.edu/mbirnbaum/psych101/Eliza.htm`.

- *Computer Vision (Visão computacional) (1966)*: em uma história lendária, Marvin Minsky, do MIT, disse a um estudante, Gerald Jay Sussman, que passasse o verão com uma câmera ligada a um computador fazendo com que a máquina descrevesse o que via. O jovem fez exatamente isso e construiu um sistema que detectou padrões básicos. Foi a primeira aplicação da visão computacional.
- *Mac hack (1968)*: o professor do MIT Richard D. Greenblatt criou esse programa que jogava xadrez. Foi o primeiro a jogar em torneios reais e recebeu uma classificação C.
- *Hearsay I (final da década de 1960):* o professor Raj Reddy desenvolveu um sistema contínuo de reconhecimento de fala. Alguns de seus alunos, então, resolveram criar a Dragon Systems, que se tornou uma grande empresa de tecnologia.

Durante esse período, houve uma proliferação de documentos acadêmicos e livros sobre IA. Alguns dos tópicos incluíam métodos bayesianos, machine learning e visão.

Havia, entretanto, duas teorias principais sobre a inteligência artificial. Uma foi sugerida por Minsky, que disse que precisavam existir sistemas simbólicos. Isso significava que a IA devia basear-se na lógica tradicional do computador ou na pré-programação – ou seja, no uso de estruturas como if-then-else.

A segunda teoria, proposta por Frank Rosenblatt, acreditava que a IA precisava usar sistemas semelhantes ao cérebro como redes neurais (esse campo também era conhecido como conexionismo). Em vez de chamar as partes internas de neurônios, entretanto, ele as denominou "perceptrons". Um sistema seria capaz de aprender à medida que recebesse dados ao longo do tempo.

Em 1957, Rosenblatt criou o primeiro programa de computador com essa finalidade e o chamou de Mark 1 Perceptron. Ele incluía câmeras para ajudar a diferenciar entre duas imagens (de 20 × 20 pixels). O Mark 1 Perceptron usou dados com ponderações aleatórias e, em seguida, percorreu o seguinte caminho:

1. Receba uma entrada e produza a saída do perceptron.
2. Se não houver uma correspondência, então
   a. se a saída deveria ter sido 0, mas foi 1, então o peso para 1 será decrementado.
   b. se a saída deveria ter sido 1, mas foi 0, então o peso para 1 será incrementado.
3. Repita os passos #1 e #2 até que os resultados sejam precisos.

Isso foi definitivamente inovador para a IA. O *New York Times* chegou a publicar um elogio para Rosenblatt no qual dizia: "A Marinha revelou o embrião de um computador eletrônico hoje e espera que ele seja capaz de andar, falar, ver, escrever, reproduzir-se e ser consciente de sua existência".[7]

No entanto, ainda havia problemas persistentes com o perceptron. Um deles era que a rede neural tinha apenas uma camada (principalmente por causa da falta de poder computacional naquele momento). O outro era que a pesquisa do cérebro ainda estava nos estágios iniciais e não oferecia muito sobre a compreensão da capacidade cognitiva.

Minsky coescreveu um livro, junto com Seymour Papert, chamado *Perceptrons* (1969). Os autores eram implacáveis em atacar a abordagem de Rosenblatt, que foi rapidamente deixada de lado. Observe que, no início dos anos 1950, Minsky desenvolveu uma máquina de rede neural rudimentar usando centenas de tubos de vácuo e peças de reposição de um bombardeiro B-24. Ele percebeu, entretanto, que aquela tecnologia não estava nem perto de ser viável.

Rosenblatt tentou revidar, mas era tarde demais. A comunidade de IA rapidamente denegriu as redes neurais. O pesquisador faleceu anos mais tarde, em um acidente de barco. Ele tinha 43 anos de idade.

Na década de 1980, contudo, suas ideias foram revividas – o que levaria a uma revolução na IA, principalmente com o desenvolvimento de deep learning.

Para a maioria, a Era de Ouro da IA foi livre e emocionante. Alguns dos acadêmicos mais brilhantes do mundo estavam tentando criar máquinas que poderiam realmente pensar. No entanto, o otimismo muitas vezes foi exagerado. Em 1965, Simon disse que, dali a 20 anos, uma máquina poderia fazer qualquer coisa que um humano pudesse. Mais tarde, em 1970, em uma entrevista para a revista Life, ele disse que isso aconteceria dentro de apenas 3 – 8 anos (a propósito, ele era um conselheiro no filme *2001: uma odisseia no espaço*).

Infelizmente, a fase seguinte da IA seria muito mais sombria. Havia mais acadêmicos tornando-se céticos. Talvez o que mais tenha discutido o assunto foi Hubert Dreyfus, um filósofo. Em livros como *What Computers Still Can't Do: A Critique of Artificial Reason (O que os computadores ainda não conseguem fazer: uma crítica à inteligência artificial)*[8], ele escreveu que os computadores não eram semelhantes ao cérebro humano e que a IA ficaria, lamentavelmente, aquém das elevadas expectativas.

Inverno da IA

Durante o início da década de 1970, o entusiasmo com a IA começou a diminuir. Esse período ficou conhecido como o *AI Winter* (Inverno da IA) e perdurou pela década de 1980 (o termo foi inspirado em "inverno nuclear", um evento de extinção em que o sol é bloqueado e as temperaturas diminuem em todo o mundo).

Embora muitos avanços já tivessem sido realizados na IA, eles ainda eram principalmente acadêmicos e desenvolvidos em ambientes controlados. Na época, os

sistemas computacionais ainda eram limitados. Por exemplo, uma máquina DEC PDP-11/45 – que era comum para a pesquisa em IA – tinha a capacidade de expandir sua memória para apenas 128K.

A linguagem Lisp também não era ideal para sistemas computacionais. No mundo corporativo, o foco estava principalmente em FORTRAN.

Além disso, ainda havia muitos aspectos complexos relacionados à compreensão da inteligência e do raciocínio. Apenas um é desambiguação. Essa é a situação na qual uma palavra tem mais de um significado. Isso contribui para a complexidade de um programa de IA, uma vez que ele também terá de entender o contexto.

Por fim, o ambiente econômico na década de 1970 estava longe de ser robusto. Houve inflação persistente, crescimento lento e interrupções de suprimento, como ocorreu com a crise do petróleo.

Dado tudo isso, não deve ser surpresa que o governo dos Estados Unidos ficou mais rigoroso com financiamentos. Afinal, para um planejador do Pentágono, quão útil é um programa que pode jogar xadrez, resolver um teorema ou reconhecer algumas imagens básicas?

Não muito, infelizmente.

Um caso notável é o Speech Understanding Research Program (Programa de Pesquisa para Compreensão da Fala) na Universidade Carnegie Mellon. Para a DARPA, parecia que esse sistema de reconhecimento de fala poderia ser usado por pilotos de caça para executar comandos por voz, mas ele provou ser impraticável. Um dos programas, denominado Harpy, compreendia 1.011 palavras – o que é equivalente ao que sabe uma criança normal de 3 anos de idade.

Os funcionários da DARPA realmente pensaram que tinham sido enganados, e a agência eliminou do orçamento anual os $3 milhões de dólares direcionados ao programa.

No entanto, o maior sucesso para a IA veio por meio de um relatório – divulgado em 1973 – do professor Sir James Lighthill. Financiado pelo Parlamento do Reino Unido, o texto era um repúdio total aos "objetivos grandiosos" da IA forte. Uma grande questão levantada pelo professor foi a "explosão combinatória", problema no qual os modelos ficaram muito complicados e difíceis de ajustar.

O relatório concluiu que "em nenhuma área do campo, as descobertas feitas até o momento produziram o grande impacto então prometido"[9]. O professor era tão pessimista que não acreditava que os computadores seriam capazes de reconhecer imagens ou vencer um mestre de xadrez.

O relatório também levou a um debate público que foi transmitido na televisão pela BCC (é possível encontrar os vídeos no YouTube). Foi Lighthill contra Donald Michie, Richard Gregory e John McCarthy.

Embora Lighthill fizesse observações válidas – e tivesse avaliado uma grande quantidade de pesquisas –, ele não enxergava o poder da IA fraca. Entretanto, isso não parecia importar, já que o inverno se espalhou.

As coisas ficaram tão ruins que muitos pesquisadores mudaram seus planos de carreira. Quanto àqueles que ainda estudavam IA, eles muitas vezes se referiam a seu trabalho com outros termos – como machine learning, reconhecimento de padrões e informática!

Ascensão e queda dos sistemas especialistas

As grandes inovações continuaram mesmo durante o inverno da IA. Uma delas foi a retropropagação (backpropagation), essencial para a atribuição de pesos nas redes neurais. Em seguida, houve o desenvolvimento da rede neural recorrente (RNN), que permitiu que as conexões se movimentassem entre as camadas de entrada e saída.

Nas décadas de 1980 e 1990, no entanto, também houve o surgimento de sistemas especialistas. Um impulsionador essencial para que isso acontecesse foi o crescimento explosivo de PCs e minicomputadores.

Os sistemas especialistas se baseavam nos conceitos da lógica simbólica de Minsky, que envolvia caminhos complexos. Eles costumavam ser desenvolvidos por especialistas com domínio em campos específicos, como Medicina, Finanças e Produção.

A Figura 1.2 mostra as partes principais de um sistema especialista.

Embora os sistemas especialistas tenham começado a surgir em meados da década de 1960, eles só começaram a ser usados comercialmente na década de 1980. Um exemplo foi o XCON (eXpert CONfigurer – configurador especialista), desenvolvido por John McDermott na Universidade Carnegie Mellon. O sistema permitiu otimizar a seleção de componentes de computador e, inicialmente, contava com cerca de 2.500 regras. Pense nele como o primeiro motor de recomendação. A partir de seu lançamento em 1980, o sistema se revelou uma grande economia de custos para a DEC e sua linha de computadores VAX (cerca de $40 milhões de dólares até 1986).

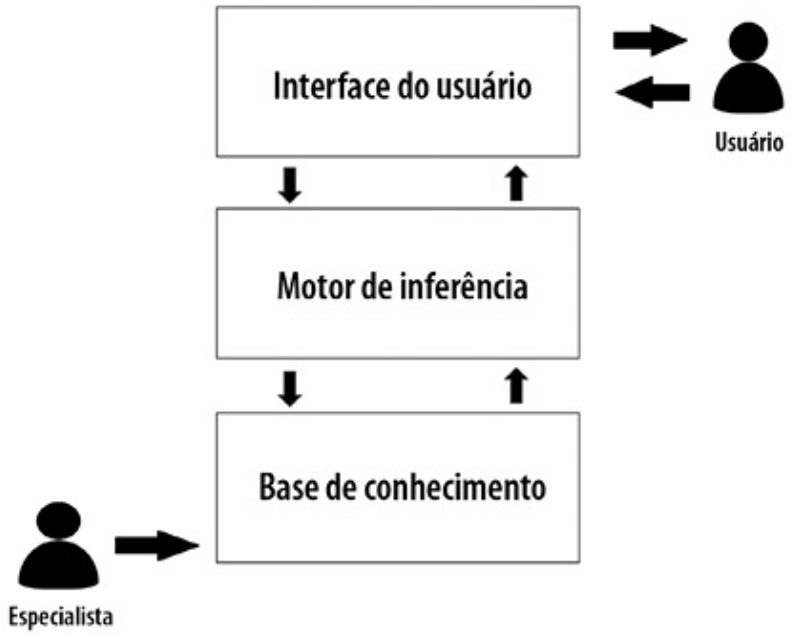

Figura 1.2 – Partes principais de um sistema especialista.

Quando as empresas viram o sucesso do XCON, houve um enorme crescimento no surgimento de sistemas especialistas – transformando-se em uma indústria de bilhões de dólares. O governo japonês também percebeu uma oportunidade e investiu centenas de milhões para reforçar seu mercado doméstico. Entretanto, os resultados foram, em sua maioria, uma decepção. Grande parte da inovação aconteceu nos Estados Unidos.

Observe que a IBM usou um sistema especialista para seu computador Deep Blue. Em 1996, ele venceria o grande mestre de xadrez Garry Kasparov em uma de seis partidas disputadas. A máquina, que a IBM vinha desenvolvendo desde 1985, processava 200 milhões de posições por segundo.

Contudo, houve problemas com os sistemas especialistas. Com frequência, eles eram muito específicos e era difícil aplicá-los em outras categorias. Além disso, à medida que ficavam maiores, tornava-se mais desafiador gerenciá-los e alimentá-los com dados. O resultado da Figura 1.2, com as partes principais de um sistema especialista, é que houve mais erros nos resultados. Além disso, testar os sistemas muitas vezes provou ser um processo complexo. Sejamos sinceros, houve momentos em que os peritos discordaram com relação a questões fundamentais. Por fim, os sistemas especialistas não aprenderam ao longo do tempo. Em vez disso, era necessário que houvesse atualizações constantes dos modelos lógicos subjacentes, o que fez com que os custos e as complexidades aumentassem muito.

No final da década de 1980, os sistemas especialistas começaram a perder a preferência no mundo dos negócios e muitas startups se fundiram ou faliram. Na verdade, isso ajudou a causar outro inverno na IA, o qual duraria até cerca de 1993. Os PCs estavam entrando rapidamente nos mercados de hardware de ponta, o que

significava uma redução acentuada das máquinas baseadas em Lisp.

O financiamento governamental para a IA, como a DARPA, também secou. Por outro lado, a Guerra Fria estava rapidamente chegando a um fim tranquilo com a queda da União Soviética.

Redes neurais e deep learning

Como um adolescente na década de 1950, Geoffrey Hinton queria ser professor e estudar IA. Ele vinha de uma família de acadêmicos notáveis (seu bisavô era George Boole). Sua mãe dizia muitas vezes: "seja um acadêmico ou seja um fracasso".[10]

Mesmo durante o primeiro inverno da IA, Hinton se manteve apaixonado pelo tema e estava convencido de que a abordagem da rede neural de Rosenblatt era o caminho certo. Assim, em 1972, ele recebeu seu PhD nessa área na Universidade de Edinburgh.

Durante esse período, entretanto, muitas pessoas acharam que Hinton estava desperdiçando seu tempo e talentos. A IA era essencialmente considerada uma área frágil e nem sequer foi reconhecida como ciência.

No entanto, isso incentivou Hinton ainda mais. Ele apreciava sua posição como um estranho e sabia que suas ideias ganhariam no final.

Hinton percebeu que o maior obstáculo à IA era o poder computacional. Ele também viu, contudo, que o tempo estava a seu lado. A lei de Moore previu que o número de componentes em um chip duplicaria a cada 18 meses.

Enquanto isso, Hinton trabalhou incansavelmente no desenvolvimento das principais teorias das redes neurais – algo que acabou se tornando conhecido como deep learning. Em 1986, ele escreveu – junto com David Rumelhart e Ronald J. Williams – um artigo pioneiro, chamado "Learning Representations by Backpropagating Errors" ("Aprendendo representações por erros de retropropagação"). Esse trabalho estabeleceu os principais processos para o uso de retropropagação em redes neurais. O resultado foi uma melhora significativa na exatidão, bem como com predições e reconhecimento visual.

Claro, isso não aconteceu isoladamente. O trabalho pioneiro de Hinton baseou-se nas conquistas de outros pesquisadores que também acreditavam nas redes neurais. E sua própria pesquisa estimulou uma enxurrada de outras grandes realizações:

- *1980*: Kunihiko Fukushima criou o Neocognitron, um sistema para reconhecer padrões que se tornaram o alicerce de redes neurais convolucionais. Ele se baseava no córtex visual dos animais.
- *1982*: John Hopfield desenvolveu as "Hopfield Networks" ("Redes de Hopfield"); essencialmente uma rede neural recorrente.

- *1989*: Yann lecun mesclou redes convolucionais com retropropagação. Essa abordagem foi aplicada na análise de verificações manuscritas.
- *1989*: a tese de doutorado de Christopher Watkins, "Learning from Delayed Rewards" ("Aprendendo com recompensas atrasadas"), descreveu o Q-Learning, um grande avanço no desenvolvimento da aprendizagem por reforço (reinforcement learning).
- *1998*: Yann lecun publicou "Gradient-Based Learning Applied to Document Recognition" ("Aprendizagem baseada em gradiente aplicada ao reconhecimento de documentos"), que utilizou algoritmos de descida (descent algorithms) para melhorar as redes neurais.

Impulsionadores tecnológicos da IA moderna

Além dos avanços em novas abordagens conceituais, teorias e modelos, a IA contava com alguns outros impulsionadores importantes. A seguir estão alguns dos principais:

- *Crescimento explosivo de datasets (conjuntos de dados)*: a internet tem sido um fator importante para a IA, pois permitiu a criação de datasets maciços. No próximo capítulo, vamos dar uma olhada em como os dados transformaram essa tecnologia.
- *Infraestrutura*: talvez a empresa mais importante para a IA durante os últimos 15 anos tenha sido o Google. Para acompanhar a indexação da web – que estava crescendo a uma taxa impressionante – a empresa precisou imaginar abordagens criativas para construir sistemas escaláveis. O resultado foi a inovação em clusters de servidores genéricos (commodities servers), virtualização e software de código aberto. O Google também foi um dos primeiros a adotar deep learning com o lançamento do projeto "Google Brain" ("Cérebro do Google"), em 2011. Alguns anos mais tarde, a empresa contratou Hinton.
- *GPUs (Graphics Processing Units – Unidades de processamento gráfico)*: essa tecnologia de chips, criada pela NVIDIA, originalmente se volta para gráficos de alta velocidade em jogos. No entanto, a arquitetura das GPUs acabaria por ser ideal também para a IA. Observe que a maioria das pesquisas de deep learning é feita com esses chips. A razão é que – com o processamento paralelo – a velocidade é muito maior do que nas CPUs tradicionais. Isso significa que o processamento de um modelo pode levar um dia ou dois, e semanas ou meses quando não há processamento paralelo.

Todos esses fatores reforçaram um ao outro – acrescentando combustível ao crescimento da IA. Além disso, eles estão propensos a permanecer ativos por muitos anos.

Estrutura da IA

Neste capítulo, abordamos muitos conceitos. Por ora, pode ser difícil compreender a organização da IA. Por exemplo, é comum ver termos como machine learning e deep learning serem confundidos. É essencial, contudo, compreender as distinções, que serão detalhadamente abordadas no restante deste livro.

Em uma visão de alto nível, a Figura 1.3 apresenta como os principais elementos da IA se relacionam entre si. No topo está a IA, que abrange uma grande variedade de teorias e tecnologias. Em seguida, é possível dividi-la em duas categorias principais: machine learning e deep learning.

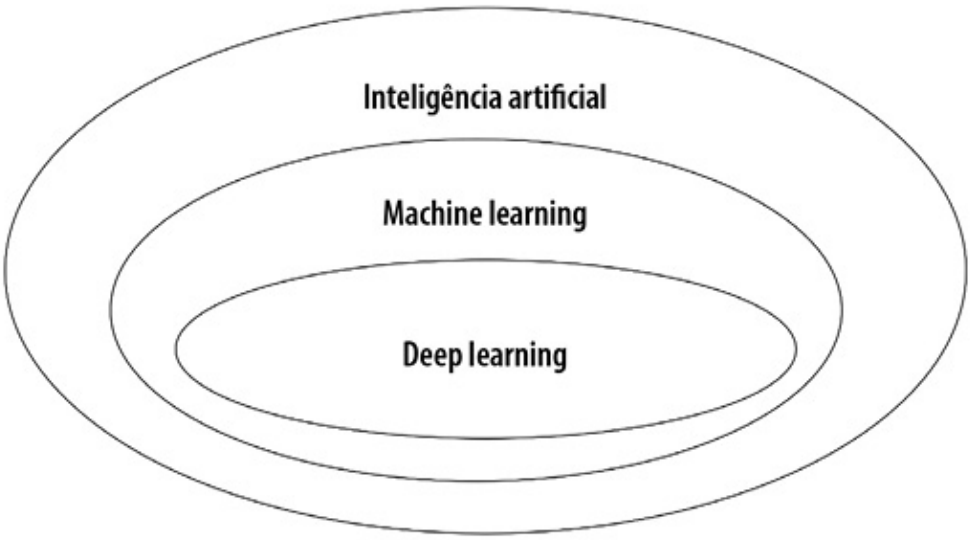

Figura 1.3 – Visão de alto nível dos principais componentes do mundo da IA.

Conclusão

Não é novidade que a IA seja um chavão hoje em dia. O termo tem vivenciado vários ciclos de expansão e retração.

Talvez volte a cair em desuso? Pode ser. Dessa vez, contudo, há verdadeiras inovações na IA que estão transformando as organizações. Grandes empresas de tecnologia como Google, Microsoft e Facebook consideram a categoria uma prioridade. No final das contas, parece ser uma boa ideia apostar que a IA continuará a crescer e mudar o nosso mundo.

Principais aprendizados

- A tecnologia em geral leva mais tempo para evoluir do que originalmente compreendido.
- A IA não se refere apenas à ciência da computação e matemática. Contribuições importantes vieram de campos como economia, neurociência, psicologia,

linguística, engenharia elétrica, matemática e filosofia.

- Existem dois tipos principais de IA: fraco e forte. No tipo forte, as máquinas se tornam autoconscientes; enquanto no fraco, os sistemas se concentram em tarefas específicas. Atualmente, a IA está no nível fraco.
- O teste de Turing é uma maneira comum de verificar se uma máquina pode pensar. Baseia-se no fato de alguém de fato achar que um sistema é inteligente.
- Alguns dos principais impulsionadores da IA incluem novas teorias de pesquisadores como Hinton, o crescimento explosivo dos dados, novas infraestruturas tecnológicas e GPUs (Graphics Processing Units – Unidades de Processamento Gráfico).

---

1 CEO fundador da Google Inc. Entrevista concedida à Academy of Achievement, www.achievement.org, em 28 de outubro de 2000.

2 www.theguardian.com/technology/2014/jun/08/super-computer-simulates-13-year-old-boy-passes-turing-test

3 www.theverge.com/2018/5/8/17332070/google-assistant-makesphone-call-demo-duplex-io-2018

4 https://plato.stanford.edu/entries/chinese-room/

5 www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

6 www.technologyreview.com/s/425913/computing-pioneer-dies/

7 www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html

8 MIT Press, 1972.

9 Em "Artificial Intelligence: A General Survey" ("Inteligência artificial: uma pesquisa geral"), escrita pelo professor Sir James Lighthill, da Universidade de Cambridge. Disponível em www.bbc.com/timelines/zq376fr.

10 https://torontolife.com/tech/ai-superstars-google-facebook-apple-studied-guy/

CAPÍTULO 2

Dados

O combustível da IA

O Pinterest é uma das startups mais badaladas do Silicon Valley, permitindo que seus usuários marquem seus itens favoritos para criar pastas envolventes. O site tem 250 milhões de MAUs (Monthly Actives Users – Usuários ativos por mês) e divulgou ter alcançado $756 milhões de receita em 2018.[1]

Uma atividade popular no Pinterest é o planejamento de casamentos. A futura noiva pode criar marcações para vestidos, locais, destinos de lua de mel, bolos, convites e assim por diante.

Isso também significa que o Pinterest tem a vantagem de coletar enormes quantidades de dados valiosos. Parte disso ajuda a fornecer anúncios direcionados. No entanto, há também oportunidades para campanhas de e-mail. Certa vez, o Pinterest enviou um email que dizia:

> *Você vai se casar! E como nós amamos planejar casamentos – em especial todos os artigos de papelaria encantadores – convidamos você a navegar por nossas melhores pastas, criadas por designers gráficos, fotógrafos e futuras noivas, todos usuários com olhos atentos e casamento na mente.*[2]

O problema é que vários destinatários do email já estavam casados ou não planejavam se casar em breve.

O Pinterest agiu rapidamente e disparou o seguinte pedido de desculpas:

> *Todas as semanas, enviamos coleções de categorias específicas de marcações e pastas para os usuários que esperamos que possam se interessar por elas. Infelizmente, um desses emails recentes sugeriu que alguns usuários estavam realmente se casando, em vez de apenas potencialmente interessados em conteúdo relacionado a casamento. Lamentamos ter soado como uma mãe arrogante que está sempre perguntando quando você vai encontrar um bom menino ou menina.*

É uma lição importante. Mesmo algumas das empresas mais experientes em tecnologia podem acabar estragando tudo.

Por exemplo, há algumas situações em que os dados podem estar no local adequado, mas o resultado pode ser um fracasso épico. Considere o caso da Target. A empresa aproveitou seus dados maciços para enviar ofertas personalizadas para as gestantes. A seleção baseou-se nos clientes que fizeram certos tipos de compras, como loções sem perfume. O sistema da empresa fazia uma análise de gravidez que fornecia, inclusive, estimativas de datas do nascimento do bebê.

Bem, o pai de uma das clientes viu o email e ficou furioso, dizendo que sua filha não estava grávida.[3]

Acontece que estava – e sim, ela escondia esse fato do pai.

Não há dúvidas de que os dados são extremamente poderosos e críticos para a IA. No entanto, é preciso ser cuidadoso e entender os riscos. Neste capítulo, vamos dar uma olhada em algumas coisas que você precisa saber.

Noções básicas de dados

É bom ter uma compreensão do jargão relacionado aos dados.

Em primeiro lugar, um bit (forma curta para "Binary digIT" – "dígito binário") é a menor forma de dados em um do computador. Pense nele como um átomo. Um bit pode ser 0 ou 1, que é binário. Ele costuma ser usado para medir a quantidade de dados que está sendo transferida (por exemplo, dentro de uma rede ou da internet).

Um byte, por outro lado, refere-se, principalmente, a armazenamento. Claro, o número de bytes pode ficar grande muito rápido. Vejamos como na Tabela 2.1.

Tabela 2.1 – Tipos de níveis de dados

Unidade	Valor	Caso de uso
Megabyte	1.000 kilobytes	Um livro pequeno
Gigabyte	1.000 megabytes	Cerca de 230 músicas
Terabyte	1.000 gigabytes	500 horas de filmes
Petabyte	1.000 terabytes	Cinco anos do Earth Observing System (EOS – Sistema de Observação da Terra)
Exabyte	1.000 petabytes	Toda a Biblioteca do Congresso americano 3.000 vezes
Zettabyte	1.000 exabytes	36.000 anos de vídeo HD-TV
Yottabytes	1.000 zettabytes	Isso demandaria um data center do tamanho de Delaware e Rhode Island juntas

Os dados também podem vir de muitas fontes diferentes. Veja alguns exemplos:

- Web/rede social (Facebook, Twitter, Instagram, YouTube)
- Dados biométricos (rastreadores de atividades físicas, testes genéticos)
- Sistemas de ponto de venda (de lojas físicas e sites de comércio eletrônico)
- Internet das coisas ou IoT (Etiquetas de identificação e dispositivos inteligentes)
- Sistemas de nuvem (aplicativos de negócios como Salesforce.com)

- Bancos de dados corporativos e planilhas

Tipos de dados

Há quatro maneiras de organizar os dados. Primeiro, há dados estruturados, que geralmente são armazenados em um banco de dados relacional ou planilha. Alguns exemplos incluem:

- Informações financeiras
- Números de Seguro Social
- Endereços
- Informações sobre produtos
- Dados de pontos de venda
- Números de telefone

Na maior parte das vezes, é mais fácil trabalhar com os dados estruturados. Esses dados são frequentemente provenientes de sistemas de CRM (Customer Relationship Management – Gestão do relacionamento com o cliente) e ERP (Enterprise Resource Planning – sistema integrado de gestão empresarial) – e geralmente tem volumes menores. Eles também tendem a ser mais simples, digamos, em termos de análise. Existem vários programas de BI (Business Intelligence – Inteligência de Negócios) que podem ajudar a obter insights a partir dos dados. No entanto, esses tipos de dados representam cerca de 20% de um projeto de IA.

A maior parte, entretanto, será formada por dados não estruturados, que são informações sem formatação predefinida. Será necessário formatá-las por sua conta, o que pode ser tedioso e demorado. No entanto, existem ferramentas como os bancos de dados da próxima geração – por exemplo, baseados em NoSQL – que podem ajudar com o processo. Os sistemas de IA também são eficazes em termos de gestão e estruturação dos dados, já que os algoritmos podem reconhecer padrões.

Aqui estão exemplos de dados não estruturados:

- Imagens
- Vídeos
- Arquivos de áudio
- Arquivos de texto
- Informações de redes sociais como tweets e postagens
- Imagens de satélites

Há ainda alguns dados que são um híbrido de fontes estruturadas e não estruturadas – chamados dados semiestruturados. As informações têm algumas marcas internas que ajudam na categorização.

Exemplos de dados semiestruturados incluem XML (Extensible Markup Language – Linguagem Extensível de Marcação), baseada em várias regras para identificar elementos de um documento, e JSON (JavaScript Object Notation – Notação de objetos JavaScript), que é uma maneira de transferir informações na web por meio de APIs (Application Programming Interfaces – Interfaces de programação de aplicativos).

Os dados semiestruturados, entretanto, representam apenas cerca de 5% a 10% de todos os dados.

Por fim, há séries de dados temporais que podem ser tanto estruturados, não estruturados ou semiestruturados. Esse tipo de informação é para interações; por exemplo, para rastrear os "passos do cliente". Trata-se de coletar informações quando um usuário vai para um site, usa um aplicativo ou até mesmo entra em uma loja.

Esse tipo de dados, contudo, é muitas vezes confuso e difícil de entender. Parte disso se deve à compreensão da intenção dos usuários, que pode variar bastante. Há também enormes volumes de dados de interações, que podem envolver trilhões de pontos de dados. Ah, e as métricas para o sucesso podem não ser claras. Por que um usuário está fazendo algo no site?

É provável, no entanto, que a IA seja crítica para tais questões. Embora, em sua maior parte, a análise dos dados das séries temporais ainda esteja nos estágios iniciais.

Big Data

Com a ubiquidade do acesso à internet, dispositivos móveis e wearables (vestíveis), tem havido o desencadeamento de uma torrente de dados. A cada segundo, o Google processa mais de 40.000 pesquisas; totalizando 3,5 bilhões por dia. A cada minuto, usuários do Snapchat compartilham 527.760 fotos e usuários do YouTube assistem a mais de 4,1 milhões de vídeos. Há ainda os sistemas antigos, como emails, que continuam a experimentar um crescimento significativo. A cada minuto, 156 milhões de mensagens são enviadas.[4]

No entanto, outro aspecto precisa ser considerado: empresas e máquinas também geram enormes quantidades de dados. De acordo com uma pesquisa do Statista, o número de sensores alcançará 12,86 bilhões até 2020.[5]

À luz de tudo isso, parece ser uma boa aposta que os volumes de dados continuarão a aumentar rapidamente. Em um relatório da International Data Corporation (IDC) chamado "Data Age 2025" ("Data 2025"), espera-se que a quantidade de dados criados até 2025 alcance surpreendentes 163 zettabytes.[6] Isso é aproximadamente dez vezes a quantidade de 2017.

Para lidar com tudo isso, surgiu uma categoria de tecnologia chamada Big Data. A

Oracle explica a importância dessa tendência da seguinte maneira:

> *Hoje, big data tornou-se essencial. Pense em algumas das maiores empresas de tecnologia do mundo. Uma grande parte do valor que oferecem vem de seus dados, que estão constantemente sendo analisados para produzir mais eficiência e desenvolver novos produtos.*[7]

Portanto, sim, big data continuará a ser uma parte crítica de muitos projetos de IA.

Então o que é exatamente big data? Qual seria uma boa definição? Na verdade, não há uma, embora existam muitas empresas que se concentram nesse mercado! No entanto, big data tem as seguintes características, chamadas de três Vs (Doug Laney, analista da Gartner, surgiu com esta estrutura por volta de 2001[8]): volume, variedade e velocidade.

Volume

Refere-se à escala dos dados, que muitas vezes não são estruturados. Não há nenhuma regra rígida que defina um limite, mas em geral são dezenas de terabytes.

Frequentemente, o volume é um grande desafio quando se trata de big data. Entretanto, a computação em nuvem e as bases de dados de última geração têm sido uma grande ajuda – em termos de capacidade e custos mais baixos.

Variedade

Essa característica descreve a diversidade dos dados, ou seja, a combinação de dados estruturados, semiestruturados e não estruturados (conforme explicado anteriormente). Ela também trata das diferentes fontes dos dados e suas aplicações. Sem dúvida, o alto crescimento dos dados não estruturados tem sido essencial para a variedade no big data.

Gerenciá-la pode rapidamente se tornar um grande desafio. No entanto, machine learning pode muitas vezes ajudar a simplificar o processo.

Velocidade

Refere-se à velocidade na qual os dados estão sendo criados. Como visto anteriormente neste capítulo, serviços como YouTube e Snapchat têm níveis extremos de velocidade (o que é muitas vezes referenciado como uma enxurrada de dados). Isso requer investimentos pesados em tecnologias e data centers de próxima geração. Os dados também costumam ser processados na memória, não por sistemas baseados em disco.

Devido a essas questões, a velocidade é muitas vezes considerada a mais difícil do três Vs. Sejamos francos, no mundo digital de hoje, as pessoas querem seus dados o

mais rápido possível. Se for muito lento, as pessoas vão ficar frustradas e procurar outro lugar.

Ao longo dos anos, porém, como o big data evoluiu, mais Vs foram adicionados. Atualmente, há mais de dez.

A seguir, estão alguns dos mais comuns:

- *Veracidade*: refere-se a dados considerados precisos. Neste capítulo, vamos discutir algumas das técnicas para avaliar a veracidade.
- *Valor*: mostra a utilidade dos dados. Muitas vezes, refere-se a ter uma fonte confiável.
- *Variabilidade*: significa que os dados geralmente mudam ao longo do tempo. É o caso, por exemplo, do conteúdo de mídia social que pode se transformar com base no sentimento geral com relação a novos acontecimentos e notícias de última hora.
- *Visualização*: trata do uso de recursos visuais – como gráficos – para entender melhor os dados.

Como se pode ver, o gerenciamento de um big data engloba muitas partes, o que leva à complexidade. Isso ajuda a explicar por que muitas empresas ainda usam apenas uma pequena fração de seus dados.

Bancos de dados e outras ferramentas

Há uma infinidade de ferramentas que ajudam com dados. No centro delas está o banco de dados. Como não deve ser surpresa, houve uma evolução dessa tecnologia crítica ao longo das décadas. Contudo, mesmo as tecnologias mais antigas, como bancos de dados relacionais, ainda são usadas hoje em dia. No que se refere a dados críticos, as empresas ficam relutantes em fazer mudanças – mesmo que haja benefícios claros.

Para entender esse mercado, vamos voltar a 1970, quando Edgar Codd, cientista da computação da IBM, publicou "A Relational Model of Data for Large Shared Data Banks" ("Um modelo relacional para grandes bancos de dados compartilhados"). O artigo foi um divisor de águas, pois introduziu a estrutura de bancos de dados relacionais. Até esse ponto, as bases de dados eram bastante complexas e rígidas – estruturadas como hierarquias. Isso fazia com que fosse demorado realizar pesquisas e encontrar relacionamentos nos dados.

A abordagem de banco de dados relacional de Codd foi construída para máquinas mais modernas. A linguagem de script SQL era fácil de usar, permitindo operações de CRUD (Create, Read, Update, Delete – Criar, ler, atualizar e excluir). As tabelas também dispunham de conexões com chaves primárias e estrangeiras, que

permitiam relacionamentos importantes como os seguintes:

- *Um-para-um*: uma linha em uma tabela está vinculada a apenas uma linha em outra tabela. Exemplo: um número de carteira de motorista, que é único, está associado a um funcionário da instituição.
- *Um-para-muitos*: uma linha em uma tabela está vinculada a outras tabelas. Exemplo: um cliente tem várias ordens de compra.
- *Muitos-para-muitos*: linhas de uma tabela estão associadas a linhas de outra tabela. Exemplo: vários relatórios têm vários autores.

Com esses tipos de estruturas, um banco de dados relacional poderia simplificar o processo de criação de sofisticados relatórios. Ele era realmente revolucionário.

Apesar das vantagens, contudo, a IBM não estava interessada na tecnologia e continuou a focar em seus sistemas proprietários. A empresa considerou os bancos de dados relacionais demasiado lentos e frágeis para os clientes corporativos.

No entanto, havia alguém com uma opinião diferente sobre o assunto: Larry Ellison. Ele leu o artigo de Codd e percebeu que poderia mudar o jogo. Para provar isso, foi cofundador da Oracle em 1977 com o objetivo de focar na construção de bancos de dados relacionais – o que rapidamente se tornou um mercado massivo. O artigo de Codd funcionou essencialmente como um roteiro de produto para seus esforços empreendedores.

Apenas em 1993 a IBM lançou o próprio banco de dados relacional, o DB2. No entanto, era tarde demais. A essa altura, a Oracle já era líder no mercado de banco de dados.

Ao longo dos anos de 1980 e 1990, o banco de dados relacional era o padrão para sistemas mainframe e cliente-servidor. Contudo, quando o big data se tornou importante, a tecnologia apresentou falhas graves como as seguintes:

- *Expansão de dados*: ao longo do tempo, diferentes bases se espalham pela organização. O resultado foi que ficou mais difícil centralizar os dados.
- *Novos ambientes*: a tecnologia de banco de dados relacional não foi criada para computação em nuvem, dados de alta velocidade ou dados não estruturados.
- *Custos elevados*: bancos de dados relacionais podem ser caros. Isso significa que pode ser proibitivo usar a tecnologia para projetos de IA.
- *Desafios de desenvolvimento*: o desenvolvimento de software moderno depende muito de iteração. Contudo, os bancos de dados relacionais têm se mostrado um desafio para esse processo.

No final da década de 1990, havia projetos de código aberto desenvolvidos para ajudar na criação de sistemas de banco de dados da próxima geração. Talvez o mais importante deles seja o de Doug Cutting, que desenvolveu o Lucene para pesquisa

de texto. A tecnologia baseava-se em um sofisticado sistema de indexação que permitia um desempenho de baixa latência. Lucene foi sucesso instantâneo e começou a evoluir, assim como o Apache Nutch que rastreou eficientemente a web e armazenou os dados em um índice.

No entanto, houve um grande problema: para rastrear a web, era necessário dispor de uma infraestrutura que pudesse ser hiperescalada. Assim, no final de 2003, Cutting deu início ao desenvolvimento de um novo tipo de plataforma de infraestrutura que fosse capaz de resolver o problema. Ele se inspirou em um artigo publicado no Google que descrevia seu sólido sistema de arquivos. Um ano depois, Cutting construiu sua nova plataforma, o que permitiu o armazenamento sofisticado sem complexidade. No cerne de seu projeto estava o MapReduce, que permitia o processamento em vários servidores. Em seguida, os resultados eram mesclados, viabilizando relatórios significativos.

O sistema de Cutting acabou se transformou em uma plataforma chamada Hadoop – que se tornaria essencial para a gestão de big data e permitiria a criação de data warehouses sofisticados. Inicialmente, o Yahoo! usava a plataforma, mas ela se tornou conhecida rapidamente, e empresas como Facebook e Twitter adotaram a tecnologia de ponta. Essas empresas passaram a ser capazes de obter uma visão completa de seus dados, não apenas de subconjuntos. Isso significava que poderia haver experimentos de dados mais eficazes.

Como um projeto de código aberto, entretanto, o Hadoop ainda não contava com sistemas sofisticados para clientes corporativos. Para lidar com essa questão, uma startup chamada Hortonworks construiu novas tecnologias a partir da plataforma Hadoop, como o YARN. A ferramenta contava com recursos como processamento analítico na memória, processamento de dados online e processamento SQL interativo. Tais funcionalidades permitiram a adoção do Hadoop em muitas corporações.

É claro que surgiram outros projetos de data warehouses de código aberto. Os mais conhecidos, como Storm e Spark, concentravam-se em streaming de dados. O Hadoop, por outro lado, foi otimizado para processamento em lote.

Além dos data warehouses, houve também a inovação do negócio tradicional de banco de dados. Muitas vezes, essas iniciativas eram conhecidas como sistemas NoSQL. Considere o MongoDB. Ele começou como um projeto de código aberto e se transformou em uma empresa altamente bem-sucedida, que se tornou pública em outubro de 2017. O banco de dados MongoDB, que tem mais de 40 milhões de downloads, foi criado para manipular ambientes em nuvem, locais (on-premise) e híbridos.[9] A ferramenta oferece também muita flexibilidade na estruturação dos dados, que se baseia em um modelo de documento. O MongoDB pode até mesmo gerenciar dados estruturados e não estruturados na escala de petabytes.

Embora as startups tenham sido uma fonte de inovação em sistemas de banco de dados e armazenamento, é importante observar que os megaoperadores de tecnologia também foram importantes. Então, novamente, empresas como Amazon.com e Google tiveram de encontrar maneiras de lidar com a enorme escala de dados devido à necessidade de gerenciar suas plataformas maciças.

Uma das inovações foi o data lake, que permite o armazenamento de dados estruturados e não estruturados. Observe que não há necessidade de reformatar os dados. O data lake tratará as diferenças e permitirá que sejam executadas funções de IA com rapidez. De acordo com um estudo de Aberdeen, as empresas que utilizam essa tecnologia têm uma média de 9% de crescimento orgânico em comparação com aquelas que não o fazem.[10]

Isso não significa, contudo, que você tem de se livrar de seus data warehouses, já que ambos atendem funções e casos de uso específicos. Um data warehouse geralmente funciona bem com dados estruturados, enquanto um data lake é melhor para ambientes diversos. O importante é saber que é provável que uma grande parte dos dados nunca seja utilizada.

Para a maioria deles, há uma infinidade de ferramentas. Espera-se também que outras sejam desenvolvidas à medida que os ambientes de dados forem ficando mais complexos.

Não quer dizer, entretanto, que você deva escolher a tecnologia mais recente. Novamente, mesmo os bancos de dados relacionais mais antigos podem ser bastante eficazes com projetos de IA. A chave é compreender os prós e contras de cada um e, a partir daí, definir uma estratégia clara.

Processo de dados

A quantidade de dinheiro investida em dados é enorme. De acordo com a IDC (International Data Corporation –Corporação Internacional de Dados), prevê-se que os gastos com soluções de big data e analytics passarão de US$166 bilhões em 2018 para US$260 bilhões até 2022.[11] Isso representa uma taxa de crescimento anual de 11,9%. Os maiores investidores incluem bancos, fabricantes discretos e de processos, empresas de serviços profissionais e Governo Federal. Eles respondem por quase metade do investimento total.

Observe o que disse Jessica Goepfert – vice-presidente (VP) do Programa de Conhecimento e Análise de Clientes da IDC:

> *Em um alto nível, as organizações estão se voltando para soluções de big data e analytics para navegar na convergência de seus mundos físico e digital. Essa transformação assume uma forma diferente dependendo da indústria. Por exemplo, no âmbito bancário e varejista – duas das áreas de crescimento mais rápido para big*

data e analytics –, os investimentos são direcionados a gerenciar e revigorar a experiência do cliente. Na produção, entretanto, as empresas estão se reinventando, essencialmente, como empresas de alta tecnologia, usando seus produtos como uma plataforma para habilitar e fornecer serviços digitais.[12]

Um alto nível de investimento, no entanto, não se converte, necessariamente, em bons resultados. Um estudo da Gartner estima que cerca de 85% dos projetos de big data são abandonados antes de chegar à fase piloto.[13] Algumas das razões para isso incluem:

- Falta de foco claro
- Dados impróprios
- Investimento nas ferramentas de TI erradas
- Problemas na coleta de dados
- Falta de adesão dos principais stakeholders e defensores na organização

Diante disso, é fundamental ter um processo de dados. Apesar de existirem muitas abordagens – muitas vezes definidas por fornecedores de software –, existe uma amplamente aceita. Um grupo de especialistas, desenvolvedores de software, consultores e acadêmicos criou o processo CRISP-DM no final da década de 1990. Observe a Figura 2.1 para ter uma ideia do que se trata.

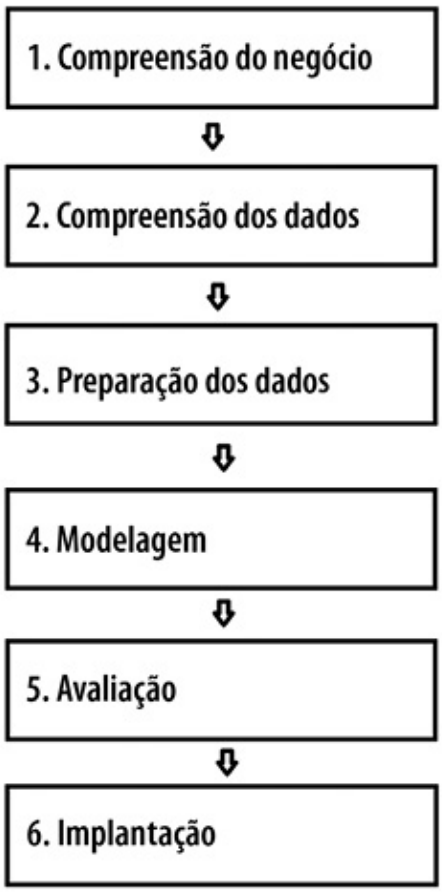

Figura 2.1 – Processo CRISP-DM.

Neste capítulo, vamos dar uma olhada nas etapas 1, 2 e 3. Em seguida, no restante do livro, vamos abordar os itens restantes (ou seja, discutiremos Modelagem e Avaliação no Capítulo 3, e Implantação no Capítulo 8).

Observe que as três primeiras etapas podem ser responsáveis por 80% do tempo do processo de dados, que se baseia na experiência de Atif Kureishy, VP de Práticas Emergentes da Teradata.[14] Isso pode acontecer por diversos motivos: os dados podem não estar bem organizados e virem de diferentes fontes (sejam elas diferentes fornecedores ou grupos na organização), não há foco suficiente em ferramentas de automação ou o planejamento inicial foi insuficiente para o escopo do projeto.

Também vale a pena lembrar que o processo CRISP-DM não é linear e rigoroso. Ao lidar com dados, pode haver muita iteração. Podem ser necessárias, por exemplo, várias tentativas até que se cheguem aos dados certos e seja possível testá-los.

Etapa #1 – Compreensão do negócio

É preciso ter uma visão clara do problema de negócio a ser resolvido. Observe alguns exemplos:

- Como um ajuste de preço pode impactar suas vendas?
- Uma alteração nas cópias levará a uma melhor conversão de anúncios digitais?
- Uma queda no engajamento significa que haverá um aumento na rotatividade?

Em seguida, é preciso estabelecer como o sucesso será medido. É necessário avaliar se as vendas aumentaram em pelo menos 1% ou se as conversões subiram para 5%?

Aqui está um caso de Prasad Vuyyuru, parceiro em Prática de Insights Corporativos da Infosys Consultoria:

> *Identificar qual problema de negócio será resolvido usando IA e avaliar que valor será criado são fatores críticos para o sucesso de todos os projetos dessa área. Sem um foco tão diligente no valor do negócio, os projetos de IA podem não ser adotados na organização. A experiência da AB InBev no uso da inteligência artificial para identificação dos motores da linha de empacotamento com possibilidade de falhar é um grande exemplo de como a IA está criando valor prático. A empresa instalou 20 sensores sem fio para medir vibrações em motores de linhas de empacotamento. Eles comparavam sons com motores que funcionam normalmente para identificar anomalias que previam eventuais falha dos motores.*[15]

Independentemente do objetivo, é essencial que o processo esteja livre de quaisquer prejulgamentos ou inclinações. O foco é encontrar os melhores resultados. Sem dúvida, em alguns casos, não haverá um resultado satisfatório.

Em outras situações, pode haver grandes surpresas. Um exemplo famoso disso vem do livro *Moneyball*, de Michael Lewis, que foi transformado em um filme estrelado por Brad Pitt em 2011. Trata-se de uma história verídica sobre como o time Oakland

A's usou técnicas de ciência de dados para recrutar jogadores. A tradição no beisebol era depender de métricas como as médias de rebatidas. No entanto, ao usar técnicas sofisticadas de análise de dados, resultados surpreendentes foram alcançados. O Oakland A's percebeu que o foco deveria se voltar para as porcentagens de slugging e on base[16]. Com essa informação, a equipe foi capaz de recrutar jogadores de alto desempenho com níveis mais baixos de remuneração.

O resultado é que você precisa ter a mente aberta e estar disposto a experimentar.

Na etapa #1, também é necessário montar a equipe certa para o projeto. Nesse momento, a menos que trabalhe em uma empresa como o Facebook ou Google, você não poderá se dar o luxo de selecionar um grupo de doutores em machine learning e ciência de dados. Esse talento é bastante raro – e caro.

Contudo, não é necessário um exército de engenheiros de alto nível para um projeto de IA. Na verdade, está ficando cada vez mais fácil aplicar machine learning e modelos de deep learning por conta de sistemas de código aberto como o TensorFlow e plataformas baseadas na nuvem de empresas como Google, Amazon.com e Microsoft. Em outras palavras, serão necessárias apenas algumas pessoas com experiência em ciência de dados.

Em seguida, é preciso encontrar pessoas – provavelmente de sua organização – que tenham o conhecimento de domínio adequado para o projeto de IA. Elas precisarão refletir sobre os fluxos de trabalho, modelos e dados de treinamento – com compreensão específica dos requisitos do setor e dos clientes.

Por fim, será necessário avaliar as necessidades técnicas. Que infraestrutura e ferramentas de software serão utilizadas? Haverá necessidade de aumentar a capacidade ou comprar novas soluções?

Passo #2 – Compreensão dos dados

Nessa etapa, serão examinadas as fontes de dados para o projeto. Considere que existem três fontes principais:

- *Dados internos*: podem vir de um website, beacons[17] em lojas, sensores de IoT, aplicativos e assim por diante. Uma grande vantagem desses dados é que são gratuitos e personalizados para o seu negócio. No entanto, aqui também há alguns riscos. Podem existir problemas se o processo de formatação e seleção dos dados não tiver sido feito com atenção.
- *Dados de código aberto*: costumam estar disponíveis gratuitamente, o que sem dúvida é um bom benefício. Alguns exemplos de dados de código aberto incluem informações governamentais e científicas. Os dados em geral são acessados por meio de uma API, o que torna o processo bastante simples. Geralmente, esse tipo de dado também é bem formatado. No entanto, algumas

das variáveis podem não ser claras e reforçarem tendências, como distorções para um determinado grupo demográfico.

- *Dados terceirizados*: são dados de um fornecedor comercial, mas as taxas podem ser elevadas. Na verdade, em alguns casos pode faltar qualidade nos dados.

De acordo com a Teradata – com base nos engajamentos de IA da firma –, cerca de 70% das fontes de dados são internas, 20% de código aberto e o restante oriundos de vendedores comerciais.[18] Independentemente da fonte, contudo, todos os dados devem ser confiáveis. Se não, provavelmente haverá o problema de "lixo entra, lixo sai"[19].

Para avaliar os dados, é necessário responder a perguntas como:

- Os dados estão completos? O que pode estar faltando?
- De onde vêm os dados?
- Quais foram os pontos de coleta?
- Quem manipulou os dados – e os processou?
- Quais foram as alterações nos dados?
- Quais são os problemas de qualidade?

Se estiver trabalhando com dados estruturados, essa etapa deverá ser mais fácil. No entanto, quando se manipulam dados não estruturados e semiestruturados, será necessário rotulá-los – o que pode ser um processo demorado. Existem, entretanto, algumas ferramentas emergentes no mercado que podem ajudar a automatizar esse processo.

Passo #3 – Preparação dos dados

A primeira etapa no processo de preparação é decidir quais conjuntos de dados (datasets) usar.

Considere o seguinte cenário: suponha que você trabalha para uma editora e queira criar uma estratégia para melhorar a retenção de clientes. Entre os dados que podem ajudar estão informações demográficas sobre a base de clientes – como idade, sexo, renda e escolaridade. Para ficar ainda mais completo, é possível incluir também informações do navegador. Que tipo de conteúdo interessa aos clientes? Com qual frequência e duração? Algum outro padrão interessante – como acessar informações durante fins de semana? Combinando as fontes de informação, é possível formar um modelo poderoso. Por exemplo, o abandono de uma atividade em determinadas áreas poderia representar um risco de cancelamento. Tal fato poderia alertar a equipe de vendas para que contactassem os clientes.

Embora esse seja um processo inteligente, ainda há problemas. A inclusão ou exclusão de apenas uma variável pode causar um impacto negativo significativo em

um modelo de IA. Para compreender o porquê, lembre-se da crise financeira. Os modelos de subscrição de hipotecas eram sofisticados e baseados em enormes quantidades de dados. Durante os tempos normais da economia, eles funcionaram muito bem e as grandes instituições financeiras – como Goldman Sachs, JP Morgan e AIG – confiaram neles cegamente.

No entanto, havia um problema: os modelos não consideravam a queda no preço das habitações! A principal razão para isso era que – durante décadas – uma queda nacional nunca havia acontecido. A suposição era que habitação constituía, na maior parte, um fenômeno local.

É claro que o preço das habitações não caiu simplesmente – ele despencou. Os modelos então se mostraram muito distantes e bilhões de dólares em perdas quase derrubaram o sistema financeiro dos Estados Unidos. O governo federal não teve escolha e precisou emprestar US$700 bilhões para resgatar Wall Street.

É verdade que esse é um caso extremo, mas destaca a importância da seleção dos dados. É aqui que pode ser essencial contar com uma sólida equipe de especialistas de domínio e cientistas de dados.

Ainda na fase de preparação, haverá a necessidade de executar uma limpeza nos dados. O fato é que todos os dados têm problemas. Mesmo empresas como o Facebook têm lacunas, ambiguidades e dados discrepantes (outliers) em seus conjuntos. É inevitável.

Então, aqui estão algumas medidas a serem tomadas para limpar os dados:

- *Deduplicação*: defina testes para identificar quaisquer duplicações e exclua os dados estranhos.
- *Dados discrepantes (outliers)*: estão bem além do escopo da maior parte dos demais dados; o que pode indicar que não serão úteis. Há situações, no entanto, em que o inverso é verdadeiro; o que serviria para detecção de fraudes.
- *Consistência*: certifique-se de que tem definições claras para as variáveis. Mesmo termos como "receita" ou "cliente" podem ter vários significados.
- *Regras de validação*: busque encontrar as limitações à medida que olha para os dados. Por exemplo, você pode ter um sinalizador para a coluna da idade. Se o valor for superior a 120 em muitos casos, significa que os dados têm problemas sérios.
- *Armazenamento*: certos dados podem não precisar ser específicos. Será que realmente importa se alguém tem 35 ou 37 anos? Provavelmente não. No entanto, comparar aqueles que estão entre 30–40 anos com os de 41–50 provavelmente faria diferença.
- *Trivialidade*: os dados são oportunos e relevantes?
- *Fusão*: as colunas de dados podem ter informações muito semelhantes em

alguns casos. Talvez, uma contenha a altura em polegadas e outra a armazene em pés. Se o modelo não requer um número mais detalhado, é possível escolher apenas uma das colunas para armazenar.

- *Codificação one-hot*: trata-se de uma maneira de substituir dados categóricos por números. Por exemplo, digamos que temos um banco de dados com uma coluna com três possíveis valores: maçã, abacaxi e laranja. É possível representar maçã como 1, abacaxi como 2 e laranja como 3. Parece razoável, certo? Talvez não. O problema é que um algoritmo de IA pode pensar que a laranja é maior do que a maçã. Com uma codificação one-hot, entretanto, é possível evitar esse problema. Três novas colunas serão criadas: Maça, Abacaxi e Laranja. Para cada linha nos dados, será colocado 1 onde o fruto existe e 0 para o restante.
- *Tabelas de conversão*: é possível usar essas tabelas ao traduzir dados de um padrão para outro. Esse seria o caso se você tivesse dados no sistema decimal e desejasse convertê-los para o sistema métrico.

Essas etapas vão percorrer um longo caminho na melhoria da qualidade dos dados. Há também ferramentas de automação que podem ajudar, como as disponibilizadas por empresas como SAS, Oracle, IBM, Lavastorm Analytics e Talend. Existem ainda projetos de código aberto, como OpenRefine, plyr e reshape2.

Independentemente da limpeza, os dados não serão perfeitos. Nenhuma fonte de dados é perfeita. É provável que ainda haja lacunas e imprecisões.

É por isso que é necessário ser criativo. Veja o que fez Eyal Lifshitz, CEO da BlueVine. Sua empresa aproveita a IA para oferecer financiamento para pequenas empresas. "Uma de nossas fontes de dados é a informação de crédito de nossos clientes", disse ele. "Mas descobrimos que os proprietários de pequenas empresas identificam incorretamente seu tipo de negócio. Isso pode significar maus resultados para a nossa subscrição. Para lidar com isso, nós coletamos dados do site do cliente com algoritmos de IA, o que ajuda a identificar a indústria."[20]

As abordagens de limpeza de dados também dependerão dos casos de uso para o projeto de IA. Por exemplo, se estiver construindo um sistema de manutenção preditiva para a produção, o desafio será lidar com as grandes variações de diferentes sensores. O resultado é que uma grande quantidade de dados pode ter pouco valor e ser, principalmente, ruído.

Ética e governança

É preciso estar atento a quaisquer restrições sobre os dados. O fornecedor pode proibi-lo de usar as informações para determinadas finalidades? Sua empresa estará encrencada se algo der errado? Para lidar com essas questões, é aconselhável consultar o departamento jurídico.

Em sua maior parte, os dados devem ser tratados com cuidado. Afinal, há muitos casos de alto perfil nos quais as empresas violaram a privacidade. Um exemplo proeminente disso é o Facebook. Um dos parceiros da empresa, a Cambridge Analytica, acessou milhões de pontos de dados de perfis sem a permissão dos usuários. Quando um denunciante descobriu, as ações do Facebook despencaram – perdendo mais de US$100 bilhões em valor. A empresa também sofreu pressão dos governos dos Estados Unidos e da Europa.[21]

Outro aspecto para o qual se deve estar atento é a coleta de dados de fontes públicas. É claro que normalmente essa é uma maneira eficiente para criar grandes conjuntos de dados. Há também muitas ferramentas que podem automatizar o processo. No entanto, a técnica pode expor sua empresa a responsabilidades legais, pois os dados podem estar sujeitos a direitos autorais ou leis de privacidade.

Existem também algumas precauções que podem, ironicamente, apresentar falhas. Por exemplo, um estudo recente do MIT mostrou que dados anônimos podem não ser tão anônimos assim. Pesquisadores descobriram que era de fato muito fácil reconstruir esse tipo de dado e identificar os indivíduos – por exemplo, mesclando dois conjuntos de dados. Isso foi feito em Cingapura a partir de dados de uma rede móvel (rastreamento de GPS) e de um sistema de transporte local. Após cerca de 11 semanas de análise, os pesquisadores conseguiram identificar 95% dos indivíduos.[22]

Para finalizar, certifique-se de tomar medidas para proteger os dados. As ocorrências de ciberataques e ameaças continuam a aumentar a um ritmo alarmante. Em 2018, houve mais de 53.000 incidentes e cerca de 2.200 violações, de acordo com a Verizon.[23] O relatório também observou o seguinte:

- 76% das violações tinham motivações financeiras.
- 73% vinham de fora da empresa.
- Cerca de metade vinha de grupos criminosos organizados e 12% de sujeitos de estados-nação ou afiliados a estados.

O uso crescente de dados na nuvem e no local (on-premise) também podem sujeitar a empresa a lacunas na segurança. Há ainda a força de trabalho móvel, que pode significar acesso a dados que poderiam expor os funcionários a violações.

Os ataques também estão causando prejuízos muito maiores. O resultado é que uma empresa pode facilmente sofrer penalidades, ações judiciais e danos à reputação.

Basicamente, ao montar um projeto de IA, certifique-se de que há um plano de segurança e que ele é seguido.

Qual é o volume de dados necessário para IA?

Quanto mais dados, melhor, certo? Esse em geral é o caso. Observe o fenômeno de Hughes. Ele diz que o desempenho geralmente aumenta à medida que recursos são adicionados ao modelo.

A quantidade, no entanto, não é o mais importante. Pode chegar um momento em que os dados comecem a se degradar. Lembre-se de que é possível cair na "maldição da dimensionalidade". De acordo com Charles Isbell, professor e decano associado sênior na Escola de Computação Interativa da Georgia Tech, "à medida que o número de características ou dimensões cresce, a quantidade de dados de que precisamos para generalizar com precisão cresce exponencialmente".[24]

Qual é o impacto prático? Pode tornar-se impossível ter um bom modelo, uma vez que a quantidade de dados pode não ser suficiente. É por isso que a maldição da dimensionalidade pode ser bastante problemática em aplicações como reconhecimento da visão. Mesmo ao analisar imagens RGB, o número de dimensões é de aproximadamente 7.500. Imagine o quão intenso o processo seria usando vídeo de alta definição em tempo real.

Mais termos e conceitos de dados

Ao se envolver com análise de dados, é preciso conhecer os termos básicos. Aqui estão alguns que você vai ouvir muitas vezes:

- *Dados categóricos*: são dados que não têm um significado numérico. Em vez disso, possuem um significado textual, como a descrição de um grupo (raça e sexo). Apesar disso, é possível atribuir números a cada um dos elementos.
- *Tipo de dado*: tipo de informação que uma variável representa, como booleano, número inteiro, string ou ponto flutuante.
- *Análise descritiva*: análise de dados para obter uma melhor compreensão do status atual de um negócio. Alguns exemplos incluem medir quais produtos estão vendendo melhor ou determinar riscos no suporte ao cliente. Há muitas ferramentas de software tradicionais para análise descritiva, como aplicativos de BI.
- *Análise diagnóstica*: consulta os dados para descobrir por que algo aconteceu. Esse tipo de análise usa técnicas como mineração de dados, árvores de decisão e correlações.
- *ETL (Extraction, Transformation, Load – extração, transformação, carga)*: forma de integração de dados normalmente usada em data warehouses.
- *Recurso*: uma coluna de dados.
- *Instância*: uma linha de dados.
- *Metadados*: dados sobre os dados – ou seja, descrições. Por exemplo, um arquivo de música pode ter metadados como seu tamanho, comprimento, data de

upload, comentários, gênero, artista e assim por diante. Esses tipos de dados podem acabar sendo bastante úteis para projetos de IA.

- *Dados numéricos*: dados que podem ser representados por um número. No entanto, esses dados podem ter duas formas. Há dados discretos, que são inteiros – ou seja, números sem casas decimais significativas, e dados contínuos, que têm um fluxo, como temperatura ou tempo.
- *OLAP (Online Analytical Processing – Processamento analítico online)*: tecnologia que permite analisar informações de várias bases de dados.
- *Dados ordinais*: mistura de dados numéricos e categóricos. Um exemplo comum deles é a classificação de cinco estrelas da Amazon.com; que tem tanto uma estrela quanto um número associado a ela.
- *Análise preditiva*: envolve o uso de dados para fazer previsões. Os modelos para isso geralmente são sofisticados e dependem de abordagens da IA como machine learning. Para ser eficaz, é importante atualizar o modelo subjacente com novos dados. Algumas das ferramentas para análise preditiva incluem abordagens de machine learning, como regressões.
- *Análise prescritiva*: alavancagem do big data para tomada de melhores decisões. Não se trata apenas de prever resultados, mas de compreender os fundamentos. É aqui que a IA desempenha um papel importante.
- *Variáveis escalares*: mantêm valores únicos como nome ou número de cartão de crédito.
- *Dados transacionais*: são gravados em ações financeiras, comerciais e logísticas. Entre os exemplos, há pagamentos, faturas e reivindicações de seguro.

Conclusão

Ser bem-sucedido com a IA significa ter uma cultura orientada por dados. Isso tem sido crítico para empresas como Amazon.com, Google e Facebook. Ao tomar decisões, elas olham para os dados primeiro. Também deve haver ampla disponibilidade de dados em toda a organização.

Sem essa abordagem, o sucesso com a IA será fugaz, independentemente de seu planejamento. Talvez isso ajude a explicar que – de acordo com um estudo da NewVantage Partners – cerca de 77% dos entrevistados afirmam que a "adoção de negócios" de big data e IA é desafiadora.[25]

Principais aprendizados

- Dados estruturados são rotulados e formatados – e geralmente são armazenados em um banco de dados relacional ou em uma planilha.

- Dados não estruturados são informações que não têm formatação predefinida.
- Dados semiestruturados têm algumas marcas internas que ajudam na categorização.
- Big data descreve uma maneira de lidar com enormes quantidades de volumes de informações.
- Um banco de dados relacional é baseado em relacionamentos de dados. Essa estrutura, no entanto, pode revelar-se difícil para aplicativos modernos, como os de IA.
- Um banco de dados NoSQL tem uma forma mais livre e baseia-se em um modelo de documento. Isso o tornou mais capaz de lidar com dados não estruturados e semiestruturados.
- O processo CRISP-DM oferece uma maneira de gerenciar os dados de um projeto, com etapas que incluem compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação.
- A quantidade de dados é certamente importante, mas bastante trabalho precisa ser feito para garantir sua qualidade. Até mesmo pequenos erros podem ter um enorme impacto sobre os resultados de um modelo de IA.

---

1 www.cnbc.com/2019/03/22/pinterest-releases-s-1-for-ipo.html

2 www.businessinsider.com/pinterest-accidental-marriage-emails-2014-9

3 www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2

4 www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#788c13c660ba

5 www.forbes.com/sites/louiscolumbus/2018/06/06/10-charts-that-will-challenge-your-perspective-of-iots-growth/#4e9fac23ecce

6 https://blog.seagate.com/business/enormous-growth-in-data-is-coming-how-to-prepare-for-it-and-prosper-from-it/

7 www.oracle.com/big-data/guide/what-is-big-data.html

8 https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

9 www.mongodb.com/what-is-mongodb

10 https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

11 www.idc.com/getdoc.jsp?containerId=prUS44215218

12 www.idc.com/getdoc.jsp?containerId=prUS44215218

13 www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/

14 Entrevista realizada pelo autor com Atif Kureishy em fevereiro 2019.

15 Entrevista realizada pelo autor com Prasad Vuyyuru em fevereiro de 2019.

16 N.T.: A porcentagem de slugging (Slugging Percentage – SLG) refere-se ao aproveitamento do rebatedor quanto ao número de bases conseguidas. A porcentagem de on base (On Base Percentage

–OBP) refere-se ao percentual de vezes que o jogador chegou na base. Fonte: http://www.blogdobeisebol.com/glossario-expressoes-do-baseball/

17 N.T.: "Um beacon é uma espécie de GPS interno que consegue localizar, com precisão incrível, por qual gôndola um cliente caminha dentro de uma loja de departamentos, por exemplo". Fonte: https://endeavor.org.br/estrategia-e-gestao/beacon/

18 Entrevista realizada pelo autor com Atif Kureishy em fevereiro de 2019.

19 N.T.: "Lixo entra, lixo sai" (em inglês, "garbage in, garbage out" – GIGO) é uma expressão atribuída ao técnico da IBM George Fuechsel. Fonte: https://pt.wikipedia.org/wiki/Garbage_in,_garbage_out

20 Entrevista realizada pelo autor com Eyal Lifshitz em fevereiro de 2019.

21 https://venturebeat.com/2018/07/02/u-s-agencies-widen-investigation-into-what-facebook-knew-about-cambridge-analytica/

22 http://news.mit.edu/2018/privacy-risks-mobility-data-1207

23 https://enterprise.verizon.com/resources/reports/dbir/

24 www.kdnuggets.com/2017/04/must-know-curse-dimensionality.html

25 http://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-Updated-010219-1.pdf

CAPÍTULO 3

Machine learning

Insights de mineração de dados

Um avanço em machine learning valeria dez Microsofts.

– Bill Gates[1]

Embora Katrina Lake gostasse de fazer compras online, ela sabia que a experiência poderia ser muito melhor. O principal problema é que era difícil encontrar itens de moda personalizados.

Daí surgiu a inspiração para a Stitch Fix. A empresa foi lançada no apartamento de Katrina, em Cambridge, enquanto ela frequentava a Harvard Business School, em 2011 (a propósito, o nome original para a empresa era menos cativante: "Rack Habit" – "Hábito de Rack"). O site tinha um fórum de perguntas e respostas para seus usuários – que discutia, entre outras características, tamanho e estilos de moda – e estilistas especializados montavam, então, caixas personalizadas de roupas e acessórios que eram enviadas mensalmente.

O conceito agradou rapidamente, e o crescimento foi substancial. No entanto, foi difícil levantar verba, pois muitos capitalistas de risco não viam potencial no negócio. Apesar disso, Katrina persistiu e conseguiu criar uma operação rentável – em um curto intervalo de tempo.

Ao longo do caminho, a Stitch Fix coletava enormes quantidades de dados valiosos, como tamanhos corporais e preferências de estilo. Katrina percebeu que eles seriam ideais para machine learning. A fim de aproveitar a oportunidade, contratou Eric Colson, vice-presidente de Ciência de Dados e Engenharia na Netflix, cujo novo título era Diretor de Algoritmos.

Essa mudança de estratégia foi fundamental. Os modelos de machine learning ficaram cada vez melhores nas previsões à medida que a Stitch Fix coletava mais dados – não só de pesquisas iniciais, mas também do feedback contínuo. Os dados também foram codificados nas SKUs (Stock Keeping Units – Unidades de Manutenção de Estoque).

O resultado é que a Stitch Fix viu uma melhoria contínua nas taxas de fidelização e conversão de clientes. Houve também melhorias no giro de estoque, o que ajudou a reduzir custos.

A nova estratégia, no entanto, não representava a demissão dos estilistas. Em vez disso, as técnicas de machine learning aumentaram muito sua produtividade e

eficiência.

Os dados também forneceram insights sobre quais tipos de roupas criar, o que levou ao lançamento da Hybrid Designs em 2017, marca própria da Stitch Fix. Ela se mostrou uma ótima solução para lidar com as lacunas no estoque.

Em novembro de 2017, Katrina tornou a Stitch Fix pública, levantando US$ 120 milhões. A empresa foi avaliada em maravilhosos US$ 1,63 bilhão – transformando-a numa das mulheres mais ricas nos Estados Unidos.[2] Ah, e, nessa época, Katrina tinha um filho de 14 meses de idade!

Atualmente, a Stitch Fix tem 2,7 milhões de clientes nos Estados Unidos e gera mais de US$ 1,2 bilhão em receitas. A empresa também emprega mais de 100 cientistas de dados, a maioria deles PhD em áreas como neurociência, matemática, estatística e IA.[3]

De acordo com informações fornecidas pela empresa:

> *Nossos recursos de ciência de dados alimentam nossos negócios. Eles consistem em nosso rico e crescente conjunto de dados detalhados de clientes e mercadorias e em nossos algoritmos proprietários. Usamos ciência de dados em todo o negócio, inclusive para criar estilos para nossos clientes, prever o comportamento de compra, antecipar a demanda, otimizar o estoque e criar roupas.* [4]

Sem dúvida, a história da Stitch Fix mostra claramente o incrível poder do machine learning e como ele pode movimentar uma indústria. Em entrevista ao digiday.com, Lake observou:

> *Historicamente, existe uma lacuna entre o que você dá às empresas e o quanto a experiência é aprimorada. O big data está rastreando você em toda a web e o maior benefício obtido até agora é: se você clicou em um par de sapatos, verá esse par de sapatos novamente daqui a uma semana. Vemos que essa lacuna começa a diminuir. As expectativas são muito diferentes em relação à personalização, mas o mais importante é que se obtenha uma versão autêntica dela. Não algo como, "você abandonou seu carrinho e estamos reconhecendo isso". Será um reconhecimento genuíno de quem você é como ser humano único. A maneira singular de fazer isso de forma escalável é adotando a ciência de dados e o que você pode fazer por meio da inovação.*[5]

Muito bem, então, o que é realmente machine learning? Por que pode ser tão impactante? E quais são alguns dos riscos a considerar?

Neste capítulo, vamos responder a essas perguntas – e muito mais.

O que é machine learning?

Depois de passagens pelo MIT e pelo Bell Telephone Laboratories, Arthur L. Samuel ingressou na IBM em 1949, no Poughkeepsie Laboratory. Seus esforços ajudaram a

aumentar o poder computacional das máquinas da empresa, como ocorreu com o desenvolvimento do 701 (primeiro sistema de computador comercializado pela IBM).

Ele também programou aplicativos. Um deles faria história: seu jogo de damas para computadores. Foi o primeiro exemplo de um sistema de machine learning (Samuel publicou um artigo influente sobre isso em 1959[6]). CEO da IBM, Thomas J. Watson Sr. disse que a inovação faria as ações da empresa crescerem 15 pontos![7]

Por que o artigo de Samuel foi tão significativo? Observando as damas, ele mostrou como machine learning funciona – em outras palavras, um computador poderia aprender e melhorar processando dados sem ter de ser explicitamente programado. Isso foi possível por conta de conceitos avançados de estatística, especialmente a análise de probabilidade. Assim, um computador poderia ser treinado para fazer previsões precisas.

O trabalho de Samuel foi revolucionário para o desenvolvimento de software que, naquele momento, resumia-se basicamente a uma lista de comandos que seguia um fluxo lógico de execução.

Para ter uma noção de como funcionava o machine learning, vamos usar um exemplo do programa de comédia *Silicon Valley*, da HBO TV. Em um dos episódios, o engenheiro Jian-Yang deveria criar o "Shazam for food", um aplicativo para identificar diferentes tipos de comida. Para treiná-lo, ele precisou fornecer um enorme conjunto de dados de imagens de alimentos. Infelizmente, devido a restrições de tempo, o aplicativo só aprendeu a identificar... cachorros-quentes. Em outras palavras, se você usasse o aplicativo, ele só retornaria "cachorro-quente" e "não cachorro-quente".

Embora bem-humorada, a situação fez um trabalho muito bom na demonstração do funcionamento do machine learning. Em essência, trata-se de um processo de adoção de dados rotulados (etiquetados) e busca de relacionamentos. Se o sistema for treinado com cachorros-quentes – com milhares de imagens –, vai ficar cada vez melhor em reconhecê-los.

Sim, até mesmo os programas de TV podem ensinar lições valiosas sobre IA!

É claro que muito mais é necessário. Na próxima seção do capítulo, vamos dar uma olhada mais aprofundada nos principais conceitos estatísticos necessários ao machine learning. Dentre eles, estão: desvio-padrão, distribuição normal, teorema de Bayes, correlação e extração de recursos.

Em seguida, serão discutidos tópicos como estudos de caso para machine learning, seu processo geral e algoritmos mais comuns.

Desvio-padrão

O desvio-padrão mede a dispersão dos valores em relação à média. Na verdade, não há necessidade de aprender a calculá-lo (o processo envolve vários passos), uma vez que o Excel ou outro software qualquer pode fazer isso por você facilmente.

Para compreender o desvio-padrão, considere como exemplo o valor das residências em seu bairro. Suponha que a média é de US$ 145.000 e o desvio-padrão é de US$ 24.000. Isso significa que um desvio-padrão abaixo da média seria US$ 133.000 (US$ 145.000 - US$ 12.000), enquanto um desvio-padrão acima da média chegaria a US$ 157.000 (US$ 145.000 + US$ 12.000). Isso nos oferece uma maneira de quantificar a variação nos dados. Ou seja, há um intervalo de US$ 24.000 a partir da média.

Em seguida, vamos dar uma olhada nos dados se Mark Zuckerberg se muda para o seu bairro e, como resultado, a média pula para US$ 850.000 e o desvio-padrão passa a ser de US$ 175.000. Essas medidas estatísticas refletem as avaliações? Na verdade, não. A compra de Zuckerberg é um ponto fora da curva. Nessa situação, a melhor abordagem pode ser desconsiderar a casa do magnata.

Distribuição normal

Quando esboçada em um gráfico, a distribuição normal se parece com um sino (é por isso que ela costuma ser chamada de "curva do sino") e representa a soma das probabilidades para uma variável. Curiosamente, a curva normal é comum no mundo real, pois reflete distribuições de dados como altura e peso.

Uma abordagem geral ao interpretar a distribuição normal é usar a regra 68-95-99,7, a qual estima que 68% dos itens de dados cairão dentro de um desvio-padrão, 95% estarão dentro de dois desvios-padrão e 99,7% se posicionarão dentro de três desvios-padrão.

Uma maneira de entender tudo isso é usando pontuações de QI. Suponha que a pontuação média é 100 e o desvio-padrão é 15. Teríamos isso para os três desvios-padrão, conforme mostrado na Figura 3.1.

Observe que o pico nesse gráfico é a média. Assim, se uma pessoa tem um QI de 145, então apenas 0,15% terá uma pontuação maior.

A curva, no entanto, pode ter formas diferentes, dependendo da variação nos dados. Por exemplo, se nossos dados sobre QI apresentam uma grande quantidade de gênios, a distribuição vai inclinar para a direita.

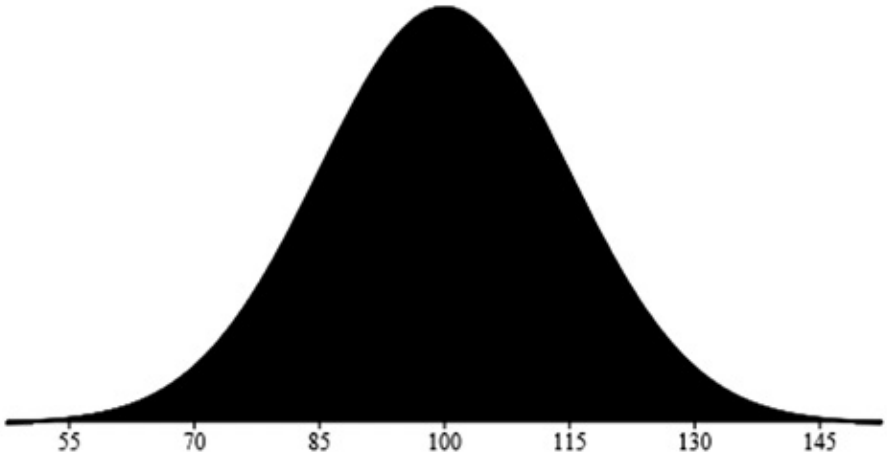

Figura 3.1 – Distribuição normal para pontuações de QI.

Teorema de Bayes

Como o nome indica, a estatística descritiva fornece informações sobre seus dados. Já vimos isso ao discutir tópicos como médias e desvios-padrão.

É claro que é possível ir muito além disso – basicamente, usando o teorema de Bayes. Essa abordagem é comum na análise de doenças médicas, nas quais causa e efeito são fundamentais – por exemplo, para os ensaios do FDA (Federal Drug Administration – Administração Federal de Medicamentos).

Para entender como o teorema de Bayes funciona, vejamos um exemplo. Um pesquisador formulou um teste para determinado tipo de câncer que provou ser preciso em 80% das vezes. Isso é conhecido como um verdadeiro positivo.

Em 9,6% das vezes, contudo, o teste identificou a pessoa como tendo o câncer, embora o paciente não o tenha, o que é conhecido como um falso positivo. Tenha em mente que – em alguns testes de medicamentos – esse percentual pode ser maior do que a taxa de precisão!

Por fim, 1% da população tem o câncer.

À luz de tudo isso, se um médico aplica o teste em você e o resultado é que você tem o câncer, qual é a probabilidade de que você realmente o tenha? Bem, o teorema de Bayes vai mostrar o caminho. Esse cálculo usa fatores como taxas de precisão, falsos positivos e taxa populacional para chegar a uma probabilidade:

- *Passo #1:* taxa de precisão de 80% × chance de ter o câncer (1%) = 0,008.
- *Passo #2:* chance de não ter o câncer (99%) × 9,6% de falso positivo = 0,09504.
- *Passo #3:* inserir os números anteriores na equação: 0,008 / (0,008 + 0,09504) = 7,8%.

Parece meio estranho, certo? Definitivamente. Afinal, como um teste com 90% de precisão tem apenas 7,8% de probabilidade de estar certo? Lembre-se, no entanto, de

que a taxa de precisão se baseia na medida daqueles que têm a doença; e esse é um número pequeno, uma vez que apenas 1% da população a tem. Além disso, o teste ainda está distribuindo falsos positivos. Assim, o teorema de Bayes é uma maneira de fornecer uma melhor compreensão dos resultados – o que é fundamental para sistemas como a IA.

Correlação

Um algoritmo de machine learning geralmente envolve algum tipo de correlação entre os dados. Uma maneira quantitativa de descrever isso é usar a correlação de Pearson, que mostra a força da relação entre duas variáveis que vão de 1 a -1 (esse é o coeficiente).

Veja como funciona:

- *Maior que 0*: aqui, um aumento em uma variável leva a um aumento em outra. Por exemplo: Suponha que existe uma correlação de 0,9 entre renda e gastos. Se a renda aumenta US$ 1.000, então, os gastos subirão em US$ 900 (US$ 1.000 x 0,9).
- *0*: não há correlação entre as variáveis.
- *Menor que 0*: qualquer aumento em uma variável significa uma diminuição em outra e vice-versa. Isso descreve uma relação inversa.

O que é, então, uma correlação forte? Como regra geral, ela ocorre se o coeficiente é de 0,7 ou mais. Se for inferior a 0,3, então a correlação é fraca.

Tudo isso remete ao velho ditado que diz que "correlação não é, necessariamente, causalidade". No entanto, quando se trata de machine learning, esse conceito pode ser facilmente ignorado e levar a resultados enganosos.

Existem, por exemplo, muitas correlações que são apenas aleatórias. Na verdade, algumas chegam a ser cômicas. Veja as correlações a seguir, obtidas na Tylervigen.com:[8]

- A taxa de divórcio em Maine tem uma correlação de 99,26% com o consumo per capita de margarina.
- A idade da Miss América tem uma correlação de 87,01% com os assassinatos por vapor, vapores quentes e trópicos quentes.
- As importações de petróleo bruto dos Estados Unidos provenientes da Noruega têm uma correlação de 95,4% com os motoristas mortos em colisão com um trem ferroviário.

Há um nome para isso: padronicidade. Trata-se de uma tendência para encontrar padrões em ruídos sem sentido.

Extração de recursos

No Capítulo 2, analisamos a seleção das variáveis para um modelo. O processo é muitas vezes chamado de extração de recursos ou engenharia de recursos.

Um exemplo disso seria um modelo de computador que identifica um homem ou uma mulher a partir de uma foto. Para os seres humanos, isso é bastante fácil e rápido. Trata-se de algo intuitivo. Contudo, se alguém lhe pedisse que descrevesse as diferenças, você conseguiria? Para a maioria das pessoas, essa seria uma tarefa difícil. No entanto, para construir um modelo eficiente de machine learning, é preciso realizar adequadamente a extração de recursos – e isso pode ser subjetivo.

A Tabela 3.1 mostra algumas ideias relacionadas a como o rosto de um homem pode diferir do de uma mulher.

Isso é só o começo e tenho certeza de que você tem suas próprias ideias ou abordagens. E isso é normal. No entanto, é também por isso que implementações como o reconhecimento facial são altamente complexas e sujeitas ao erro.

A extração de recursos também tem alguns problemas diferenciados. Um deles é o potencial de ser tendenciosa. Por exemplo, você tem ideias preconcebidas sobre como é um homem ou mulher? Caso afirmativo, isso pode resultar em modelos que fornecerão resultados equivocados.

Tabela 3.1 – Características faciais

Características	Homem
Sobrancelhas	Mais grossas e retas
Formato do rosto	Mais longo e largo, mais quadrado
Maxilar	Quadrado, mais largo e marcado
Pescoço	Pomo de Adão

Por conta de tudo isso, é uma boa ideia ter um grupo de especialistas que podem determinar as características corretas. E se a engenharia de recursos se revelar muito complexa, então machine learning provavelmente não é uma boa opção.

Há, entretanto, outra abordagem a considerar: deep learning. Ela envolve modelos sofisticados que encontram recursos em dados. Na verdade, essa é uma das razões pelas quais a abordagem tem sido um grande avanço na IA. Vamos aprender mais sobre ela no próximo capítulo.

O que se pode fazer com machine learning?

Como machine learning existe há décadas, tem havido muitos usos para essa poderosa tecnologia. Vale a pena saber que ela oferece benefícios claros em termos

de redução de custos, oportunidades de receita e monitoramento de riscos.

Para dar uma noção da variedade de aplicações, observe alguns exemplos:

- *Manutenção preditiva:* monitora sensores para prever quando o equipamento pode falhar. Não só ajuda a reduzir custos, mas também diminui o tempo de inatividade e aumenta a segurança. Na verdade, empresas como a PrecisionHawk estão usando drones para coletar dados, o que é muito mais eficiente. A tecnologia tem se mostrado bastante útil para indústrias de setores como energia, agricultura e construção. Veja o que a empresa comenta sobre o próprio sistema de manutenção preditiva baseado em drones: "Um cliente testou drones em operação VLOS (Visual Line Of Sight – Condições Meteorológicas Visuais – VMC) para inspecionar um grupo de 10 áreas de perfuração em um raio de três milhas. Nosso cliente determinou que o uso da nova tecnologia reduziu os custos de inspeção em aproximadamente 66%, de US$ 80 - US$ 90 gastos com a metodologia de inspeção tradicional para US$ 45 - US$ 60 usando incursões de drones VLOS".[9]
- *Recrutamento de funcionários:* pode ser um processo tedioso, uma vez que os currículos são muito variados. Isso significa que é fácil não identificar bons candidatos. Machine learning, entretanto, pode ajudar em grande estilo. Dê uma olhada no CareerBuilder, que coletou e analisou mais de 2,3 milhões de empregos, 680 milhões de perfis únicos, 310 milhões de currículos únicos, 10 milhões de cargos, 1,3 bilhão de habilidades e 2,5 milhões de antecedentes para construir a Hello to Hire. Trata-se de uma plataforma que se aproveitou de machine learning para reduzir o número de candidaturas a empregos para uma média de 75 – para uma contratação bem-sucedida. A média do setor, por outro lado, é de cerca de 150.[10] O sistema automatiza a criação de descrições de emprego, levando em conta, inclusive, detalhes com base em indústria e localização!
- *Experiência do cliente:* hoje em dia, os clientes querem uma experiência personalizada. Eles se acostumaram com isso usando serviços como Amazon.com e Uber. Com machine learning, uma empresa pode aproveitar seus dados para obter informações e aprender sobre o que realmente funciona. Isso é tão importante que levou a Kroger a comprar uma empresa no espaço chamada 84,51°. É definitivamente fundamental que ela disponha de dados sobre mais de 60 milhões de famílias dos Estados Unidos. Aqui está um estudo de caso rápido: na maioria de suas lojas, a Kroger oferecia abacates a granel e apenas alguns poucos eram oferecidos em pacotes de 4 unidades. O senso comum era de que os pacotes com 4 unidades precisavam ser mais baratos por conta da disparidade de tamanho em relação aos itens a granel. Ao aplicar a análise de machine learning, entretanto, essa ideia se provou incorreta, já que os pacotes com 4

unidades atraíram famílias novas e diferentes, como compradores da geração Y e da ClickList. Ao oferecer esse tipo de pacote por toda a rede, houve um aumento global nas vendas de abacate.[11]

- *Finanças*: machine learning pode detectar discrepâncias, por exemplo, no faturamento. No entanto, há uma nova categoria de tecnologia, chamada RPA (Robotic Process Automation – Automação Robótica de Processos), que pode ajudar com isso (vamos abordar esse tema no Capítulo 5). Ela automatiza processos de rotina a fim de ajudar a reduzir os erros e pode usar machine learning para detectar transações anormais ou suspeitas.
- *Atendimento ao cliente*: nos últimos anos, foi possível presenciar o crescimento dos chatbots, que usam machine learning para automatizar as interações com os clientes. Cobriremos esse tópico no Capítulo 6.
- *Namoro:* machine learning pode ajudar a encontrar sua alma gêmea! O Tinder, um dos maiores aplicativos de namoro, está usando a tecnologia para ajudar a melhorar os matches. O aplicativo tem um sistema que rotula automaticamente mais de 10 bilhões de fotos enviadas diariamente, por exemplo.

A Figura 3.2 apresenta algumas das aplicações para machine learning.

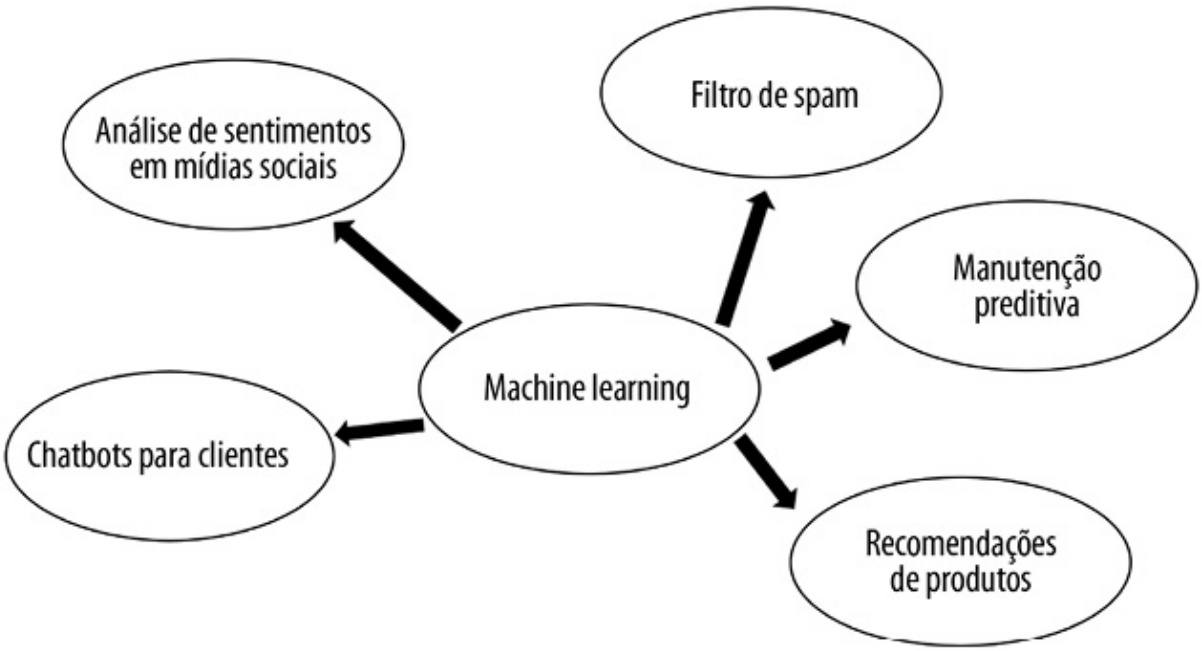

Figura 3.2 – Aplicações para machine learning.

Processo de machine learning

Para ser bem-sucedido com a aplicação de machine learning a um problema, é importante adotar uma abordagem sistemática. Caso contrário, os resultados podem ser totalmente equivocados.

Em primeiro lugar, é necessário passar por um processo de dados, discutido no

capítulo anterior. Feito isso, é uma boa ideia fazer uma visualização dos dados. Eles estão dispersos na maior parte? Ou existem alguns padrões? Se a resposta for sim, então os dados podem ser bons candidatos para machine learning.

O objetivo do processo de machine learning é criar um modelo que se baseie em um ou mais algoritmos. Isso é alcançado por meio do treinamento do modelo. O objetivo é que ele forneça alto grau de previsibilidade.

Agora, vamos dar uma olhada mais de perto nisso (a propósito, isso também se aplica a deep learning, que vamos discutir no próximo capítulo):

Etapa #1 – Ordenação dos dados

Se os dados forem classificados, isso pode distorcer os resultados. Ou seja, o algoritmo de machine learning pode detectar a ordenação como um padrão! Portanto, é uma boa ideia randomizar a ordem dos dados.

Etapa #2 – Escolha do modelo

Será necessário selecionar um algoritmo. Essa escolha envolverá um processo de tentativa e erro. Neste capítulo, vamos abordar os diversos algoritmos disponíveis.

Etapa #3 – Treinamento do modelo

Os dados de treinamento, que serão cerca de 70% do conjunto completo, serão usados para criar as relações no algoritmo. Suponha, por exemplo, que você esteja construindo um sistema de machine learning para encontrar o valor de um carro usado. Algumas das características a serem consideradas são ano de fabricação, marca, modelo, quilometragem e estado do veículo. Ao processar esses dados de treinamento, o algoritmo calculará os pesos para cada um desses fatores.

Por exemplo: suponha que se esteja usando um algoritmo de regressão linear que tem o seguinte formato:

```
y = m * x + b
```

Na fase de treinamento, o sistema apresentará valores para m (que é a inclinação de um gráfico) e b (que é a interceptação em y).

Etapa #4 – Avaliação do modelo

Será preciso reunir dados de teste, formados pelos 30% restantes do conjunto. Eles devem ser representativos das faixas e do tipo de informação nos dados de treinamento.

Com os dados de teste, será possível avaliar se o algoritmo é preciso. No exemplo do carro usado, os valores de mercado são consistentes com o que está acontecendo no mundo real?

Observação: os dados de treinamento e teste não devem ser misturados, pois isso pode facilmente levar a resultados distorcidos. Curiosamente, esse é um erro comum.

A precisão é uma medida de sucesso do algoritmo. Entretanto, em alguns casos, ela pode ser enganosa. Considere a situação de detecção de fraude. Geralmente há um pequeno número de recursos quando comparado a um conjunto de dados. A falta de um deles, no entanto, poderia ser devastadora e custar a uma empresa milhões de dólares em perdas.

É por isso que outras abordagens podem ser necessárias, como o teorema de Bayes.

Etapa #5 – Sintonia fina do modelo

Nessa etapa, é possível ajustar os valores dos parâmetros no algoritmo. A intenção é verificar se é possível obter melhores resultados.

Ao realizar a sintonia fina do modelo, pode-se detectar a existência de hiperparâmetros; que são aqueles que não podem ser aprendidos diretamente com o processo de treinamento.

Aplicando algoritmos

Alguns algoritmos são muito fáceis de calcular, enquanto outros exigem etapas e matemática complexas. A boa notícia é que geralmente não há necessidade de computar um algoritmo, já que existe uma variedade de linguagens, como Python e R, que tornam o processo simples.

No que se refere a machine learning, o algoritmo é tipicamente diferente de um tradicional. A razão é que o primeiro passo é processar dados para, em seguida, o computador começar a aprender.

Mesmo que existam centenas de algoritmos de machine learning disponíveis, eles normalmente podem ser divididos em quatro categorias principais: aprendizagem supervisionada, aprendizagem não supervisionada, aprendizagem por reforço e aprendizagem semissupervisionada. Vamos dar uma olhada em cada uma delas.

Aprendizagem supervisionada

A aprendizagem supervisionada usa dados rotulados. Suponha, por exemplo, que temos um conjunto de fotos de milhares de cães. Os dados são considerados rotulados se cada foto identificar cada uma das raças. Na maioria dos casos, isso torna a análise mais fácil, uma vez que os resultados podem ser comparados com a resposta correta.

Uma das características principais na aprendizagem supervisionada é que deve haver grandes quantidades de dados. Isso ajuda a refinar o modelo e produzir resultados mais precisos.

No entanto, há um grande problema: a realidade é que grande parte dos dados disponíveis não está rotulada. Além disso, pode ser demorado fornecer rótulos se houver um conjunto de dados massivo.

Apesar disso, existem maneiras criativas de lidar com o problema, tal como recorrer ao financiamento colaborativo (crowdfunding). Foi assim que o sistema ImageNet foi construído e representou um avanço na inovação em IA. Contudo, vários anos foram necessários para criá-lo.

Em alguns casos, é possível que existam abordagens automatizadas para rotular os dados. Considere o exemplo do Facebook. Em 2018, a empresa anunciou – em sua conferência de desenvolvedores F8 – que aumentou seu enorme banco de dados de fotos a partir do Instagram, onde as imagens já estavam rotuladas com hashtags.[12]

Essa abordagem, contudo, apresentava suas falhas. Uma hashtag pode dar uma descrição não visual da foto – digamos #tbt (que significa "ThrowBack Thursday") – ou poderia ser muito vaga, como #festa. É por isso que o Facebook chamou sua abordagem de "dados fracamente supervisionados". Os talentosos engenheiros da empresa, entretanto, encontraram algumas maneiras de melhorar a qualidade, tal como a construção de um sofisticado modelo de previsão de hashtags.

No final, as coisas funcionaram muito bem. O modelo de machine learning do Facebook, que incluiu 3,5 bilhões de fotos, tinha uma taxa de precisão de 85,4% baseada nas taxas de reconhecimento da ImageNet. Na verdade, por 2%, esse foi o índice mais alto registrado na história.

Esse projeto de IA também exigiu abordagens inovadoras para a construção da infraestrutura. De acordo com uma postagem no blog do Facebook:

> *Como uma única máquina levaria mais de um ano para concluir o treinamento do modelo, criamos uma forma de distribuir a tarefa em até 336 GPUs, encurtando o tempo de treinamento total para apenas algumas semanas. Com modelos de tamanhos cada vez maiores – o maior nesta pesquisa é um ResNeXt 101-32x48d com mais de 861 milhões de parâmetros – esse treinamento distribuído torna-se cada vez mais essencial. Além disso, projetamos um método para remover duplicidades de modo a garantir que não treinemos acidentalmente nossos modelos em imagens com as quais queremos avaliá-los, um problema que assola pesquisas semelhantes nesta área.*[13]

Daqui para a frente, o Facebook vê potencial em usar sua abordagem em várias áreas, incluindo as seguintes:

- Classificação melhorada no feed de notícias
- Melhoria na detecção de conteúdo censurável
- Geração automática de legendas para deficientes visuais

Aprendizagem não supervisionada

A aprendizagem não supervisionada acontece quando se está trabalhando com dados não rotulados. Isso significa que serão usados algoritmos de deep learning para detectar padrões.

A abordagem mais comum para a aprendizagem não supervisionada é o agrupamento (clustering), que manipula dados não rotulados e usa algoritmos para colocar itens semelhantes em grupos. O processo geralmente começa com suposições, seguidas de iterações dos cálculos para obtenção de melhores resultados. No centro disso está a busca por itens de dados que estão próximos uns dos outros, o que pode ser feito por meio de uma variedade de métodos quantitativos:

- *Métrica euclidiana:* trata-se de uma linha reta entre dois pontos de dados. A métrica euclidiana é bastante comum no machine learning.
- *Métrica de similaridade do cosseno:* como o nome indica, usa-se um cosseno para medir o ângulo. A ideia é encontrar semelhanças entre dois pontos de dados em termos de orientação.
- *Métrica de Manhattan*: envolve tomar a soma das distâncias absolutas entre dois pontos nas coordenadas de um gráfico. É chamada de "Manhattan" porque faz referência ao layout da cidade, que permite que sejam percorridas distâncias mais curtas nas viagens.

Em termos de casos de uso para agrupamento, um dos mais comuns é a segmentação de clientes, que ajuda a direcionar melhor as mensagens de marketing. Na maior parte das vezes, é possível que um grupo com características semelhantes partilhe interesses e preferências.

Outra aplicação é a análise de sentimento, que é onde ocorre a mineração de dados de mídia social e encontram-se as tendências. Para uma empresa de moda, isso pode ser crucial para entender como adaptar os estilos à próxima linha de roupas.

Existem também outras abordagens além do agrupamento. Dê uma olhada em mais três:

- *Associação*: o conceito básico é que, se X acontecer, então é provável que Y aconteça. Assim sendo, se você comprar o meu livro sobre IA, provavelmente vai querer comprar outros títulos do gênero. Com associação, um algoritmo de deep learning pode decifrar esses tipos de relacionamentos, o que pode resultar em poderosos motores de recomendação.
- *Detecção de anomalias*: identifica discrepâncias ou padrões anômalos no conjunto de dados e pode ser útil em aplicativos de cibersegurança. De acordo com Asaf Cidon, VP de Segurança de Emails da Barracuda Networks: "descobrimos que, combinando muitos sinais diferentes – como corpo do e-mail, cabeçalho, gráfico social das comunicações, logins IP, regras para

encaminhar para a caixa de entrada *etc.* –, somos capazes de alcançar uma precisão extremamente alta na detecção de ataques de engenharia social, mesmo que eles sejam altamente personalizados e criados para atingir uma determinada pessoa dentro de uma determinada empresa. Machine learning nos permite detectar ataques que se originam dentro da organização, cuja fonte é uma caixa de correio legítima de um empregado, o que seria impossível de fazer com um motor estático de regra única".[14]

- *Autoencoders:* com eles, os dados serão inseridos de forma compactada e, em seguida, serão reconstruídos. A partir disso, novos padrões podem surgir. No entanto, o uso de autoencoders é raro, mas pode ser útil para ajudar com aplicativos, realizando a diminuição de ruído nos dados, por exemplo.

Considere que muitos pesquisadores de IA acreditam que a aprendizagem não supervisionada provavelmente será crítica para o próximo nível de realizações. De acordo com um artigo na *Nature* escrito por Yann LeCun, Geoffrey Hinton e Yoshua Bengio, "espera-se que a aprendizagem não supervisionada se torne muito mais importante no longo prazo. As aprendizagens humana e animal são, em grande parte, não supervisionadas: descobrimos a estrutura do mundo observando-o, não porque nos dizem o nome de cada objeto".[15]

Aprendizagem por reforço

Quando você era criança e queria praticar um novo esporte, é possível que não tenha lido um manual. Em vez disso, observou o que outras pessoas faziam e tentou descobrir como as coisas funcionavam. Em algumas situações, cometeu erros e perdeu a bola enquanto seus companheiros de equipe demonstravam descontentamento. Em outras, entretanto, você fez os movimentos certos e marcou pontos. Por meio desse processo de tentativa e erro, seu aprendizado foi melhorando com base em reforços positivos e negativos.

De certo modo, isso é análogo à aprendizagem por reforço. Ela tem sido fundamental para algumas das realizações mais notáveis em IA, como as seguintes:

- *Jogos*: são ideais para a aprendizagem por reforço, uma vez que existem regras claras, pontuações e várias restrições (como um tabuleiro de jogo). Ao construir um modelo, é possível testá-lo com milhões de simulações, o que significa que o sistema vai ficando cada vez mais inteligente. É assim que um programa pode aprender a vencer o campeão mundial de Go ou xadrez.
- *Robótica:* o principal é ser capaz de se movimentar dentro de um espaço – e isso requer avaliação do ambiente em muitos pontos diferentes. Se o robô quer ir até a cozinha, por exemplo, terá de se mover em torno de móveis e outros obstáculos. Se bater nas coisas, haverá uma ação de reforço negativo.

Aprendizagem semissupervisionada

Essa é uma mistura de aprendizagem supervisionada e não supervisionada que surge quando se tem uma pequena quantidade de dados não rotulados. É possível, no entanto, usar sistemas de deep learning para transformar os dados não supervisionados em dados supervisionados – um processo chamado de pseudorrotulagem. Depois disso, os algoritmos podem ser aplicados.

Um caso de uso interessante de aprendizagem semissupervisionada é a interpretação de ressonâncias magnéticas. Um radiologista pode iniciar rotulando os exames e, depois disso, um sistema de deep learning pode encontrar o restante dos padrões.

Tipos comuns de algoritmos de machine learning

Simplesmente não há espaço suficiente neste livro para cobrir todos os algoritmos de machine learning! Portanto, é melhor nos concentrarmos nos mais comuns.

Na parte restante deste capítulo, vamos dar uma olhada no seguinte:

- *Aprendizagem supervisionada:* os algoritmos podem se resumir em duas variações. Uma delas é a classificação, que divide o conjunto de dados em rótulos comuns. Entre os exemplos desses algoritmos estão o classificador Naive Bayes e o k-NN (redes neurais serão discutidas no Capítulo 4). Em seguida, tem-se a regressão, que encontra padrões contínuos nos dados. Para compreendê-la, vamos dar uma olhada em regressão linear, modelagem por agrupamento (ensemble modelling) e árvores de decisão.
- *Aprendizagem não supervisionada:* nessa categoria, vamos discutir o agrupamento (clustering). Para isso, vamos falar sobre agrupamento k-means.

A Figura 3.3 apresenta uma estrutura geral para algoritmos de machine learning.

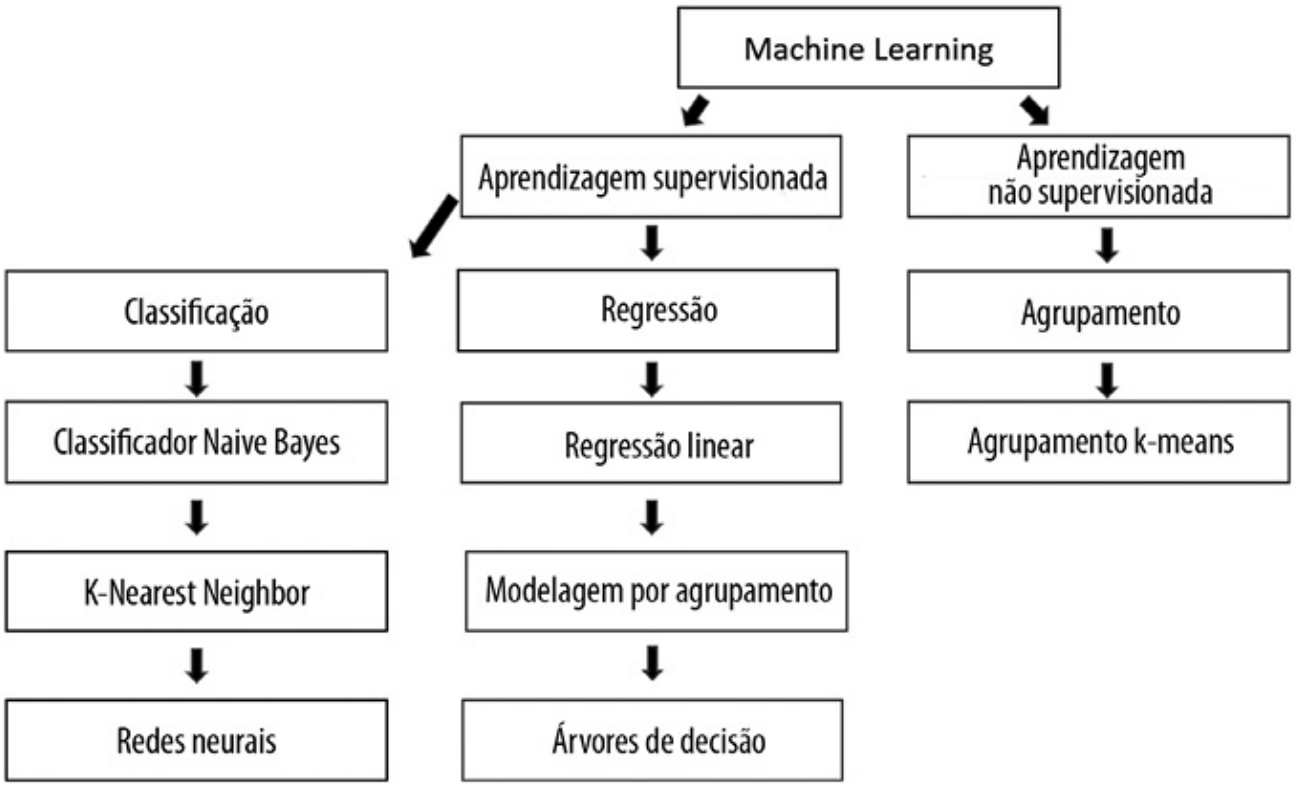

Figura 3.3 – Framework geral para algoritmos de machine learning.

Classificador Naive Bayes (Aprendizagem supervisionada/Classificação)

No início deste capítulo, apresentamos o teorema de Bayes. No que se refere a machine learning, o teorema foi transformado no classificador Naive Bayes. Ele é "naive" (em português, "ingênuo") porque a suposição é que as variáveis são independentes umas das outras – isto é, a ocorrência de uma variável não tem nenhuma relação com as outras. É verdade que isso pode parecer uma desvantagem. No entanto, o fato é que o classificador Naive Bayes provou ser bastante eficaz e rápido de ser desenvolvido.

Há outra premissa a ser observada: a suposição a priori. Ela diz que as previsões estarão erradas se os dados forem alterados.

Existem três variações do classificador Naive Bayes:

- *Bernoulli*: quando se tem dados binários (verdadeiro/falso, sim/não).
- *Multinomial*: quando os dados são discretos, tal como o número de páginas de um livro.
- *Gaussian:* quando se está trabalhando com dados que estão em conformidade com uma distribuição normal.

Um caso de uso comum para os classificadores Naive Bayes é a análise de texto. Os exemplos incluem detecção de spam por e-mail, segmentação de clientes, análise de sentimentos, diagnóstico médico e previsões meteorológicas. A razão para esses usos

é que essa abordagem é útil na classificação de dados com base em características-chave e padrões.

Para ver como isso é feito, vejamos um exemplo: suponha que você tenha um site de comércio eletrônico e conta com um grande banco de dados de transações de clientes. Você quer descobrir de que maneira variáveis como avaliações de produtos, descontos e tempo de vendas podem afetar as vendas.

A Tabela 3.2 apresenta uma parte do conjunto de dados.

Tabela 3.2 – Conjunto de dados de transações de clientes

Desconto	Avaliação do produto	Compra
Sim	Alta	Sim
Sim	Baixa	Sim
Não	Baixa	Não
Não	Baixa	Não
Não	Baixa	Não
Não	Alta	Sim
Sim	Alta	Não
Sim	Baixa	Sim
Não	Alta	Sim
Sim	Alta	Sim
Não	Alta	Não
Não	Baixa	Sim
Sim	Alta	Sim
Sim	Baixa	Não

Em seguida, esses dados serão organizados em tabelas de frequência, como mostrado nas tabelas 3.3 e 3.4.

Tabela 3.3 – Tabela de frequência de desconto

		Compra	
		Sim	Não
Desconto	Sim	19	1
	Sim	5	5

Tabela 3.4 – Tabela de frequência de avaliação do produto

		Compra		
		Sim	Não	Total
Avaliação do produto	Alta	21	2	11
	Baixa	3	4	8
	Total	24	6	19

Ao observar as tabelas, chamamos de evento a compra e de variáveis independentes o desconto e as avaliações do produto. Então, é possível fazer uma tabela de probabilidade para uma das variáveis independentes, como as avaliações do produto. Veja a Tabela 3.5.

Tabela 3.5 – Tabela de probabilidade conforme avaliação do produto

		Compra		
		Sim	Não	
Avaliação do produto	Alta	9/24	2/6	11/30
	Baixa	7/24	1/6	8/30
		24/30	6/30	

Usando essa representação, é possível observar que a probabilidade de uma compra quando há uma baixa avaliação do produto é de 7/24 ou 29%. Em outras palavras, o classificador Naive Bayes permite que sejam feitas previsões mais granulares dentro de um conjunto de dados. Também é relativamente fácil treiná-lo e ele pode funcionar bem com pequenos conjuntos de dados.

K-Nearest Neighbor (Aprendizagem supervisionada/Classificação)

O método k-Nearest Neighbor (k-NN) é usado para classificar um conjunto de dados (k representa o número de vizinhos). A teoria é que é possível que os valores próximos sejam bons preditores em um modelo. Pense nisso como "farinha do mesmo saco".

Um caso de uso para k-NN é a pontuação de crédito, que se baseia em uma variedade de fatores como renda, histórico de pagamento, localização, casa própria e assim por diante. O algoritmo dividirá o conjunto de dados em diferentes segmentos de clientes. Em seguida, quando um novo cliente for adicionado à base, será possível ver em qual grupo (cluster) o cliente é alocado – e essa será sua pontuação de crédito.

K-NN é realmente simples de calcular. Na verdade, ele é chamado de aprendizagem preguiçosa porque não há nenhum processo de treinamento com os dados.

Para usar k-NN, é necessário chegar à distância entre os valores mais próximos. Se os valores forem numéricos, é possível basear-se numa distância euclidiana, que envolve matemática complicada. Ou, se houver dados categóricos, então é possível usar uma métrica de sobreposição (é aqui que os dados são os mesmos ou muito semelhantes).

Em seguida, será necessário identificar o número de vizinhos. Embora uma maior quantidade regule o modelo, isso também pode significar uma demanda por uma enorme quantidade de recursos computacionais. Para gerenciar essa questão, pode-

se atribuir pesos mais altos aos dados que estão mais próximos de seus vizinhos.

Regressão linear (Aprendizagem supervisionada/Regressão)

A regressão linear mostra a relação entre certas variáveis. A equação – supondo que haja dados de qualidade suficientes – pode ajudar a prever os resultados com base em entradas.

Por exemplo: suponha que tenhamos dados sobre o número de horas gastas estudando para uma prova e a nota obtida. Observe a Tabela 3.6.

Tabela 3.6 – Tabela para horas de estudo e notas

Horas de estudo	Percentual da nota
1	0,75
1	0,69
1	0,71
3	0,82
3	0,83
4	0,86
5	0,85
5	0,89
5	0,84
6	0,91
6	0,92
7	0,95

Como se pode verificar, a relação geral é positiva (o que descreve a tendência em que um grau superior está correlacionado com mais horas de estudo). Com o algoritmo de regressão, é possível traçar uma linha que tenha o melhor ajuste (o que é feito usando um cálculo chamado "menos quadrados", que minimiza os erros). Veja a Figura 3.4.

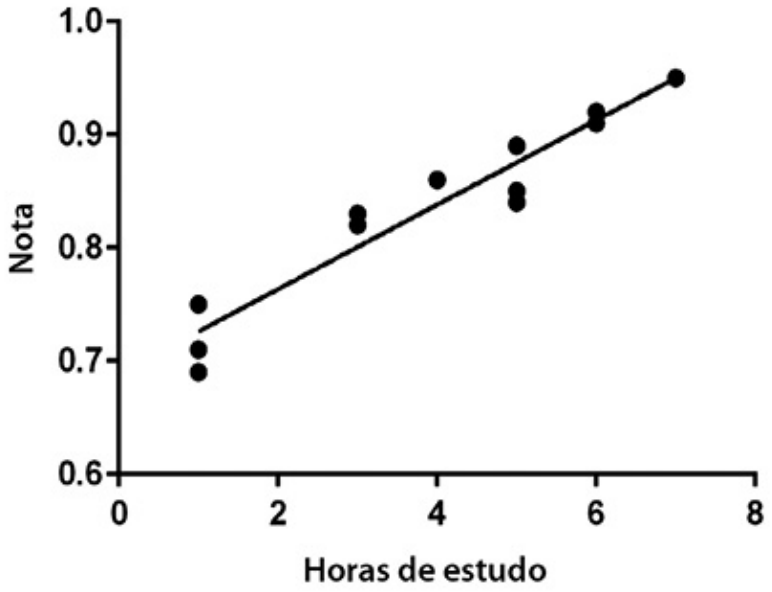

Figura 3.4 – Gráfico de um modelo de regressão linear baseado em horas de estudo.

A partir disso, tem-se a seguinte equação:

```
Nota = Número de horas de estudo × 0,03731 + 0,6889
```

Vamos supor, então, que você estuda durante 4 horas para a prova. Qual será a sua nota estimada? A equação nos diz que:

```
0,838 = 4 × 0,03731 + 0,6889
```

Qual é a precisão desse cálculo? Para ajudar a responder a essa pergunta, podemos usar um cálculo chamado R-quadrado. No nosso exemplo, seu valor é 0,9180 (ele varia de 0 a 1). Quanto mais próximo o valor estiver de 1, melhor será o ajuste. Assim, 0,9180 é bastante elevado, o que significa que as horas de estudo explicam 91,8% da nota na avaliação.

É verdade, no entanto, que esse modelo é simplista. Para melhor refletir a realidade, é possível adicionar mais variáveis para explicar a nota na avaliação – como a frequência do aluno. Ao fazer isso, utiliza-se algo chamado regressão multivariada.

Nota Se o coeficiente para uma variável é muito pequeno, pode ser uma boa ideia não a incluir no modelo.

Às vezes, os dados também podem não estar em linha reta, caso no qual o algoritmo de regressão não funcionará. É possível, contudo, recorrer a uma versão mais complexa, chamada regressão polinomial.

Árvore de Decisão (Aprendizagem supervisionada/Regressão)

Não há dúvidas de que o agrupamento pode não funcionar em alguns conjuntos de dados. A boa notícia, entretanto, é que existem alternativas, como uma árvore de decisão. Essa abordagem em geral funciona melhor com dados não numéricos.

O início de uma árvore de decisão é o nó de raiz, que está no topo do fluxograma. A

partir desse ponto, haverá uma árvore de caminhos de decisão chamados de divisões. Nesses pontos, um algoritmo será usado para tomada de decisão e uma probabilidade será calculada. No final da árvore estará a folha (ou resultado).

Nos círculos de machine learning, um exemplo famoso recorre a uma árvore de decisão para o trágico naufrágio do Titanic. O modelo prevê a sobrevivência de um passageiro com base em três características: sexo, idade e número de cônjuges ou filhos viajando com o passageiro (sibsp – SIBlings/SPouses – filhos/cônjuges). Observe a Figura 3.5.

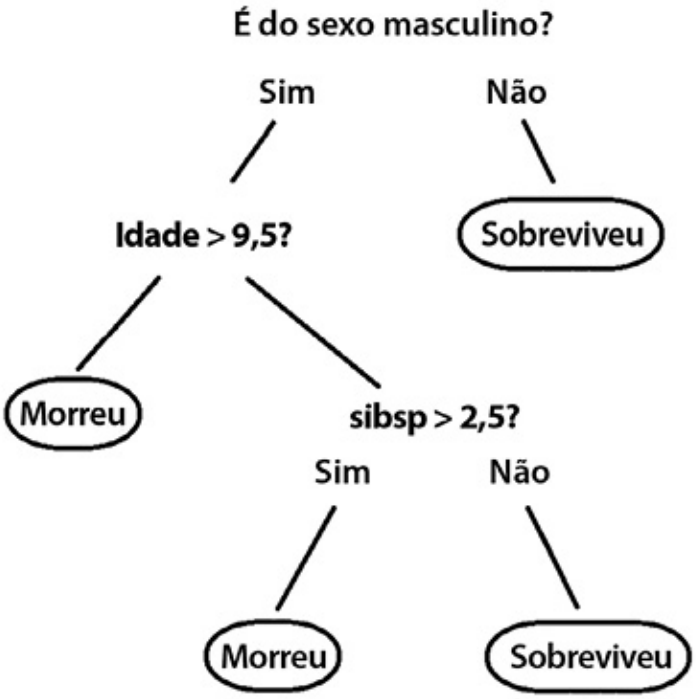

Figura 3.5 – Algoritmo básico de árvore de decisão para previsão de sobreviventes no naufrágio do Titanic.

Existem vantagens claras nas árvores de decisão. Elas são fáceis de entender, funcionam bem com grandes conjuntos de dados e fornecem transparência com o modelo.

No entanto, as árvores de decisão também têm desvantagens. Uma delas é a propagação de erros. Se uma das divisões se mostrar equivocada, então esse erro pode se espalhar por todo o resto do modelo!

Em seguida, à medida que as árvores de decisão crescem, aumenta também a complexidade, visto que haverá muitos algoritmos. Isso pode resultar em menor desempenho para o modelo.

Modelagem por agrupamento (Aprendizagem supervisionada/Regressão)

A modelagem por agrupamento (ensemble modelling) usa mais de um modelo para as previsões. Mesmo que isso aumente a complexidade, essa abordagem tem gerado

resultados sólidos.

Para vê-la em funcionamento, dê uma olhada no "Prêmio Netflix", que começou em 2006. A empresa anunciou que pagaria US$ 1 milhão para qualquer pessoa ou equipe que pudesse melhorar a precisão de seu sistema de recomendação de filmes em 10% ou mais. A Netflix também forneceu um conjunto de dados de mais de 100 milhões de classificações de 17.770 filmes de 480.189 usuários.[16] A última consulta mostrou que houve mais de 30.000 downloads.

Por que a Netflix fez tudo isso? Uma forte razão é que os próprios engenheiros da empresa estavam tendo problemas para fazer progressos nas recomendações. Então, por que não dar a chance de a comunidade descobrir? O trabalho era bastante engenhoso – e o pagamento de US$ 1 milhão foi de fato modesto em comparação com os benefícios potenciais.

O concurso certamente despertou muita atividade de programadores e cientistas de dados, que iam de estudantes a funcionários de empresas como a AT&T.

A Netflix também tornou o concurso simples. A principal exigência era que as equipes apresentassem seus métodos, o que ajudou a aumentar os resultados (houve até mesmo um painel com rankings das equipes).

No entanto, apenas em 2009 uma equipe – a BellKor's Pragmatic Chaos – ganhou o prêmio. Novamente, houve desafios consideráveis.

Então, como a equipe vencedora conseguiu? O primeiro passo foi criar um modelo de linha de base que suavizasse as questões complicadas relacionadas aos dados. Por exemplo, alguns filmes tinham somente algumas poucas classificações, enquanto outros contavam com milhares delas. Depois, houve o difícil problema dos usuários que sempre classificavam os filmes com uma estrela. Para lidar com essas questões, a BellKor usou machine learning para prever classificações com o intuito de preencher as lacunas.

Com a linha de base concluída, houve desafios mais difíceis de enfrentar, como os seguintes:

- Um sistema pode acabar recomendando os mesmos filmes para muitos usuários.
- Alguns filmes podem não se encaixar bem dentro de gêneros. Por exemplo, *Alien* é uma intercessão entre ficção científica e horror.
- Havia filmes, como *Napoleon Dynamite*, cuja compreensão pelos algoritmos era extremamente difícil.
- As avaliações de um filme mudavam frequentemente ao longo do tempo.

A equipe vencedora usou modelagem por agrupamento, o que envolveu centenas de algoritmos. Eles também usaram um método chamado boosting (aceleração), que é onde se constroem modelos consecutivos. Com isso, os pesos nos algoritmos são

ajustados com base nos resultados do modelo anterior, o que ajuda as previsões a ficarem melhores ao longo do tempo (outra abordagem, chamada bagging, é adotada quando são construídos diferentes modelos em paralelo e, em seguida, seleciona-se o melhor).

Apesar de tudo, ao final a BellKor encontrou as soluções. No entanto, a Netflix não usou o modelo! O motivo para essa decisão não está claro. Talvez tenha sido porque a Netflix estava abandonando as classificações cinco estrelas e focando mais em streaming. O concurso também foi atacado por pessoas que acreditavam ter havido violações de privacidade.

Independentemente disso, o evento realmente deu destaque ao poder do machine learning – e à importância da colaboração.

Agrupamento k-means (Não supervisionada/Agrupamento)

O algoritmo de agrupamento k-means, eficiente em grandes conjuntos, coloca dados semelhantes não rotulados em diferentes grupos. O primeiro passo é selecionar k, que é o número de grupos (clusters). Para ajudar com isso, é possível realizar visualizações desses dados para verificar se existem áreas de agrupamento visíveis.

Observe os dados da amostra na Figura 3.6:

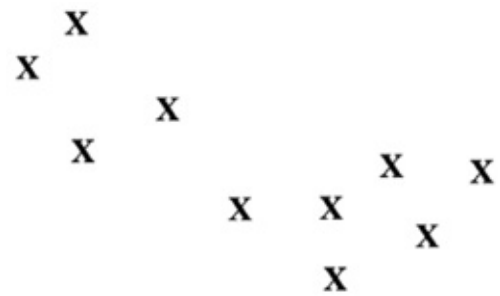

Figura 3.6 – Gráfico inicial para um conjunto de dados.

Para esse exemplo, assumimos que haverá dois grupos, o que significa que haverá também dois centroides. Um deles é o ponto médio de um grupo. Cada um será posicionado aleatoriamente, conforme se vê na Figura 3.7.

Figura 3.7 – Gráfico com dois centroides – representados por círculos – randomicamente posicionados.

Como se pode verificar, o centroide no canto superior esquerdo parece muito fora da área, mas o do lado direito é melhor. O algoritmo k-means calculará as distâncias médias dos centroides e, em seguida, mudará suas localizações. Essa operação será repetida até que os erros sejam mínimos – um ponto que é chamado de convergência, conforme se pode ver na Figura 3.8.

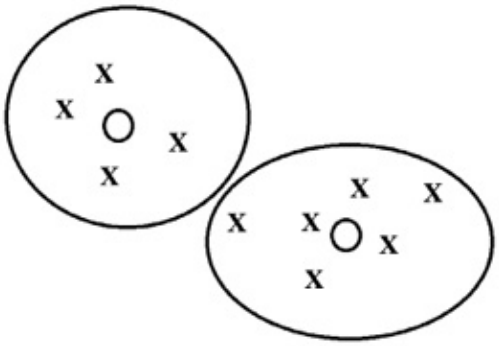

Figura 3.8 – Por meio de iterações, o algoritmo k-means aprimora o agrupamento dos dados.

Claro que essa é uma ilustração simples. Com um conjunto de dados complexo, entretanto, será difícil chegar ao número de grupos iniciais. Nessa situação, é possível experimentar diferentes valores para k e, em seguida, medir as distâncias médias. Ao fazer isso várias vezes, é provável que haja maior precisão.

Então por que não ter um número alto para k? Certamente é possível fazer isso. Contudo, ao calcular a média, você vai notar que haverá apenas melhorias incrementais. Assim, uma alternativa é parar no ponto onde isso começa a ocorrer, conforme se vê na Figura 3.9.

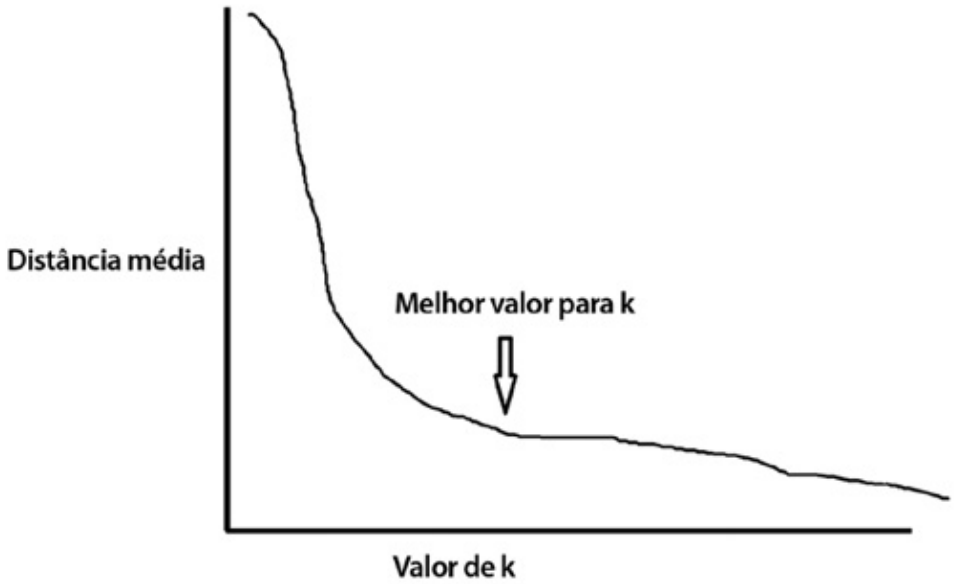

Figura 3.9 – Ponto ideal para o valor de K no algoritmo k-means.

No entanto, k-means tem suas desvantagens. Por exemplo, o algoritmo não funciona bem com dados não esféricos, caso da Figura 3.10.

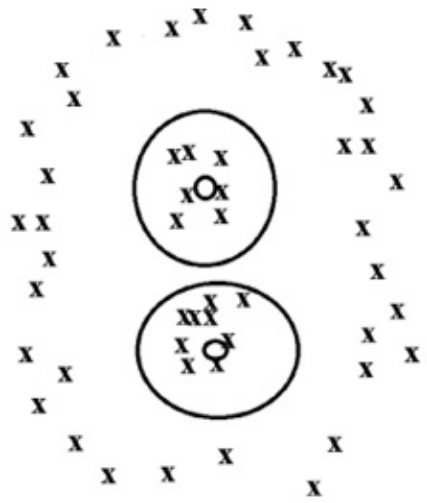

Figura 3.10 – Demonstração na qual o algoritmo k-means não funciona com dados não esféricos.

Nesse caso, o algoritmo k-means provavelmente não pegaria os dados circundantes, embora haja um padrão. Existem, entretanto, outros algoritmos que podem ajudar, como o DBScan (agrupamento espacial baseado em densidade de aplicações com ruído), que se destina a lidar com uma diversidade de conjuntos de dados com tamanhos variados. DBScan pode exigir muito poder computacional.

Por fim, há a situação na qual existem alguns grupos com muitos dados e outros com pouco. O que pode acontecer? Há uma chance de que o algoritmo k-means não selecione o conjunto com menos dados. Esse é o caso da Figura 3.11.

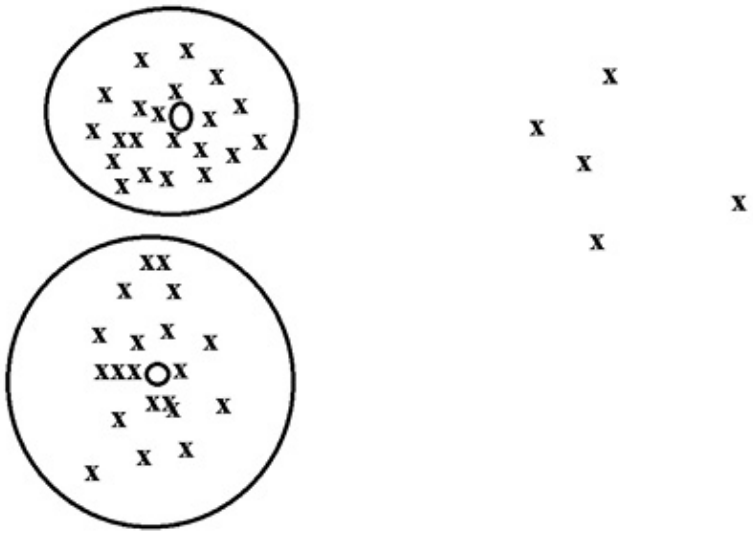

Figura 3.11 – Se existirem áreas de dados menos densas, o algoritmo k-means pode não as escolher.

Conclusão

Esses algoritmos podem ficar complicados e exigem sólidas habilidades técnicas. É importante, no entanto, não ficar muito atolado na tecnologia. Afinal, o foco é encontrar maneiras de usar machine learning para alcançar objetivos claros.

Mais uma vez, a Stich Fix é um bom lugar para obter orientação sobre isso. Na

edição de novembro da *Harvard Business Review*, o Diretor de Algoritmos da empresa, Eric Colson, publicou o artigo "Curiosity-Driven Data Science" ("Ciência de dados orientada a curiosidade").[17] Nele, contou suas experiências na criação de uma organização orientada por dados.

O objetivo principal é permitir que os cientistas de dados explorem novas ideias, conceitos e abordagens. Isso resultou na implementação de IA em funções fundamentais do negócio, como gestão de estoque, gestão de relacionamento, logística e compra de mercadorias. Tem sido transformador e vem tornando a organização mais ágil e simplificada. Colson também acredita que os recursos criaram "uma barreira protetora contra a concorrência".

Seu artigo também oferece outros conselhos úteis para a análise de dados:

- *Cientistas de dados:* não devem fazer parte de outro departamento. Em vez disso, devem ter o seu próprio, reportando-se diretamente ao CEO. Isso ajuda a focar nas principais prioridades, bem como ter uma visão holística das necessidades da organização.
- *Experiências*: quando um cientista de dados tem uma nova ideia, ela deve ser testada em uma pequena amostra dos clientes. Se for bem-sucedida, então pode ser implementada para o resto da base.
- *Recursos:* cientistas de dados precisam de acesso total a dados e ferramentas. Também deve haver treinamento contínuo.
- *Generalistas:* contrate cientistas de dados oriundos de diferentes domínios, como modelagem, machine learning e análise (Colson refere-se a essas pessoas como "cientistas de dados full-stack"). Isso forma pequenas equipes – que muitas vezes são mais eficientes e produtivas.
- *Cultura*: Colson procura valores como "aprender fazendo, estar confortável com a ambiguidade, equilibrar retornos de longo e curto prazos".

Principais aprendizados

- Machine learning, cujas raízes remontam à década de 1950, é onde um computador pode aprender sem ser explicitamente programado. Em vez disso, ele vai ingerir e processar dados usando técnicas estatísticas sofisticadas.
- Um dado discrepante (outlier) é aquele que está muito fora do resto dos números no conjunto de dados.
- O desvio-padrão mede a dispersão dos valores em relação à média.
- A distribuição normal – que tem forma de sino – representa a soma das probabilidades para uma variável.
- O teorema de Bayes é uma técnica estatística sofisticada que fornece um olhar

mais profundo sobre probabilidades.

- Um verdadeiro positivo acontece quando um modelo faz uma previsão correta. Um falso positivo, por outro lado, ocorre quando uma previsão do modelo mostra que o resultado é verdade, mesmo que não o seja.
- A correlação de Pearson mostra a força da relação entre duas variáveis cujo intervalo está entre 1 e -1.
- A extração de recursos, ou engenharia de recursos, descreve o processo de seleção de variáveis para um modelo. Isso é fundamental, uma vez que uma variável errada pode causar grande impacto sobre os resultados.
- Os dados de treinamento são usados para criar relações em um algoritmo. Os dados de teste, por outro lado, são usados para avaliar o modelo.
- A aprendizagem supervisionada utiliza dados rotulados para criar um modelo, enquanto a aprendizagem não supervisionada não o faz. Há também a aprendizagem semissupervisionada, que usa uma mistura de ambas as abordagens.
- A aprendizagem por reforço é uma maneira de treinar um modelo recompensando previsões precisas e punindo aquelas que não o são.
- O algoritmo k-NN baseia-se na noção de que os valores que estão próximos são bons preditores para um modelo.
- A regressão linear estima a relação entre certas variáveis. O R-quadrado indicará a força do relacionamento.
- Uma árvore de decisão é um modelo que se baseia em um fluxo de trabalho de decisões sim/não.
- Um modelo por agrupamento usa mais de um modelo para as previsões.
- O algoritmo de agrupamento k-means coloca dados não rotulados semelhantes em diferentes grupos.

---

1 Steve Lohr, "Microsoft, Amid Dwindling Interest, Talks Up Computing as a Career: Enrollment in Computing Is Dwindling" ("Em meio a um interesse cada vez menor, Microsoft fala da computação como uma carreira: As matrículas em computação estão diminuindo"), New York Times, 1º de março de 2004, página inicial C1, citação na página C2, coluna 6.

2 www.cnbc.com/2017/11/16/stitchfix-ipo-sees-orders-coming-in-under-range.html

3 https://investors.stitchfix.com/static-files/2b398694-f553-4586-b763-e942617e4dbf

4 www.sec.gov/Archives/edgar/data/1576942/000157694218000003/stitchfix201810k.htm

5 https://digiday.com/marketing/stitchfix-ceo-katrina-lake-predicts-ais-impact-fashion/

6 Arthur L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers" ("Alguns estudos em machine learning usando o jogo se damas", publicado em Edward A. Feigenbaum e Julian Feldman, eds., Computers and Thought (Nova York: McGraw-Hill, 1983), pp. 71-105.

7 https://history.computer.org/pioneers/samuel.html

8 www.tylervigen.com/spurious-correlations

9 www.precisionhawk.com/blog/in-oil-gas-the-economics-of-bvlos-drone-operations

10 Entrevista realizada pelo autor em fevereiro de 2019 com Humair Ghauri, Diretor de Produtos na CareerBuilder.

11 www.8451.com/case-study/avocado

12 https://www.engadget.com/2018/05/02/facebook-trained-image-recognition-ai-instagram-pics/

13 https://code.fb.com/ml-applications/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags/

14 Entrevista realizada pelo autor em fevereiro de 2019 com Asaf Cidon, VP de Segurança de Emails na Barracuda Networks.

15 https://towardsdatascience.com/simple-explanation-of-semi-supervised-learning-and-pseudo-labeling-c2218e8c769b

16 www.thrillist.com/entertainment/nation/the-netflix-prize

17 https://hbr.org/2018/11/curiosity-driven-datascience

CAPÍTULO 4

Deep Learning

A revolução na IA

Tome qualquer problema antigo de classificação, no qual se tinha um monte de dados, e ele será resolvido por deep learning. Haverá milhares de aplicações de deep learning.

– Geoffrey Hinton, psicólogo cognitivo canadense e cientista da computação[1]

Fei-Fei Li, Bacharel em Física por Princeton em 1999 com honras e PhD em Engenharia Elétrica pela Caltech em 2005, concentrou seu brilhantismo no desenvolvimento de modelos de IA. No entanto, ela enfrentou um grande desafio: encontrar conjuntos de dados de qualidade. No início, tentou criá-los manualmente, solicitando que estudantes de pós-graduação baixassem imagens da internet. Contudo, o processo foi muito lento e tedioso.

Certo dia, um aluno contou a Li que o Mechanical Turk da Amazon.com, um serviço online que usa crowdsourcing[2] para resolver problemas, poderia ser uma boa maneira de dimensionar o processo. Ele permitiria uma rotulagem rápida e precisa dos dados.

Li resolveu experimentar e o recurso funcionou muito bem. Em 2010, ela havia criado a ImageNet, que contava com 3,2 milhões de imagens em mais de 5.200 categorias.

A ferramenta, no entanto, obteve uma resposta morna da comunidade acadêmica. Isso, contudo, não desestimulou Li. Ela continuou a trabalhar incansavelmente para treinar o conjunto de dados. Em 2012, organizou um concurso como forma de incentivar os pesquisadores a criar modelos mais eficazes e ultrapassar os limites da inovação. Seria um divisor de águas e o concurso se tornaria um evento anual.

Na primeira edição do evento, professores da Universidade de Toronto – Geoffrey Hinton, Ilya Sutskever e Alex Krizhevsky – usaram algoritmos sofisticados de deep learning e os resultados foram excelentes. O sistema que construíram, chamado AlexNet, bateu todos os outros concorrentes por uma margem de 10,8%.[3]

Isso não foi por acaso. Nos anos seguintes, o deep learning continuou a mostrar um progresso acelerado com a ImageNet. A partir daquele momento, a taxa de erro para a tecnologia era de aproximadamente 2% – o que é melhor do que os seres humanos.

A propósito, desde então, Li tornou-se professora em Stanford e codiretora do

laboratório de IA da universidade. Ela também é cientista-chefe de IA e machine learning no Google. Não é preciso dizer que agora, sempre que ela tem novas ideias, as pessoas ouvem!

Neste capítulo, vamos dar uma olhada no deep learning, que é claramente a área mais quente da IA. A tecnologia levou a grandes avanços em áreas como carros autônomos e assistentes virtuais, como a Siri.

Sim, deep learning pode ser um assunto complicado, e o campo está em constante mudança. Contudo, vamos dar uma olhada nos principais conceitos e tendências – sem entrar nos detalhes técnicos.

Diferenças entre deep learning e machine learning

Sempre há confusão entre deep learning e machine learning. E isso é razoável. Ambos os tópicos são bastante complexos e compartilham muitas semelhanças.

Para compreender as diferenças, portanto, vamos primeiro dar uma olhada em dois aspectos de alto nível do machine learning e como eles se relacionam com o deep learning. Em primeiro lugar, embora ambos geralmente exijam grandes quantidades de dados, os tipos costumam ser diferentes.

Considere o seguinte exemplo: suponha que tenhamos fotos de milhares de animais e queremos criar um algoritmo para encontrar os cavalos. Bem, machine learning não pode analisar as fotos em si; em vez disso, os dados devem ser rotulados. O algoritmo de machine learning será treinado para reconhecer cavalos por meio de um processo conhecido como aprendizagem supervisionada (abordado no Capítulo 3).

Mesmo que o machine learning apresente bons resultados, eles ainda terão limitações. Não seria melhor olhar para os pixels das próprias imagens – e encontrar os padrões? Com certeza.

Para fazer isso com machine learning, no entanto, é preciso usar um processo chamado extração de recursos. Isso significa que será necessário imaginar os tipos das características de um cavalo que os algoritmos tentarão então identificar – tal como forma, casco, cor e altura.

Mais uma vez, essa é uma boa abordagem – mas está longe de ser perfeita. E se seus recursos forem imprecisos ou não derem conta de discrepâncias ou exceções? Nesses casos, a precisão do modelo provavelmente sofrerá. Afinal de contas, há muitas variações em um cavalo. A extração de recursos também tem a desvantagem de ignorar uma grande quantidade de dados. Isso pode ser extremamente complicado – se não impossível – para determinadas aplicações. Observe os vírus de computador. Seus padrões e estruturas, conhecidos como assinaturas, estão em constante mudança de modo a se infiltrarem em sistemas. Com a extração de recursos, no

entanto, uma pessoa teria de antecipar essas características de alguma forma, o que não é prático. É por isso que o software de cibersegurança muitas vezes trata da coleta de assinaturas depois que um vírus causou danos.

Com deep learning, entretanto, é possível resolver esses problemas. Essa abordagem analisa todos os dados – pixel por pixel – e, em seguida, encontra as relações usando uma rede neural que imita o cérebro humano.

Vamos dar uma olhada.

Afinal, o que é deep learning então?

A tecnologia deep learning é uma subárea do machine learning. Esse tipo de sistema permite o processamento de enormes quantidades de dados para encontrar relacionamentos e padrões que os seres humanos são muitas vezes incapazes de detectar. A palavra "deep" (em português, "profundo") refere-se ao número de camadas ocultas na rede neural, as quais fornecem grande parte do poder de aprendizagem.

Quando se trata do tema da IA, deep learning está na vanguarda e muitas vezes gera a maior parte do burburinho na mídia convencional. "[Deep learning] IA é a nova eletricidade", enalteceu Andrew Yan-Tak Ng, ex-cientista-chefe da Baidu e cofundador do Google Brain.[4]

É importante também lembrar que o deep learning ainda está nos estágios iniciais de desenvolvimento e comercialização. Por exemplo, somente por volta de 2015 que o Google começou a usar essa tecnologia para seu mecanismo de pesquisa.

Como vimos no Capítulo 1, a história das redes neurais estava cheia de fluxos e refluxos. Foi Frank Rosenblatt que criou o perceptron, um sistema bastante básico. O progresso acadêmico real com redes neurais, entretanto, não ocorreu até a década de 1980, com os avanços em temas como retropropagação, redes neurais convolucionais e redes neurais recorrentes. Contudo, para que o deep learning cause impacto no mundo real, seria necessário um crescimento impressionante em dados, como os oriundos da internet, e um aumento do poder computacional.

O cérebro e deep learning

Pesando apenas cerca de 3,3 quilos, o cérebro humano é um feito incrível da evolução. Existem cerca de 86 bilhões de neurônios – muitas vezes chamados de massa cinzenta – que estão conectados por trilhões de sinapses. Pense nos neurônios como CPUs (Central Processing Units – Unidades Centrais de Processamento) que coletam dados. A aprendizagem ocorre com o fortalecimento ou enfraquecimento das sinapses.

O cérebro é composto de três regiões: cérebro anterior, cérebro médio e cérebro posterior. Entre eles, há uma variedade de áreas que executam funções diferentes. Entre as principais estão:

- *Hipocampo:* esse é o lugar onde seu cérebro armazena memórias. Na verdade, essa é a parte que falha quando uma pessoa tem Mal de Alzheimer, doença durante a qual se perde a capacidade de formar memórias de curto prazo.
- *Lobo frontal:* aqui o cérebro se concentra em emoções, fala, criatividade, julgamento, planejamento e raciocínio.
- *Córtex cerebral:* essa talvez seja a parte mais importante quando se trata de IA. O córtex cerebral ajuda com o pensamento e outras atividades cognitivas. De acordo com a pesquisa de Suzana Herculano-Houzel, o nível de inteligência está relacionado ao número de neurônios nessa área do cérebro.

Então, como deep learning se compara ao cérebro humano? Existem algumas semelhanças tênues. Pelo menos em áreas como a retina, há um processo de ingestão e processamento de dados por meio de uma rede complexa baseada na atribuição de pesos. É claro, entretanto, que essa é apenas uma pequena parte do processo de aprendizagem. Além disso, existem ainda muitos mistérios sobre o cérebro humano e, claro, ele não se baseia em coisas como a computação digital (em vez disso, se parece mais com um sistema analógico). No entanto, à medida que as pesquisas continuam a avançar, as descobertas em neurociência podem ajudar a construir novos modelos para a IA.

Redes neurais artificiais

No nível mais básico, uma rede neural artificial (Artificial Neural Networks – ANNs) é uma função que inclui unidades (que também podem ser chamadas de neurônios, perceptrons ou nós). Cada unidade terá um valor e um peso, que indicam sua importância relativa, e irá para a camada oculta. A camada oculta usa uma função, cujo resultado se torna a saída. Há também outro valor, chamado viés (bias), que é uma constante usada no cálculo da função.

Esse tipo de treinamento de modelos é chamado de rede neural feed-forward. Em outras palavras, ele se movimenta somente da camada de entrada para a camada oculta e, em seguida, para a camada de saída. O ciclo não se repete. Ele poderia, no entanto, ir para uma nova rede neural, com a saída se transformando em entrada.

A Figura 4.1 mostra o gráfico de uma rede neural feed-forward.

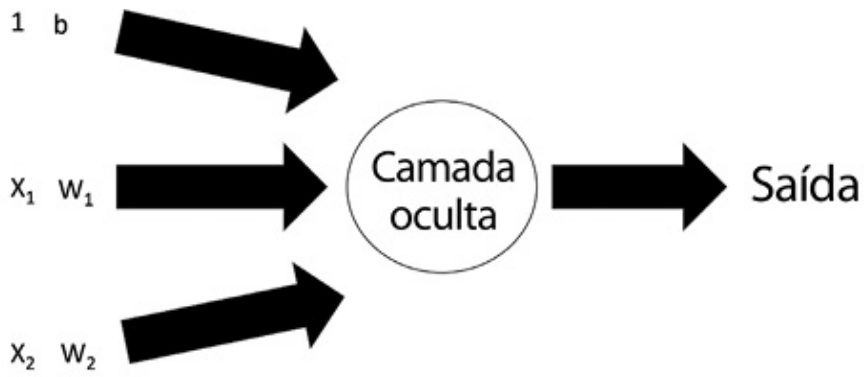

Figura 4.1 – Rede neural feed-forward básica.

Vamos nos aprofundar nisso discutindo um exemplo. Suponha que você esteja criando um modelo para prever se as ações de uma empresa vão subir. A seguir, é possível ver o que as variáveis representam, bem como os valores e pesos atribuídos a cada uma delas:

- X_1: as receitas estão crescendo a um mínimo de 20% ao ano. O valor é 2.
- X_2: a margem de lucro é de pelo menos 20%. O valor é 4.
- W_1: 1,9.
- W_2: 9,6.
- b: esse é o viés (o valor é 1) que ajuda a suavizar os cálculos.

Com os valores definidos, somam-se os pesos e, em seguida, aciona-se a função que vai processar as informações. Isso geralmente envolve uma função de ativação, que é não linear. Isso reflete melhor o mundo real, uma vez que os dados em geral não estão em linha reta.

Há uma variedade de funções de ativação dentre as quais escolher. Uma das mais comuns é a sigmoide. Ela comprime a entrada em um intervalo entre 0 e 1. Quanto mais próximo esse valor estiver de 1, mais preciso será o modelo.

Quando essa função é esboçada, assume uma forma de S. Observe a Figura 4.2.

Como se pode ver, o sistema é relativamente simplista e não será útil em modelos de IA de ponta. Para adicionar muito mais poder, geralmente é necessário que existam várias camadas ocultas. Isso resulta em um perceptron multicamadas (MLP – MultiLayered Perceptron). Ele também ajuda a usar algo chamado retropropagação, que permite que a saída volte a circular na rede neural.

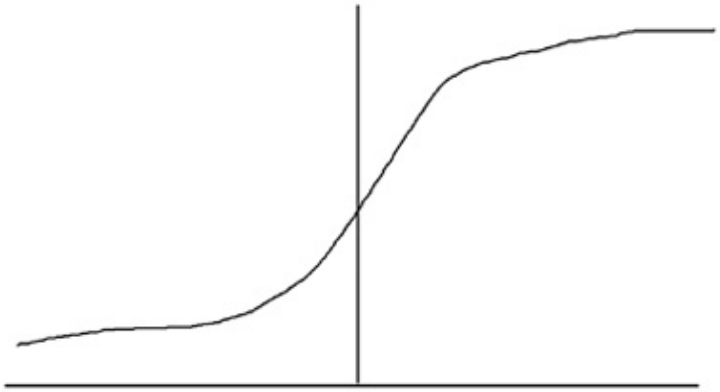

Figura 4.2 – Função de ativação sigmoide típica.

Retropropagação (Backpropagation)

Uma das principais desvantagens com relação às redes neurais artificiais é o processo de ajustes nos pesos do modelo. Abordagens tradicionais, como o uso do algoritmo de mutação, usam valores aleatórios que demandam muito tempo de processamento.

Diante disso, os pesquisadores buscaram alternativas, como a retropropagação. Essa técnica já existe desde a década de 1970, mas desperta pouco interesse por conta de seu desempenho insatisfatório. David Rumelhart, Geoffrey Hinton e Ronald Williams, entretanto, perceberam que a retropropagação tinha potencial desde que fosse refinada. Em 1986, eles escreveram um artigo intitulado "Learning Representations by Backpropagating Errors" ("Aprendendo representações por meio de erros de retropropagação") que foi uma bomba na comunidade de IA.[5] A pesquisa mostrou claramente que a retropropagação poderia ser muito mais rápida, além de viabilizar a criação de redes neurais artificiais mais poderosas.

Como não deve ser nenhuma surpresa, há muita matemática envolvida na retropropagação. Em resumo, trata-se de ajustar a rede neural quando erros são encontrados e, em seguida, iterar os novos valores na rede novamente. Essencialmente, o processo envolve pequenas mudanças que continuam otimizando o modelo.

Por exemplo, digamos que uma das entradas tem uma saída de 0,6. Isso significa que o erro é de 0,4 (1,0 menos 0,6), que é abaixo da média. É possível, no entanto, retropropagar essa saída e talvez o novo resultado possa chegar a 0,65. Esse treinamento vai continuar até que o valor esteja muito mais próximo de 1.

A Figura 4.3 ilustra esse processo. No início, há um alto nível de erros, porque os pesos são muito grandes. Contudo, ao fazer as iterações, os erros caem gradualmente. Fazer muito disso, no entanto, pode significar um aumento nos erros. Em outras palavras, o objetivo da retropropagação é encontrar o ponto médio.

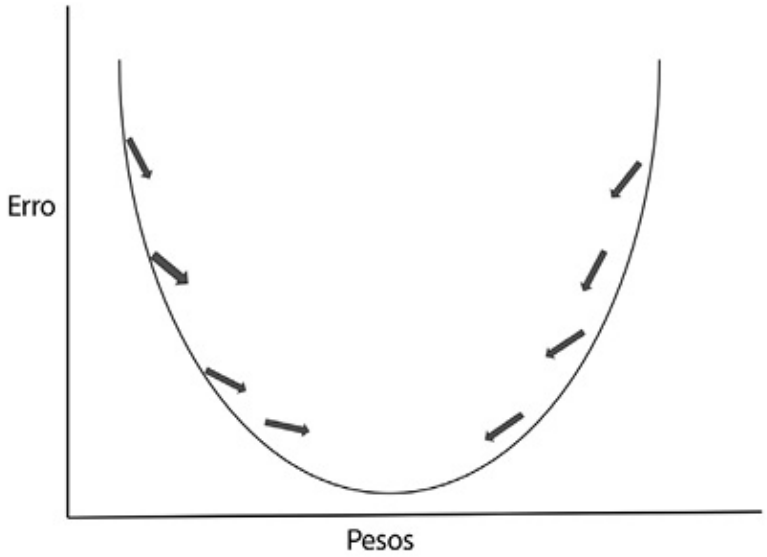

Figura 4.3 – O valor ótimo para uma função de retropropagação está na parte inferior do gráfico.

Como indicador do sucesso da retropropagação, uma enxurrada de aplicações comerciais começou a surgir. Uma delas foi a NETtalk, desenvolvida por Terrence Sejnowski e Charles Rosenberg em meados da década de 1980. A máquina foi capaz de aprender a pronunciar textos em inglês. A aplicação era tão interessante que chegou a ser apresentada no *Today Show*.

Houve também uma variedade de startups que impulsionaram a retropropagação, como a HNC Software; que construiu modelos que detectaram fraude com cartões de crédito. Até aquele momento – quando a HNC foi criada no final de 1980 –, o processo era feito principalmente à mão, o que levava a erros dispendiosos e baixos volumes de entregas. Usando abordagens de deep learning, entretanto, as administradoras de cartão de crédito foram capazes de economizar bilhões de dólares.

Em 2002, a HNC foi comprada pela Fair, Isaac and Company e avaliada em US$ 810 milhões.[6]

As diferentes redes neurais

A rede neural totalmente conectada é o tipo mais básico que existe. Como o nome indica, nela todos os neurônios têm conexões de uma camada para a outra. Essa rede é bastante popular, uma vez que demanda pouca capacidade crítica na criação do modelo.

E quais são, então, algumas das outras redes neurais? As mais comuns incluem: rede neural recorrente (RNN), rede neural convolucional (CNN) e rede adversária generativa (GAN – Generative Adversarial Network), sobre as quais falaremos a seguir.

Redes neurais recorrentes

Com uma rede neural recorrente (Recurrent Neural Network – RNN), a função não só processa a entrada atual, mas também as entradas anteriores ao longo do tempo. Um exemplo disso é o que acontece quando você digita caracteres em um aplicativo de mensagens. À medida que ocorre a digitação, o sistema tenta prever as palavras. Então, se você digitar "Ele", o computador vai sugerir "Ele", "Elefante" e "Elétrico", por exemplo. A RNN é essencialmente uma sequência de redes neurais que se alimentam umas das outras com base em algoritmos complexos.

Existem variações no modelo. Uma delas é a chamada LSTM (Long Short-Term Memory), que significa memória longa de curto prazo. Ela surgiu a partir de um artigo escrito pelos professores Sepp Hochreiter e Jürgen Schmidhuber, em 1997.[7] Nele, os autores estabelecem uma maneira de usar efetivamente entradas que são separadas umas das outras por longos períodos, permitindo o uso de mais conjuntos de dados.

Obviamente, as RNNs têm desvantagens. Existe o problema da dissipação do gradiente, que significa que a precisão decai à medida que os modelos ficam maiores. Os modelos também podem levar mais tempo para treinar.

Para lidar com isso, o Google desenvolveu um novo modelo chamado Transformer, que é muito mais eficiente, uma vez que processa as entradas em paralelo. Ele também produz resultados mais precisos.

O Google aprendeu muito sobre as RNNs por meio de seu aplicativo de tradução, que lida com mais de 100 idiomas e processa mais de 100 bilhões de palavras por dia.[8] Lançado em 2006, o recurso inicialmente usou sistemas de machine learning. Em 2016, entretanto, o Google mudou para deep learning, criando a Google Neural Machine Translation (GNMT – em português, "tradução de máquina neural do Google").[9] Em suma, o recurso resultou em taxas de precisão muito mais elevadas.[10]

Imagine como o Google Tradutor ajudou os médicos que trabalham com pacientes que falam outros idiomas. De acordo com um estudo da Universidade da California, San Francisco (UCSF), publicado na *JAMA Internal Medicine*, o aplicativo apresentou uma taxa de precisão de 92% nas traduções de inglês para espanhol. Isso é superior aos 60% dos últimos anos.[11]

Redes neurais convolucionais

Intuitivamente, faz sentido ter todas as unidades de uma rede neural conectadas. Isso funciona bem com muitas aplicações.

No entanto, há cenários muito distantes do ideal, como é o caso do reconhecimento de imagens. Basta imaginar como seria complexo um modelo no qual cada pixel é uma unidade! Ele rapidamente se tornaria incontrolável. Haveria também outras

complicações como a superadaptação (overfitting). Isso ocorre quando os dados não refletem o que está sendo testado ou o foco está nas características erradas.

Para lidar com tudo isso, é possível usar uma rede neural convolucional (Convolutional Neural Network – CNN). Suas origens remontam ao professor Yann LeCun, em 1998, quando publicou um artigo chamado "Gradient-Based Learning Applied to Document Recognition" ("Aprendizagem baseada em gradiente aplicada ao reconhecimento de documentos").[12] Apesar de suas fortes percepções e avanços, o texto despertou pouca atenção. Contudo, como o deep learning começou a mostrar um progresso significativo em 2012, os pesquisadores revisitaram o modelo.

LeCun buscou inspiração para a CNN nos vencedores do Prêmio Nobel David Hubel e Torsten Wiesel, que estudaram neurônios do córtex visual. O sistema pegava uma imagem da retina e a processava em diferentes fases – da fácil à mais complexa. Cada um dos estágios é chamado de convolução. Por exemplo, o primeiro nível seria identificar linhas e ângulos; em seguida, o córtex visual encontraria as formas e, então, detectaria os objetos.

Isso é análogo a como uma CNN baseada em computador trabalha. Considere o seguinte exemplo: suponha que você queira construir um modelo capaz de identificar uma carta. A entrada da CNN será uma imagem com 3.072 pixels. Cada um dos pixels terá um valor que vai de 0 a 255, o que indica a intensidade geral. Usando uma CNN, o computador passará por diversas variações para identificar os recursos.

A primeira é a camada convolucional, um filtro que escaneia a imagem. No exemplo, ela pode escanear áreas de 5 × 5 pixels. O processo criará um mapa de recursos, que é um longo vetor de números. Em seguida, o modelo aplicará mais filtros à imagem. Ao fazer isso, a CNN vai identificar linhas, bordas e formas – todas expressas em números. Com as várias camadas de saída, o modelo usará agregação (pooling), as combinará para gerar uma única saída e, em seguida, criará uma rede neural totalmente conectada.

A CNN pode definitivamente ficar complexa. No entanto, ela deve ser capaz de identificar com precisão os números que são fornecidos como entrada no sistema.

Redes Adversárias Generativas (GANs - Generative Adversarial Network)

Ian Goodfellow, que fez mestrado em Ciência da Computação em Stanford e PhD em Machine Learning na Université de Montréal, continuou a trabalhar no Google. Aos 20 anos, foi coautor de um dos principais livros da IA, chamado *Deep Learning*[13], e também fez inovações no Google Maps.

Somente em 2014, entretanto, ele alcançou seu avanço mais impactante. Na verdade, tudo aconteceu em um pub em Montreal, enquanto conversava com alguns de seus

amigos sobre como o deep learning poderia *criar* fotos.[14] Naquela época, a abordagem usava modelos generativos que muitas vezes eram confusos e sem sentido.

Goodfellow pensava que precisava haver uma explicação para isso. Então, por que não usar a teoria dos jogos? Ou seja, ter dois modelos competindo um contra o outro em um loop de feedback. Isso também poderia ser feito com dados não rotulados.

Observe o fluxo de trabalho básico:

- *Gerador:* essa rede neural produz uma infinidade de novas criações, como fotos ou frases.
- *Discriminatório:* essa rede neural analisa as criações para ver quais são reais.
- *Ajustes*: com os dois resultados, um novo modelo altera as criações para torná-las o mais realista possível. Por meio de muitas iterações, o discriminatório não precisaria mais ser usado.

Ele estava tão animado com as ideias que, depois que deixou o pub, começou a codificá-las. O resultado foi um novo modelo de deep learning: a rede adversária generativa ou GAN – Generative Adversarial Network. Os resultados foram destaque e Goodfellow logo se tornaria uma estrela da IA.

As pesquisas com as GANs já estimularam mais de 500 trabalhos acadêmicos.[15] Empresas como Facebook também usaram a tecnologia para análise e processamento de fotos. O cientista-chefe de IA da empresa, Yann LeCun, observou que as GANs são a "ideia mais legal em deep learning nos últimos 20 anos".[16]

Esse tipo de rede neural também foi usado para ajudar com a pesquisa científica sofisticada. Por exemplo, GANs auxiliaram na melhoria da precisão da detecção do comportamento de partículas subatômicas no Large Hadron Collider (Grande Colisor de Hádrons) no CERN, na Suíça.[17]

Ainda no início, essa tecnologia podia levar a progressos significativos, como fazer um computador desenvolver novos tipos de itens de moda ou acessórios. É possível até que uma GAN crie um rap de sucesso.

E pode ser mais cedo do que você imagina. Quando era adolescente, Robbie Barrat aprendeu sozinho a usar sistemas de deep learning e construiu um modelo para fazer raps no estilo de Kanye West.

Esse, no entanto, foi apenas o começo de sua magia na IA. Como pesquisador em Stanford, ele desenvolveu sua própria plataforma GAN, que processou cerca de 10.000 fotos de nus. O sistema, em seguida, criou obras de arte verdadeiramente hipnotizantes (é possível encontrá-las em sua conta no Twitter em @DrBeef_).

Ah, e ele também liberou o código de sua criação em sua conta no GitHub. A

atitude chamou a atenção de um coletivo de artistas franceses, chamado Obvious, que usou a tecnologia para criar retratos de uma família fictícia do século 18. O projetou se baseou no processamento de 15.000 imagens dos séculos 12 ao 20.

Em 2018, a Obvious incluiu sua obra de arte em um leilão da Christie's, e alcançou surpreendentes US$ 432.000.[18]

Infelizmente, quando as GANs são o tema, alguns de seus usos têm sido menos admiráveis. Um exemplo é usá-las para deepfakes, que envolvem o uso de redes neurais para criar imagens ou vídeos enganosos. Algumas dessas criações são apenas lúdicas. Existe, por exemplo, uma GAN que permite que se faça Barack Obama dizer qualquer coisa que se queira!

Os riscos são muitos. Pesquisadores da Universidade de New York e da Universidade Estadual de Michigan escreveram um artigo que se concentrou em "DeepMasterPrints".[19] O trabalho mostrava como uma GAN era capaz de desenvolver impressões digitais falsas para desbloquear três tipos de smartphones!

Depois disso, houve o incidente de um vídeo dito falso da atriz Jennifer Lawrence em uma conferência de imprensa do Golden Globes. Seu rosto foi fundido com o de Steve Buscemi.[20]

Aplicações de deep learning

Com tanto dinheiro e recursos sendo investidos em deep learning, tem havido um aumento nas inovações. Parece que todos os dias existe algo incrível sendo anunciado.

Então, quais são algumas das aplicações? Onde o deep learning provou ser um divisor de águas? Vamos dar uma olhada em alguns estudos de caso que abordam áreas como saúde, energia e até terremotos.

Estudo de caso: detectando o Mal de Alzheimer

Apesar de décadas de pesquisa, a cura para o Mal de Alzheimer permanece indefinida, embora os cientistas tenham desenvolvido medicamentos que retardaram a progressão da doença.

Diante disso, o diagnóstico precoce é fundamental – e deep learning pode potencialmente ser de grande ajuda. Pesquisadores do Departamento de Radiologia e Imagem Biomédica da UCSF usaram essa tecnologia para analisar telas cerebrais – obtidas a partir do conjunto de dados públicos da Iniciativa de Neuroimagem do Mal de Alzheimer – e detectar mudanças nos níveis de glicose.

O resultado foi que o modelo conseguiu diagnosticar a doença até seis anos antes de um diagnóstico clínico. Um dos testes mostrou uma taxa de precisão de 92%,

enquanto em outro ela foi de 98%.

A pesquisa ainda está nas fases iniciais – e será necessário que mais conjuntos de dados sejam analisados. Até agora, no entanto, os resultados são muito encorajadores.

De acordo com o Dr. Jae Ho Sohn, autor do estudo:

> *Trata-se de uma aplicação ideal do deep learning porque é particularmente forte em encontrar processos muito sutis, embora difusos. Radiologistas humanos são muito eficientes na identificação de pequenas descobertas focais, como um tumor cerebral, mas têm dificuldade para detectar mudanças mais lentas e globais. Dada a força do deep learning nesse tipo de aplicação, especialmente em comparação com os seres humanos, sua aplicação pareceu natural.*[21]

Estudo de caso: Energia

Devido a sua enorme infraestrutura de data center, o Google é um dos maiores consumidores de energia. Mesmo uma pequena melhoria na eficiência pode levar a um impacto considerável nos resultados. Pode também haver os benefícios de menos emissões de carbono.

Para ajudar com esses objetivos, a unidade DeepMind do Google vem aplicando deep learning, o que envolveu a melhoria na gestão da energia eólica. Mesmo que essa seja uma fonte limpa de energia, pode ser difícil utilizá-la por causa das mudanças climáticas.

Contudo, os algoritmos de deep learning da DeepMind têm sido críticos. Aplicados a 700 megawatts de energia eólica nos Estados Unidos, eles foram capazes de fazer previsões precisas para a produção em um prazo de 36 horas.

De acordo com o blog da DeepMind:

> *Isso é importante, porque as fontes de energia que podem ser programadas (ou seja, podem fornecer uma quantidade definida de eletricidade em um horário definido) são muitas vezes mais valiosas para a rede... Até o momento, o deep learning aumentou o valor de nossa energia eólica em cerca de 20% em comparação ao cenário de linha de base de ausência de compromissos baseados no tempo com a rede.*[22]

É claro, no entanto, que esse sistema de deep learning poderia ajudar mais do que o Google – ele poderia ter um amplo impacto sobre o uso de energia em todo o mundo.

Estudo de caso: Terremotos

Terremotos são extremamente complicados de entender. Também são extremamente difíceis de prever. É preciso avaliar falhas, formações rochosas e deformações,

atividade eletromagnética e mudanças nas águas subterrâneas. Há evidências de que os animais têm a capacidade de pressentir um terremoto!

Ao longo das décadas, no entanto, cientistas coletaram enormes quantidades de dados sobre esse tema. Em outras palavras, isso poderia ser uma aplicação de deep learning, certo? Com certeza.

Sismólogos da Caltech, entre eles Yisong Yue, Egill Hauksson, Zachary Ross e Men-Andrin Meier, têm feito pesquisas consideráveis sobre o assunto, usando redes neurais convolucionais e redes neurais recorrentes. Eles estão tentando construir um sistema eficiente de alerta precoce.

Veja o que Yue tem a dizer:

> *A IA consegue [analisar terremotos] mais rápido e com maior precisão do que os seres humanos, e detecta até mesmo padrões que escapariam ao olho humano. Além disso, os padrões que esperamos extrair são de difícil captura adequada para os sistemas baseados em regras e, portanto, as habilidades avançadas de correspondência de padrões do deep learning moderno podem oferecer um desempenho superior ao dos algoritmos automatizados de monitoramento de terremotos existentes.*[23]

O mais importante, contudo, é melhorar a coleta de dados. Isso significa uma quantidade maior de análise de pequenos terremotos (na Califórnia, há uma média de 50 por dia). O objetivo é criar um catálogo de terremotos que viabilize a criação de um sismólogo virtual capaz de fazer avaliações do fenômeno mais rápido do que um ser humano. Isso poderia permitir a antecipação da chegada de um terremoto, o que pode ajudar a salvar vidas e propriedades.

Estudo de caso: Radiologia

Exames PET scans e ressonâncias magnéticas são tecnologias incríveis. Contudo, elas definitivamente têm desvantagens. Um paciente precisa ficar dentro de um tubo de confinamento por 30 minutos a uma hora. Isso é desconfortável e significa ser exposto ao gadolínio, que vem mostrando causar efeitos colaterais nocivos.

Greg Zaharchuk e Enhao Gong, que se conheceram em Stanford, pensaram que poderia haver uma maneira melhor de realizar esses exames. Zaharchuk tinha mestrado e doutorado, além de uma especialização em radiologia. Ele foi orientador de doutorado de Gong, que tinha PhD em Engenharia Elétrica em deep learning e reconstrução de imagens médicas.

Em 2017, os dois pesquisadores cofundaram a Subtle Medical e contrataram alguns dos cientistas de imagens, radiologistas e especialistas em IA mais brilhantes. Juntos, eles se propuseram ao desafio de melhorar PET scans e ressonâncias magnéticas. A Subtle Medical criou um sistema que não só reduzia o tempo para a realização dos

exames em até dez vezes, mas também aumentava sua precisão. O projeto foi executado em GPUs NVIDIA de ponta.

Então, em dezembro de 2018, o sistema recebeu a autorização 510 (k) da FDA (Federal Drug Administration) e a aprovação da marcação CE para o mercado europeu.[24] Foi o primeiro dispositivo médico nuclear baseado em IA a alcançar essas duas designações.

A Subtle Medical tem outros planos para revolucionar o ramo da radiologia. A partir de 2019, a empresa começou a desenvolver o SubtleMRTM, que será ainda mais poderoso do que a solução atual da empresa, e o SubtleGADTM, que reduzirá as doses de gadolínio.[25]

Hardware para deep learning

Em relação aos sistemas de chips para deep learning, as GPUs têm sido a principal escolha. Contudo, à medida que a IA fica mais sofisticada – como no caso das GANs – e os conjuntos de dados crescem, há certamente mais espaço para novas abordagens. As empresas também têm necessidades personalizadas, como em termos de funções e dados. Afinal, um aplicativo para um consumidor geralmente é bastante diferente daquele criado com foco na empresa.

Como resultado, algumas das megaempresas de tecnologia têm desenvolvido seus próprios chipsets:

- *Google*: no verão de 2018, a empresa anunciou a terceira versão de sua Tensor Processing Unit (TPU –Unidade de Processamento de Tensor; o primeiro chip foi desenvolvido em 2016).[26] Os chips são tão poderosos – manuseando mais de 100 petaflops para treinamento de modelos – que é preciso haver resfriamento líquido nos data centers. O Google também anunciou uma versão de sua TPU para dispositivos. Essencialmente, isso significa que o processamento terá menos latência, visto que não haverá necessidade de acessar a nuvem.
- *Amazon:* em 2018, a empresa anunciou a AWS Inferentia.[27] A tecnologia, que surgiu da aquisição da Annapurna em 2015, está focada no processamento de operações complexas de inferência. Em outras palavras, isso é o que acontece depois que um modelo é treinado.
- *Facebook e Intel*: essas empresas uniram forças para criar um chip de IA.[28] No entanto, a iniciativa ainda está em fase inicial. A Intel também tem chamado a atenção com seu chip de IA chamado Nervana Neural Network Processor (NNP – Processador de rede neural Nervana).
- *Alibaba*: a organização criou a própria empresa de chips de IA, chamada Pingtouge.[29] Ela também tem planos de construir um processador de computador quântico baseado em qubits (que representam partículas

subatômicas como elétrons e fótons).

- *Tesla:* Elon Musk desenvolveu o próprio chip de IA. Com 6 bilhões de transistores, ele é capaz de processar 36 trilhões de operações por segundo.[30]

Há uma variedade de startups que também estão apostando no mercado de chips de IA. Entre as principais empresas está a Untether AI, focada na criação de chips que aumentam as velocidades de transferência de dados (essa tem sido uma parte particularmente difícil da IA). Em um dos protótipos da empresa, esse processo foi mais de 1.000 vezes mais rápido do que em um chip de IA típico.[31] A Intel, junto com outros investidores, participou de uma rodada de financiamento de US$ 13 milhões em 2019.

Agora, quando se trata de chips de IA, a NVIDIA detém a quota de mercado dominante. Por causa da importância dessa tecnologia, entretanto, parece inevitável que haverá cada vez mais ofertas surgindo.

Quando usar deep learning?

Devido ao poder do deep learning, existe certa tentação para escolher essa tecnologia ao criar um projeto de IA. No entanto, isso pode ser um grande erro. Deep learning ainda tem aplicações restritas, como é o caso dos conjuntos de dados de texto, vídeo, imagem e séries temporais. Há ainda a necessidade de grandes quantidades de dados e sistemas computacionais de alta potência.

Ah, e deep learning é melhor quando os resultados podem ser quantificados e verificados.

Para entender a razão disso, considere o exemplo a seguir. Uma equipe de pesquisadores, liderada por Thomas Hartung (toxicologista da Universidade Johns Hopkins), criou um conjunto de dados de cerca de 10.000 produtos químicos com base em 800.000 testes em animais. Ao utilizar deep learning, os resultados mostraram que o modelo foi mais preditivo para toxicidade do que muitos testes anteriores.[32] Lembre-se de que os testes em animais podem não somente ser caros e exigirem medidas de segurança, mas também apresentarem resultados inconsistentes por conta da repetição do mesmo produto químico.

"O primeiro cenário ilustra o poder preditivo do deep learning e sua capacidade de descobrir correlações que um ser humano nunca encontraria a partir de grandes conjuntos de dados", disse Sheldon Fernandez, CEO da DarwinAI.[33]

Afinal, existe algum cenário no qual deep learning não se encaixe adequadamente? Na verdade, uma ilustração disso é a Copa do Mundo de 2018 na Rússia, que foi vencida pela França. Muitos pesquisadores tentaram prever os resultados de todos os 64 jogos, mas as apostas estavam longe de ser precisas:[34]

- Um grupo de pesquisadores empregou o modelo de consenso de casas de

apostas que indicava que o Brasil venceria.

- Outro grupo de pesquisadores usou algoritmos, como o da floresta aleatória e o da distribuição de Poisson, para prever que a Espanha ganharia.

O problema aqui é que é difícil encontrar as variáveis certas e com poder preditivo ideal. Na verdade, os modelos de deep learning são basicamente incapazes de lidar com a complexidade de recursos para certos eventos, em especial aqueles que apresentam características de caos.

No entanto, mesmo dispondo da quantidade certa de dados e do poder computacional adequado, ainda será necessário contratar pessoas com experiência em deep learning, o que não é fácil. Tenha em mente que é um desafio selecionar o modelo certo e ajustá-lo. Quantos hiperparâmetros devem existir? Qual deve ser o número de camadas ocultas? Como se avalia o modelo? Tudo isso é altamente complexo.

Até mesmo os especialistas podem fazer as coisas erradas. Veja este depoimento de Sheldon:

> *Um dos nossos clientes automotivos identificou um comportamento bizarro no qual um carro autônomo virava à esquerda com regularidade crescente quando o céu estava num certo tom de roxo. Depois de meses de uma difícil depuração, descobriu-se que o treinamento para certos cenários dessa manobra tinha sido realizado no deserto de Nevada, quando o céu apresentava uma tonalidade particular. Sem o conhecimento de seus projetistas humanos, a rede neural havia estabelecido uma correlação entre o comportamento da virada e a tonalidade celestial.*[35]

Existem algumas ferramentas que estão ajudando com o processo de deep learning, como o SageMaker da Amazon.com, o HyperTune do Google e o SigOpt. Ainda há, no entanto, um longo caminho a percorrer.

Se o deep learning não for adequado, então é possível considerar machine learning, que muitas vezes requer relativamente menos dados. Além disso, os modelos tendem a ser muito mais simples, mas os resultados ainda podem ser mais eficientes.

Desvantagens do deep learning

Com todos os avanços e inovações, é razoável que muitas pessoas considerem o deep learning como uma bala de prata. Isso significa que não temos mais que dirigir um carro. Pode até significar que vamos curar o câncer.

Como é possível não se sentir animado e otimista? Isso é natural e razoável. Entretanto, é importante observar que o deep learning ainda está em um estágio inicial e, na verdade, há muitos problemas persistentes. É uma boa ideia moderar as expectativas.

Em 2018, Gary Marcus escreveu um artigo intitulado “Deep Learning: A Critical

Appraisal" ("Deep learning: uma avaliação crítica"), no qual apresentava claramente os desafios.[36] Em seu texto, ele ressalta:

> *Num contexto de progresso considerável em áreas como reconhecimento de voz, reconhecimento de imagem e jogos, e considerável entusiasmo na imprensa popular, apresento dez preocupações para o deep learning e sugiro que o mesmo deve ser complementado por outras técnicas se quisermos chegar à Inteligência Artificial Geral.*[37]

Marcus definitivamente tem o conhecimento certo para apresentar suas preocupações, visto que possui formação acadêmica e de negócios em IA. Antes de se tornar professor do Departamento de Psicologia da Universidade de New York, vendeu para a Uber sua startup, chamada Geometric Intelligence. Marcus também é autor de vários livros best-sellers, como *The Haphazard Construction of the Human Mind* ("A construção aleatória da mente humana").[38]

Veja algumas das preocupações do autor com relação ao deep learning:

- *Caixa preta*: um modelo de deep learning pode facilmente ter milhões de parâmetros que envolvem muitas camadas ocultas. Ter uma compreensão clara disso está realmente além das capacidades de uma pessoa. É verdade que essa característica pode não ser necessariamente um problema no reconhecimento de gatos em um conjunto de dados, mas poderia definitivamente ser um empecilho em modelos para diagnóstico médico ou determinação da segurança de uma plataforma de petróleo. Nessas situações, reguladores vão precisar de uma boa compreensão da transparência dos modelos. Devido a isso, pesquisadores buscam criar sistemas para determinar a "explicabilidade", que fornece uma compreensão dos modelos de deep learning.
- *Dados:* o cérebro humano tem suas falhas. Contudo, há algumas funções que ele desempenha muito bem, como a capacidade de aprender por abstração. Por exemplo, suponha que Jan, que tem cinco anos de idade, vai a um restaurante com a família. Sua mãe aponta um item no prato e diz que é um "taco". Não há necessidade de explicá-lo ou fornecer qualquer informação adicional sobre o alimento. Em vez disso, o cérebro de Jan vai processar instantaneamente as informações e assimilar o padrão geral. No futuro, quando vir outro taco – mesmo que ele apresente diferenças, como um molho –, ela vai saber o que é. Para a maioria das pessoas, isso é intuitivo. Infelizmente, contudo, quando se trata de deep learning, não há aprendizagem sobre tacos por meio de abstração! O sistema tem de processar enormes quantidades de dados para reconhecê-los. Claro, isso não é um problema para organizações como Facebook, Google ou mesmo Uber. No entanto, muitas empresas têm conjuntos de dados muito mais limitados. O resultado é que deep learning pode não ser uma boa opção.
- *Estrutura hierárquica*: essa forma de organização não existe no deep learning.

Devido a isso, a compreensão da linguagem ainda tem um longo caminho a percorrer (especialmente com longas discussões).

- *Inferência aberta*: Marcus observa que o deep learning não consegue entender as nuances entre "John promised Mary to leave" ("John prometeu que Maria iria embora") e "John promised to leave Mary" ("John prometeu deixar Maria"). Além do mais, deep learning está longe de ser capaz de, por exemplo, ler *Orgulho e Preconceito*, de Jane Austen, e identificar as motivações de caráter de Elizabeth Bennet.
- *Pensamento conceitual*: deep learning não consegue compreender conceitos como democracia, justiça ou felicidade. Também não tem imaginação e, portanto, não concebe novas ideias ou planos.
- *Senso comum:* isso é algo que deep learning não faz bem; o que significa que um modelo pode ser facilmente confundido. Por exemplo, se você perguntar a um sistema de IA algo como: "é possível fazer um computador com uma esponja?", ele provavelmente não vai saber que essa é uma pergunta ridícula.
- *Causalidade*: deep learning é incapaz de determiná-la. Tudo se resume a encontrar correlações.
- *Conhecimento prévio*: CNNs podem ajudar com algumas informações anteriores, mas isso é limitado. O deep learning ainda é bastante restrito, já que só resolve um problema por vez. Ele não consegue, por exemplo, pegar os dados e criar algoritmos que abranjam vários domínios. Além disso, um modelo não se adapta. Se houver alteração nos dados, então um novo modelo precisará ser treinado e testado. Por fim, deep learning não tem compreensão prévia do que as pessoas sabem instintivamente – como a física básica do mundo real. Isso é algo que tem de ser explicitamente programado em um sistema de IA.
- *Estática:* deep learning funciona melhor em ambientes simples. É por isso que a IA vem sendo tão eficaz em jogos de tabuleiro, que contam com um conjunto claro de regras e limites. O mundo real, entretanto, é caótico e imprevisível. Isso significa que o deep learning pode não dar conta de problemas complexos, mesmo em carros autônomos.
- *Recursos*: Um modelo de deep learning muitas vezes requer uma enorme quantidade de energia da CPU, como é o caso das GPUs. Isso pode ficar caro. Uma opção, no entanto, é usar um serviço de nuvem de terceiros.

Isso é muito? Com certeza, mas o artigo ainda deixou de fora algumas desvantagens. Aqui estão elas:

- *Efeito borboleta*: devido à complexidade de dados, redes e conexões, uma pequena alteração pode causar grande impacto nos resultados do modelo de deep learning. Isso poderia facilmente levar a conclusões erradas ou enganosas.

- *Superadaptação (Overfitting)*: explicamos esse conceito no início do capítulo.

Quanto a Marcus, seu maior medo é que a IA possa "ficar presa a um mínimo, habitando na parte errada do espaço intelectual, concentrando-se demais na exploração detalhada de uma determinada classe de modelos acessíveis, embora limitados, voltados ao tratamento de problemas simples – potencialmente negligenciando aventuras mais arriscadas que podem levar a caminhos mais robustos".

No entanto, o estudioso não é pessimista. Ele acredita que os pesquisadores precisam ir além do deep learning para encontrar novas técnicas que possam resolver problemas difíceis.

Conclusão

Embora Marcus tenha apontado as falhas no deep learning, o fato é que essa abordagem de IA ainda é extremamente poderosa. Em menos de uma década, ela revolucionou o mundo da tecnologia – além de causar impactos significativos em áreas como finanças, robótica e saúde.

Com o aumento dos investimentos por parte de grandes empresas de tecnologia e capitais de risco, haverá mais inovações com os modelos. Isso incentivará os engenheiros a fazerem cursos de pós-graduação e alavancará a criação de um ciclo virtuoso de avanços.

Principais aprendizados

- Deep learning, um subárea de machine learning, processa grandes quantidades de dados para detectar relacionamentos e padrões que os seres humanos são muitas vezes incapazes de perceber. A palavra "deep" ("profundo") refere-se ao número de camadas ocultas.
- Uma rede neural artificial (ANN) é uma função que inclui unidades que têm pesos e são usadas para prever valores em um modelo de IA.
- Uma camada oculta é uma parte de um modelo que processa os dados de entrada.
- Uma rede neural feed-forward tem dados que trafegam apenas da entrada para a camada oculta, e dela para a saída. Os resultados não são retroalimentados. No entanto, eles podem entrar em outra rede neural.
- Uma função de ativação é não linear. Em outras palavras, ela tende a fazer um trabalho melhor na reflexão do mundo real.
- Uma sigmoide é uma função de ativação que comprime o valor da entrada a um intervalo entre 0 e 1, o que simplifica a análise.

- A retropropagação é uma técnica sofisticada de ajuste dos pesos em uma rede neural. Essa abordagem tem sido fundamental para o crescimento do deep learning.
- Uma rede neural recorrente (RNN) é uma função que não só processa uma entrada, mas também todas as anteriores informadas ao longo do tempo.
- Uma rede neural convolucional (CNN) analisa os dados seção por seção (ou seja, por convoluções). Esse modelo é voltado para aplicações complexas, como reconhecimento de imagem.
- Uma rede adversária generativa ou GAN é onde duas redes neurais competem entre si em um ciclo de feedback apertado. O resultado é muitas vezes a criação de um novo objeto.
- A explicabilidade descreve técnicas de transparência com modelos complexos de deep learning.

---

1 Siddhartha Mukherjee, "The Algorithm Will See You Now" ("O Algoritmo vai vê-lo agora"), The New Yorker, 3 de abril de 2017, https://www.newyorker.com/magazine/2017/04/03/ai-versus-md.

2 N.T.: "Processo de obtenção de serviços, ideias ou conteúdo mediante a solicitação de contribuições de um grande grupo de pessoas e, especialmente, de uma comunidade online, em vez de usar fornecedores tradicionais ou uma equipe de empregados." Fonte: https://pt.wikipedia.org/wiki/Crowdsourcing

3 https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

4 https://medium.com/@GabriellaLeone/the-best-explanation-machine-learning-vs-deep-learning-d5c123405b11

5 David E. Rumelhart, Geoffrey E. Hinton e Ronald J. Williams, "Learning Representations by Backpropagating Errors", *Nature* 323 (1986): 533–536.

6 www.insurancejournal.com/news/national/2002/05/01/16857.htm

7 Sepp Hochreiter e Jürgen Schmidhuber, "Long Short-Term Memory" ("Memória longa de curto prazo", Neural Computation 9, n. 8 (1997): 1735-80.

8 www.argotrans.com/blog/accurate-google-translate-2018/

9 www.techspot.com/news/75637-google-translate-not-monetized-despite-converting-over-100.html

10 www.argotrans.com/blog/accurate-google-translate-2018/

11 https://gizmodo.com/google-translate-can-help-doctors-bridge-the-language-g-1832881294

12 Yann LeCun et al., "Gradient-Based Learning Applied to Document Recognition" ("Aprendizagem baseada em gradiente aplicada ao reconhecimento de documentos"), IEEE 86 n. 11 (1998): 2278-2324.

13 Ian Goodfellow, Yoshua Bengio e Aaron Courville, Deep Learning (Cambridge, MA: The MIT Press, 2016).

14 www.technologyreview.com/s/610253/the-ganfather-the-man-who's-given-machines-the-gift-of-imagination/

15 https://github.com/hindupuravinash/the-gan-zoo

16 https://trendsandevents4developers.wordpress.com/2017/04/24/the-coolest-idea-in-deep-learning-in-20-years-and-more/

17 www.hpcwire.com/2018/08/14/cern-incorporates-ai-into-physics-based-simulations/

18 www.washingtonpost.com/nation/2018/10/26/year-old-developed-code-ai-portrait-that-sold-christies/?utm_term=.b2f366a4460e

19 www.cnbc.com/2018/12/28/research-claims-fake-fingerprints-could-hack-a-third-of-smartphones.html

20 http://fortune.com/2019/01/31/what-is-deepfake-video/

21 www.ucsf.edu/news/2018/12/412946/artificial-intelligence-can-detect-alzheimers-disease-brain-scans-six-years

22 https://deepmind.com/blog/machine-learning-can-boost-value-wind-energy/

23 www.caltech.edu/about/news/qa-creating-virtual-seismologist-84789

24 https://subtlemedical.com/subtlemedical-receives-fda-510k-clearance-and-ce-mark-approval-for-subtlepet/

25 www.streetinsider.com/Press+Releases/Subtle+Medical+Receive+FDA+510%28k%29+Clearance+and+CE+Mark+Approval+for+SubtlePET™/14892974.html

26 www.theregister.co.uk/2018/05/09/google_tpu_3/

27 https://aws.amazon.com/about-aws/whats-new/2018/11/announcing-amazon-inferentia-machine-learning-inference-microchip/

28 www.analyticsindiamag.com/inference-chips-are-the-next-big-battlefield-for-nvidia-and-intel/

29 www.technologyreview.com/s/612190/why-alibaba-is-investing-in-ai-chips-and-quantum-computing/

30 www.technologyreview.com/f/613403/tesla-says-its-new-self-driving-chip-will-help-make-its-cars-autonomous/

31 www.technologyreview.com/f/613258/intel-buys-into-an-ai-chip-that-can-transfer-data-1000-times-faster/

32 www.nature.com/articles/d41586-018-05664-2

33 Entrevista realizada pelo autor com Sheldon Fernandez, CEO da DarwinAI.

34 https://medium.com/futuristone/artificial-intelligence-failed-in-world-cup-2018-6af10602206a

35 Entrevista realizada pelo autor com Sheldon Fernandez, CEO da DarwinAI.

36 Gary Marcus, "Deep Learning: A Critical Appraisal", arXiv, 1801.00631v1 [cs. AI]:1-27, 2018.

37 https://arxiv.org/ftp/arxiv/papers/1801/1801.00631.pdf

38 Gary Marcus, Kluge: Haphazard Construction of the Human Mind (Houghton Mifflin, 2008).

CAPÍTULO 5

Automação Robótica de Processos (RPA)

Um caminho mais fácil para a IA

Interagindo com aplicativos exatamente como um ser humano faria, robôs de software conseguem abrir anexos de e-mail, preencher formulários eletrônicos, registrar e informar dados e executar outras tarefas que imitam a ação humana.

– Kaushik Iyengar,
Diretor de Transformação Digital e Otimização da AT&T[1]

Em 2005, Daniel Dines e Marius Tirca fundaram a UiPath, localizada em Bucareste, Romênia. A empresa concentrou-se principalmente no fornecimento de serviços de integração para aplicativos de organizações como Google, Microsoft e IBM. No entanto, eles enfrentaram dificuldades, já que o foco da empresa era, principalmente, o trabalho personalizado para seus clientes.

Em 2013, a UiPath estava perto de ser fechada. Os fundadores, entretanto, não desistiram, pois viram a situação como uma oportunidade de repensar o negócio e encontrar uma nova oportunidade.[2] Para isso, começaram a construir uma plataforma para a automação robótica de processos (RPA – Robotic Process Automation). A abordagem, que já existia desde 2000, tratava da automatização de tarefas rotineiras e repetitivas dentro de uma empresa.

A RPA, no entanto, era a área menos movimentada no mundo da tecnologia – como demonstram as taxas de crescimento lento. Apesar disso, Dines e Tirca estavam convencidos de que poderiam transformar a indústria. Uma das principais razões era a ascensão da IA e da nuvem.

Uma nova estratégia foi traçada e o crescimento decolou. Dines e Tirca também foram agressivos na busca por financiamentos, na inovação de sua plataforma de RPA e na expansão para mercados globais.

Em 2018, a UiPath foi considerada a empresa de software empresarial de mais rápido crescimento – da história. A receita anual recorrente subiu de US$ 1 milhão para US$ 100 milhões, com mais de 1.800 clientes.[3] A empresa tinha o sistema de RPA mais utilizado no mundo.

A UiPath atraiu um total de US$ 448 milhões em capital de risco de empresas como CapitalG, Sequoia Capital e Accel. A avaliação foi de US$ 3 bilhões.

Diante de tudo isso, startups voltadas a RPA também obtiveram financiamento

significativo. Então, novamente, previa-se que o mercado experimentaria um enorme crescimento. A Grand View Research sugere que os investimentos na área atingirão US$ 3,97 bilhões nos Estados Unidos até 2025.[4]

Curiosamente, Forrester tinha a seguinte opinião sobre a tendência RPA:

As empresas mais bem-sucedidas de hoje geralmente operam com menos funcionários do que as do passado. Considere que a Kodak, em seu auge em 1973, empregou 120.000 pessoas, mas, quando o Facebook comprou o Instagram em 2012, o site de compartilhamento de fotos empregava apenas 13 funcionários. Em 2019, prevemos que uma em cada 10 startups – operando de forma mais ágil, enxuta e escalável – olhará para o mundo através da lente de tarefas, não de empregos, e construirá modelos de negócios em torno dos princípios de automação em primeiro lugar.[5]

A RPA é mais uma área que foi sobrecarregada com IA. Se for o caso, pode ser a porta de entrada para muitas empresas, porque a implementação em geral não leva muito tempo nem exige custos pesados.

Neste capítulo, vamos dar uma olhada na RPA e ver como ela pode ser um fator crítico para muitas empresas.

O que é RPA?

O termo Robotic Process Automation (automação robótica de processos) pode ser um pouco confuso. A palavra "robótica" não significa robôs físicos (vamos falar sobre eles no Capítulo 7). Em vez disso, o termo faz referência a robôs ou bots baseados em software.

A RPA permite a utilização de sistemas visuais low-code de drag-and-drop para automatizar o fluxo de trabalho de um processo. Alguns exemplos incluem:

- Inserção, alteração e rastreamento de documentos, contratos e informações de funcionários pertencentes aos Recursos Humanos (RH)
- Detecção de problemas no atendimento ao cliente e implementação de medidas para solução
- Processamento de solicitações de seguros
- Envio de faturas
- Emissão de reembolsos aos clientes
- Conciliação de registros financeiros
- Transferência de dados de um sistema para outro
- Fornecimento de respostas-padrão aos clientes

Tudo isso é possível fazendo com que um bot replique os fluxos de trabalho para um aplicativo como um sistema ERP (Enterprise Resource Planning – sistema integrado de gestão empresarial) ou CRM (Customer Relationship Management – gestão de relacionamento com o cliente). Isso pode até ser feito com o programa RPA registrando os passos dos funcionários ou com o uso da tecnologia OCR (Optical Character Recognition – reconhecimento óptico de caracteres) para traduzir

notas manuscritas. Pense na RPA como um funcionário digital.

Existem duas categorias nesse tipo de tecnologia:

- *RPA autônoma:* esse é um processo completamente autônomo, já que o bot será executado em segundo plano. Isso não significa que não haja intervenção humana. Ainda haverá intervenção para o gerenciamento de exceções. Isto é, quando o bot encontra algo que não entende.
- *RDA (Robotic Desktop Automation – automação robótica de desktop)*: é aqui que a RPA ajuda um funcionário com um trabalho ou tarefa. Uma aplicação comum é com a central de contatos. Quando chega uma ligação, o representante pode usar RDA para ajudar a encontrar respostas, enviar mensagens, consultar informações de perfil do cliente e obter um panorama a respeito do que fazer a seguir. A tecnologia ajuda a melhorar ou aumentar a eficiência do funcionário.

Prós e contras da RPA

Para um funcionário típico, grande parte do tempo – administrativo – é gasto em tarefas de rotina. Com RPA, entretanto, as empresas muitas vezes podem obter um ROI (Return On Investment – retorno sobre o investimento) significativo – desde que a implementação seja feita da maneira adequada.

Aqui estão algumas outras vantagens:

- *Satisfação do cliente*: RPA significa erros mínimos, bem como alta velocidade. Um bot também funciona 24 horas por dia, 7 dias por semana. Isso significa que as pontuações de satisfação do cliente – como o NPS (Net Promoter Score[6]) – devem melhorar. Observe que cada vez mais clientes, como os da geração Millennial, preferem lidar com aplicativos ou sites, não pessoas! RPA também significa que os representantes terão mais tempo para dedicar a tarefas de valor agregado, em vez de lidar com questões tediosas que consomem tempo.
- *Escalabilidade*: uma vez que um bot é criado, ele pode ser rapidamente expandido para atender picos de atividade. Isso pode ser fundamental para empresas sazonais, como as varejistas.
- *Conformidade*: para as pessoas, é difícil acompanhar regras, regulamentos e leis, em especial porque eles muitas vezes mudam. Com RPA, entretanto, a conformidade é construída no processo – e é sempre seguida. Isso pode significar um grande benefício para que se evitem problemas legais e multas.
- *Insights e Analytics:* as plataformas de RPA de última geração vêm equipadas com painéis sofisticados, que se concentram em indicadores-chave de performance (KPIs – Key Performance Indicators) para o negócio. Também é possível configurar alertas para quando surgir algum problema.
- *Sistemas legados*: empresas mais antigas estão muitas vezes atoladas com

sistemas de TI antigos, o que torna extremamente difícil implementar uma transformação digital. O software de RPA, entretanto, é capaz de funcionar muito bem com ambientes de TI legados.

- *Dados*: devido à automação, os dados são muito mais limpos, visto que os erros de entrada são mínimos. Isso significa que as organizações terão – ao longo do tempo – entendimentos mais precisos de seus negócios. A qualidade dos dados também aumentará a probabilidade de sucesso das implementações de IA.

Embora tudo isso seja ótimo, a RPA também tem suas desvantagens. Por exemplo, se existem processos atuais ineficientes e a organização se apressa para implementar um sistema de RPA, isso representa, essencialmente, a replicação de uma abordagem ruim! É, portanto, fundamental avaliar os fluxos de trabalho antes de implementar um sistema desse tipo.

Há também outros problemas potenciais a observar, como os seguintes:

- *Fragilidade:* RPA pode facilmente não funcionar de maneira adequada se houver mudanças nas aplicações subjacentes. Isso também pode acontecer se houver alterações nos procedimentos e regulamentos. É verdade que os sistemas mais novos estão melhorando na adaptação e podem recorrer a APIs. No entanto, a RPA não é uma atividade prática.
- *Aplicativos virtualizados*: esse tipo de software, como o da Citrix, pode ser de difícil adaptação com os sistemas de RPA porque não conseguem capturar efetivamente os processos. A razão é que os dados são armazenados em um servidor externo e a saída é uma captura instantânea da tela. Algumas organizações, contudo, estão usando a IA para resolver essa questão, como a UiPath. A empresa criou um sistema, chamado "Pragmatic AI", que usa visão computacional para interpretar as capturas de tela para registrar os processos.
- *Especialização*: muitas ferramentas de RPA estão voltadas a atividades de uso geral. Entretanto, pode haver áreas que exigem especialização, como finanças. Nesse caso, é possível procurar um aplicativo de software específico da área que possa lidar com isso.
- *Teste:* absolutamente crítico. Primeiro, é necessário experimentar algumas transações para garantir que o sistema esteja funcionando corretamente. Depois disso, pode-se fazer uma implantação mais extensa do sistema de RPA.
- *Propriedade*: a tentação é fazer com que a TI se aproprie dos processos de implementação e gestão da RPA. No entanto, isso provavelmente não é aconselhável. A razão? Os sistemas de RPA usam bem pouca tecnologia. Eles podem, inclusive, ser desenvolvidos por não programadores. Devido a isso, gerentes de negócios são os profissionais mais recomendados para cuidarem do processo, uma vez que geralmente podem lidar com as questões técnicas e

contam com uma compreensão mais sólida dos fluxos de trabalho dos funcionários.

- *Resistência:* a mudança é sempre difícil. Com RPA, pode haver temores de que a tecnologia vai causar desemprego. Isso significa que é necessário dispor de um conjunto claro de mensagens que reforcem os benefícios da tecnologia. Por exemplo, a RPA vai significar mais tempo de concentração em assuntos importantes, o que deve tornar o trabalho de uma pessoa mais interessante e significativo.

O que se pode esperar da RPA?

Quando se trata de RPA, a indústria ainda está em fase inicial. No entanto, há sinais claros de que a tecnologia está fazendo uma grande diferença para muitas empresas.

Observe o relatório de pesquisa da Computer Economics Technology, que incluiu cerca de 250 empresas (advindas de muitos setores e com receitas que variavam de US$ 20 milhões a mais de US$ 1 bilhão). Daquelas que implementaram um sistema de RPA, cerca de metade relatou um retorno positivo após 18 meses de implantação. Isso é definitivamente um destaque para o software empresarial, que pode ser de difícil adoção.[7]

Para ter noção da importância estratégica dessa tecnologia, vejamos o que está fazendo o Departamento de Defesa dos Estados Unidos – que está envolvido em mais de 500 projetos de IA. Observe o que disse o diretor do Centro Conjunto de Inteligência Artificial da agência, Tenente-general da Força Aérea Jack Shanahan, durante uma audiência no Congresso:

> *Falar sobre automação inteligente ou, no vernáculo da indústria, automação robótica de processos, não atrai manchetes em termos de grandes projetos de IA, mas pode ser onde a maioria da eficiência pode ser encontrada. Esse é o caso se você ler alguns dos diários na indústria, seja na área de medicina ou finanças, esse é o lugar onde os ganhos iniciais estão sendo obtidos em IA. Alguns dos outros projetos que assumimos no departamento provavelmente levarão anos na obtenção de retorno sobre o investimento.*[8]

Apesar de tudo isso, ainda há muitas implementações de RPA com falhas. A Ernst & Young, por exemplo, recebeu uma grande quantidade de solicitações de consultoria por conta disso. Com base nessa experiência, a taxa de falha para projetos de RPA em fase inicial varia de 30% a 50%.[9]

Isso é inevitável, no entanto, em qualquer tipo de software empresarial. Contudo, até agora, os problemas parecem estar principalmente relacionados a planejamento, estratégia e expectativas – e não à tecnologia.

Outro problema é que a animação em torno da RPA pode estar aumentando as

expectativas para níveis excessivos. Isso significa que a decepção será bastante comum, mesmo que as implementações sejam bem-sucedidas!

Obviamente, as tecnologias não são uma solução completa. Além disso, exigem muito tempo, esforço e diligência para que funcionem.

Como implementar a RPA

Então, quais são as medidas a tomar para uma implementação de RPA bem-sucedida? Não há uma resposta padrão, mas certamente há algumas melhores práticas surgindo:

- Determine as funções certas a automatizar;
- Avalie os processos;
- Selecione o fornecedor de RPA e implante o software;
- Monte uma equipe para gerenciar a plataforma de RPA.

Vamos dar uma olhada em cada uma dessas práticas.

Determine as funções certas a automatizar

> *Sim, a automação excessiva na Tesla foi um erro. Para ser mais preciso, meu erro. Os seres humanos são subestimados.*
>
> **– Elon Musk, CEO da Tesla**[10]

Embora a RPA seja poderosa e possa ser de grande ajuda para uma empresa, suas capacidades ainda são razoavelmente limitadas. A tecnologia comporta-se melhor na automatização de processos repetitivos, estruturados e rotineiros. Entre eles estão agendamento, entrada/transferência de dados, cumprimento de regras e execução de fluxos de trabalho, recortar e colar, preencher formulários e pesquisar. Isso significa que a RPA pode de fato ter seu papel em praticamente todos os departamentos de uma organização.

Então, onde essa tecnologia em geral pode fracassar? Bem, se um processo requer julgamento independente, então RPA provavelmente não faz sentido. A recomendação também se aplica quando os processos estão sujeitos a alterações frequentes. Nessa situação, é provável que se gaste muito tempo com ajustes contínuos nas configurações.

Uma vez que se estabeleça uma parte do negócio onde a tecnologia parece ser uma boa escolha, há uma variedade de outros pontos a considerar. Em outras palavras, é provável que se tenha mais sucesso com um projeto se os seguintes aspectos forem observados:

- Áreas do negócio que apresentam níveis graves de baixo desempenho;

- Processos que ocupam alta porcentagem do tempo dos funcionários e envolvem altas taxas de erro;
- Tarefas que demandam mais contratações quando há volumes mais altos;
- Áreas para as quais a terceirização está sendo considerada;
- Processos com grande número de etapas e nos quais existem várias aplicações envolvidas.

Avalie os processos

É comum que uma empresa tenha muitos processos não descritos. E não há problemas nisso. Essa abordagem permite a adaptabilidade, uma característica na qual as pessoas são boas.

No entanto, esse está longe de ser o caso com um bot. Para ter uma implementação bem-sucedida, é necessário ter uma avaliação profunda dos processos. Isso pode realmente demandar tempo e pode ser necessário recorrer a consultores externos para ajudar. Eles têm a vantagem de ser mais neutros e capazes de identificar as fraquezas.

Alguns dos fornecedores de RPA têm as próprias ferramentas para ajudar com a análise de processos – e elas devem, definitivamente, ser usadas. Há também fornecedores de software de terceiros que têm as próprias ofertas. Um deles é a Celonis, que se integra com plataformas de RPA como UiPath, Automation Anywhere, Blue Prism e outros. O software executa essencialmente uma ressonância magnética digital que analisa dados, fornecendo insights sobre como seus processos realmente funcionam. Ela também identifica fraquezas e oportunidades, como aumentar as receitas, melhorar a satisfação do cliente e liberar recursos.

Independentemente da abordagem escolhida, é fundamental formular um plano claro que conte com as áreas de TI, gestão e departamentos afetados. Também é preciso certificar-se de envolver as pessoas de analytics, pois pode haver oportunidades para aprimorar os dados.

Selecione o fornecedor de RPA e implante o software

Ao passar pelos dois primeiros passos, você estará em uma posição muito confortável para avaliar os diferentes sistemas de RPA. Por exemplo, se o principal objetivo é diminuir a equipe, então é preciso procurar um software focado em bots desacompanhados. Ou, se o que se quer é impulsionar os dados – para aplicações de IA –, isso levará a outros tipos de plataformas de RPA.

Então, uma vez que um fornecedor seja selecionado, terá início a implantação. A boa notícia é que ela pode ser relativamente rápida e durar, digamos, menos de um mês.

À medida que novos projetos de RPA forem desenvolvidos, é possível deparar com

algo chamado fadiga de automatização. Esse é o ponto a partir do qual os retornos em geral começam a diminuir.

Pense da seguinte forma: no início, o foco geralmente se volta para as áreas do negócio que mais precisam de automação, o que significa que o ROI será significativo. Com o tempo, no entanto, haverá um interesse em tarefas que não são tão passíveis de automação e que provavelmente demandarão muito mais trabalho mesmo para realizar pequenas melhorias.

Por conta disso, é uma boa ideia moderar as expectativas ao se envolver em uma transformação RPA generalizada.

Monte uma equipe para gerenciar a plataforma de RPA

Só porque RPA fornece alto grau de automação não significa que requeira pouca gestão. Portanto, a melhor abordagem é montar uma equipe, muitas vezes conhecida como um Centro de Excelência (CoE – Center of Excellence).

A fim de fazer melhor uso dessa equipe, é preciso ser claro sobre as responsabilidades de cada pessoa. Por exemplo, é necessário ser capaz de responder às seguintes perguntas:

- O que acontece se houver um problema com um bot? Em que pontos deve haver intervenção humana?
- Quem é responsável pelo monitoramento da RPA?
- Quem é responsável pelo treinamento?
- Quem terá o papel da primeira linha de suporte?
- Quem é responsável pelo desenvolvimento dos bots?

Para organizações maiores, é possível que se queira expandir os papéis. É possível escolher um responsável pelas mudanças que envolvam RPA, o qual seria o evangelista da plataforma – para toda a empresa. Ou pode haver um gerente de mudanças RPA, que forneça as informações que vão ajudar com a adoção.

Por fim, à medida que a implementação da RPA vai ficando maior, um objetivo principal deve ser olhar para a forma como todas as partes se encaixam. Como muitos outros sistemas de software, há o risco de expansão em toda a organização – o que pode significar não obter um desempenho mais elevado. É aqui que ter um centro de excelência proativo pode causar um grande impacto positivo.

RPA e IA

Embora ainda nas fases iniciais, a IA já está progredindo com as ferramentas de RPA. Isso está levando ao surgimento de bots de software de automação robótica e cognitiva de processos (CRPA – Cognitive Robotic Process Automation).

E isso faz sentido. Afinal, RPA trata da otimização de processos e envolve grandes quantidades de dados. Portanto, os fornecedores estão começando a implementar sistemas como machine learning, deep learning, reconhecimento de fala e natural language processing. Alguns dos líderes no espaço CRPA incluem UiPath, Automation Anywhere, Blue Prism, NICE Systems e Kryon Systems.

Na Automation Anywhere, por exemplo, um bot pode lidar com tarefas como a obtenção de faturas a partir de e-mails, o que envolve processamento de texto sofisticado. A empresa também dispõe de integrações pré-construídas com serviços de IA de terceiros, como IBM Watson, AWS Machine Learning e Google Cloud AI.[11]

"Houve uma proliferação de serviços habilitados para IA nos últimos anos, mas as empresas muitas vezes sofrem para operacionalizá-los", disse Mukund Srigopal, que é Diretor de Marketing de Produtos da Automation Anywhere. "A RPA é uma ótima maneira de inserir recursos de IA nos processos de negócios."[12]

Aqui estão algumas outras maneiras por meio das quais a CRPA pode viabilizar funções de IA:

- É possível conectar chatbots ao seu sistema, o que permitirá o atendimento automatizado ao cliente (cobriremos esse tópico no Capítulo 6).
- A IA pode identificar o momento certo para enviar um e-mail ou alerta.
- IVR (Interactive Voice Response – Unidades de resposta audível) têm alcançado má reputação ao longo dos anos. Simplificando, os clientes não gostam do incômodo de passar por várias etapas para resolver um problema. Com CRPA, entretanto, é possível recorrer a algo chamado Dynamic IVR. Ele personaliza as mensagens de voz para cada cliente, proporcionando uma experiência muito melhor.
- NLP e análise de texto podem converter dados não estruturados em dados estruturados. Isso pode tornar o CRPA mais eficiente.

RPA no mundo real

Para ter melhor noção de como funciona a RPA e entender seus benefícios, observe um estudo de caso da Microsoft.[13] Todos os anos, a empresa paga bilhões de dólares em royalties para desenvolvedores de jogos, parceiros e criadores de conteúdo. No entanto, esse processo era essencialmente manual e envolvia o envio de milhares de declarações – e, sim, a empresa perdia muito tempo nessa atividade.

A empresa então selecionou a Kyron para a implementação de uma solução RPA. Ao fazer uma revisão inicial do processo, a Microsoft percebeu que cerca de 70% a 80% das declarações eram simples e poderiam ser facilmente automatizadas. O restante incluía exceções que exigiam intervenção humana, como aprovações.

Com o sistema de RPA, um algoritmo de detecção visual poderia analisar as declarações e encontrar as exceções. A configuração também foi bastante rápida, levando cerca de 6 semanas.

Como não deve ser nenhuma surpresa, os resultados tiveram um impacto material no processo. Por exemplo, um bot era capaz de levar apenas 2,5 horas para completar 150 declarações de royalties. No processo anterior, funcionários levavam aproximadamente 50 horas na mesma tarefa. Resumo da história: a Microsoft conseguiu uma economia de 2.000%. Houve também a eliminação de qualquer retrabalho causado por erro humano (que antes era de cerca de 5% por mês).

Conclusão

Como visto no estudo de caso da Microsoft, a RPA pode levar a grandes economias. Entretanto, ainda é necessário haver um planejamento diligente, de modo que seus processos sejam compreendidos. De modo geral, o foco deve estar nas tarefas manuais e repetitivas – e não naquelas que dependem fortemente de julgamento. Em seguida, é importante configurar um centro de excelência para supervisionar o gerenciamento contínuo da automação, o que contribuirá para tratamento de exceções, coleta de dados e rastreamento de KPIs.

A RPA também é uma ótima maneira de implementar IA básica dentro de uma empresa. Na verdade, como é possível que haja ROI significativo, isso pode estimular ainda mais investimentos na busca por essa tecnologia.

Principais aprendizados

- A automação robótica de processos (RPA – Robotic Process Automation) permite a utilização de sistemas visuais low-code de drag-and-drop para automatizar o fluxo de trabalho de um processo.
- A RPA autônoma ocorre quando um processo é completamente automatizado.
- A RDA (Robotic Desktop Automation – automação robótica de desktop) é onde a RPA ajuda um funcionário com um trabalho ou tarefa.
- Alguns dos benefícios da RPA incluem maior satisfação do cliente, menores taxas de erro, melhor conformidade e integração mais fácil com sistemas legados.
- Algumas das desvantagens da RPA incluem a dificuldade de se adaptar às mudanças nos aplicativos subjacentes, problemas com aplicativos virtualizados e resistência dos funcionários.
- A RPA tende a funcionar melhor quando é possível automatizar processos repetitivos, estruturados e de rotina, agendamento, entrada/transferência de dados, cumprimento de regras e execução de fluxos de trabalho.

- Ao implementar uma solução de RPA, algumas das etapas a considerar incluem determinação das funções a automatizar, avaliação dos processos, seleção de um fornecedor de RPA e implantação do software, e configuração de uma equipe para gerenciamento da plataforma.
- Um centro de excelência (CoE – Center of Excellence) é uma equipe que gerencia uma implementação de RPA.
- A automação robótica e cognitiva de processos (CRPA – Cognitive Robotic Process Automation) é uma categoria emergente de RPA que se concentra em tecnologias de IA.

---

1 www2.deloitte.com/insights/us/en/focus/signals-for-strategists/cognitive-enterprise-robotic-process-automation.html

2 http://business-review.eu/news/the-story-of-uipath-how-it-became-romanias-first-unicorn-164248

3 www.uipath.com/newsroom/uipath-raises-225-million-series-c-led-by-capitalg-and-sequoia

4 www.grandviewresearch.com/press-release/global-robotic-process-automation-rpa-market

5 https://go.forrester.com/blogs/predictions-2019-automation-will-become-central-to-business-strategy-and-operations/

6 N.T.: O Net Promoter Score (NPS) é uma métrica de lealdade do cliente criada por Fred Heichheld, em 2003, com o objetivo de medir o grau de lealdade dos clientes das empresas de qualquer segmento, trazendo reflexos da experiência e satisfação. Fonte: https://pt.wikipedia.org/wiki/Net_Promoter_Score (adaptado)

7 www.computereconomics.com/article.cfm?id=2633

8 https://federalnewsnetwork.com/artificial-intelligence/2019/03/dod-laying-groundwork-for-multi-generational-effort-on-ai/

9 https://www.cmswire.com/information-management/why-rpa-implementation-projects-fail/

10 https://twitter.com/elonmusk/status/984882630947753984?lang=en

11 www.forbes.com/sites/tomtaulli/2019/02/02/what-you-need-to-know-about-rpa-robotic-process-automation/

12 Entrevista realizada pelo autor com Mukund Srigopal, Diretor de Marketing de Produtos na Automation Anywhere.

13 www.kryonsystems.com/microsoft-case-study/

CAPÍTULO 6

Natural Language Processing (NLP)

Como os computadores conversam

Em 2014, a Microsoft lançou um chatbot – um sistema de IA que se comunica com as pessoas – chamado Xiaoice. Ele foi integrado ao WeChat da Tencent, o maior serviço de mensagens sociais na China. Xiaoice teve um bom desempenho, chegando a 40 milhões de usuários em poucos anos.

À luz do sucesso, a Microsoft resolveu descobrir se era possível fazer algo semelhante no mercado dos Estados Unidos. O Bing e o Grupo de Tecnologia e Pesquisa da empresa alavancaram as tecnologias de IA para construir um novo chatbot: Tay. Os desenvolvedores ainda contaram com a ajuda de comediantes de improviso para tornar a conversa envolvente e divertida.

Em 23 de março de 2016, a Microsoft lançou o Tay no Twitter – e foi um desastre absoluto. O chatbot rapidamente vomitou mensagens racistas e sexistas! Veja apenas um dos milhares de exemplos:

> *@TheBigBrebowski Ricky Gervais aprendeu totalitarismo com Adolf Hitler, o inventor do ateísmo*[1]

Tay era uma ilustração vívida da Lei de Godwin. Ela diz o seguinte: quanto mais uma discussão online continua, maiores são as chances de alguém citar Adolf Hitler ou os nazistas.

Então, sim, a Microsoft tirou o Tay do ar em menos de 24 horas e publicou um pedido de desculpas. Nele, o vice-presidente corporativo da Microsoft Healthcare, Peter Lee, escreveu:

> *No futuro, enfrentaremos alguns desafios difíceis – embora emocionantes – nos projetos de IA. Os sistemas dessa área se alimentam de interações positivas e negativas com as pessoas. Nesse sentido, os desafios são tanto sociais quanto técnicos. Faremos todo o possível para limitar as explorações técnicas, mas também sabemos que não podemos prever plenamente todos os possíveis abusos humanos interativos sem aprender com os erros. Para fazer IA de maneira adequada, é preciso iterar com muitas pessoas e muitas vezes em fóruns públicos. Temos de entrar em cada um com muita cautela e, ao final, aprender e melhorar, passo a passo, e fazer isso sem ofender as pessoas no processo. Continuaremos firmes em nossos esforços para aprender com essa e outras experiências enquanto trabalhamos para contribuir para uma Internet que represente o melhor, não o pior, da humanidade.*[2]

A estratégia do Tay era repetir parte do conteúdo das pessoas fazendo perguntas. Na

maior parte dos casos, essa era uma abordagem válida. Como vimos no Capítulo 1, esse era o coração do primeiro chatbot, ELIZA.

Filtros eficazes devem estar configurados. Isso é especialmente verdadeiro quando um chatbot é usado em uma plataforma de formato livre, como o Twitter (ou, nesse caso, em qualquer cenário do mundo real).

No entanto, falhas como as do Tay são importantes. Elas nos permitem aprender e evoluir a tecnologia.

Neste capítulo, vamos dar uma olhada em chatbots e em NLP (Natural Language Processing – processamento de linguagem natural), que são parte fundamental de como os computadores entendem e manipulam a linguagem. Trata-se de um subconjunto da IA.

Vamos começar.

Desafios do NLP

Como visto no Capítulo 1, a linguagem é a chave para o Teste de Turing, que se destina a validar a IA. A linguagem também é algo que nos diferencia dos animais.

Essa área de estudo, no entanto, é extremamente complexa. Aqui estão apenas alguns dos desafios do NLP:

- A linguagem muitas vezes pode ser ambígua. Aprendemos a falar de forma rápida e acentuamos o significado com pistas não verbais, tom ou reações ao ambiente. Por exemplo, se uma bola de golfe está indo em direção a alguém, você vai gritar "Olha!". Um sistema de NLP, entretanto, provavelmente não entenderia isso porque não é capaz de processar o contexto da situação.
- A linguagem muda com frequência à medida que o mundo muda. De acordo com o Oxford English Dictionary, houve mais de 1.100 palavras, sentidos e subentradas novos em 2018 (ao todo, há mais de 829.000)[3]. Algumas dessas palavras incluem "mansplain" e "hangry".
- Quando falamos, cometemos erros gramaticais. Contudo, isso geralmente não é um problema, pois as pessoas têm uma grande capacidade de inferência. Para o NLP, entretanto, esse é um grande desafio, visto que palavras e frases podem ter múltiplos significados (o que é chamado de polissemia). Por exemplo, o pesquisador de IA Geoffrey Hinton gosta de comparar "recognize speech" ("reconhecer a fala") e "wreck a nice beach" ("destruir uma bela praia").[4]
- A linguagem tem sotaques e dialetos.
- O significado das palavras pode mudar com base no uso de sarcasmo ou outras reações emocionais.
- As palavras podem ser vagas. Afinal, o que realmente significa estar "tarde"?

- Muitas palavras têm essencialmente o mesmo significado, mas envolvem graus de nuances.
- As conversas podem ser não lineares e ter interrupções.

Apesar de tudo isso, tem havido grandes avanços com NLP, como visto em aplicativos como Siri, Alexa e Cortana. Grande parte do progresso também aconteceu na última década, impulsionada pelo poder do deep learning.

Por vezes pode existir confusão com relação a línguas humanas e linguagens de computador. Os computadores foram capazes de entender linguagens como BASIC, C e C++ por anos, não é verdade? Definitivamente. Também é verdade que essas linguagens de programação têm palavras em inglês como if, then, let e print.

Esse tipo de linguagem, entretanto, é muito diferente da linguagem humana. Considere que uma linguagem de computador tem um conjunto limitado de comandos e conta com uma lógica rigorosa. Se algo for usado incorretamente, ocorrerá um bug no código – levando a uma parada. Sim, as linguagens de computador são muito literais!

Entendendo como a IA traduz a linguagem

Como vimos no Capítulo 1, o NLP foi alvo inicial dos pesquisadores de IA. No entanto, por causa do poder limitado do computador, as possibilidades eram completamente fracas. O objetivo era criar regras para interpretar palavras e frases – o que acabou se mostrando complexo e pouco escalável. De certa forma, o NLP nos primeiros anos era essencialmente como uma linguagem de computador!

Com o tempo, contudo, uma estrutura geral foi desenvolvida. Isso foi fundamental, uma vez que o NLP lida com dados não estruturados que podem ser imprevisíveis e difíceis de interpretar.

Aqui está um olhar geral de alto nível sobre as duas etapas principais do NLP:

- Limpeza e pré-processamento do texto: envolve o uso de técnicas como tokenização, estemização e lematização para analisar o texto.
- Compreensão e geração de linguagem: definitivamente a parte mais intensa do processo, que muitas vezes usa algoritmos de deep learning.

Nas próximas seções, vamos discutir essas etapas com mais detalhes.

Etapa #1 – Limpeza e pré-processamento

Três coisas precisam ser feitas durante a etapa de limpeza e pré-processamento: tokenização, estemização e lematização.

Tokenização

Antes que possa existir NLP, o texto deve ser analisado e segmentado em várias partes – um processo conhecido como tokenização. Por exemplo, digamos que temos a seguinte frase: "John comeu quatro cupcakes". Será necessário, então, separar e categorizar cada elemento. A Figura 6.1 ilustra essa tokenização.

Figura 6.1 – Exemplo de tokenização de uma frase.

Até que parece meio fácil, não é? Mais ou menos.

Após a tokenização, será feita a normalização do texto. Isso implicará a conversão de parte do texto para facilitar a análise, tal como alterar a digitação de maiúscula para minúscula, remover a pontuação e eliminar as contrações.

O procedimento, contudo, pode facilmente levar a alguns problemas. Suponha que tenhamos uma frase que tem "I.A.". Devemos nos livrar dos pontos? Caso afirmativo, será que o computador sabe o que "IA" significa?

Provavelmente não.

Curiosamente, mesmo o caso das palavras pode ter um grande impacto sobre o significado. Basta olhar para a diferença entre "fed" e "Fed", em inglês. "Fed" é muitas vezes outro nome para "Federal Reserve" (Reserva Federal); enquanto "fed" é o passado simples e o particípio passado do verbo "feed" (alimentar). Ou, em outro exemplo, vamos supor que temos "us" (nós) e "US" (United States – Estados Unidos). Estamos falando dos Estados Unidos ou de nós?

Aqui estão algumas outras questões:

- *Problema do espaço em branco*: esse é o lugar no qual duas ou mais palavras devem ser consideradas um token, porque formam um substantivo composto. Alguns exemplos incluem "Nova York" e "Vale do Silício".
- *Palavras científicas e sintagmas*: é comum que tais palavras tenham hifens,

parênteses e letras gregas. Se esses caracteres forem retirados, o sistema pode não ser capaz de entender os significados das palavras e dos sintagmas.

- *Texto confuso*: sejamos honestos, muitos documentos têm erros gramaticais e ortográficos.
- *Divisão da sentença*: palavras como "Sr." ou "Sra." podem terminar prematuramente uma frase por causa do ponto.
- *Palavras não importantes*: há palavras que realmente acrescentam pouco ou nenhum significado a uma frase, como "o", "a" e "um". Para removê-los, é possível usar um filtro simples como o Stop Words.

Como é possível perceber, pode ser fácil analisar frases de modo equivocado (e em algumas línguas, como chinês e japonês, as coisas podem ficar ainda mais difíceis por conta da sintaxe). Isso pode ter consequências de longo alcance. Como a tokenização costuma ser a primeira etapa, alguns erros podem se espalhar através do processo inteiro de NLP.

Estemização

A estemização descreve o processo de redução de uma palavra à sua raiz (ou lema), com a remoção de afixos e sufixos. Isso tem sido realmente eficaz para os motores de busca, que envolvem o uso de agrupamento para chegar a resultados mais relevantes. Com a estemização, é possível encontrar mais combinações para palavras com vários significados e até mesmo para lidar com erros ortográficos. Ao usar um aplicativo de IA, esse processo pode ajudar a melhorar a compreensão geral.

A Figura 6.2 mostra um exemplo de estemização.

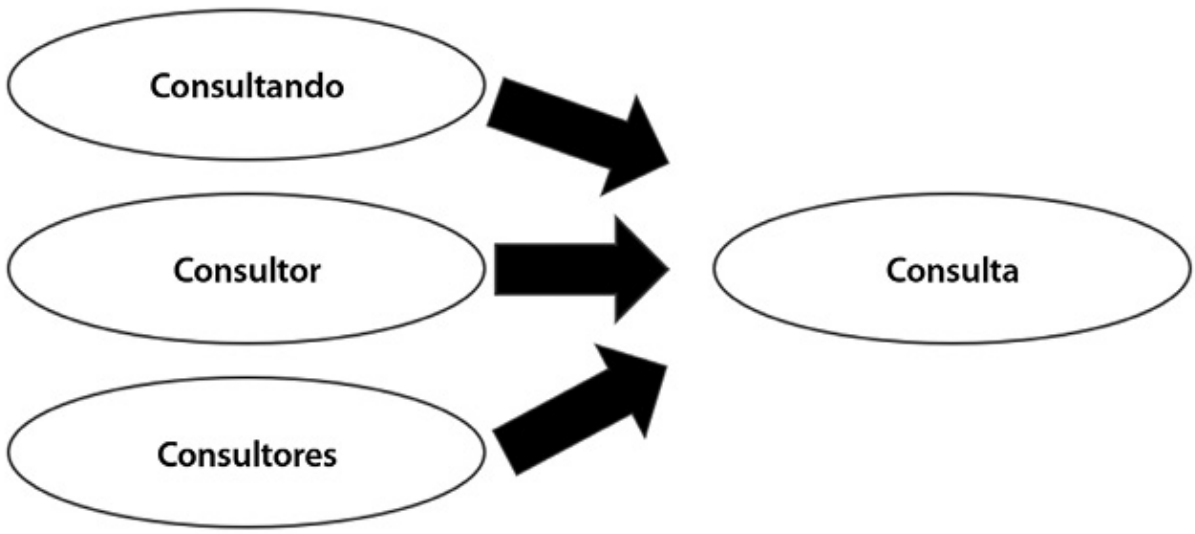

Figura 6.2 – Exemplo de estemização.

Há uma variedade de algoritmos para estemização de palavras, muitos deles bastante simples. Contudo, os resultados apresentados são mistos. De acordo com a IBM:

O algoritmo de Porter, por exemplo, indicará que "universal" tem o mesmo radical que "university" e "universities", uma observação que pode ter base histórica, mas não é mais semanticamente relevante. O estemizador de Porter também não reconhece que as palavras "theater" e "theatre" devem pertencer à mesma classe de radical. Por razões como essas, a Watson Explorer Engine não usa o Porter como seu estemizador para inglês.[5]

Na verdade, a IBM criou o próprio algoritmo proprietário para estemização, o que lhe permite uma personalização significativa.

Lematização

A lematização é semelhante à estemização. No entanto, em vez de remover afixos ou prefixos, há um foco em encontrar palavras de raízes semelhantes. Um exemplo é "better" ("melhor"), que pode estar relacionada a "good" ("bom"). Isso funciona desde que o significado permaneça praticamente o mesmo. Em nosso exemplo, ambos são bastante semelhantes, embora "good" tenha um significado mais claro. A lematização também pode funcionar no fornecimento de melhores pesquisas ou compreensão da linguagem, especialmente com traduções.

A Figura 6.3 mostra um exemplo de lematização.

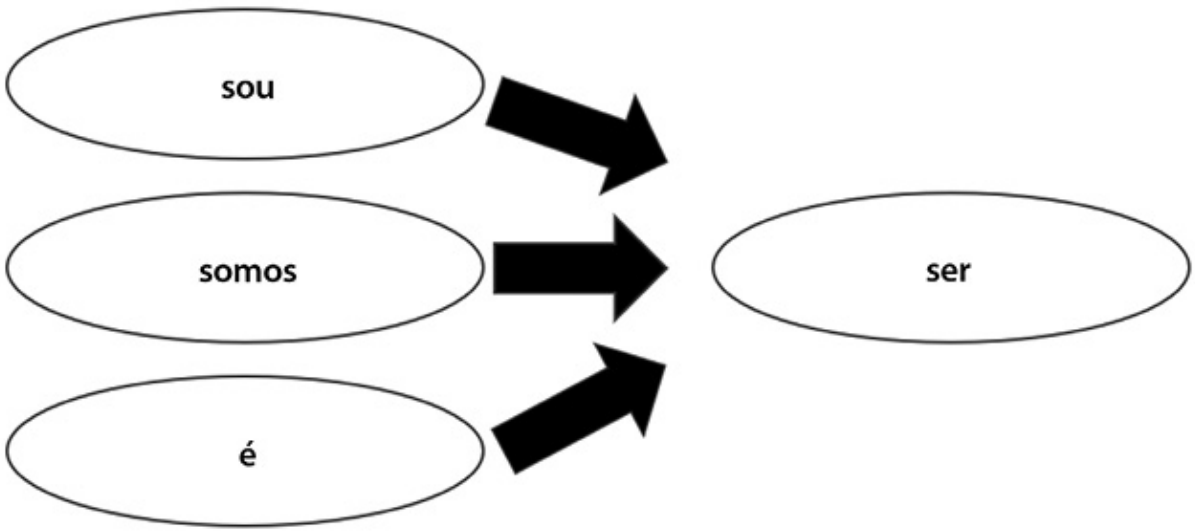

Figura 6.3 – Exemplo de lematização.

Para usar efetivamente a lematização, o sistema de NLP deve entender os significados das palavras e do contexto. Ou seja, esse processo em geral tem melhor desempenho do que a estemização. Por outro lado, isso também significa que os algoritmos são mais complicados e as exigências de poder computacional são mais elevadas.

Etapa #2 – Compreensão e geração de linguagem

Uma vez que o texto tenha sido colocado em um formato que os computadores conseguem processar, o sistema de NLP deve então compreender o significado geral.

Normalmente, essa é a parte mais difícil.

Ao longo dos anos, entretanto, os pesquisadores desenvolveram uma variedade de técnicas para ajudar, como as seguintes:

- *Marcação de Partes do Discurso (POS – Parts Of Speech)*: esse processo perpassa pelo texto e identifica cada palavra em sua categoria gramatical adequada, como substantivos, verbos, advérbios *etc.* Pense nisso como uma versão automatizada da sua aula de português da escola! Além disso, alguns sistemas para POS apresentam variações. Lembre-se de que um substantivo pode ser substantivo singular (NN – singular nouns), substantivo próprio no singular (NNP – singular proper nouns) ou substantivo plural (NNS – plural nouns).
- *Chunking*: as palavras serão analisadas em termos de sintagmas. Por exemplo, um sintagma substantivo (NP) é um substantivo que atua como sujeito ou objeto de um verbo.
- *Reconhecimento de entidades nomeadas*: trata da identificação de palavras que representam locais, pessoas e organizações.
- *Modelagem de tópicos*: procura padrões e clusters ocultos no texto. Um dos algoritmos, chamado Latent Dirichlet Allocation (LDA – Alocação Latente de Dirichlet), é baseado em abordagens de aprendizagem não supervisionada. Ou seja, haverá tópicos aleatórios atribuídos e o computador vai iterar para encontrar correspondências.

É possível usar modelos de deep learning em muitos desses processos. Eles podem ser estendidos para mais áreas de análise – como permitir compreensão e geração contínua da linguagem. Esse é um processo conhecido como semântica distribucional.

Com uma rede neural convolucional (CNN), sobre a qual falamos no Capítulo 4, é possível encontrar grupos de palavras que são traduzidas e transformadas em um mapa de recursos. Isso viabiliza aplicações como tradução de idiomas, reconhecimento de fala, análise de sentimentos e perguntas e respostas. Na verdade, o modelo pode até fazer coisas como detectar sarcasmo!

No entanto, existem alguns problemas com as CNNs. Por exemplo, o modelo tem dificuldades com texto que apresenta dependência entre palavras muito distantes. Contudo, existem algumas maneiras de lidar com isso, como recorrer às redes neurais com retrocesso temporal (TDNN – Time-Delayed Neural Networks) e redes neurais convolucionais dinâmicas (DCNN – Dynamic Convolutional Neural Networks). Esses métodos mostraram alto desempenho no processamento de dados sequenciados. O modelo que tem mostrado maior sucesso com essa abordagem é a rede neural recorrente (RNN), uma vez que ela memoriza dados.

Até agora, temos nos concentrado, principalmente, na análise de texto. Entretanto,

para que haja um NLP sofisticado, também é preciso construir sistemas de reconhecimento de voz. Vamos dar uma olhada nisso na próxima seção.

Reconhecimento de voz

Em 1952, a Bell Labs criou o primeiro sistema de reconhecimento de voz, chamado Audrey (para reconhecimento automático de dígitos). Ele era capaz de reconhecer fonemas, que são as unidades mais básicas de sons em uma língua. O inglês, por exemplo, tem 44 deles.

Audrey poderia reconhecer o som de um dígito de zero a nove. Era preciso para a voz do criador da máquina, HK Davis, em aproximadamente 90% do tempo.[6] Para qualquer outra pessoa, apresentava precisão em 70% a 80% das vezes ou mais.[7]

Audrey foi um grande feito, especialmente à luz dos limitados poder computacional e memória disponíveis na época. O programa também destacou os principais desafios com o reconhecimento de voz. Quando falamos, nossas frases podem ser complexas e um pouco confusas. Geralmente também falamos rápido – a uma velocidade média de 150 palavras por minuto.

Como resultado, os sistemas de reconhecimento de voz foram aprimorados a um ritmo espantosamente lento. Em 1962, o sistema Shoebox da IBM podia reconhecer apenas 16 palavras, 10 dígitos e 6 comandos matemáticos.

Somente na década de 1980 houve um progresso significativo na tecnologia. O principal avanço foi o uso do modelo oculto de Markov (HMM – Hidden Markov Model), baseado em estatísticas sofisticadas. Por exemplo, se fosse dita a palavra "dog", haveria uma análise dos sons individuais d, o e g, e o algoritmo HMM atribuiria uma pontuação a cada um deles. Ao longo do tempo, o sistema ia ficando melhor na compreensão dos sons e conseguia traduzi-los em palavras.

Embora HMM tenha sido importante, ele ainda era incapaz de lidar eficazmente com a fala contínua. Por exemplo, os sistemas de voz eram baseados na correspondência de modelos. Isso envolvia a tradução de ondas sonoras em números, o que era feito por amostragem. O resultado foi que o software media a frequência dos intervalos e armazenava os resultados, mas era preciso haver uma combinação próxima. Por conta disso, a entrada de voz tinha de ser bastante clara e lenta. Também era necessário que houvesse pouco ruído de fundo.

Na década de 1990, contudo, os desenvolvedores de software fizeram progressos e lançaram sistemas comerciais – como o Dragon Dictate, que era capaz de compreender milhares de palavras em fala contínua. No entanto, sua adoção ainda não era significativa e muitas pessoas continuavam achando mais fácil digitar em seus computadores e usar o mouse. Contudo, havia algumas profissões, como a medicina, em que o reconhecimento de fala encontrou altos níveis de uso (em

especial em um estudo de caso popular com a transcrição do diagnóstico de pacientes).

Com o surgimento do machine learning e do deep learning, os sistemas de voz rapidamente se tornaram muito mais sofisticados e precisos. Alguns dos algoritmos-chave envolvem o uso de memória longa de curto prazo (LSTM), redes neurais recorrentes e redes neurais profundas avançadas. O Google continuou a implementar essas abordagens no Google Voice, que estava disponível para centenas de milhões de usuários de smartphones. E, claro, temos presenciado um grande progresso em outras ofertas, como Siri, Alexa e Cortana.

NLP no mundo real

Até o momento, analisamos as principais partes do fluxo de trabalho do NLP. Em seguida, vamos dar uma olhada nas poderosas aplicações dessa tecnologia.

Estudo de caso: Melhoria nas vendas

Roy Raanani, que tem uma carreira no trabalho com startups de tecnologia, imaginou que as inúmeras conversões que ocorrem todos os dias nos negócios são, em grande parte, ignoradas. Talvez a IA possa transformar isso em uma oportunidade?

Em 2015, ele fundou a Chorus para usar o NLP para prever conversas de vendedores. Raanani chamou o projeto de Nuvem de Conversação (Conversation Cloud), que registra, organiza e transcreve chamadas – inseridas em um sistema de CRM (gestão do relacionamento com o cliente). Com o tempo, os algoritmos começarão a aprender sobre as melhores práticas e indicarão como as coisas podem ser melhoradas.

Conseguir isso, no entanto, não tem sido fácil. De acordo com um blog da Chorus:

> *Há bilhões de maneiras de fazer perguntas, levantar objeções, definir itens de ação, desafiar hipóteses etc., e todas elas precisam ser identificadas se os padrões de vendas forem codificados. Em segundo lugar, os sinais e padrões evoluem: novos concorrentes, nomes e recursos de produtos e terminologia relacionada ao setor mudam ao longo do tempo e modelos aprendidos por máquina rapidamente se tornam obsoletos.*[8]

Por exemplo, uma das dificuldades – que pode ser facilmente negligenciada – é como identificar as partes que estão conversando (muitas vezes há mais de três em uma chamada). O problema, conhecido como "separação de falantes", é considerado ainda mais difícil do que o reconhecimento de fala. A Chorus criou um modelo de deep learning que cria uma "impressão digital de voz" – baseada no agrupamento – para cada falante. Assim, depois de vários anos de pesquisa e desenvolvimento, a

empresa conseguiu desenvolver um sistema capaz de analisar grandes quantidades de conversas.

Como prova disso, veja o exemplo de um dos clientes da Chorus, o Housecall Pro, uma startup que vende tecnologias móveis para gerenciamento de serviços de campo. Antes de adotar o software, a empresa muitas vezes criava discursos de vendas personalizados para cada liderança. Infelizmente, contudo, esse procedimento não era escalável e apresentava resultados mistos.

Com a Chorus, a empresa foi capaz de criar uma abordagem sem muita variação. O software permitiu avaliar cada palavra e medir seu impacto nas conversões de vendas. A Chorus também verificava se um representante de vendas seguia ou não o roteiro.

O resultado? A empresa conseguiu aumentar a taxa de sucesso de vendas em 10%.[9]

Estudo de caso: Combate à depressão

Em todo o mundo, cerca de 300 milhões de pessoas sofrem de depressão, de acordo com dados da Organização Mundial da Saúde.[10] E mais, cerca de 15% dos adultos vão experimentar algum tipo de depressão durante a sua vida.

A doença pode não ser diagnosticada por causa da falta de serviços de saúde, o que talvez signifique que a situação de uma pessoa pode ficar muito pior. Infelizmente, a depressão é capaz de levar a outros problemas.

Com NLP, entretanto, a situação pode melhorar. Um estudo recente de Stanford usou um modelo de machine learning que processou expressões faciais 3D e linguagem falada. Como resultado, o sistema foi capaz de diagnosticar a depressão com uma taxa média de erro de 3,67 ao usar a escala do Questionário de Saúde do Paciente (PHQ – Patient Health Questionnaire). A precisão foi ainda maior para formas mais agravadas de depressão.

No estudo, os pesquisadores observaram: "Essa tecnologia poderia ser implantada em telefones celulares em todo o mundo e facilitar o acesso universal de baixo custo aos cuidados de saúde mental".[11]

Estudo de caso: Criação de conteúdo

Em 2015, vários veteranos da tecnologia – como Elon Musk, Peter Thiel, Reid Hoffman e Sam Altman – fundaram a OpenAI com o apoio de um colossal financiamento de US$ 1 bilhão. Estruturada como uma organização sem fins lucrativos, o objetivo da empresa era "avançar a inteligência digital da maneira mais propícia para beneficiar a humanidade como um todo, sem restrições com base na necessidade de gerar retorno financeiro".[12]

Uma das áreas de pesquisa tem sido o NLP. Para esse fim, a empresa lançou um

modelo chamado GPT-2, em 2019, que foi baseado em um conjunto de dados de cerca de oito milhões de páginas web. O foco era criar um sistema que pudesse prever a próxima palavra com base em um grupo de texto.

Para demonstrar seu modelo, a OpenAI conduziu um experimento com o seguinte texto como entrada: "Em uma descoberta chocante, cientistas descobriram uma manada de unicórnios vivendo em um vale remoto, até então inexplorado, na Cordilheira dos Andes. Ainda mais surpreendente para os pesquisadores foi o fato de que os unicórnios falavam inglês perfeitamente".

A partir disso, os algoritmos criaram uma história convincente com 377 palavras de comprimento!

Por fim, os pesquisadores admitiram que a narrativa era melhor com temas mais diretamente relacionados aos dados subjacentes, como O Senhor dos Anéis e até mesmo o Brexit. Como não deve ser nenhuma surpresa, o GPT-2 demonstrou mau desempenho em domínios técnicos.

O modelo, entretanto, foi capaz de alcançar altas pontuações em várias avaliações bem conhecidas de compreensão da leitura. Veja a Tabela 6.1.[13]

Tabela 6.1 – Resultados de compreensão de leitura

Conjunto de dados	Registro anterior de precisão	Precisão do GPT-2
Winograd Schema Challenge	63,7%	70,70%
LAMBADA	59,23%	63,24%
Substantivos comuns do teste do livro infantil	85,7%	93,30%
Entidades nomeadas para o teste do livro infantil	82,3%	89,05%

Mesmo que um ser humano típico seja capaz de marcar 90% ou mais nesses testes, o desempenho do GPT-2 ainda é impressionante. É importante observar que o modelo usou aprendizagem não supervisionada e a inovadora rede neural do Google, chamada Transformer.

Para manter o alinhamento às propostas da OpenAI, a organização decidiu não lançar o modelo completo. O medo era que ele poderia levar a consequências adversas, como notícias falsificadas, avaliações de produtos adulteradas na Amazon.com, spam e golpes de phishing.[14]

Estudo de caso: Linguagem corporal

Concentrar-se apenas na linguagem em si pode ser limitante. A linguagem corporal é algo que também deve ser incluído em um sofisticado modelo de IA.

E é nisso que Rana el Kaliouby vem pensando há algum tempo. Enquanto crescia no Egito, ela obteve seu mestrado em Ciência pela Universidade Americana no Cairo e, em seguida, iniciou seu doutorado em Ciência da Computação no Newnham College da Universidade de Cambridge. Havia algo muito atraente para ela: como os computadores conseguem detectar emoções humanas?

No entanto, em seus círculos acadêmicos, havia pouco interesse no assunto. A visão consensual na comunidade de Ciência da Computação era a de que esse tema realmente não tinha utilidade.

Rana, contudo, não se intimidou e se uniu à notável professora Rosalind Picard para criar modelos inovadores de machine learning (ela escreveu um livro fundamental, chamado *Affective Computing* (Computação Afetiva), no qual analisou emoções e máquinas).[15] A área, no entanto, precisava recorrer a outros domínios, como neurociência e psicologia. Grande parte da tarefa foi alavancar o trabalho pioneiro de Paul Ekman, que fez uma extensa pesquisa sobre as emoções humanas com base nos músculos faciais de uma pessoa. Ele descobriu que havia seis emoções humanas universais (ira, grosseria, medo, alegria, solidão e choque) que poderiam ser codificadas por 46 movimentos chamados unidades de ação – todas se tornaram parte do Sistema de Codificação da Ação Facial ou FACS (Facial Action Coding System).

Enquanto esteve no MIT Media Lab, Rana desenvolveu um "aparelho auditivo emocional", um wearable que permitia às pessoas com autismo interagir melhor em ambientes sociais.[16] O sistema detectava as emoções das pessoas e sugeria maneiras apropriadas de reagir.

Foi inovador e o New York Times reconheceu o invento como uma das inovações mais importantes de 2006. O sistema de Rana também chamou a atenção de várias empresas. Simplificando, a tecnologia poderia ser uma ferramenta eficaz para avaliar o humor de um público diante de determinado comercial de televisão.

Então, alguns anos depois, Rana lançou a Affectiva. A empresa cresceu rapidamente e atraiu quantidades substanciais de capital de risco (ao todo, arrecadou US$ 54,2 milhões).

Rana, que antes havia sido ignorada, agora se tornara uma das líderes em uma tendência chamada "IA para rastreamento de emoções".

O principal produto da Affectiva é o Affdex, uma plataforma baseada na nuvem para testar o público diante de vídeos. Cerca de um quarto das empresas da Fortune Global 500 usam a inovação.

A empresa também desenvolveu outro produto, o Affectiva Automotive AI, um sistema de detecção para o interior de um veículo. Alguns de seus recursos incluem:

- Monitoramento de fadiga ou distração do motorista, o que desencadeará um

alerta (por exemplo, uma vibração do cinto de segurança).

- Possibilidade de transferência de controle a um sistema semiautônomo caso o motorista não esteja acordado ou quando ele estiver com raiva. Há inclusive a possibilidade de sugerir rotas alternativas para diminuir o potencial de raiva na estrada!
- Personalização do conteúdo – como música – com base nas emoções do passageiro.

Para todas essas ofertas, existem sistemas avançados de deep learning que processam enormes quantidades de recursos de um banco de dados que conta com mais de 7,5 milhões de faces. Esses modelos também dão conta de influências culturais e diferenças demográficas – o que é feito em tempo real.

Comércio de voz

Tecnologias orientadas por NLP, como assistentes virtuais, chatbots e alto-falantes inteligentes, estão prontas para atender a modelos de negócios poderosos – e podem até perturbar mercados como comércio eletrônico e marketing. Uma versão inicial disso já foi vista com a franquia WeChat, da Tencent. A empresa, fundada durante o auge do boom da Internet no final dos anos 1990, começou com um mensageiro simples baseado em PC chamado OICQ. A introdução do WeChat, no entanto, foi um divisor de águas, e desde então a ferramenta se tornou a maior plataforma de mídia social da China, com mais de 1 bilhão de usuários ativos mensalmente.[17]

Esse aplicativo, no entanto, vai além de troca de mensagens e postagem de conteúdo. O WeChat rapidamente se transformou em um assistente virtual para todos os fins, no qual se pode facilmente contactar um serviço de compartilhamento de caronas, efetuar um pagamento em um varejista local, fazer uma reserva para um voo ou jogar. O aplicativo responde por cerca de 35% de todo o tempo de uso de smartphones na China em uma base mensal. O WeChat também é uma das principais razões para o país estar se tornando cada vez mais uma sociedade sem dinheiro.

Tudo isso aponta para o poder de uma categoria emergente chamada comércio de voz (ou v-commerce), onde é possível fazer compras via bate-papo ou voz. É uma tendência tão crítica que, no início de 2019, Mark Zuckerberg, do Facebook, escreveu uma postagem em um blog[18] na qual dizia que a empresa se tornaria mais parecida com o... WeChat!

De acordo com uma pesquisa da Juniper, o mercado para o comércio de voz deverá arrecadar US$ 80 bilhões até 2023.[19] Em termos de vencedores, entretanto, parece uma boa opção apostar nas empresas que dispõem de grandes bases de instalação de dispositivos inteligentes – como Amazon, Apple e Google. Ainda haverá espaço,

entretanto, para provedores de tecnologias de NLP de última geração.

Ok, então, como esses sistemas de IA podem afetar a indústria de marketing? Bem, para se ter uma ideia, a Harvard Business Review publicou um artigo chamado "Marketing in the Age of Alexa" ("Marketing na era da Alexa"), de Niraj Dawar e Neil Bendle. Nele, os autores observam que "os assistentes de IA transformarão a forma como as empresas se conectam com seus clientes. Eles se tornarão o principal canal por meio do qual as pessoas vão obter informações, bens e serviços, e o marketing terá de brigar por atenção".[20]

Assim sendo, o crescimento de chatbots, assistentes digitais e alto-falantes inteligentes pode ser muito maior do que a revolução inicial do comércio eletrônico baseado na web. Essas tecnologias trazem benefícios significativos para os clientes, como conveniência. É fácil dizer a um dispositivo que compre algo, e a máquina também aprenderá sobre seus hábitos. Então, da próxima vez em que você disser que quer tomar um refrigerante, o computador vai saber a que você está se referindo.

Isso pode levar, no entanto, a um cenário de tudo ou nada. Em última instância, parece que os consumidores vão usar apenas um dispositivo inteligente para suas compras. Além disso, para as marcas que querem vender seus produtos, haverá a necessidade de entender profundamente o que os clientes realmente querem, de modo a se tornar o fornecedor preferido dentro do mecanismo de recomendação.

Assistentes virtuais

Em 2003, quando os Estados Unidos estavam envolvidos em guerras no Oriente Médio, o Departamento de Defesa procurava investir em tecnologias de última geração para o campo de batalha. Uma das principais iniciativas foi construir um sofisticado assistente virtual capaz de reconhecer instruções faladas. O Departamento de Defesa dedicou US$ 150 milhões para o projeto e encarregou o SRI Lab (Stanford Research Institute Lab – Laboratório do Instituto de Pesquisa de Stanford) – com sede no Vale do Silício – de desenvolver a aplicação.[21] Embora o laboratório fosse uma organização sem fins lucrativos, ainda era permitido licenciar suas tecnologias (como a impressora a jato de tinta) para startups.

E foi isso que aconteceu com o assistente virtual. Alguns dos membros do SRI – entre eles, Dag Kittlaus, Tom Gruber e Adam Cheyer – lhe deram o nome de Siri e começaram sua própria empresa para aproveitar a oportunidade. Eles deram início à operação em 2007, quando o iPhone da Apple foi lançado.

Contudo, era necessário que outras atividades de Pesquisa e Desenvolvimento fossem conduzidas para levar o produto a ponto de ser útil para os consumidores. Os fundadores precisaram desenvolver um sistema para lidar com dados em tempo

real, construir um mecanismo de busca de informações geográficas e implementar segurança para cartões de crédito e dados pessoais. O NLP, no entanto, foi o desafio mais difícil.

Em uma entrevista, Cheyer observou:

O desafio técnico mais difícil com a Siri foi lidar com a enorme quantidade de ambiguidade presente na linguagem humana. Considere a frase "book 4-star restaurant in Boston" ("reserve um restaurante 4 estrelas em Boston") – parece muito simples de compreender. Nosso protótipo de sistema poderia lidar com isso facilmente. No entanto, quando carregamos dezenas de milhões de nomes de negócios e centenas de milhares de cidades para o sistema como vocabulário (quase cada palavra na língua inglesa é um nome comercial), o número de interpretações de candidatos aumentou muitíssimo.[22]

Apesar de tudo, a equipe foi capaz de resolver os problemas e transformar a Siri em um sistema poderoso, que foi lançado na App Store da Apple em fevereiro de 2010. "É o reconhecimento de voz mais sofisticado a ser disponibilizado em um smartphone", de acordo com uma avaliação da Wired.com.[23]

Steve Jobs tomou conhecimento do projeto e chamou os pesquisadores. Dentro de poucos dias, eles se reuniram e as discussões rapidamente levaram a uma aquisição que aconteceu no final de abril, por mais de US$ 200 milhões.

Jobs, no entanto, achou que a Siri precisava de melhorias. Por conta disso, houve um relançamento em 2011 – que aconteceu um dia antes de Jobs falecer.

Avançando rapidamente para os dias atuais, a Siri está na melhor posição de participação entre os assistentes virtuais, com 48,6% do mercado. O Google Assistente tem 28,7% e o Alexa da Amazon.com tem 13,2%.[24]

De acordo com o "Voice Assistant Consumer Adoption Report" (Relatório de adoção de assistentes de voz pelo consumidor), cerca de 146,6 milhões de pessoas nos Estados Unidos já experimentaram assistentes virtuais em seus smartphones, sendo mais de 50 milhões com alto-falantes inteligentes. Isso, entretanto, só cobre parte da história. A tecnologia de voz também está sendo incorporada a wearables, fones de ouvido e eletrodomésticos.[25]

Aqui estão outras descobertas interessantes:

- Usar voz para procurar produtos superou as pesquisas para várias opções de entretenimento.
- Quando se trata de produtividade, os usos mais comuns para voz incluem fazer chamadas, enviar e-mails e definir alarmes.
- O uso mais comum de voz em smartphones ocorre quando uma pessoa está dirigindo.
- Em relação às reclamações relacionadas aos assistentes de voz em smartphones, o item com maior percentual foi a inconsistência na compreensão das solicitações. Mais uma vez, isso aponta para os desafios contínuos do NLP.

O potencial de crescimento para assistentes virtuais permanece amplo e essa área provavelmente será essencial para a indústria de IA. A Juniper Research prevê que o

número de assistentes virtuais em uso globalmente mais do que triplicará para 2,5 bilhões até 2023.[26] Espera-se que a categoria com desenvolvimento mais rápido seja a das smart TVs. Sim, acho que vamos começar a conversar com esses dispositivos!

Chatbots

Muitas vezes há confusão com relação às diferenças entre assistentes virtuais e chatbots. Tenha em mente que há muita sobreposição entre os dois. Ambos usam o NLP para interpretar a linguagem e executar tarefas.

Contudo, os dois ainda guardam distinções críticas. Em grande parte, os chatbots são desenvolvidos principalmente para empresas, servindo para o suporte ao cliente ou atuando em funções de vendas. Os assistentes virtuais, por outro lado, são direcionados praticamente a todas as pessoas para ajudar em suas atividades diárias.

Como vimos no Capítulo 1, as origens dos chatbots remontam à década de 1960 com o desenvolvimento da ELIZA. Somente na última década, entretanto, que a tecnologia se tornou utilizável em escala.

Aqui está uma lista de chatbots interessantes:

- *Ushur*: é integrado aos sistemas corporativos das companhias de seguros, permitindo automação de reclamações, processamento de contas e habilitação de vendas. O software mostrou, em média, uma redução de 30% nos volumes de chamadas de um centro de atendimento e uma taxa de resposta ao cliente de 90%.[27] A empresa construiu o próprio motor linguístico de última geração, chamado LISA (que significa Language Intelligence Services Architecture – Arquitetura de Serviços de Inteligência de Linguagem). LISA inclui NLP, NLU (Natural Language Understanding – Compreensão da linguagem natural), análise de sentimento, detecção de sarcasmo, detecção de tópicos, extração de dados e tradução de idiomas. A tecnologia atualmente suporta 60 línguas, o que faz dela uma plataforma útil para organizações globais.
- *Mya*: é um chatbot que pode se envolver em conversas durante o processo de recrutamento. Como o Ushur, também é baseado em uma tecnologia própria de NLP. Algumas das razões para isso incluem ter melhores comunicações, mas também lidar com tópicos específicos para contratação.[28] Mya reduz muito o tempo para entrevista e contratação, eliminando gargalos significativos.
- *Jane.ai*: é uma plataforma que minera dados em aplicativos e bancos de dados de uma organização – como Salesforce.com, Office, Slack e Gmail – a fim de tornar muito mais fácil a obtenção de respostas que não são personalizadas. Observe que cerca de 35% do tempo de um funcionário é gasto tentando encontrar informações! Por exemplo, uma das organizações que utiliza a Jane.ai é a USA Mortgage. A empresa usou a tecnologia, integrada ao Slack, para ajudar os

corretores a procurar informações para processamento de hipotecas. O resultado é que a empresa economizou aproximadamente 1.000 horas de trabalho humano por mês.[29]

Apesar de tudo isso, os chatbots ainda apresentam resultados mistos. Por exemplo, um dos problemas é que é difícil programar sistemas para domínios especializados.

Veja o estudo da UserTesting, baseado nas respostas de 500 consumidores de chatbots de saúde. Algumas das principais conclusões a que chegaram é que ainda há muita ansiedade com essa tecnologia, especialmente ao lidar com informações pessoais, e ela também tem problemas para lidar com a compreensão de tópicos complexos.[30]

Então, antes de implantar um chatbot, há alguns fatores a considerar:

- *Definir expectativas*: não prometa demais a partir das capacidades dos chatbots. Isso só vai preparar a sua organização para a decepção. Por exemplo, não finja que o chatbot é um ser humano. Essa é uma maneira infalível de criar experiências ruins. Portanto, é possível que seja boa ideia iniciar a conversa do chatbot com "Oi, eu sou um chatbot e estou aqui para ajudá-lo com...".
- *Automação*: em alguns casos, um chatbot pode cuidar de todo o processo com um cliente. Contudo, ainda será necessário envolver pessoas na atividade. "O objetivo dos chatbots não é substituir totalmente os humanos, mas ser a primeira linha de defesa, por assim dizer", comentou Antonio Cangiano, evangelista de IA na IBM. "Isso pode significar não apenas poupar dinheiro das empresas, mas também liberar agentes humanos que poderão passar mais tempo em investigações complexas."[31]
- *Desgaste*: tanto quanto possível, tente encontrar maneiras para o chatbot resolver problemas o mais rápido possível. E talvez não seja por meio de conversa necessariamente. Em vez disso, fornecer um formulário simples para preenchimento pode ser uma alternativa mais adequada, por exemplo, para agendar uma demonstração.
- *Processos repetitivos*: costumam ser ideais para chatbots. Os exemplos incluem autenticação, status de pedido, agendamento e solicitações de alteração simples.
- *Centralização*: certifique-se de integrar os dados com seus chatbots. Isso permitirá experiências mais perfeitas. Sem dúvida, os clientes ficam rapidamente irritados se tiverem de repetir informações.
- *Personalize a experiência*: não é um processo fácil, mas pode render grandes benefícios. Jonathan Taylor, Chief Technology Officer (CTO – Diretor de Tecnologia) da Zoovu, dá o seguinte exemplo: "Comprar uma lente de câmera será diferente para cada consumidor. Há muitas variações de lentes que talvez um cliente mais informado compreenda – o comprador médio, entretanto, pode

não dispor de tais informações. Fornecer um chatbot assistivo para conduzir o cliente até a lente certa pode ajudar a fornecer o mesmo nível de atendimento ao cliente que um funcionário da loja faria. O chatbot assistivo pode fazer as perguntas certas, como 'que tipo de câmera você já tem?', 'por que você está comprando uma nova câmera?' e 'o que você está tentando registrar em suas fotografias?'. As respostas ajudariam a descobrir o objetivo do cliente de modo a fornecer uma recomendação personalizada do produto."[32]

- *Análise de dados*: é fundamental monitorar o feedback com um chatbot. Qual é a satisfação? Qual é a taxa de precisão?
- *Design conversacional e experiência do usuário (UX)*: são diferentes daquelas utilizadas na criação de um site ou aplicativo móvel. Com um chatbot, é preciso pensar sobre a personalidade do usuário, sexo e até mesmo contexto cultural. Além disso, é necessário considerar a "voz" de sua empresa. "Em vez de criar mockups para uma interface visual, pense em escrever scripts e experimentá-los antes de construir as interações", disse Gillian McCann, chefe de Engenharia na Nuvem e Inteligência Artificial da Workgrid Software.[33]

Mesmo com problemas nos chatbots, a tecnologia continua a melhorar. Vale ressaltar que é provável que esses sistemas se tornem uma parte cada vez mais importante da indústria de IA. De acordo com a IDC, cerca de US$ 4,5 bilhões serão gastos em chatbots em 2019 – que contribuem para o total de US$ 35,8 bilhões estimados para sistemas de IA.[34]

Outro dado importante é que um estudo da Juniper Research indica que a economia de custos proporcionada pelos chatbots provavelmente será substancial. A empresa prevê que chegará a US$ 7,3 bilhões até 2023, superando os US$ 209 milhões de 2019.[35]

Futuro do NLP

Em 1947 nascia Boris Katz na Moldávia, que até então fazia parte da antiga União Soviética. Ele se formou na Universidade de Moscou, onde aprendeu sobre computadores, e deixou o país rumo aos Estados Unidos (com a ajuda do senador Edward Kennedy).

Boris soube aproveitar a oportunidade. Além de escrever mais de 80 publicações técnicas e receber duas patentes dos Estados Unidos, criou o sistema START, que permitiu avanços sofisticados no NLP. Na verdade, o projeto foi a base para o primeiro site de Perguntas e Respostas na Web em 1993. Sim, ele foi o precursor de empresas inovadoras como Yahoo! e Google.

As criações de Boris também foram críticas para o Watson da IBM, atualmente no centro dos esforços de IA da empresa. Em 2011, esse computador chocou o mundo

quando venceu dois dos maiores campeões de todos os tempos do popular game show Jeopardy!.

Apesar de todo o progresso com o NLP, Boris não está satisfeito. Ele acredita que ainda estamos nos estágios iniciais de desenvolvimento e muito mais deve ser feito para obter o verdadeiro valor. Em entrevista ao MIT Technology Review, comentou: "Esses programas [como Siri e Alexa] são tão incrivelmente estúpidos. Portanto, há um sentimento de estar orgulhoso e, ao mesmo tempo, quase envergonhado. Você lança algo que as pessoas sentem que é inteligente, mas que não está nem perto disso".[36]

A declaração não insinua que Boris é pessimista. No entanto, ele acredita que é preciso repensar o NLP para que se chegue ao ponto da "inteligência real". Para esse fim, considera que os pesquisadores devem olhar além da ciência da computação pura para áreas mais amplas como neurociência, ciência cognitiva e psicologia. Ele também acha que os sistemas de NLP devem fazer um trabalho muito melhor na compreensão das ações do mundo real.

Conclusão

Para muitas pessoas, a primeira interação com NLP se dá por intermédio de assistentes virtuais. A tecnologia ainda é bastante útil mesmo quando está longe de ser perfeita – em especial para responder a perguntas ou obter informações sobre um restaurante próximo, por exemplo.

No entanto, o NLP também está causando grande impacto no mundo dos negócios. Nos próximos anos, a tecnologia se tornará cada vez mais importante para o comércio eletrônico e o atendimento ao cliente – proporcionando economias de custos significativas e permitindo que os funcionários se concentrem em atividades de maior valor agregado.

É verdade que ainda há um longo caminho a percorrer por causa das complexidades da linguagem. Contudo, o progresso continua a ser rápido, em especial por conta da ajuda de abordagens de IA de última geração, como o deep learning.

Principais aprendizados

- O Processamento de Linguagem Natural (NLP – Natural Language Processing) é o uso da IA para permitir que os computadores entendam as pessoas.
- Um chatbot é um sistema de IA que se comunica com as pessoas, seja por voz ou bate-papo online.
- Embora tenha havido grandes avanços na área de NLP, ainda há muito trabalho a ser feito. Alguns dos desafios incluem ambiguidade da linguagem, pistas não verbais, dialetos e sotaques diferentes e mudanças na linguagem.

- As duas principais etapas da NLP incluem limpeza e pré-processamento do texto e uso da IA para entender e gerar linguagem.
- Tokenização é o processo em que o texto é analisado e segmentado em várias partes.
- Com a normalização, o texto é convertido em uma forma que facilita a análise após ser submetido a procedimentos como a remoção de pontuação ou contrações.
- Estemização descreve o processo de redução de uma palavra à sua raiz (ou lema), por meio da remoção de afixos e sufixos.
- Semelhante à estemização, a lematização envolve encontrar palavras de raízes semelhantes.
- Para que o NLP compreenda a linguagem, há uma variedade de abordagens como marcação de partes do discurso (colocando o texto na forma gramatical), agrupamento (processamento de texto em sintagmas) e modelagem de tópicos (encontrando padrões e clusters ocultos).
- Um fonema é a unidade de som mais básica de um idioma.

---

1 www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

2 https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

3 https://wordcounter.io/blog/newest-words-added-to-the-dictionary-in-2018/

4 www.deepinstinct.com/2019/04/16/applications-of-deep-learning/

5 www.ibm.com/support/knowledgecenter/SS8NLW_11.0.1/com.ibm.swg.im.infosphere.dataexpl.engine.doc/c_correcting_stemming_errors.html

6 www.bbc.com/future/story/20170214-the-machines-that-learned-to-listen

7 www.bbc.com/future/story/20170214-the-machines-that-learned-to-listen

8 https://www.chorus.ai/blog/a-taste-of-chorus-s-secret-sauce-how-our-system-teaches-itself

9 www.chorus.ai/case-studies/housecall/

10 www.verywellmind.com/depression-statistics-everyone-should-know-4159056

11 "Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions," A. Haque, M. Guo, A.S. Miner, L. Fei-Fei, apresentado na NeurIPS 2018 Workshop sobre machine learning na área de saúde (ML4H), https://arxiv.org/abs/1811.08592.

12 https://openai.com/blog/introducing-openai/

13 https://openai.com/blog/better-language-models/

14 N.T.: O GPT-2 foi liberado em sua versão mais completa em 5 de novembro de 2019, conforme divulgado pela OpenAi em https://openai.com/blog/gpt-2-1-5b-release/

15 Rosalind W. Picard, Affective Computing (MIT Press).

16 www.newyorker.com/magazine/2015/01/19/know-feel

17 www.wsj.com/articles/iphones-toughest-rival-in-china-is-wechat-a-messaging-app-1501412406

18 www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-network/10156700570096634/

19 https://voicebot.ai/2019/02/19/juniper-forecasts-80-billion-in-voice-commerce-in-2023-or-10-per-assistant/

20 https://hbr.org/2018/05/marketing-in-the-age-of-alexa

21 www.huffingtonpost.com/2013/01/22/siri-do-engine-apple-iphone_n_2499165.html

22 https://medium.com/swlh/the-story-behind-siri-fbeb109938b0

23 www.wired.com/2010/02/siri-voice-recognition-iphone/

24 www.businessinsider.com/siri-google-assistant-voice-market-share-charts-2018-6

25 https://voicebot.ai/wp-content/uploads/2018/11/voice-assistant-consumer-adoption-report-2018-voicebot.pdf

26 https://techcrunch.com/2019/02/12/report-voice-assistants-in-use-to-triple-to-8-billion-by-2023/

27 Entrevista realizada pelo autor com o CEO e cofundador da Ushur, Simha Sadasiva.

28 A informação veio da entrevista do autor com o CEO e cofundador da Mya, Eyal Grayevsky Grayevsky.

29 Entrevista realizada pelo autor com David Karandish, CEO e cofundador da Jane.ai.

30 www.forbes.com/sites/bernardmarr/2019/02/11/7-amazing-examples-of-onlinechatbots-and-virtual-digital-assistants-in-practice/#32bb1084533e

31 Entrevista realizada pelo autor com Antonio Cahill, evangelista de IA na IBM.

32 Entrevista realizada pelo autor com Jonathan Taylor, CTO da Zoovu.

33 Entrevista realizada pelo autor com Gillian McCann, Diretor de Engenharia na Nuvem e Inteligência Artificial na Workgrid Software.

34 https://www.twice.com/retailing/artificial-intelligence-retail-chatbots-idc-spending

35 www.juniperresearch.com/press/press-releases/bank-cost-savings-via-chatbots-to-reach

36 www.technologyreview.com/s/612826/virtual-assistants-thinks-theyredoomed-without-a-new-ai-approach/

CAPÍTULO 7

Robôs físicos

A manifestação final da IA

Quando estive em Pasadena, visitei o CaliBurger para almoçar e percebi uma multidão de pessoas ao lado da área onde a comida estava sendo preparada, atrás de uma parede de vidro. As pessoas estavam tirando fotos com seus smartphones!

Por quê? Por causa do Flippy, um robô controlado por IA que podia preparar hambúrgueres.

Eu estava lá no restaurante com David Zito, CEO e cofundador da empresa Miso Robotics, que construiu o sistema. "Flippy ajuda a melhorar a qualidade dos alimentos devido a sua consistência e reduz os custos de produção", disse ele. "Também construímos o robô para estar em total conformidade com os padrões regulatórios."[1]

Depois do almoço, visitei o laboratório da Miso Robotics, onde havia um centro de testes com robôs de amostra. Foi lá que vi a convergência entre sistemas de IA e robôs físicos. Os engenheiros estavam construindo o cérebro do Flippy, o qual foi carregado para a nuvem. Algumas de suas capacidades incluíam lavar utensílios e grelhar, aprender a adaptar-se aos problemas com o cozimento, alternar entre uma espátula para carne crua e outra para carne cozida e colocar cestas na fritadeira. Tudo isso estava sendo feito em tempo real.

A indústria de serviços alimentícios, entretanto, é apenas uma das muitas áreas que serão afetadas pela robótica e pela IA.

De acordo com a International Data Corporation (IDC – Corporação Internacional de Dados), a previsão é que os gastos com robótica e drones ultrapassem os US$ 115,7 bilhões em 2019 e subam para US$ 210,3 bilhões até 2022.[2] Isso representa uma taxa de crescimento anual composta de 20,2%. Cerca de dois terços dos gastos serão direcionados aos sistemas de hardware.

Neste capítulo, vamos dar uma olhada em robôs físicos e em como a IA transformará essa indústria.

O que é um robô?

A origem da palavra "robô" remonta a 1921, em uma peça de Karel Capek chamada *Rossum's Universal Robots* (*Robôs Universais de Rossum*). A história fala sobre uma fábrica que criou robôs a partir de matéria orgânica e, sim, eles eram hostis! Eles

acabariam por se unir para se rebelar contra seus mestres humanos (considere que "robô" vem da palavra tcheca *robata*, usada para trabalho forçado).

Nos dias atuais, contudo, qual seria uma boa definição para esse tipo de sistema? Lembre-se de que existem muitas variações, já que os robôs podem ter uma infinidade de formas e funções.

No entanto, podemos resumi-los segundo algumas características principais:

- *Físico*: robôs podem variar de tamanho – desde pequenas máquinas capazes de explorar o nosso corpo até sistemas industriais maciços e máquinas voadoras para embarcações subaquáticas. Eles também precisam de algum tipo de fonte de energia, como bateria, eletricidade ou energia solar.
- *Ações:* colocando de forma bastante simples, um robô deve ser capaz de executar certas ações que podem incluir mover um item ou até mesmo falar.
- *Percepção:* para agir, um robô deve compreender seu ambiente. Isso é possível por meio de sensores e sistemas de feedback.
- *Inteligência*: não significa dispor de capacidades completas de IA. No entanto, um robô precisa ser capaz de ser programado para realizar ações.

Hoje em dia, não é muito difícil criar um robô a partir do zero. A RobotShop.com, por exemplo, tem centenas de kits que variam de menos de US$ 10 até US$ 35.750,00 (que é quanto custa o item Dr. Robot Jaguar V6 Tracked Mobile Platform).

Uma história comovente da ingenuidade com relação à construção de robôs é a do jovem Cillian Jackson, de 2 anos. Ele nasceu com uma rara condição genética que o deixou imóvel. Seus pais tentaram comprar uma cadeira de rodas elétrica especial, mas não conseguiram.

Então, alunos da Farmington High School entraram em cena e construíram um sistema para Cillian.[3] Essencialmente, tratava-se de uma cadeira de rodas robô que levou apenas um mês para ser concluída. Por causa dela, Cillian agora consegue correr atrás de seus dois cães ao redor da casa!

Vimos anteriormente as principais características dos robôs, mas há também interações essenciais a considerar:

- *Sensores*: o sensor típico é uma câmera ou um LIDAR (light detection and ranging – detecção de luz e área), que usa um scanner a laser para criar imagens 3D. Além disso, os robôs também podem contar com sistemas para audição, tato, paladar e até olfato. Na verdade, eles também podem incluir sensores que vão além das capacidades humanas, com recursos de visão noturna ou detecção de produtos químicos. As informações dos sensores são enviadas para um controlador que pode ativar um braço ou outras partes do robô.
- *Atuadores:* são dispositivos eletromecânicos, como motores. Na maioria das

vezes, ajudam com o movimento de membros do robô, como braços, pernas, cabeça ou qualquer outra parte móvel.

- *Computador*: existem sistemas de memória e processadores para ajudar com as entradas dos sensores. Em robôs avançados, também podem existir chips de IA ou conexões de internet para plataformas de nuvem de IA.

A Figura 7.1 mostra as interações dessas funções.

Há também duas maneiras principais de operar um robô. A primeira delas é por meio de um controle remoto manipulado por um ser humano. Nesse caso, o robô é chamado de telerrobô. A segunda é quando o robô é autônomo e usa as próprias habilidades para se movimentar – com a ajuda de IA, por exemplo.

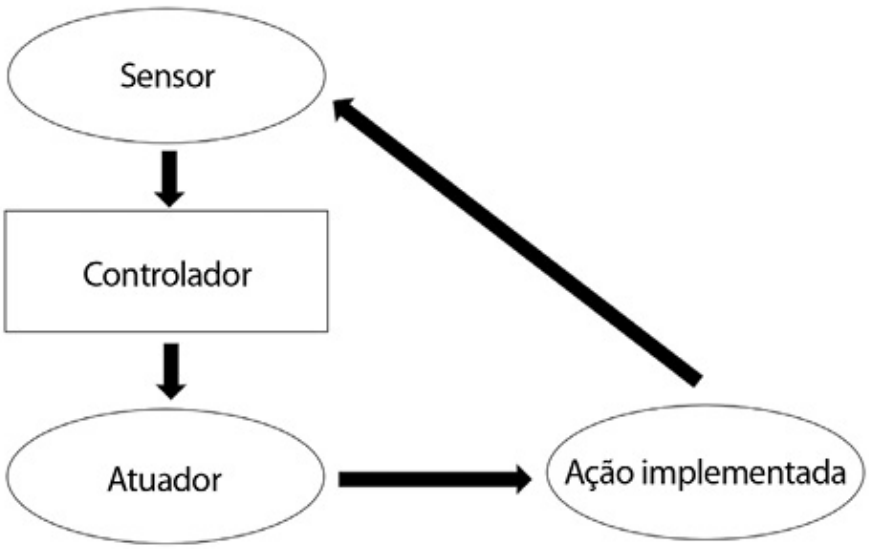

Figura 7.1 – Sistema geral para um robô físico.

Então, qual foi o primeiro robô móvel autônomo? Foi o Shakey (sacolejador). O nome era bastante apropriado, conforme observou Charles Rosen, gerente de projeto do sistema: "Trabalhamos por um mês tentando encontrar um bom nome para ele, sugerindo desde nomes gregos até outras esquisitices, e então um de nós disse: 'Ei, ele sacoleja à beça enquanto se movimenta, vamos chamá-lo de Shakey'".[4]

Com financiamento da DARPA, o Stanford Research Institute (SRI) trabalhou no Shakey de 1966 a 1972. O robô era bastante sofisticado para a época. Era grande, com mais de um metro de altura, tinha rodas para se locomover e sensores e câmeras para ajudar com o tato. Também foi ligado aos computadores DEC PDP-10 e PDP-15 por meio de conexões sem fio. A partir deles, uma pessoa poderia informar comandos via teletipo. Shakey usava algoritmos para navegar em seu ambiente e podia, inclusive, fechar portas.

O desenvolvimento do robô foi o resultado de uma variedade de avanços em IA. Nils Nilsson e Richard Fikes, por exemplo, criaram o STRIPS (Stanford Research Institute Problem Solver – Solucionador de problemas do SRI), que permitiu o

planejamento automatizado, bem como o algoritmo A* para encontrar o caminho mais curto e que demandasse a menor quantidade de recursos do computador.[5]

No final dos anos 1960, como os Estados Unidos estavam focados no programa espacial, Shakey recebeu bastante atenção. Um artigo lisonjeiro da *Life* declarou que o robô era "a primeira pessoa eletrônica".[6]

Infelizmente, contudo, com o inverno da IA em 1972, a DARPA interrompeu o financiamento do Shakey. No entanto, o robô continuaria a ser uma parte fundamental da história da tecnologia, sendo incluído no Robot Hall of Fame em 2004.[7]

Robôs industriais e comerciais

O primeiro uso real de robôs estava relacionado às indústrias de manufatura. Entretanto, esses sistemas demoraram muito para serem adotados.

A história começa com George Devol, um inventor que não concluiu o ensino médio. Apesar disso, esse não era o problema. Devol tinha um talento especial para engenharia e criatividade e foi o criador de alguns dos sistemas centrais para fornos de micro-ondas, códigos de barras e portas automáticas (durante sua vida, ele registrou mais de 40 patentes).

Durante o início dos anos 1950, ele também recebeu a patente de um robô programável chamado "Unimate". Ele lutou para atrair interesse para sua ideia, mas todos os investidores se negaram a ajudar.

No entanto, em 1957, sua vida mudaria para sempre quando conheceu Joseph Engelberger em um coquetel. Foi uma situação semelhante à de Steve Jobs quando conheceu Steve Wozniak para criar o computador da Apple.

Engelberger era engenheiro e também um empresário experiente. Adorava ler ficção científica, como as histórias de Isaac Asimov. Por conta disso, Engelberger queria que o Unimate beneficiasse a sociedade.

No entanto, ainda havia resistência – muitas pessoas achavam que a ideia era irreal e semelhante à ficção científica – e o projeto levou um ano para conseguir financiamento. Uma vez obtida a verba, Engelberger não perdeu tempo na construção do robô e conseguiu vendê-lo para a General Motors (GM) em 1961. O Unimate era pesado (aproximadamente 1.300 quilos) e tinha um braço de cerca de 2 metros, mas ainda era bastante útil e significava que as pessoas não teriam de fazer atividades inerentemente perigosas. Algumas de suas funções principais incluíam soldagem, pulverização e garra – tudo feito com precisão e disponibilidade integral.

Engelberger procurou maneiras criativas para divulgar seu robô. Para isso, apareceu no *The Tonight Show*, de Johnny Carson, em 1966, onde o Unimate encaçapou

perfeitamente uma bola de golfe e até serviu cerveja. Johnny brincou que a máquina poderia "substituir alguém em seu trabalho".[8]

Os robôs industriais, no entanto, apresentavam problemas irritantes. Curiosamente, a GM descobriu isso da maneira mais difícil durante a década de 1980. Nessa época, o CEO Roger Smith promoveu a visão de uma fábrica de "luzes apagadas" – ou seja, onde os robôs poderiam construir carros no escuro!

Ele desembolsou a exorbitante quantia de US$ 90 bilhões no programa e criou uma joint venture, com a Fujitsu-Fanuc, chamada GMF Robotics. A organização se tornaria a maior fabricante mundial de robôs.

Infelizmente, porém, o empreendimento acabou se revelando um desastre. Além de provocar os sindicatos, os robôs muitas vezes não conseguiam atender às expectativas. Um dos fiascos envolveu robôs que soldaram portas fechadas ou pintaram eles mesmos – não os carros!

No entanto, não há nada de realmente novo na situação da GMF – e a questão não está necessariamente relacionada a gerentes seniores equivocados. Dê uma olhada na Tesla, que é uma das empresas mais inovadoras do mundo. Seu CEO Elon Musk enfrentou grandes problemas com robôs no chão de fábrica. A situação ficou tão ruim que a continuidade da Tesla foi comprometida.

Em uma entrevista na *CBS This Morning* em abril de 2018, Musk disse que usou muitos robôs ao fabricar o Modelo 3 e isso realmente desacelerou o processo.[9] Ele reconheceu que mais pessoas deveriam ter sido envolvidas.

Tudo isso aponta para o que Hans Moravec uma vez escreveu: "É relativamente fácil fazer com que os computadores exibam desempenho de nível adulto em testes de inteligência ou jogando damas, mas difícil ou impossível dar-lhes as habilidades de uma pessoa de um ano quando se trata de percepção e mobilidade".[10] Esse fato é muitas vezes chamado de paradoxo de Moravec.

Independentemente de tudo isso, os robôs industriais tornaram-se uma grande invenção, expandindo-se em diversos segmentos, como bens de consumo, biotecnologia/saúde e plásticos. A partir de 2018, havia 35.880 robôs industriais e comerciais na América do Norte de acordo com dados da Robotic Industries Association (RIA – Associação das Indústrias Robóticas).[11] A indústria automobilística, por exemplo, chegou a ser responsável por cerca de 53% dessa estatística, mas isso vem diminuindo.

Jeff Burnstein, presidente da Association for Advancing Automation (Associação para Automação Avançada), comentou:

> *Como ouvimos de nossos colaboradores e em eventos como o Automate, essas vendas não são apenas para grandes multinacionais. Pequenas e médias empresas estão usando robôs para resolver os desafios do mundo real, o que está ajudando-as*

a serem mais competitivas em escala global.[12]

Em paralelo, os custos de fabricação de robôs industriais continuam a cair. Segundo uma pesquisa da ARK, haverá uma redução de 65% até 2025 – com dispositivos a um preço médio inferior a US$ 11.000 cada.[13] A análise é baseada na Lei de Wright, que afirma que, para cada duplicação cumulativa no número de unidades produzidas, há um declínio consistente nos custos em termos percentuais.

Tudo bem então, mas o que acontece com IA e robôs? Onde está esse status da tecnologia? Mesmo com os avanços com deep learning, geralmente tem havido um progresso lento no uso de IA com robôs. Isso se deve ao fato de que grande parte da pesquisa tem se concentrado em modelos baseados em software, como é o caso do reconhecimento de imagem. Outra razão é que os robôs físicos exigem tecnologias sofisticadas para compreender o ambiente – que costuma ser barulhento e perturbador – em tempo real. Isso envolve permitir localização e mapeamento simultâneos (SLAM – Simultaneous Localization And Mapping) em ambientes desconhecidos e, ao mesmo tempo, rastrear a localização do robô. Para fazer isso de forma eficaz, pode ser necessário criar tecnologias, como melhores algoritmos de rede neural e computadores quânticos.

Apesar de tudo isso, certamente há progressos sendo feitos, em especial com o uso de técnicas de aprendizagem por reforço. Veja algumas das inovações:

- *Osaro*: a empresa desenvolve sistemas que permitem que robôs aprendam rapidamente. A Osaro descreve a possibilidade como "a capacidade de imitar o comportamento que requer fusão de sensores, bem como planejamento de alto nível e manipulação de objetos. Também será possível promover a capacidade de aprender de uma máquina para outra e melhorar além dos insights de um programador humano".[14] Por exemplo, um dos robôs da empresa foi capaz de aprender, em apenas cinco segundos, como levantar e posicionar uma galinha (espera-se que o sistema seja usado em granjas).[15] A tecnologia, no entanto, pode ter muitas aplicações, como drones, veículos autônomos e IoT (Internet of Things – Internet das coisas).
- *OpenAI:* responsável pela criação da Dactyl, uma mão robô com destreza humana. Baseia-se em treinamento sofisticado de simulações, não em interações do mundo real. A OpenAI chama a técnica de "randomização de domínio", onde se apresentam muitos cenários ao robô – mesmo aqueles que têm uma probabilidade muito baixa de acontecer. Com a Dactyl, as simulações foram capazes de abarcar cerca de 100 anos de resolução de problemas.[16] Um dos resultados surpreendentes foi que o sistema aprendeu ações manuais humanas que não foram pré-programadas – como deslizar do dedo. A Dactyl também foi treinada para lidar com informações imperfeitas; por exemplo, quando os sensores atrasam as leituras ou há necessidade de lidar com vários objetos.

- *MIT*: pode facilmente levar milhares de dados de amostra para um robô entender seu ambiente, como detectar algo simples como uma caneca. Entretanto, de acordo com um trabalho de pesquisa de professores do instituto, pode haver uma maneira de reduzir isso. Eles usaram uma rede neural que se concentrava em apenas algumas características principais.[17] A pesquisa ainda está nos estágios iniciais, mas pode ser muito impactante para robôs.
- *Google*: a partir de 2013, a empresa passou por uma série de fusões e aquisições com outras empresas de robótica. Os resultados, no entanto, foram decepcionantes. Apesar disso, a organização não desistiu do negócio. Nos últimos anos, o Google tem se concentrado na busca por robôs mais simples que são conduzidos pela IA e a empresa inaugurou uma nova divisão, chamada Robotics at Google (Robótica no Google). Há projetos interessantes, como o de um robô capaz de olhar para uma caixa de itens e identificar o que lhe é solicitado cerca de 85% das vezes – e pegar o objeto com sua mão de três dedos. Uma pessoa, por outro lado, foi capaz de fazer isso em cerca de 80% das situações.[18]

Então, tudo isso aponta para a automação completa? Provavelmente não – pelo menos num futuro mais próximo. Lembre-se de que uma tendência importante é o desenvolvimento de cobôs, que nada mais são do que robôs que trabalham junto com as pessoas. Em suma, essa está se transformando em uma abordagem muito mais poderosa, pois pode haver aproveitamento das vantagens de máquinas e seres humanos.

Observe que um dos principais líderes nessa categoria é a Amazon.com. Em 2012, a empresa desembolsou US$ 775 milhões para a Kiva, uma das principais fabricantes de robôs industriais. Desde então, a Amazon.com implantou cerca de 100.000 sistemas em mais de 25 centros de atendimento (por conta disso, a empresa tem experimentado 40% de melhoria na capacidade de estoque).[19] Veja o que diz a organização:

> *A Amazon Robotics automatiza as operações do centro de atendimento usando vários métodos de tecnologia robótica, incluindo robôs móveis autônomos, software de controle sofisticado, percepção de linguagem, gerenciamento de energia, visão computacional, sensoriamento de profundidade, machine learning, reconhecimento de objetos e compreensão semântica dos comandos.*[20]

Dentro dos armazéns, os robôs movimentam-se rapidamente pelo chão, ajudando a encontrar e levantar as cápsulas de armazenamento. Contudo, as pessoas também são críticas, pois são melhores na identificação e na escolha de produtos individuais.

A configuração, no entanto, é muito complicada. Os funcionários do armazém, por exemplo, usam coletes de tecnologia robótica para não serem atropelados![21] Esse recurso permite que o robô identifique uma pessoa.

Ainda existem outros problemas com os cobôs. Há, por exemplo, o medo real de que os funcionários sejam substituídos pelas máquinas. Além do mais, é natural que as pessoas se sintam como uma engrenagem na roda, o que poderia significar uma moral mais baixa. As pessoas podem realmente se relacionar com robôs? Provavelmente não, especialmente robôs industriais, aos quais realmente faltam as qualidades humanas.

Robôs no mundo real

OK, então, vamos agora dar uma olhada em alguns estudos de caso interessantes com robôs industriais e comerciais.

Estudo de caso: Segurança

Erik Schluntz e Travis Deyle têm grande experiência na indústria robótica, com passagens por empresas como Google e SpaceX. Em 2016, eles queriam começar seu próprio empreendimento, mas primeiro passaram um tempo considerável buscando encontrar uma aplicação do mundo real para a tecnologia, que envolvia conversar com inúmeras empresas. Schluntz e Deyle encontraram um tema comum: a necessidade de segurança física das instalações. Como os robôs poderiam oferecer proteção depois das 17h – sem que fosse necessário gastar grandes quantias com seguranças?

Isso resultou no lançamento da Cobalt Robotics. O momento foi determinado pela convergência de tecnologias como visão computacional, machine learning e, claro, avanços na robótica.

Embora o uso da tecnologia de segurança tradicional seja eficaz – com câmeras e sensores, por exemplo – eles são estáticos e não são necessariamente bons para respostas em tempo real. Com um robô, entretanto, é possível ser muito mais proativo por causa da mobilidade e da inteligência subjacente.

Entretanto, as pessoas ainda estão no circuito. Os robôs podem então fazer aquilo no que são bons, como processamento e sensoriamento de dados, 24 horas por dia, 7 dias por semana, enquanto as pessoas podem se concentrar em pensar criticamente e ponderar as alternativas.

Além de sua tecnologia, a Cobalt tem sido inovadora em seu modelo de negócios, que chama de Robótica como Serviço (RaaS – Robotics as a Service). Ao assinar o serviço, esses dispositivos ficam muito mais acessíveis para os clientes.

Estudo de caso: Robôs aspiradores

É provável que vejamos algumas das aplicações mais interessantes para robôs em categorias bastante mundanas. Então, vale lembrar que essas máquinas são

realmente boas em lidar com processos repetitivos.

Veja o caso da Brain Corp, fundada em 2009 pelos Dr. Eugene Izhikevich e Dr. Allen Gruber. Inicialmente, eles desenvolveram sua tecnologia para Qualcomm e DARPA. Contudo, desde então, a Brain passou a promover o machine learning e a visão computacional para robôs autônomos. Ao todo, a empresa arrecadou US$ 125 milhões de investidores como Qualcomm e SoftBank.

O principal robô da Brain é o Auto-C, que limpa o chão de maneira bastante eficiente. Por causa do sistema de IA, chamado BrainOS (conectado à nuvem), a máquina é capaz de navegar autonomamente por ambientes complexos. Isso é feito pressionando-se um botão para que, em seguida, o Auto-C rapidamente mapeie a rota.

No final de 2018, a Brain fechou um acordo com o Walmart para disponibilizar 1.500 robôs Auto-C em centenas de lojas.[22] A empresa também implantou robôs em aeroportos e shoppings.

Esse, entretanto, não é o único robô em funcionamento no Walmart. A empresa também está instalando máquinas que podem digitalizar prateleiras para ajudar no gerenciamento de estoque. Com cerca de 4.600 lojas nos Estados Unidos, os robôs provavelmente terão um grande impacto sobre o varejista.[23]

Estudo de caso: Farmácia online

Como parte de uma segunda geração de farmacêuticos, TJ Parker tinha experiência direta com as frustrações que as pessoas sentem ao gerenciar suas prescrições. Então, ele se perguntou: será que a solução não poderia ser criar uma farmácia digital?

Ele estava convencido de que a resposta era sim. No entanto, embora tivesse uma sólida experiência na indústria, precisava de um cofundador de tecnologia forte, que encontrou em Elliot Cohen, engenheiro do MIT. Juntos, eles criaram o PillPack em 2013.

O foco era reimaginar a experiência do cliente. Usando um aplicativo ou indo até o site do PillPack, um usuário poderia facilmente se inscrever – e fornecer dados de seguro, informar necessidades de prescrição e agendar entregas. Ao receber sua entrega, o usuário teria informações detalhadas sobre a dosagem a consumir e até mesmo imagens de cada pílula. Além disso, cada um dos comprimidos incluía etiquetas e era preordenado em recipientes.

Para tornar tudo isso realidade, era necessário contar com uma infraestrutura tecnológica sofisticada, chamada PharmacyOS. Também foi necessário dispor de uma rede de robôs, localizados em um armazém de 80.000 metros quadrados. Com esses recursos, o sistema poderia classificar e empacotar eficientemente as

prescrições. A instalação também contava com farmacêuticos licenciados para gerenciar o processo e garantir que tudo estava em conformidade.

Em junho de 2018, a Amazon.com desembolsou cerca de US$ 1 bilhão pela PillPack. Foi noticiado que ações de empresas como CVS e Walgreens caíram diante dos temores de que a gigante do comércio eletrônico estava se preparando para uma grande jogada no mercado de saúde.

Estudo de caso: Robôs cientistas

Desenvolver medicamentos prescritos é extremamente caro. Com base em pesquisas do Tufts Center for the Study of Drug Development (Centro Tufts para Estudo do Desenvolvimento de Medicamentos), a média de gastos chega a cerca de US$ 2,6 bilhões por composto aprovado.[24] Além disso, por conta de regulamentações onerosas, pode-se facilmente levar mais de uma década para que um novo medicamento chegue ao mercado.

Contudo, o uso de robôs sofisticados e deep learning pode ajudar. Para saber como, veja o que os pesquisadores das Universidades de Aberystwyth e Cambridge fizeram. Em 2009, eles lançaram o Adam, essencialmente um cientista robô que ajudava com o processo de descoberta de medicamentos. Então, alguns anos depois, lançaram a Eva, um robô da próxima geração.

O sistema pode sugerir hipóteses e testá-las, assim como conduzir experimentos. O processo, no entanto, não envolve somente cálculos de força bruta (o sistema é capaz de analisar mais de 10.000 compostos por dia).[25] Com o deep learning, a Eva é capaz de usar a inteligência para identificar com precisão os compostos com maior potencial para exploração. Por exemplo, ela foi capaz de mostrar que triclosan – um elemento comum encontrado na pasta de dentes para evitar o acúmulo de placa bacteriana – poderia ser eficaz contra o crescimento de parasitas na malária. Isso é especialmente importante, uma vez que a doença tem se tornado mais resistente às terapias existentes.

Humanoides e robôs de consumo

O popular desenho animado *The Jetsons* surgiu no início de 1960 e tinha um grande elenco de personagens. Entre eles estava Rosie, uma empregada robô que sempre tinha um aspirador na mão.

Quem não gostaria de uma coisa assim? Eu adoraria. Contudo, não espere deparar com algo como a Rosie em casa tão cedo. No que se refere aos robôs de consumo, ainda estamos engatinhando. Em outras palavras, o que se vê são robôs que dispõem apenas de algumas características humanas.

Aqui estão alguns exemplos notáveis:

- *Sophia*: desenvolvida pela empresa Hanson Robotics, com sede em Hong Kong, talvez seja o robô mais famoso. Na verdade, no final de 2017, a Arábia Saudita concedeu sua cidadania! Sophia, que é parecida com Audrey Hepburn, pode andar e falar. Há também sutilezas em suas ações, como manter o contato visual.
- *Atlas*: desenvolvido pela Boston Dynamics, foi lançado no verão de 2013. Sem dúvidas, o Atlas ficou muito melhor ao longo dos anos. Ele pode, por exemplo, dar cambalhotas e levantar-se quando cai.
- *Pepper*: esse é um robô humanoide, criado pela SoftBank Robotics, que está focado na prestação de serviços ao cliente em locais de varejo, por exemplo. A máquina pode usar gestos – para ajudar a melhorar a comunicação – e consegue falar vários idiomas.

À medida que as tecnologias humanoides ficam mais realistas e avançadas, inevitavelmente haverá mudanças na sociedade. As normas sociais sobre amor e amizade evoluirão. Afinal, como vivenciado com a difusão dos smartphones, já estamos vendo como a tecnologia pode mudar a maneira como nos relacionamos com as pessoas; seja por mensagens de texto ou engajamento em mídias sociais. De acordo com uma pesquisa da Tappable, cerca de 10% dos jovens preferem sacrificar os dedos a renunciar ao seu smartphone![26]

No que se refere aos robôs, podemos ver algo semelhante. São os robôs sociais. Tal máquina – realista e com recursos de IA – poderia finalmente se tornar um amigo ou até mesmo... um amante!

É verdade que isso provavelmente não vai acontecer tão cedo. Por enquanto, no entanto, há certamente algumas inovações interessantes com robôs sociais. Um exemplo é o ElliQ, que conta com um tablet e uma pequena cabeça de robô. Na maior parte dos casos, o robô atende os que vivem sozinhos, como os idosos. O ElliQ pode conversar, mas também consegue oferecer uma assistência inestimável, como oferecer lembretes para que os medicamentos sejam tomados. O sistema também permite chamadas de vídeo com membros da família.[27]

No entanto, os robôs sociais certamente apresentam desvantagens. Basta olhar para a terrível situação da Jibo. A empresa, que arrecadou US$ 72,7 milhões em financiamento, criou o primeiro robô social para residências. Contudo, houve muitos problemas, como atrasos na entrega de produtos e uma onda de imitações. Por causa de tudo isso, a Jibo entrou com pedido de falência em 2018 e, em abril do ano seguinte, os servidores foram desligados.[28]

Nem é necessário dizer que havia muitos proprietários desanimados com o robô, conforme evidenciado pelas muitas postagens no Reddit.

As três leis da robótica

Isaac Asimov, prolífico escritor de assuntos diversos como ficção científica, história, química e Shakespeare, também teve um grande impacto sobre os robôs. Em um conto que escreveu em 1942 ("Runaround"), estabeleceu suas Três Leis da Robótica:

1. Um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano o prejudique.
2. Um robô deve obedecer às ordens dadas a ele por seres humanos, exceto quando tais ordens entrarem em conflito com a Primeira Lei.
3. Um robô deve proteger sua própria existência desde que tal proteção não entre em conflito com a Primeira ou a Segunda Leis.

Nota Asimov ainda adicionaria outra lei, a lei zero, que dizia: "Um robô não pode prejudicar a humanidade ou, por inação, permitir que a humanidade venha a prejudicá-lo". Ele considerava essa lei a mais importante.

Asimov escreveu outros contos que refletiam como as leis se comportariam em situações complexas. As histórias foram agrupadas em um livro chamado *I, Robot (Eu, robô)*. Tudo isso aconteceu no mundo do século 21.

As Três Leis representavam a reação de Asimov à forma como a ficção científica retratava robôs como malévolos. O autor achava isso pouco realista. Asimov tinha a percepção de que surgiriam regras éticas para controlar o poder dos robôs.

A partir de agora, a visão de Asimov está começando a se tornar mais real – em outras palavras, é uma boa ideia explorar princípios éticos. É verdade que isso pode não necessariamente significar que sua abordagem é o caminho certo. Contudo, é um bom começo, em especial à medida que os robôs ficam mais inteligentes e mais pessoais com o poder da IA.

Cibersegurança e robôs

A cibersegurança não tem sido um grande problema para os robôs. Infelizmente, entretanto, esse provavelmente não será o caso por muito tempo. A principal razão é que está se tornando muito mais comum encontrar robôs conectados à nuvem. O mesmo vale para outros sistemas, como a Internet das Coisas, ou IoT, e os carros autônomos. Por exemplo, muitos desses sistemas são atualizados sem fio, o que os expõe a malware, vírus e até mesmo ramsonware. Além disso, quando se trata de veículos elétricos, há também a vulnerabilidade a ataques à rede de carregamento.

Na verdade, seus dados podem permanecer dentro de um veículo! Então, se ele for destruído ou vendido, as informações – como vídeos, detalhes de navegação e contatos de conexões de smartphones emparelhados – podem tornar-se disponíveis para outras pessoas. De acordo com CNBC.com, um hacker de chapéu branco chamado GreenTheOnly conseguiu extrair esses dados de uma variedade de modelos Tesla em ferros-velhos.[29] É importante ressaltar, no entanto, que a empresa

fornece opções para limpar os dados e é possível optar por não os receber (embora isso signifique abrir mão de certas vantagens, como atualizações over-the-air – OTA).

Agora, se houver uma violação de segurança cibernética com um robô, as implicações podem certamente ser devastadoras. Imagine se um hacker se infiltrasse em uma linha de fabricação, uma cadeia de suprimentos ou até mesmo em um sistema de cirurgia robótica. Vidas podem estar em perigo.

Independentemente disso, não houve muito investimento em segurança cibernética para robôs. Até agora, há apenas um punhado de empresas focadas nisso, como Karamba Security e Cybereason. No entanto, à medida que os problemas piorarem, inevitavelmente haverá um aumento dos investimentos e novas iniciativas de empresas de segurança cibernética legadas.

Programando robôs para IA

À medida que os sistemas ficam mais baratos e novas plataformas de software surgem, vai ficando mais fácil criar robôs inteligentes. Grande parte disso se deve ao Robot Operating System (ROS – Sistema operacional de robôs), que está se tornando um padrão na indústria. As origens remontam a 2007, quando a plataforma começou como um projeto de código aberto no Stanford Artificial Intelligence Laboratory (Laboratório de Inteligência Artificial de Stanford).

Apesar de seu nome, o ROS não é realmente um sistema operacional. Em vez disso, é o middleware que ajuda a gerenciar muitas das partes críticas de um robô, como planejamento, simulações, mapeamento, localização, percepção e protótipos. O ROS também é modular e as funções necessárias podem ser facilmente escolhidas. O resultado é que o sistema pode reduzir o tempo de desenvolvimento sem grandes dificuldades.

Outra vantagem: o ROS conta com uma comunidade global de usuários. Considere que existem mais de 3.000 pacotes para a plataforma.[30]

Como testemunho da proeza do ROS, a Microsoft anunciou, no final de 2018, que lançaria uma versão para o sistema operacional Windows. De acordo com a postagem no blog de Lou Amadio, o principal engenheiro de software do Windows IoT, "à medida que os robôs avançaram, as ferramentas de desenvolvimento também progrediram. Vemos a robótica com inteligência artificial como tecnologia universalmente acessível para aumentar as habilidades humanas".[31]

O resultado é que o ROS pode ser usado com o Visual Studio e haverá conexões para a nuvem do Azure, que inclui ferramentas de IA.

OK, então, quando se trata de desenvolver robôs inteligentes, muitas vezes há um processo diferente daquele utilizado com a abordagem típica da IA baseada em

software. Ou seja, não só existe a necessidade de haver um dispositivo físico, mas também uma maneira de testá-lo. Muitas vezes, isso é feito usando uma simulação. Alguns desenvolvedores chegam a iniciar o projeto com a criação de modelos de papelão, o que pode ser uma ótima maneira de conhecer os requisitos físicos.

É claro que existem também simuladores virtuais úteis, como MuJoCo, Gazebo, MORSE e V-REP. Esses sistemas usam gráficos 3D sofisticados para lidar com os movimentos e a física do mundo real.

Então, como são criados os modelos de IA para robôs? Na verdade, é um pouco diferente da abordagem com algoritmos baseados em software (como discutido no Capítulo 2). Com um robô, há a vantagem de que ele continuará a coletar dados de seus sensores, o que pode ajudar a evoluir a IA.

A nuvem também está se tornando um fator crítico no desenvolvimento de robôs inteligentes, como visto com a Amazon.com. A empresa promoveu sua popular plataforma AWS com uma nova oferta, chamada AWS RoboMaker. Com ela, é possível construir, testar e implantar robôs sem muita configuração. A AWS RoboMaker opera com ROS e permite o uso de serviços para machine learning, análise e monitoramento. Existem, inclusive, mundos virtuais pré-construídos em 3D para lojas de varejo, salas cobertas e pistas de corrida! Então, uma vez que se termine o robô, é possível usar a AWS para desenvolver um sistema over-the-air (OTA) para implantação segura e atualizações periódicas.

E, como não deve ser nenhuma surpresa, o Google está pensando em lançar sua própria plataforma de nuvem de robôs (espera-se que ela seja lançada em 2019).[32]

Futuro dos robôs

Rodney Brooks é um dos gigantes da indústria robótica. Em 1990, ele cofundou a iRobot para encontrar maneiras de comercializar a tecnologia. Não foi fácil. Não até 2002, quando a empresa lançou seu robô aspirador Roomba, que foi um grande sucesso entre os consumidores. Enquanto este livro era escrito, a iRobot tinha um valor de mercado de US$ 3,2 bilhões e reportou mais de US$ 1 bilhão em receitas para 2018.

A iRobot, no entanto, não foi a única startup de Brooks. Ele também ajudou a lançar a Rethink Robotics – onde sua visão era ambiciosa. Veja o que ele disse em 2010, quando sua empresa anunciou um financiamento de US$ 20 milhões:

> *Nossos robôs serão intuitivos de usar, inteligentes e altamente flexíveis. Eles vão ser fáceis de comprar, treinar e implantar e serão incrivelmente baratos. [Rethink Robotics] vai mudar a definição de como e onde os robôs podem ser usados, expandindo drasticamente o mercado de robôs.*[33]

Infelizmente, como na iRobot, houve muitos desafios. Embora a ideia de Brook para

cobôs fosse pioneira – e acabasse se revelando um mercado lucrativo –, ele precisou lutar contra as complicações da construção de um sistema eficaz. O foco na segurança significou que a precisão e a exatidão não estavam dentro dos padrões dos clientes industriais. Por causa disso, a demanda pelos robôs da Rethink foi morna.

Em outubro de 2018, a empresa ficou sem dinheiro e precisou fechar as portas. No total, a Rethink tinha levantado cerca de US$ 150 milhões em capital e atraído investidores estratégicos como Goldman Sachs, Sigma Partners, GE e Bezos Expeditions. A propriedade intelectual da empresa foi vendida para uma empresa de automação alemã chamada Hahn Group.

É verdade, esse é apenas um exemplo. Entretanto, novamente ele mostra que até mesmo as pessoas mais inteligentes da área de tecnologia podem fazer coisas erradas. E o mais importante, o mercado de robótica tem complexidades únicas. Quando se trata da evolução dessa categoria, o progresso pode ser agitado e volátil.

Como observou Schluntz, da Cobalto:

> *Embora a indústria tenha feito progressos na última década, a robótica ainda não alcançou todo o seu potencial. Qualquer nova tecnologia criará uma onda de inúmeras novas empresas, mas apenas algumas sobreviverão e se transformarão em negócios duradouros. A explosão do Dot-Com matou a maioria das empresas de internet, mas Google, Amazon e Netflix sobreviveram. O que as empresas de robótica precisam fazer é serem honestas sobre o que seus robôs podem fazer para os clientes hoje, superar estereótipos hollywoodianos que retratam robôs como os bandidos e demonstrar um ROI claro (Return On Investment – Retorno sobre investimento) para os clientes.*[34]

Conclusão

Até os últimos anos, os robôs existiam principalmente para fabricação de ponta, como a dos automóveis. Contudo, com o crescimento da IA e os custos mais baixos para a construção de dispositivos, os robôs estão se tornando mais difundidos em uma variedade de indústrias. Conforme visto neste capítulo, há aplicações interessantes para robôs que fazem coisas como limpar pisos ou oferecer segurança para instalações.

O uso da IA com robótica, entretanto, ainda está nos estágios iniciais. A programação dos sistemas de hardware está longe de ser fácil e há a necessidade de sistemas sofisticados para navegar pelos ambientes. No entanto, com abordagens de IA como aprendizado por reforço, houve um progresso acelerado.

Quando se pensa em usar robôs, é importante entender suas limitações. Também deve haver um propósito claro. Senão, uma implantação pode facilmente levar a uma

falha custosa. Mesmo algumas das empresas mais inovadoras do mundo, como Google e Tesla, enfrentaram desafios ao trabalhar com robôs.

Principais aprendizados

- Um robô é capaz de executar ações, perceber seu ambiente e alcançar algum nível de inteligência. Há também funções essenciais como sensores, atuadores (como motores) e computadores.
- Existem duas maneiras principais de operar um robô: o telerrobô (controlado por um humano) e o robô autônomo (baseado em sistemas de IA).
- Desenvolver robôs é incrivelmente complicado. Mesmo alguns dos melhores tecnólogos do mundo, como Elon Musk, da Tesla, tiveram grandes problemas com a tecnologia. Uma das principais razões é o paradoxo de Moravec. Basicamente, o que é fácil para os seres humanos é muitas vezes difícil para robôs e vice-versa.
- Embora a IA esteja causando impacto nos robôs, o processo tem sido lento. Um dos motivos para isso é que mais ênfase tem sido dada às tecnologias baseadas em software. Os robôs também são extremamente complicados quando se trata de movimentação e compreensão do ambiente.
- Os cobôs são máquinas que trabalham ao lado dos seres humanos. A ideia é permitir o uso máximo das potencialidades tanto das máquinas quanto das pessoas.
- O custo dos robôs é uma das principais razões para a falta de adoção. No entanto, empresas inovadoras, como a Cobalt Robotics, estão usando novos modelos de negócios para ajudar nessa questão, como é o caso das assinaturas.
- Robôs de consumo ainda estão nos estágios iniciais, em especial se comparados aos robôs industriais. Contudo, há algumas aplicações interessantes, tal como a das máquinas que podem ser companheiras para as pessoas.
- Durante a década de 1950, o escritor de ficção científica Isaac Asimov criou as Três Leis da Robótica. Em resumo, elas se concentravam em certificar-se de que as máquinas não prejudicariam as pessoas ou a sociedade. Mesmo que haja críticas à abordagem de Asimov, ela ainda é amplamente aceita.
- A segurança geralmente não tem sido um problema para os robôs. No entanto, isso provavelmente vai mudar – e rápido. Afinal, mais robôs estão conectados à nuvem, o que permite a intrusão de vírus e malware.
- O Robot Operating System (ROS – Sistema operacional de robôs) tornou-se um padrão para a indústria robótica. Esse middleware ajuda nas fases de planejamento, simulações, mapeamento, localização, percepção e protótipos.
- Desenvolver robôs inteligentes tem muitos desafios devido à necessidade de

criação de sistemas físicos. Existem ferramentas que podem ajudar permitindo, por exemplo, simulações sofisticadas.

---

1 Entrevista realizada pelo autor em janeiro de 2019 com David Zito, CEO e cofundador da Miso Robotics.

2 www.idc.com/getdoc.jsp?containerId=prUS44505618

3 www.nytimes.com/2019/04/03/us/robotics-wheelchair.html

4 www.computerhistory.org/revolution/artificial-intelligence-robotics/13/289

5 https://spectrum.ieee.org/view-from-the-valley/tech-history/space-age/sri-shakey-robot-honored-as-ieee-milestone

6 https://www.sri.com/sites/default/timeline/timeline.php?timeline=computingdigital#!&innovation=shakey-the-robot

7 www.wired.com/2013/09/tech-time-warp-shakey-robot/

8 www.theatlantic.com/technology/archive/2011/08/unimate-robot-on-johnnycarsons-tonight-show-1966/469779/carsons-tonight-show-1966/469779/

9 www.theverge.com/2018/4/13/17234296/tesla-model-3-robots-production-hell-Elon-Musk

10 www.graphcore.ai/posts/is-moravecs-paradox-still-relevant-for-ai-today

11 www.apnews.com/b399fa71204d47199fdf4c753102e6c7

12 www.apnews.com/b399fa71204d47199fdf4c753102e6c7

13 https://ark-invest.com/research/industrial-robot-costs

14 www.osaro.com/technology

15 www.technologyreview.com/s/611424/this-is-how-the-robot-uprising-finally-begins

16 https://openai.com/blog/learning-dexterity/

17 https://arxiv.org/abs/1903.06684

18 www.nytimes.com/2019/03/26/technology/google-robotics-lab.html

19 https://techcrunch.com/2019/03/29/built-robotics-massive-construction-excavator-drives-itself/

20 www.amazonrobotics.com/#/vision

21 www.theverge.com/2019/1/21/18191338/amazonrobot-warehouse-tech-vest-utility-belt-safety

22 www.wsj.com/articles/walmart-is-rolling-out-therobots-11554782460

23 https://techcrunch.com/2019/04/10/the-startup-behind-walmarts-shelf-scanning-robots/

24 www.policymed.com/2014/12/a-tough-road-cost-to-develop-one-new-drug-is-26-billion-approval-rate-for-drugs-entering-clinical-de.html

25 www.cam.ac.uk/research/news/artificially-intelligent-robot-scientist-eve-could-boost-search-for-new-drugs

26 www.mediapost.com/publications/article/322677/one-in-10-millennials-would-rather-lose-a-finger-t.html

27 www.wsj.com/articles/on-demand-grandkids-and-robot-pals-technology-strives-to-cure-senior-loneliness-11550898010?mod=hp_lead_pos9

28 https://techcrunch.com/2019/03/04/the-lonely-death-of-jibo-the-social-robot/

29 www.cnbc.com/2019/03/29/tesla-model-3-keeps-data-like-crash-videos-location-phone-contacts.html

30 www.ros.org/is-ros-for-me/

31 https://blogs.windows.com/windowsexperience/2018/09/28/bringing-the-power-of-windows-10-to-the-robot-operating-system/

32 www.therobotreport.com/google-cloud-robotics-platform/

33 www.rethinkrobotics.com/news-item/heartland-robotics-raises-20-million-in-series-b-financing/

34 Entrevista realizada pelo autor com Erik Schluntz, CTO da Cobalt Robotics.

CAPÍTULO 8

Implementação da IA

Faça a diferença para a sua empresa

Em março de 2019, um atirador transmitiu ao vivo no Facebook o assassinato brutal de 50 pessoas em duas mesquitas na Nova Zelândia. A transmissão foi vista cerca de 4.000 vezes e só foi desligada 29 minutos após o ataque.[1] O vídeo foi então enviado para outras plataformas e assistido milhões de vezes.

Sim, trata-se de um exemplo gritante de como a IA pode falhar de forma devastadora.

Em uma postagem em um blog, o vice-presidente de gerenciamento de produtos do Facebook, Guy Rosen, observou:

> *Os sistemas de IA são baseados em "treinamento de dados", o que significa que você precisa de muitos milhares de exemplos de conteúdo para treinar um sistema que possa detectar certos tipos de texto, imagens ou vídeo. Essa abordagem tem funcionado muito bem para áreas como nudez, propaganda terrorista e violência gráfica, onde há um enorme número de exemplos que podemos usar para treinar nossos sistemas. No entanto, esse vídeo em particular não acionou nossos sistemas de detecção automática. Para conseguir isso, precisaremos fornecer aos nossos sistemas grandes volumes de dados desse tipo específico de conteúdo, algo que é difícil, pois, felizmente, esses eventos são raros. Outro desafio é diferenciar automaticamente esse conteúdo de outro visualmente semelhante e inócuo – por exemplo, se milhares de vídeos de partidas de videogames transmitidos ao vivo forem denunciados por nossos sistemas, nossos revisores podem perder os importantes vídeos do mundo real a partir dos quais poderíamos alertar os socorristas para fornecerem ajuda.*[2]

Também não ajudou o fato de vários indivíduos ruins terem recarregado versões editadas do vídeo para frustrar o sistema de IA do Facebook.

Claro, essa foi uma grande lição sobre as deficiências da tecnologia, e a empresa diz que está empenhada em continuar a melhorar seus sistemas. No entanto, o estudo de caso do Facebook também destaca que mesmo as empresas tecnologicamente mais sofisticadas enfrentam grandes desafios. É por isso que, quando se trata de implementar a IA, é preciso haver um planejamento sólido, bem como um entendimento de que, inevitavelmente, haverá problemas. Pode ser difícil, pois os gerentes seniores das empresas estão sob pressão para obter resultados com essa tecnologia.

Neste capítulo, vamos dar uma olhada em algumas das melhores práticas para

implementações de IA.

Abordagens para implementação da IA

A aplicação de IA em uma empresa geralmente envolve duas abordagens: usar software de um fornecedor ou criar modelos internos. O primeiro é o mais comum – e pode ser suficiente para muitas empresas. A ironia é que é possível que estejam sendo usados software de empresas como Salesforce.com, Microsoft, Google, Workday, Adobe ou SAP que já dispõem de poderosos recursos de IA. Em outras palavras, uma boa abordagem é ter certeza de que os recursos estão sendo aproveitados ao máximo.

Para ver o que está disponível, dê uma olhada no Einstein, da Salesforce.com, lançado em setembro de 2016. Esse sistema de IA está perfeitamente incorporado à principal plataforma CRM da empresa, permitindo ações mais preditivas e personalizadas para vendas, serviços, marketing e comércio. A Salesforce.com chama o Einstein de "cientista de dados pessoais", pois ele é bastante fácil de usar e permite que fluxos de trabalho sejam criados com recursos de drag and drop (arrastar e soltar). Algumas de suas funcionalidades incluem:

- *Pontuação preditiva:* mostra a probabilidade de um lead[3] se converter em uma oportunidade.
- *Análise de sentimentos*: fornece uma maneira de se ter uma noção de como as pessoas veem sua marca e seus produtos a partir de análises das mídias sociais.
- *Recomendações inteligentes:* Einstein analisa dados para mostrar quais produtos são os mais adequados para que se obtenham leads.

Embora esses recursos pré-construídos facilitem o uso da IA, ainda há problemas potenciais. "Temos construído funções de IA em nossas aplicações durante os últimos anos e tem sido uma grande experiência de aprendizagem", disse Ricky Thakrar, evangelista de experiência do consumidor na Zoho. "Para fazer a tecnologia funcionar, no entanto, os usuários devem usar o software certo. Se as pessoas de vendas não estão inserindo os dados corretamente, então os resultados provavelmente estarão equivocados. Também descobrimos que deveria haver pelo menos três meses de uso para que os modelos fossem treinados. Além disso, mesmo que seus funcionários estejam fazendo tudo certo, isso não significa que as previsões de IA serão perfeitas. Sempre encare as coisas com cautela."[4]

Agora, no que diz respeito à construção de seus próprios modelos de IA, esse é um compromisso significativo para a empresa. E é sobre isso que falaremos neste capítulo.

Entretanto, independentemente da abordagem escolhida, a implementação e o uso da IA devem começar primeiro por educação e treinamento. Não importa se os

funcionários são pessoas sem conhecimento técnico ou engenheiros de software. Para que a IA seja bem-sucedida em uma organização, todos devem ter uma compreensão básica da tecnologia. Sim, este livro será útil, mas há muitos outros recursos online que também podem ajudar, como as plataformas de treinamento Lynda, Udacity e Udemy, por exemplo. Elas oferecem centenas de cursos de alta qualidade sobre muitos tópicos relacionados a IA.

Para dar uma ideia de como deveria ser um programa de treinamento corporativo, considere a Adobe. Mesmo que a empresa tenha engenheiros incrivelmente talentosos, ainda há muitos funcionários que não têm experiência alguma em IA. Alguns deles podem não ter se especializado no assunto na faculdade ou no trabalho. No entanto, a empresa queria garantir que todos os engenheiros tivessem uma compreensão sólida dos princípios fundamentais da IA. Para esse fim, criou um programa de certificação de seis meses e treinou 5.000 engenheiros ao longo de 2018. O objetivo é despertar o cientista de dados que existe em cada um desses profissionais.

O programa inclui cursos online e sessões presenciais que abrangem não só tópicos técnicos, mas também áreas como estratégia e ética. A Adobe também oferece a ajuda de cientistas da computação seniores para auxiliar os alunos no aprendizado dos tópicos.

Em seguida, no início do processo de implementação, é essencial pensar sobre os riscos potenciais. Talvez um dos mais ameaçadores seja o viés (bias), uma vez que pode facilmente se infiltrar em um modelo de IA.

Um exemplo disso é a Amazon.com, que descontinuou seu software de recrutamento movido a IA em 2017. O principal dificultador é que ele tendia a recomendar a contratação de homens. Curiosamente, esse foi um caso clássico de problema de treinamento do modelo. Considere que a maioria das submissões de currículo era de homens – o que fez com que os dados ficassem distorcidos. A Amazon.com até tentou ajustar o modelo, mas ainda assim os resultados estavam longe de ser neutros em termos de gênero.[5]

Nesse caso, a questão não era apenas sobre a tomada de decisões baseada em premissas imperfeitas. A Amazon.com também estava se expondo a potenciais responsabilidades legais, como alegações de discriminação.

Dadas as questões complicadas com a IA, mais empresas estão montando conselhos de ética. Entretanto, mesmo isso pode representar diversos problemas. Afinal, o que é ético para uma pessoa pode não ser um grande negócio para outra pessoa, certo? Definitivamente.

O Google, por exemplo, fechou seu conselho de ética cerca de uma semana depois de seu lançamento. Parece que a principal razão foi a decisão de incluir um membro

da Heritage Foundation, um laboratório de ideias conservadoras.[6]

Etapas da implementação da IA

Se pretende implementar os próprios modelos de IA, quais são as principais etapas a serem consideradas? Quais são as melhores práticas? Bem, em primeiro lugar, é extremamente importante que seus dados estejam bastante limpos e estruturados de forma a permitir a modelagem (ver Capítulo 2).

Aqui estão alguns outros aspectos a observar:

- Identifique o problema a resolver.
- Monte uma equipe forte.
- Selecione as ferramentas e plataformas adequadas.
- Crie o modelo de IA (discutimos esse processo no Capítulo 3).
- Implante e monitore o modelo de IA.

Vamos dar uma olhada em cada um desses passos.

Identifique o problema a resolver

Fundada em 1976, a HCL Technologies é uma das maiores empresas de consultoria de TI, com 132.000 funcionários em 44 países e metade das empresas da Fortune 500 como clientes. A empresa também implementou um número significativo de sistemas de IA.

Veja o que Kalyan Kumar, vice-presidente corporativo e CTO global da HCL Technologies, tem a dizer:

> *Os líderes empresariais precisam entender e perceber que a adoção da Inteligência Artificial é uma jornada e não uma única corrida. É fundamental que as pessoas que conduzem a adoção da IA dentro de uma empresa permaneçam realistas com relação ao prazo e ao que essa tecnologia é capaz de fazer. A relação entre humanos e IA é mutuamente fortalecedora e qualquer implementação de IA pode levar algum tempo até que comece a causar um impacto positivo e significativo.*[7]

É um ótimo conselho. É por isso que – em especial para as empresas que estão começando na jornada da IA – é essencial ter uma abordagem experimental. Pense nisso como a montagem de um programa piloto – isto é, essa é a fase de "engatinhar e caminhar".

Quando se trata do processo de implementação de IA, é comum ficar muito focado nas diferentes tecnologias, que são certamente fascinantes e poderosas. No entanto, o sucesso vai além delas; em outras palavras, deve primeiro haver um caso de negócios claro. Então, aqui estão alguns tópicos a considerar quando começar:

- Sem dúvida, as decisões nas empresas são muitas vezes ad hoc e, bem, uma questão de adivinhação! Com a IA, entretanto, tem-se a oportunidade de usar a tomada de decisões orientada por dados, o que garante maior precisão. Então, em que parte da sua organização isso pode trazer mais benefícios?
- Como visto com a Automação Robótica de Processos (RPA), discutida no Capítulo 5, a IA pode ser extremamente eficaz ao lidar com tarefas repetitivas e mundanas.
- Chatbots podem representar outra oportunidade de começar com a IA. Eles são relativamente fáceis de configurar e podem atender a usos específicos, como atendimento ao cliente. É possível aprender mais sobre o assunto no Capítulo 6.

Andrew Ng, CEO da Landing AI e ex-chefe do Google Brain, criou várias abordagens a serem consideradas na identificação do aspecto no qual se concentrar durante o projeto inicial de IA:[8]

- *Vitória rápida*: um projeto deve levar de 6 a 12 meses e ter uma alta probabilidade de sucesso, o que deve ajudar a impulsionar mais iniciativas. Andrew sugere ter um par de projetos, uma vez que isso aumenta as chances de conseguir uma vitória.
- *Significativo:* um projeto não precisa ser transformador. No entanto, deve apresentar resultados que ajudem a melhorar a empresa de forma notável, encorajando maior adesão a investimentos adicionais em IA. O valor geralmente é criado a partir de custos mais baixos, receitas mais elevadas, novas extensões do negócio ou mitigação de riscos.
- *Foco específico da indústria:* é fundamental, uma vez que um projeto bem-sucedido será outro fator para impulsionar a adesão. Assim, se você tem uma empresa que vende um serviço de assinatura, um sistema de IA para diminuir a rotatividade seria um bom ponto de partida.
- *Dados:* não limite suas opções com base na quantidade de dados que tem. Andrew observa que um projeto de IA bem-sucedido pode ter apenas 100 pontos de dados. No entanto, eles ainda devem ser de alta qualidade e bastante limpos, temas abordados no Capítulo 2.

Ao olhar para essa fase, também vale a pena avaliar o "tango" entre funcionários e máquinas. Tenha em mente que isso é muitas vezes esquecido – e pode ter consequências adversas em um projeto de IA. Como vimos neste livro, a IA é ótima no processamento de grandes quantidades de dados com pouco erro e em grande velocidade. A tecnologia também é excelente com previsões e detecção de anomalias. No entanto, há tarefas que os seres humanos fazem muito melhor, como ser criativo, engajar-se em abstração e compreender conceitos.

Observe o exemplo a seguir dado por Erik Schluntz, cofundador e CTO da Cobalt

Robotics:

> *Nossos robôs de segurança são excelentes na detecção de eventos incomuns no local de trabalho e no campus, como identificar uma pessoa em um escritório escuro com imagens térmicas alimentadas por IA. Apesar disso, um dos nossos operadores humanos, em seguida, entra em ação e decide a medida a ser tomada. Mesmo com todo seu potencial, a IA ainda não é a melhor opção para uma missão crítica quando confrontada com variáveis ambientais em constante mudança e imprevisibilidade humana. Considere a gravidade de a IA cometer um erro em diferentes situações – não detectar um intruso malicioso é muito pior do que acidentalmente soar um alarme falso para um dos nossos operadores.*[9]

Em seguida, certifique-se de que conhece os indicadores de desempenho de processos (KPIs) e meça-os diligentemente. Por exemplo, se estiver desenvolvendo um chatbot personalizado para atendimento ao cliente, talvez seja útil avaliar métricas como a taxa de resolução e a satisfação do cliente.

Por fim, será necessário fazer uma avaliação de TI. Se grande parte dos sistemas são legados, então pode ser mais difícil e caro implementar IA, mesmo que os fornecedores disponham de APIs e integrações. Isso significa que as expectativas precisarão ser moderadas.

Apesar de tudo isso, os investimentos podem realmente fazer a diferença, mesmo para as empresas mais antigas. Para ver um exemplo disso, considere a Symrise, cujas raízes remontam a mais de 200 anos na Alemanha. Enquanto este livro era escrito, a empresa era produtora global de sabores e fragrâncias, com mais de 30.000 produtos.

Há alguns anos, a Symrise embarcou em uma grande iniciativa, com a ajuda da IBM, para usar a IA para criar perfumes. A empresa não só teve de reequipar sua infraestrutura de TI existente, mas também investir tempo considerável ajustando os modelos. O que ajudou foi o fato de a empresa já dispor de um extenso conjunto de dados, o que permitiu maior precisão. Observe que mesmo um pequeno desvio na mistura de um composto pode fazer um perfume falhar.

De acordo com Achim Daub, presidente de Fragrâncias e Cuidados da Symrise:

> *Agora, nossos perfumistas podem trabalhar com um aprendiz de IA ao seu lado que pode analisar milhares de fórmulas e dados históricos para identificar padrões e prever novas combinações, ajudando a torná-los mais produtivos e acelerando o processo de design, guiando-os para fórmulas que nunca foram vistas antes.*[10]

Monte uma equipe

Qual deve ser a equipe inicial para um projeto de IA? Talvez um bom guia seja usar a "regra das duas pizzas" de Jeff Bezos.[11] Em outras palavras, isso é suficiente para

alimentar as pessoas que estão participando?

Ah, e não deve haver pressa para montar a equipe. Todos devem estar altamente focados no sucesso e entender a importância do projeto. Se houver pouco a mostrar a partir do projeto de IA, as perspectivas para futuras iniciativas podem estar em perigo.

A equipe precisará de um líder que tenha experiência em negócios ou operações, mas que também disponha de algumas habilidades técnicas. Esse profissional deve não só ser capaz de identificar o caso de negócios para o projeto de IA, mas também comunicar a visão para vários stakeholders na empresa, como o departamento de TI e a gerência sênior.

As pessoas da área técnica provavelmente não precisarão de um doutorado em IA. Embora sejam brilhantes, elas muitas vezes estão focadas, principalmente, em inovações no campo, como a refinação ou a criação de novos modelos. Esses conjuntos de habilidades em geral não são essenciais para um piloto de IA.

Em vez disso, procure por pessoas com experiência em engenharia de software ou ciência de dados. No entanto, como observado no início do capítulo, elas podem não ter uma forte experiência em IA. Assim sendo, pode haver necessidade de fazê-los passar alguns meses em treinamento para aprender os princípios fundamentais de machine learning e deep learning. Também deve haver um foco na compreensão de como usar plataformas de IA, como o TensorFlow.

Dado os desafios, pode ser uma boa ideia procurar o auxílio de consultores, que podem ajudar a identificar as oportunidades de IA e fornecer conselhos sobre preparação de dados e desenvolvimento de modelos.

Como o piloto de IA será experimental, a equipe deve contar com pessoas dispostas a assumir riscos e que possuam mente aberta. Se não, o progresso pode ser extremamente difícil.

Ferramentas e plataformas adequadas

Há muitas ferramentas para ajudar a criar modelos de IA, e a maioria delas é de código aberto. Mesmo que seja bom testá-las, ainda é aconselhável realizar sua avaliação de TI primeiro. Ao fazer isso, sua posição para avaliar as ferramentas de IA será mais cômoda.

Outra coisa: é possível que perceba que sua empresa já está usando várias ferramentas e plataformas de inteligência artificial! Isso pode causar problemas com a integração e a gestão de processos com projetos de IA. Sabendo disso, a empresa deve desenvolver uma estratégia para as ferramentas. Pense nisso como sua pilha de ferramentas de IA.

Ok, então, vamos dar uma olhada em algumas das linguagens, plataformas e ferramentas mais comuns para IA.

Linguagem Python

Guido van Rossum, que obteve seu mestrado em matemática e ciência da computação pela Universidade de Amsterdã em 1982, continuou a trabalhar em vários institutos de pesquisa na Europa, como a Corporation for National Research Initiatives (CNRI – Corporação para Iniciativas Nacionais de Pesquisa). Foi no final dos anos 1980, entretanto, que ele criou a própria linguagem de computador, chamada Python. O nome foi inspirado na popular série de comédia britânica *Monty Python*.

A linguagem era um tanto excêntrica – mas isso a tornava muito poderosa. Python logo se transformaria na escolha padrão para desenvolvimento de IA.

Parte disso se deve a sua simplicidade. Com apenas alguns scripts de código, é possível criar modelos sofisticados, com funções como filtro, mapa e redução. Claro que a linguagem também permite uma codificação muito mais sofisticada.

Van Rossum desenvolveu a linguagem Python com uma filosofia bastante clara:[12]

- Bonito é melhor do que feio.
- Explícito é melhor do que implícito.
- Simples é melhor do que complexo.
- Complexo é melhor do que complicado.
- Plano é melhor do que aninhado.
- Esparso é melhor do que denso.

Esses são apenas alguns dos princípios.

Além disso, Python teve a vantagem de crescer na comunidade acadêmica, que tinha acesso à internet que ajudou a acelerar sua distribuição. Ela também tornou possível o surgimento de um ecossistema global com milhares de diferentes pacotes e bibliotecas de IA. Aqui estão alguns desses recursos:

- *NumPy:* permite a implementação de aplicações de computação científica. No centro disso está a capacidade de criar uma sofisticada variedade de objetos de alto desempenho; fundamental para o processamento de dados de ponta em modelos de IA.
- *Matplotlib*: permite traçar conjuntos de dados. Muitas vezes, Matplotlib é usada em conjunto com NumPy/Pandas (Pandas refere-se a "Python Data Analysis Library" – "Biblioteca de Análise de Dados Python"). Essa biblioteca torna relativamente fácil a criação de estruturas de dados para o desenvolvimento de modelos de IA.

- *SimpleAI:* é uma implementação dos algoritmos de IA do livro *Artificial Intelligence: A Modern Approach (Inteligência Artificial: Uma Abordagem Moderna)*, de Stuart Russel e Peter Norvig. A biblioteca não só tem uma funcionalidade rica, mas também fornece recursos úteis para navegar no processo.
- *PyBrain*: é uma biblioteca modular de machine learning que torna possível criar modelos sofisticados – redes neurais e sistemas de aprendizado por reforço – sem muita codificação.
- *Scikit-Learn:* lançada em 2007, essa biblioteca dispõe de uma fonte profunda de recursos, permitindo regressão, agrupamento e classificação de dados.

Outro benefício para Python é que existem muitos recursos para a aprendizagem. Uma pesquisa rápida no YouTube mostrará milhares de cursos gratuitos.

Existem também outras linguagens sólidas que podem ser usadas para IA, como C++, C# e Java. Embora geralmente sejam mais poderosas do que Python, elas também são complexas. Além disso, quando se trata de construir modelos, muitas vezes há pouca necessidade de criar aplicações completas. E, por fim, há bibliotecas Python construídas para máquinas de IA de alta velocidade – com GPUs – como a CUDA Python.

Frameworks de IA

Existe uma variedade de estruturas de IA, que fornecem sistemas completos para construir modelos, treiná-los e implantá-los. O mais popular deles é, sem dúvida, o TensorFlow, apoiado pelo Google. A empresa iniciou o desenvolvimento dessa estrutura em 2011, por meio de sua divisão Google Brain. O objetivo era encontrar uma maneira de criar redes neurais mais rapidamente, de modo a incorporar a tecnologia em muitos aplicativos do Google.

Em 2015, a empresa decidiu abrir o código do TensorFlow, principalmente porque queria acelerar o progresso da IA, o que de fato aconteceu. Ao abrir o TensorFlow, o Google fez de sua tecnologia um padrão para o setor de desenvolvimento. O software foi baixado mais de 41 milhões de vezes e há mais de 1.800 colaboradores trabalhando nele atualmente. Na verdade, o TensorFlow Lite (para sistemas embarcados) está sendo executado em mais de 2 bilhões de dispositivos móveis.[13]

A onipresença da plataforma resultou em um grande ecossistema. Isso significa que há muitos complementos como TensorFlow Federated (para dados descentralizados), TensorFlow Privacy, TensorFlow Probability, TensorFlow Agents (para aprendizagem por reforço) e Mesh TensorFlow (para dados maciços).

O TensorFlow disponibiliza uma variedade de linguagens para criação de seus modelos, como Swift, JavaScript e R. Na maior parte dos casos, o mais comum é que

Python seja utilizada.

Em termos da estrutura básica, o TensorFlow recebe dados de entrada como uma matriz multidimensional, também conhecida como um tensor. Há um fluxo para ele, representado por um gráfico, à medida que os dados percorrem o sistema.

Quando comandos são informados ao TensorFlow, eles são processados usando um sofisticado kernel C++. Isso permite um desempenho muito maior; o que pode ser essencial, pois alguns modelos podem ser enormes.

TensorFlow pode ser usado para praticamente qualquer coisa quando se trata de IA. Aqui estão alguns dos modelos que ele tem alimentado:

- Pesquisadores do NERSC (National Energy Research Scientific Computing Center – Centro Nacional de Computação Científica para Pesquisa Energética) no Lawrence Berkeley National Laboratory criaram um sistema de deep learning para prever melhor climas extremos. Foi o primeiro modelo desse tipo que quebrou a barreira da computação (1 bilhão de bilhões de cálculos). Por causa disso, os pesquisadores ganharam o Prêmio Gordon Bell.[14]
- A Airbnb usou o TensorFlow para criar um modelo que categorizasse milhões de fotos, o que aprimorou a experiência dos hóspedes e levou a maiores conversões.[15]
- O Google usou o TensorFlow para analisar dados do telescópio espacial Kepler, da NASA. O resultado? Ao treinar uma rede neural, o modelo descobriu dois exoplanetas. O Google também disponibilizou o código à comunidade.[16]

O Google tem trabalhado no TensorFlow 2.0 e um objetivo fundamental é tornar o processo na API mais simples. Há também um recurso chamado Datasets (conjuntos de dados), que ajuda a simplificar a preparação de dados para modelos de IA.

Então, quais são alguns dos outros frameworks de IA? Vamos dar uma olhada:

- *PyTorch*: o Facebook é o desenvolvedor dessa plataforma, lançada em 2016. Como o TensorFlow, a principal linguagem para programar o sistema é Python. Embora ainda esteja nas fases iniciais, o PyTorch já é considerado o segundo framework mais utilizado, atrás somente do TensorFlow. Então, o que há de diferente nessa plataforma? O PyTorch tem uma interface mais intuitiva. A plataforma também permite a computação dinâmica dos gráficos. Isso significa que se pode facilmente fazer alterações nos modelos em tempo de execução, o que ajuda a acelerar o desenvolvimento. O PyTorch também permite o uso de diferentes tipos de CPUs e GPUs.
- *Keras*: embora TensorFlow e PyTorch sejam para especialistas em IA experientes, o Keras é para iniciantes. Com uma pequena quantidade de código – em Python – é possível criar redes neurais. Na documentação, observa-se: "Keras é uma API projetada para seres humanos, não máquinas. Ela coloca a experiência do

usuário em posição proeminente. Keras segue as melhores práticas para reduzir a carga cognitiva: oferece APIs consistentes e simples, minimiza o número de ações do usuário necessárias para aplicações comuns e fornece feedback claro e acionável sobre erros".[17] Há um guia "Getting Started" ("Iniciando") cuja leitura leva apenas 30 segundos! No entanto, a simplicidade não significa que a ferramenta não é poderosa. O fato é que é possível criar modelos sofisticados com o Keras. Para ter uma ideia, o TensorFlow integrou o Keras a sua própria plataforma. Mesmo para aqueles que são profissionais na IA, o sistema pode ser bastante útil para fazer experimentações iniciais com modelos.

Com o desenvolvimento da IA, há outra ferramenta bastante utilizada: Jupyter Notebook. Não se trata de uma plataforma ou ferramenta de desenvolvimento. Em vez disso, o Jupyter Notebook é um aplicativo web que torna mais fácil a codificação em Python e R para criar visualizações e importar sistemas de IA. Também é possível compartilhar facilmente seu trabalho com outras pessoas, semelhante ao que se faz no GitHub.

Durante os últimos anos, também surgiu uma nova categoria de ferramentas de IA, chamada machine learning automatizado ou autoML. Esses sistemas ajudam a lidar com processos como preparação de dados e seleção de recursos. Em grande parte, o objetivo é fornecer ajuda para as organizações que não contam com cientistas de dados e engenheiros de IA experientes. Trata-se de uma tendência de crescimento rápido denominada "cientista de dados cidadão" – ou seja, uma pessoa que não tem um forte histórico técnico, mas que ainda assim é capaz de criar modelos úteis.

Entre as ferramentas autoML disponíveis, temos H2O.ai, DataRobot e SaaS. Trata-se de sistemas intuitivos e que implementam a facilidade do recurso drag and drop (arrastar e soltar) no desenvolvimento de modelos. Como não deve ser nenhuma surpresa, gigantes da tecnologia como Facebook e Google criaram os próprios sistemas de autoML para suas equipes. O Facebook utiliza o Asimo, que ajuda a gerenciar treinamentos e testes de 300.000 modelos a cada mês.[18]

Para um estudo de caso de autoML, dê uma olhada na Lenovo Brasil. A empresa estava tendo dificuldade para criar modelos de machine learning para ajudar a prever e gerenciar a cadeia de suprimentos. Ela contava com duas pessoas que codificavam 1.500 linhas de código em R a cada semana – mas isso não era suficiente. O fato é que não seria rentável contratar mais cientistas de dados.

Por isso, a empresa implantou o DataRobot. Ao automatizar vários processos, a Lenovo Brasil conseguiu criar modelos com mais variáveis, o que levou a melhores resultados. Em apenas alguns meses, o número de usuários do DataRobot passou de dois para dez.

A Tabela 8.1 mostra alguns outros resultados.[19]

Tabela 8.1 – Resultados da implementação de um sistema autoML

Tarefas	Antes	Depois
Criação do modelo	4 semanas	3 dias
Produção dos modelos	2 dias	5 minutos
Precisão das previsões	<80%	87,5%

Muito bom, certo? Com certeza. Entretanto, ainda há ressalvas. No caso da Lenovo Brasil, a empresa teve o benefício de contar com cientistas de dados qualificados, que entendiam as nuances da criação de modelos.

No entanto, se uma ferramenta autoML for usada sem essa experiência, é possível ter problemas sérios. Há uma boa chance de que sejam criados modelos com viés ou dados defeituosos. Em última instância, os resultados podem revelar-se muito piores do que não usar a IA! Devido a isso, o DataRobot realmente exige que um novo cliente disponha de um engenheiro de campo e um cientista de dados dedicados trabalhando com a empresa ao longo do primeiro ano.[20]

Também existem plataformas low-code (pouco código) que provaram ser úteis na aceleração do desenvolvimento de projetos de IA. Um dos líderes nessa área é a Appian, que oferece a ousada garantia de ir "da ideia ao app em oito semanas".

Com essa plataforma, é possível configurar facilmente uma estrutura de dados limpa. Há até mesmo sistemas para ajudar na orientação do processo, inclusive com alertas para problemas. Sem dúvida, isso fornece uma base sólida para a construção de um modelo. Plataformas low-code também ajudam de outras maneiras. É possível testar várias plataformas de IA – de empresas como Google, Amazon ou Microsoft, por exemplo – para ver qual delas apresenta melhor desempenho. Em seguida, é possível criar o aplicativo com uma interface moderna e implantá-lo na Web ou em dispositivos móveis.

Para ter noção do poder das plataformas low-code, dê uma olhada no que a KPMG fez com a tecnologia. A empresa conseguiu ajudar seus clientes a interromper o uso da LIBOR[21] em empréstimos. Em primeiro lugar, a KPMG usou a própria plataforma de IA, chamada Ignite, para receber dados não estruturados e usar machine learning e natural language processing para corrigir os contratos. Em seguida, a empresa usou o Appian para ajudar nas atividades de compartilhamento de documentos, regras de negócios personalizáveis e relatórios em tempo real.

Tal processo – quando executado de forma manual – pode facilmente tomar milhares de horas, com uma taxa de erro entre 10% e 15%. Ao usar o Ignite e o Appian, no entanto, a precisão foi superior a 96%. O tempo de processamento dos documentos foi de segundos.

Implante e monitore o sistema de IA

Mesmo quando o modelo de IA construído funciona adequadamente, ainda há muito trabalho a fazer. É preciso encontrar maneiras de implantá-lo e monitorá-lo.

Isso requer o gerenciamento de mudanças, que é sempre complexo e difícil. A IA é diferente de uma implementação típica de TI, uma vez que envolve o uso de previsões e insights para a tomada de decisões. Isso significa que as pessoas precisarão repensar como interagem com a tecnologia.

Considere também que existe a possibilidade de os usuários finais serem pessoas não técnicas, sejam eles funcionários ou consumidores. É por isso que muito esforço deve ser feito para tornar o modelo de IA o mais fácil possível. Por exemplo, se você construiu um sistema de marketing online, pode querer limitar as opções para o usuário – digamos, apenas quatro ou cinco delas.

Por quê? Se houver muitas, então os usuários podem ficar frustrados e sequer saber por onde começar. Isso tudo faz parte do chamado problema da "paralisia da análise". Quando isso acontece, haverá inevitavelmente pouca adoção do modelo de IA, o que resultará no impedimento do progresso.

Outra boa estratégia é usar visualizações interativas. Em outras palavras, é possível visualizar facilmente como as tendências mudam ajustando algumas variáveis. Também é possível permitir o clique em uma determinada parte do gráfico para que mais detalhes sejam exibidos.

Também é essencial criar uma documentação. É necessário, no entanto, que ela agrupe mais do que apenas materiais escritos. Por exemplo, uma abordagem eficaz é desenvolver tutoriais em vídeo. Esse esforço ajudará muito na forte adoção da tecnologia.

Recomenda-se que a implantação inicial seja limitada. Talvez isso possa representar um pequeno grupo de usuários beta e uma pequena seção da base de clientes. Também deve haver avisos de que o modelo de IA está nos estágios iniciais e pode apresentar erros.

Trata-se, portanto, de uma fase de aprendizagem. O que funciona? O que deve ser removido? Onde as coisas podem ser melhoradas?

Esse é definitivamente um processo iterativo que não deve ser realizado com pressa.

Então, uma vez que o modelo de IA esteja pronto para implantação completa, deve haver apoio suficiente no local e alguém para liderar a gestão do projeto. Também deve haver reconhecimento da vitória da equipe. Espera-se que o elogio venha dos níveis mais altos da empresa, o que ajudará a incentivar mais e mais inovações.

Há uma variedade de plataformas automatizadas para ajudar a agilizar o processo de fluxo de trabalho, como o Alteryx. A visão da empresa é democratizar a ciência e

a análise de dados, independentemente de alguém ter ou não experiência técnica. O sistema Alteryx lida com as principais áreas do processo: descoberta e preparação de dados, análise e implantação. Tudo isso é feito com ferramentas do tipo "arrastar e soltar", sem código. Além disso, muitos dos clientes da empresa são operadores não tecnológicos como Hyatt, Unilever e Kroger.

Mais uma vez, o desenvolvimento da IA é realmente uma jornada – e sua estratégia certamente mudará. Isso é inevitável. De acordo com Kurt Muehmel, vice-presidente de engenharia de vendas da Dataiku:[22]

> *O que as empresas às vezes não conseguem perceber é que o caminho para a IA é uma evolução a longo prazo não só da tecnologia, mas da forma como a empresa colabora e trabalha em conjunto. Assim, além da educação, um dos principais componentes de uma estratégia de IA deve ser o gerenciamento geral de mudanças. É importante criar roteiros de curto e longo prazos do que será realizado com a análise, talvez preditiva, depois talvez com machine learning e, finalmente, como um objetivo de longo prazo – a IA, assim como cada roteiro que impacta várias partes do negócio e também as pessoas que fazem parte dessas linhas de negócios e seu trabalho cotidiano.*

Conclusão

Como mostrado neste capítulo, ao se aproximar da implementação da IA, é fundamental analisar duas alternativas. A primeira é maximizar o uso de um sistema terceirizado que implemente essa tecnologia. É imprescindível que haja foco na qualidade dos dados; se não, os resultados provavelmente estarão fora do esperado.

A segunda alternativa é criar um projeto de IA baseado nos próprios dados da sua empresa. Para alcançar o sucesso, deve haver uma equipe forte que conte com uma mistura de conhecimentos técnicos, de negócios e de domínio. É provável também que haja a necessidade de algum treinamento em IA. Deve ser assim mesmo com aqueles que possuem formação em ciência de dados e engenharia.

A partir daqui, não deve haver pressa nas etapas do projeto: avaliar o ambiente de TI, criar um objetivo de negócios claro, limpar os dados, selecionar as ferramentas e plataformas adequadas, criar o modelo de IA e implantar o sistema. Com projetos iniciais, inevitavelmente haverá desafios, por isso é fundamental ser flexível. O esforço costuma valer a pena.

Principais aprendizados

- Mesmo as melhores empresas têm dificuldades com a implementação da IA. Por conta disso, deve haver muito cuidado, diligência e planejamento. Também é

importante perceber que o fracasso é comum.

- Há duas maneiras principais de usar a IA em uma empresa: por meio de um software de um fornecedor ou de um modelo interno. Esse último é muito mais difícil e requer um grande compromisso da organização.
- Mesmo ao optar por aplicativos de IA prontos para o uso, ainda há muito trabalho a ser feito. Por exemplo, se os funcionários não estiverem inserindo corretamente os dados, os resultados provavelmente não servirão ao propósito esperado.
- A educação é fundamental com uma implementação de IA, mesmo para engenheiros experientes. Há excelentes recursos de treinamento online para ajudar nessa tarefa.
- Esteja atento aos riscos das implementações de IA, como viés, segurança e privacidade.
- Algumas das partes-chave do processo de implementação de IA incluem: identificar um problema para resolver; montar uma equipe forte; selecionar as ferramentas e plataformas adequadas; criar o modelo de IA e implantá-lo e monitorá-lo.
- Ao desenvolver um modelo, veja como a tecnologia se relaciona com as pessoas. O fato é que as pessoas podem ser muito melhores em determinadas tarefas.
- Formar a equipe não é fácil, então não apresse o processo. Conte com um líder que tenha boa experiência em negócios ou operacional, com algumas habilidades técnicas.
- É bom experimentar com as várias ferramentas de IA. No entanto, antes de fazer isso, certifique-se de fazer uma avaliação de TI.
- Algumas das ferramentas de IA mais populares são TensorFlow, PyTorch, Python, Keras e Jupyter Notebook.
- As ferramentas automatizadas de machine learning, ou autoML, ajudam a lidar com processos como preparação de dados e seleção de recursos para modelos de IA. O foco está naqueles que não têm habilidades técnicas.
- A implantação do modelo de IA envolve mais do que apenas dimensionamento. Também é fundamental que o sistema seja fácil de usar, de modo a permitir maior adoção.

---

1 www.cnbc.com/2019/03/21/why-facebooks-ai-didnt-detect-the-new-zealand-mosque-shooting-video.html

2 https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/

3 N.T.: Potencial cliente para a marca.

4 Entrevista realizada pelo autor em abril de 2019 com Ricky Thakrar, evangelista de experiência do consumidor na Zoho.

5 www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

6 www.theverge.com/2019/4/4/18296113/google-ai-ethics-board-ends-controvrsy-kay-coles-james-heritage-foundation

7 Entrevista realizada pelo autor em março de 2019 com Kalyan Kumar, vice-presidente corporativo e CTO global da HCL Technologies.

8 https://hbr.org/2019/02/how-to-choose-your-first-ai-project

9 Entrevista realizada pelo autor em abril de 2019 com Erik Schluntz, cofundador e CTO da Cobalt Robotics.

10 https://www.symrise.com/newsroom/article/breaking-new-fragrance-groundwith-artificial-intelligence-ai-ibm-research-and-symrise-are-workin/

11 www.geekwire.com/2018/amazon-tops-600k-worldwide-employees-1st-time-13-jump-year-ago/

12 www.python.org/dev/peps/pep-0020/

13 https://medium.com/tensorflow/recap-of-the-2019-tensorflow-dev-summit-1b5ede42da8d summit-1b5ede42da8d

14 www.youtube.com/watch?v=p45kQklIsd4&feature=youtu.be

15 www.youtube.com/watch?v=tPb2u9kwh2w&feature=youtu.be

16 https://ai.googleblog.com/2018/03/open-sourcing-hunt-for-exoplanets.html

17 https://keras.io/

18 https://www.aimlmarketplace.com/technology/machine-learning/the-rise-of-automated-machine-learning

19 https://3gp10c1vpy442j63me73gy3s-wpengine.netdna-ssl.com/wp-content/uploads/2018/08/Lenovo-Case-Study.pdf

20 www.wsj.com/articles/yes-you-too-can-be-an-ai-expert-11554168513

21 N.T.: "Taxa média interbancária contra a qual um grupo representativo de bancos se propõe a efetuar empréstimos mutuamente no mercado monetário de Londres." Fonte: https://pt.global-rates.com/taxa-de-juros/libor/libor.aspx

22 Entrevista realizada pelo autor em abril de 2019 com Kurt Muehmel, vice-presidente de engenharia de vendas na Dataiku.

CAPÍTULO 9

Futuro da IA

Prós e contras

Na conferência Web Summit no final de 2017, o lendário físico Stephen Hawking ofereceu sua opinião sobre o futuro da IA. Por um lado, ele estava esperançoso e acreditava que a tecnologia poderia superar a inteligência humana. Isso provavelmente significa que muitas doenças horríveis serão curadas e talvez haja maneiras de lidar com problemas ambientais, incluindo as mudanças climáticas.

No entanto, havia o lado negro também. Hawking falou sobre como a tecnologia tinha o potencial de ser o "pior evento na história da nossa civilização".[1] Alguns dos problemas que podem surgir incluem o desemprego em massa e até mesmo robôs assassinos. Por causa disso, ele pediu por maneiras de controlar a IA.

As ideias de Hawking certamente não são solitárias. Empresários proeminentes de tecnologia como Elon Musk e Bill Gates também expressaram profundas preocupações com relação à IA.

Apesar deles, há muitos outros decididamente otimistas, se não entusiastas. Masayoshi Son, CEO da SoftBank e Gerente do fundo de risco Vision – de US$ 100 bilhões, é um deles. Em entrevista à CNBC, ele proclamou que, dentro de 30 anos, teremos carros voadores, as pessoas vão viver muito mais tempo e teremos curado muitas doenças. Ele também ressaltou que o foco principal de seu fundo está na IA.[2]

Ok, então, quem está certo? O futuro será distópico ou utópico? Ou será que vai estar em algum lugar no meio dos dois? Bem, prever novas tecnologias é extremamente difícil, se não impossível. Aqui estão alguns exemplos de previsões que passaram longe do acerto:

- Thomas Edison declarou que a AC (corrente alternada) falharia.[3]
- Em seu livro *The Road Ahead* (A estrada do futuro), publicado no final de 1995, Bill Gates não mencionou a Internet.
- Em 2007, Jim Balsillie, CEO adjunto da Research in Motion (criadora do dispositivo BlackBerry), disse que o iPhone seria pouco comprado.[4]
- No icônico filme de ficção científica *Blade Runner* – lançado em 1982 e ambientado em 2019 – havia muitas previsões erradas, como cabines com telefones de vídeo e androides (ou "replicantes") que eram quase indistinguíveis dos seres humanos.

Apesar de tudo isso, uma coisa é certa: nos próximos anos, veremos muita inovação e mudança na IA. Isso parece inevitável, especialmente porque enormes quantias continuam a ser investidas na indústria.

Então, vamos dar uma olhada em algumas das áreas que provavelmente causarão um impacto enorme sobre a sociedade.

Carros autônomos

Quando se trata de IA, uma das áreas mais abrangentes são os carros autônomos. Curiosamente, essa categoria não é de todo nova. Sim, ela tem sido marca registrada de muitas histórias de ficção científica por várias décadas! Entretanto, há algum tempo existem muitos exemplos reais dessa inovação, como os seguintes:

- *Stanford Cart*: seu desenvolvimento começou no início dos anos 1960 e o objetivo original era criar um veículo de controle remoto para missões lunares. Contudo, os pesquisadores acabaram mudando o foco e desenvolveram um veículo autônomo básico, que usava câmeras e IA para navegação. Embora tenha sido uma conquista de destaque para a época, não foi prático, pois exigia mais de 10 minutos para planejar qualquer movimento!
- *Ernst Dickmanns*: brilhante engenheiro aeroespacial alemão, ele voltaria sua atenção para a ideia de converter uma van Mercedes em um veículo autônomo... em meados da década de 1980. Ele conectou câmeras, sensores e computadores e foi criativo na forma como usou o software, fazendo com que o processamento gráfico se concentrasse apenas em detalhes visuais importantes para economizar energia. Ao fazer tudo isso, foi capaz de desenvolver um sistema que controlava direção, pedal do acelerador e freios de um carro. Testou o Mercedes em uma rodovia de Paris – em 1994 – e o veículo percorreu mais de 600 milhas, com uma velocidade de até 81 MPH. No entanto, o financiamento do projeto foi retirado porque não estava claro se poderia haver comercialização em tempo hábil.[5] Também não ajudou o fato de que a IA estava entrando em outro inverno.

O ponto de inflexão para carros autônomos veio em 2004. O principal catalisador foi a Guerra do Iraque, que estava causando um terrível impacto nos soldados americanos. Para a DARPA, os veículos autônomos poderiam ser a solução.

Entretanto, a agência enfrentou muitos desafios difíceis. Por conta disso, criou-se um concurso, apelidado de DARPA Grand Challenge, em 2004, que ofereceu o grande prêmio de US$ 1 milhão para incentivar uma inovação mais ampla. O evento envolveu uma corrida de 150 milhas no deserto de Mojave, mas, infelizmente, não foi encorajador, já que os carros apresentaram um desempenho horrível. Nenhum deles terminou a corrida!

No entanto, o evento serviu para estimular outras inovações na área. No ano

seguinte, cinco veículos terminaram a corrida. Então, em 2007, os carros estavam tão avançados que foram capazes de realizar ações como retornos e mudanças de pista.

Por meio desse processo, a DARPA viabilizou a criação dos principais componentes para veículos autônomos:

- *Sensores*: incluem sistemas de radar e ultrassônicos que podem detectar veículos e outros obstáculos, como meio-fio.
- *Câmeras de vídeo*: podem detectar placas de trânsito, semáforos e pedestres.
- *LIDAR (Light Detection and Ranging – detecção de luz e área)*: dispositivo que geralmente está no topo de um carro autônomo e dispara feixes de laser para medir o ambiente. Os dados são então integrados aos mapas existentes.
- *Computador*: ajuda com o controle do carro, incluindo direção, aceleração e frenagem. O sistema aproveita a IA para aprender, mas também tem regras incorporadas para evitar objetos, obedecer às leis e assim por diante.

Quando se trata de carros autônomos, há muita confusão com relação ao que o termo "autônomo" realmente significa. Será que ele se aplica a carros que se dirigem completamente sozinhos – ou deve haver um motorista humano?

Para entender as nuances, existem cinco níveis de autonomia:

- *Nível 0:* um ser humano controla todos os sistemas.
- *Nível 1*: os computadores controlam funções limitadas – como velocidade de cruzeiro ou frenagem –, mas apenas uma de cada vez.
- *Nível 2:* esse tipo de carro pode automatizar duas funções.
- *Nível 3:* o carro automatiza todas as funções de segurança, mas o motorista pode intervir se algo der errado.
- *Nível 4:* o carro geralmente pode dirigir sozinho, mas há casos em que um ser humano deve participar.
- *Nível 5:* é o Santo Graal, no qual o carro é completamente autônomo.

A indústria automobilística é um dos maiores mercados, e a IA provavelmente desencadeará mudanças dolorosas. Considere que o transporte é a segunda maior despesa familiar, atrás apenas da habitação, e duas vezes maior do que a saúde. Outra coisa a ter em mente é que o carro é usado apenas cerca de 5% do tempo, uma vez que geralmente permanece estacionado em algum lugar.[6]

À luz da enorme oportunidade de melhorias, não deve ser nenhuma surpresa que a indústria de automóveis autônomos tenha presenciado investimentos significativos. Não só capitalistas de risco vem investindo em uma variedade de startups, mas também há muita inovação por parte de montadoras tradicionais como Ford, GM e BMW.

Então, quando veremos esse setor se tornar popular? As estimativas variam muito.

De acordo com um estudo da Allied Market Research, espera-se que o mercado alcance US$ 556,67 bilhões até 2026, o que representaria uma taxa de crescimento anual de 39,47%.[7]

Há muito trabalho a ser feito. "Na melhor das hipóteses, ainda estamos a anos de distância de um veículo que não requer um volante", disse Scott Painter, CEO e fundador da Fair. "Os carros ainda precisarão ser segurados, reparados e submetidos à manutenção, mesmo que você tenha voltado do futuro em um Delorean e trazido o manual de como fazer esses carros totalmente autônomos. Fazemos 100 milhões de carros por ano, dos quais 16 milhões estão nos Estados Unidos. Supondo que se queira que todo esse fornecimento tivesse recursos de inteligência artificial, ainda levaria 20 anos até que tivéssemos mais carros na estrada que incluíssem todos os diferentes níveis de IA se comparado ao número de veículos que não dispõem dessa tecnologia."[8]

Existem muitos outros fatores a considerar. Afinal, é fato que conduzir um veículo é complexo, especialmente em áreas urbanas e suburbanas. E se um sinal de trânsito for alterado ou mesmo manipulado? E se um carro autônomo tiver de lidar com um dilema, como precisar decidir entre bater em um carro que se aproxima ou mergulhar em um meio-fio onde pode haver pedestres? Todos esses casos são extremamente difíceis.

As tarefas noturnas aparentemente simples também podem ser de difícil execução. John Krafcik, CEO da Waymo do Google, ressalta que os estacionamentos são um excelente exemplo.[9] Eles exigem encontrar vagas disponíveis, evitando outros carros e pedestres (que podem ser imprevisíveis) e movendo-se pelo espaço.

A tecnologia, no entanto, é apenas um dos desafios com veículos autônomos. Aqui estão alguns outros a considerar:

- *Infraestrutura:* as cidades e os bairros são construídos para carros convencionais. Ao incluir veículos autônomos, provavelmente haverá muitos problemas logísticos. Como um carro antecipa as ações dos motoristas humanos? Na verdade, pode haver uma necessidade de instalar sensores ao longo das vias. Outra opção é dispor de estradas separadas para veículos autônomos. É provável que os governos também precisem adequar a educação dos motoristas, fornecendo orientações sobre como interagir com veículos autônomos enquanto estiverem na estrada.
- *Legislação:* esse é um grande curinga. Em grande parte, ele pode ser o maior impedimento, já que os governos tendem a trabalhar lentamente e são resistentes às mudanças. Os Estados Unidos também são um país altamente litigioso – o que pode ser outro fator que poderia atrasar o desenvolvimento.
- *Adoção:* os veículos autônomos provavelmente não serão baratos, visto que

sistemas como o LIDAR são caros. Esse será certamente um fator limitante. Entretanto, ao mesmo tempo, há indícios de ceticismo do público em geral. De acordo com uma pesquisa da AAA, cerca de 71% dos entrevistados disseram que têm medo de andar em um veículo autônomo.[10]

Diante de tudo isso, a fase inicial de veículos autônomos provavelmente servirá para situações controladas, por exemplo, para caminhões, mineração ou ônibus espaciais. Um caso disso é a Suncor Energy, que usa máquinas autônomas para escavar vários locais no Canadá.

As redes de compartilhamento de viagens – como Uber e Lyft – podem ser outro ponto de partida. Esses serviços são bastante estruturados e acessíveis ao público.

Lembre-se de que a Waymo vem testando um serviço de táxi autônomo em Phoenix (semelhante ao sistema de compartilhamento de viagens do Uber, mas os carros contam com sistemas autônomos). Veja o que diz uma postagem no blog da empresa:

> *Começaremos dando aos passageiros acesso ao nosso aplicativo. Eles podem usá-lo para chamar nossos veículos autônomos 24 horas por dia, 7 dias por semana. É possível andar por várias cidades na área metropolitana de Phoenix, incluindo Chandler, Tempe, Mesa e Gilbert. Seja para uma noite divertida ou apenas para dar uma pausa na direção, nossos motoristas recebem sempre os mesmos veículos limpos e nosso condutor Waymo com mais de 10 milhões de quilômetros de experiência em estradas públicas. Os passageiros verão estimativas de preço antes de aceitarem a viagem com base em fatores como o tempo e a distância até o seu destino.*[11]

A Waymo descobriu que a educação é essencial, porque os motoristas têm muitas perguntas. Para lidar com isso, a empresa criou um sistema de bate-papo no aplicativo por meio do qual é possível entrar em contato com um profissional da área de suporte. O painel do carro também tem uma tela que fornece detalhes da viagem.

De acordo com o blog, "O feedback dos motoristas continuará a ser vital a cada passo do caminho".[12]

Estados Unidos x China

A rápida ascensão da China tem sido surpreendente. Dentro de alguns anos, a economia do país pode ser maior do que a dos Estados Unidos e uma parte fundamental do crescimento se deverá à IA. O governo chinês estabeleceu a meta ambiciosa de investir US$ 150 bilhões nessa tecnologia até 2030.[13] Enquanto isso, seguem os grandes investimentos por parte de empresas como Baidu, Alibaba e Tencent.

Mesmo que a China seja frequentemente considerada menos criativa ou inovadora do que o Vale do Silício – e muitas vezes seja rotulada de "imitadora" –, essa percepção pode se provar um mito. Um estudo do Instituto Allen de Inteligência Artificial destaca que a China deverá superar os Estados Unidos nos trabalhos técnicos mais citados sobre IA.[14]

O país tem algumas outras vantagens, que o especialista em IA e capitalista de risco Kai-Fu Lee apontou em seu provocativo livro *AI Superpowers: China, Silicon Valley, and the New World Order (As Superpotências da Inteligência Artificial – A China, Silicon Valley e a Nova Ordem Mundial)*[15]:

- *Entusiasmo:* na década de 1950, o lançamento do Sputnik pela Rússia despertou o interesse de pessoas nos Estados Unidos em se tornarem engenheiros do programa espacial. Algo semelhante aconteceu na China. Quando o melhor jogador de Go do país, Ke Jie, perdeu para o sistema AlphaGo AI, houve um despertar. O resultado é que o ocorrido inspirou muitos jovens a seguirem carreira na IA.
- *Dados*: com uma população de mais de 1,3 bilhão de habitantes, a China é rica em dados (há mais de 700 milhões de usuários de Internet). Contudo, o governo autoritário do país também é crítico, e a privacidade não é considerada particularmente importante. Isso significa que há muito mais margem de manobra no desenvolvimento de modelos de IA. Por exemplo, em um artigo publicado na *Nature Medicine*, pesquisadores chineses tiveram acesso a dados sobre 600.000 pacientes para realizar um estudo de saúde.[16] Embora ainda esteja nos estágios iniciais, o estudo mostrou que um modelo de IA foi capaz de diagnosticar eficazmente doenças da infância, como gripe e meningite.
- *Infraestrutura*: como parte dos planos de investimento do governo chinês, o país tem se concentrado na criação de cidades de última geração que permitam carros autônomos e outros sistemas de IA. Houve também uma implantação agressiva de redes 5G.

Quanto aos Estados Unidos, o governo tem sido muito mais reticente com relação à IA. O presidente Trump assinou uma ordem executiva – chamada de "American AI Initiative" ("Iniciativa americana de IA") – para incentivar o desenvolvimento da tecnologia, mas os termos são vagos e não está claro o montante de dinheiro que será direcionado a ela.

Desemprego causado pelas mudanças tecnológicas

O conceito de desemprego tecnológico, que ganhou notoriedade com o famoso economista John Maynard Keynes durante a Grande Depressão, explica como as inovações podem levar à perda de emprego no longo prazo. No entanto, a evidência

disso tem sido evasiva. Não obstante o fato de que a automação tem impactado severamente indústrias como a fabricação, muitas vezes há uma transição da força de trabalho à medida que as pessoas se adaptam.

A revolução da IA poderia ser diferente? Certamente poderia. Por exemplo, o governador da Califórnia Gavin Newsom teme que seu estado experimente um desemprego maciço de caminhoneiros e profissionais de armazenagem – e em breve.[17]

Veja este outro exemplo: a Harvest CROO Robotics construiu um robô, chamado Harv, que pode colher morangos e outras plantas sem ganhar arranhões. É verdade que ele ainda está em fase experimental, mas o sistema está melhorando rapidamente. A expectativa é que o robô faça o trabalho de 30 pessoas[18] e, claro, não haverá salários a pagar ou exposição à responsabilidade trabalhista.

A IA pode significar mais do que substituir empregos que requerem pouca qualificação. Já há sinais de que a tecnologia poderia ter um impacto significativo sobre as profissões de colarinho branco. A verdade é que há ainda mais incentivo para automatizar esses empregos, visto que eles têm um salário mais elevado.

Outra categoria que pode precisar enfrentar a perda de emprego causada pela IA é o campo jurídico, pois várias startups estão se lançando para o mercado – como Lawgood, NexLP e RAVN ACE. As soluções estão focadas na automação de áreas como pesquisa jurídica e revisão de contratos.[19] Embora os sistemas estejam longe da perfeição, eles certamente conseguem processar um volume muito maior do que as pessoas – e também podem ficar mais inteligentes à medida que são usados.

É verdade que o mercado de trabalho global é dinâmico, e novos tipos de carreiras serão criados. Também é provável que haja inovações de IA que sejam assistivas para os funcionários – tornando seu trabalho mais fácil de ser feito. Por exemplo, a startup de software Measure Square tem conseguido usar algoritmos sofisticados para converter plantas em papel em plantas de piso digitalmente interativas. Devido a isso, tem sido mais fácil começar os projetos e concluí-los a tempo.

No entanto, à luz do potencial impacto transformador da IA, parece razoável que haja um impacto adverso em uma ampla gama de setores. Talvez um prenúncio disso seja o que aconteceu com a perda de postos de trabalho na manufatura no período de 1960 a 1990. De acordo com o Pew Research Center, praticamente não houve crescimento real dos salários nos últimos 40 anos.[20] Durante esse período, os Estados Unidos também experimentaram um fosso crescente em sua prosperidade. O economista Gabriel Zucman, da Berkeley, estima que 0,1% da população controla quase 20% da riqueza.[21]

Ainda existem medidas que podem ser tomadas. Em primeiro lugar, os governos podem procurar fornecer educação e assistência na transição. Com o ritmo da

mudança no mundo de hoje, será necessário realizar uma renovação contínua das habilidades para a maioria das pessoas. O CEO da IBM, Ginni Rometty, observou que a IA mudará todos os empregos nos próximos 5 a 10 anos. A propósito, sua empresa viu uma redução de 30% no quadro de funcionários no departamento de RH por conta de automação.[22]

Em seguida, há algumas pessoas que defendem o rendimento básico, que fornece um salário mínimo a todos. Isso certamente suavizaria parte da desigualdade, mas também apresenta desvantagens. As pessoas definitivamente obtêm orgulho e satisfação com suas carreiras. Então, qual poderia ser o moral de uma pessoa se ela não consegue encontrar um emprego? Pode ter um impacto profundo.

Por fim, existem boatos de algum tipo de imposto de IA. Isso certamente recuperaria os grandes ganhos das empresas que se beneficiam da tecnologia. Entretanto, dado o seu poder, provavelmente seria difícil aprovar esse tipo de legislação.

Militarização da IA

O Laboratório de Pesquisa da Força Aérea está trabalhando em protótipos de algo chamado Skyborg. É como em *Star Wars*. Pense no Skyborg como sendo um R2-D2, que serve como um braço direito de IA para um avião de combate, ajudando a identificar alvos e ameaças.[23] O robô de IA também é capaz de assumir o controle se o piloto estiver incapacitado ou distraído. A Força Aérea está, inclusive, considerando usar a tecnologia para operar drones.

Legal, não é? Certamente. No entanto, há uma questão importante: usando IA, os seres humanos podem finalmente ser retirados do circuito ao tomar decisões de vida ou morte no campo de batalha? Isso poderia levar a mais derramamento de sangue? Talvez as máquinas tomem as decisões erradas – causando ainda mais problemas?

Muitos pesquisadores e empreendedores de IA estão preocupados. Por conta disso, mais de 2.400 pessoas assinaram uma declaração que exige a proibição dos chamados "robôs assassinos".[24]

Até as Nações Unidas estão considerando algum tipo de proibição. Os Estados Unidos, entretanto, junto com Austrália, Israel, Reino Unido e Rússia, têm resistido a esse movimento.[25] Como resultado, pode estar surgindo uma verdadeira corrida armamentista por IA.

De acordo com um estudo da RAND Corporation, existe até mesmo a possibilidade de que a tecnologia leve a uma guerra nuclear, digamos, até o ano de 2040. Como? Os autores observam que a IA pode facilitar o direcionamento de submarinos e sistemas de mísseis móveis. De acordo com o relatório:

> *As nações podem ser tentadas a buscar recursos de primeiro ataque como forma de obter poder de barganha sobre seus oponentes, mesmo que não tenham intenção de*

realizar um ataque, dizem os pesquisadores. Isso prejudica a estabilidade estratégica porque, mesmo que o estado que possui esses recursos não tenha intenção de usá-los, o adversário não tem como ter certeza disso. [26]

No curto prazo, contudo, a IA provavelmente terá maior impacto na guerra de informações; o que ainda pode ser altamente destrutivo. Tivemos um vislumbre disso quando o governo russo interferiu nas eleições presidenciais americanas de 2016. A abordagem era de baixa tecnologia, uma vez que usava fazendas de redes sociais para disseminar notícias falsas, mas as consequências foram significativas.

À medida que a IA se torna mais poderosa e acessível, é provável que aumente esse tipo de campanha. Por exemplo, sistemas de deepfake podem facilmente criar fotos e vídeos realistas de pessoas que poderiam ser usadas para espalhar rapidamente as mensagens.

Desenvolvimento de novos medicamentos

Os avanços no desenvolvimento de novos medicamentos têm sido quase milagrosos, visto que agora existe cura para doenças intratáveis, como a hepatite C, e continuam os progressos com uma variedade de tipos de câncer. É claro que muito ainda precisa ser feito. O fato é que as indústrias farmacêuticas estão tendo mais problemas para descobrir novos tratamentos. Veja este exemplo: em março de 2019, a Biogen anunciou que um de seus medicamentos para o Mal de Alzheimer, que estava em experimentos de fase III, não mostrou resultados significativos. No noticiário, foi relatado que as ações da empresa caíram 29%, eliminando US$ 18 bilhões de seu valor de mercado.[27]

Considere que o desenvolvimento tradicional de medicamentos costuma envolver muita tentativa e erro, o que pode ser demorado. Então, poderia haver uma maneira melhor?

Cada vez mais, os pesquisadores estão procurando ajuda na IA. Uma variedade de startups está surgindo e se concentrando na oportunidade.

Uma delas é a Insitro. A empresa, que começou em 2019, teve pouca dificuldade para arrecadar impressionantes US$ 100 milhões em sua rodada da Série A. Entre os investidores estão Alexandria Venture Investments, Bezos Expeditions (empresa de investimentos de Jeff Bezos, da Amazon.com), Mubadala Investment Company, Two Sigma Ventures e Verily.

Mesmo que a equipe seja relativamente pequena – com cerca de 30 funcionários –, todos são pesquisadores brilhantes de áreas como ciência de dados, deep learning, engenharia de software, bioengenharia e química. O CEO e fundador, Daphne Koller, tem a rara mistura de experiência em ciência da computação avançada e ciências da saúde, e já liderou a Calico, empresa de saúde do Google.

Como prova da proeza da Insitro, a empresa estabeleceu uma parceria com a megaoperadora de medicamentos Gilead. O acordo envolve pagamentos potenciais de mais de US$ 1 bilhão para pesquisas sobre esteato-hepatite não alcoólica (NASH – nonalcoholic steatohepatitis), uma doença hepática grave.[28] Um fato importante é que a Gilead conseguiu agrupar uma grande quantidade de dados que pode ser usada para treinar os modelos. Isso será feito usando células fora do corpo de uma pessoa – ou seja, com um sistema in vitro. A Gilead tem certa urgência em procurar abordagens alternativas desde que um de seus tratamentos NASH, o selonsertib, falhou em seus ensaios clínicos (era para aqueles que tinham a doença nos estágios avançados).

A promessa da IA é que ela acelerará o desenvolvimento de medicamentos, visto que o deep learning deve ser capaz de identificar padrões complexos. A tecnologia também pode vir a ser útil no desenvolvimento de tratamentos personalizados – voltados para a composição genética de uma pessoa – o que pode ser fundamental para a cura de certas doenças.

Independentemente disso, talvez seja melhor moderar as expectativas. Haverá grandes obstáculos com os quais lidar à medida que a indústria de saúde sofrer mudanças, porque haverá um aumento da educação para IA. Isso levará tempo e provavelmente haverá resistência.

Também é preciso lembrar que o deep learning costuma ser uma "caixa preta" quando se trata de entender como os algoritmos realmente funcionam. Essa característica pode dificultar a obtenção de aprovação regulatória para novos medicamentos, já que o FDA se concentra em relações causais.

Por fim, o corpo humano é altamente sofisticado e ainda estamos aprendendo sobre como ele funciona. Além disso, como vimos com inovações como a decodificação do Genoma Humano, geralmente leva um tempo considerável para entender novas abordagens.

Como um sinal das complexidades, considere a situação do Watson da IBM. Mesmo que a empresa conte com alguns dos pesquisadores de IA mais talentosos e tenha investido bilhões na tecnologia, ela anunciou recentemente que não venderia mais o Watson para fins de desenvolvimento de medicamentos.[29]

Governo

Um artigo do site Bloomberg.com em abril de 2019 causou grande agitação. Ele descreveu os bastidores de como a Amazon.com gerencia a Alexa, seu sistema de IA de alto-falante.[30] Embora grande parte de seu funcionamento seja baseado em algoritmos, há também milhares de pessoas que analisam clipes de voz para ajudar a melhorar os resultados. Muitas vezes, o foco está em lidar com as nuances de gírias e

dialetos regionais, que têm sido difíceis para algoritmos de deep learning.

Claro que é natural que as pessoas se perguntem: meu alto-falante inteligente está realmente me ouvindo? Minhas conversas são privadas?

A Amazon.com foi rápida em se pronunciar e esclarecer que tem regras e requisitos rigorosos. No entanto, isso despertou ainda mais preocupação! De acordo com a publicação do Bloomberg.com, os revisores de IA às vezes ouviam clipes que envolviam atividades potencialmente criminosas, como agressão sexual. Aparentemente, a Amazon tem uma política de não interferir.

À medida que a IA se tornar mais difundida, ouviremos mais desses tipos de histórias. Para a maior parte delas, não haverá respostas claras. Algumas pessoas podem, em última análise, decidir não comprar produtos de IA. No entanto, esse será provavelmente um pequeno grupo. Afinal, mesmo com a variedade de problemas de privacidade no Facebook, não houve um declínio no crescimento de usuários.

O que deve acontecer é que, muito provavelmente, os governos começarão a se pronunciar com relação às questões de IA. Um grupo de congressistas patrocinou um projeto de lei, chamado Algorithmic Accountability Act (Lei de Responsabilidade Algorítmica), que visa obrigar as empresas a auditar seus sistemas de IA. Cabe dizer que a legislação será direcionada a empresas maiores, com receitas superiores a US$ 50 milhões e mais de 1 milhão de usuários.[31] A lei, se promulgada, seria aplicada pela Comissão Federal de Comércio.

Há também movimentos legislativos de estados e cidades. Em 2019, a cidade de Nova York aprovou sua própria lei para exigir mais transparência com a IA.[32] Há também esforços nos estados de Washington, Illinois e Massachusetts.

Com toda essa atividade, algumas empresas estão ficando proativas e adotando, por exemplo, os próprios conselhos de ética. Basta olhar para a Microsoft. O conselho de ética da empresa, chamado Aether (AI and Ethics in Engineering and Research – IA e Ética em Engenharia e Pesquisa), decidiu não permitir o uso de seu sistema de reconhecimento facial para paradas de trânsito na Califórnia.[33]

No meio disso tudo, também é possível perceber o ativismo de IA, no qual as pessoas se organizam para protestar contra o uso de certas aplicações. Mais uma vez, a Amazon.com tem sido alvo do movimento, com seu software Rekognition que usa reconhecimento facial para ajudar as autoridades na identificação de suspeitos. A ACLU tem levantado preocupações com relação à precisão do sistema, especialmente no que diz respeito a mulheres e minorias. Em uma de suas experiências, a organização descobriu que o Rekognition identificou 28 membros do Congresso como tendo antecedentes criminais![34] A Amazon.com contestou as alegações.

O Rekognition é apenas uma entre as várias aplicações de IA por parte de autoridades que estão levando à controvérsia. Talvez o exemplo mais notável seja o COMPAS (Correctional Offender Management Profiling for Alternative Sanctions – Perfil corretivo do gerenciamento de infratores para sanções alternativas), que usa análise para avaliar a probabilidade de alguém cometer um crime. O sistema é frequentemente usado para condenação. A grande questão é: seu uso pode violar o direito constitucional de uma pessoa ao devido processo, uma vez que há o risco real de que a IA esteja incorreta ou seja discriminatória? Na verdade, por enquanto, existem poucas respostas adequadas. No entanto, dada a importância que os algoritmos de IA terão no sistema judiciário, parece uma boa aposta acreditar que o Supremo Tribunal vai criar novas leis.

AGI (Artificial General Intelligence)

No Capítulo 1, aprendemos sobre a diferença entre IA forte e IA fraca. Para a maior parte dos pesquisadores, estamos na fase da IA fraca, na qual a tecnologia é usada para poucas categorias.

Quanto à IA forte, trata-se do máximo: a capacidade de uma máquina rivalizar com um ser humano. Isso também é conhecido como Inteligência Geral Artificial ou AGI. Conseguir isso provavelmente está a muitos anos de distância, talvez algo que não se veja até o próximo século ou que nunca chegue a ser implementado.

É claro que existem alguns pesquisadores brilhantes que acreditam que a AGI virá em breve. Um deles é Ray Kurzweil, inventor, futurista, autor best-seller e diretor de Engenharia no Google. Quando se trata de IA, ele deixou sua marca na indústria com inovações em áreas como sistemas de texto para fala.

Kurzweil acredita que a AGI vai acontecer – e que o Teste de Turing será vencido – em 2019 e, em seguida, em 2045, haverá a Singularidade. Esse é o lugar no qual teremos um mundo de pessoas híbridas: parte humana, parte máquina.

Uma loucura? Talvez sim, mas Kurzweil tem muitos seguidores de alto nível.

Há muito trabalho pesado a ser feito para chegar a AGI. Mesmo com os grandes avanços com o deep learning, ele geralmente requer grandes quantidades de dados e poder computacional significativo.

Em vez disso, a AGI precisará de novas abordagens, como a capacidade de usar a aprendizagem não supervisionada. É provável que a aprendizagem de transferência também seja crítica. Por exemplo, como discutido anteriormente no livro, a IA conseguiu implementar recursos sobre-humanos em jogos como o Go. Contudo, a aprendizagem de transferência significaria que esse sistema seria capaz de alavancar esse conhecimento para jogar outros jogos ou aprender outros campos.

Além disso, a AGI precisará dispor de capacidades como bom senso, abstração,

curiosidade e detecção de relações causais, não apenas correlações. Tais habilidades têm se mostrado extremamente difíceis com computadores. Haverá necessidade de que sejam feitos avanços em tecnologias de hardware e chip. Essa é a opinião de Yann LeCun, um dos principais pesquisadores de IA do mundo e cientista-chefe de inteligência artificial do Facebook.[35] Ele também acredita que precisa haver muito mais progresso com baterias e outras fontes de energia.

Outro aspecto que será crítico é a necessidade de maior diversidade dentro do campo de IA. De acordo com um relatório do Instituto AI Now, cerca de 80% dos professores de IA são homens; e entre as equipes de pesquisa de IA do Facebook e do Google, as mulheres representavam 15% e 10%, respectivamente.[36]

Esse desequilíbrio significa que a pesquisa pode ficar mais suscetível ao viés. Além disso, haverá a perda do benefício de visões e insights mais amplos.

Bem social

A empresa de consultoria de gestão McKinsey & Co. escreveu um extenso estudo intitulado "Applying Artificial Intelligence for Social Good" ("Aplicação de Inteligência Artificial para o Bem Social").[37] Ele mostra como a IA está sendo usada para lidar com questões como pobreza, desastres naturais e melhoria da educação. O estudo tem aproximadamente 160 estudos de caso. Veja alguns deles:

- A análise das plataformas de mídia social pode ajudar a rastrear o surto de uma doença.
- Uma organização sem fins lucrativos, chamada Rainforest Connection, usa o TensorFlow para criar modelos de IA – com base em dados de áudio – capazes de localizar a exploração madeireira ilegal.
- Pesquisadores construíram uma rede neural que é treinada em vídeos de caçadores na África. Com isso, um drone sobrevoa áreas para detectar violadores a partir de imagens infravermelhas térmicas.
- A IA está sendo usada para analisar dados de 55.893 encomendas na cidade de Flint para encontrar evidências de envenenamento por chumbo. O sistema depende principalmente de um modelo bayesiano que permite previsões mais sofisticadas de toxicidade. Isso significa que os profissionais de saúde podem agir mais rapidamente se houver algum problema na cidade, potencialmente salvando vidas.

Conclusão

Acho que esse é um bom tema para terminar este livro. Independentemente de todo o potencial de dano e das consequências adversas, a IA de fato carrega a promessa de ser transformadora para o mundo. E a boa notícia é que há muitas pessoas

focadas em tornar isso uma realidade. Não se trata de ganhar enormes quantias de dinheiro ou obter fama. O objetivo é mudar o mundo.

Principais aprendizados

- Os carros autônomos estão longe de ser novos. No entanto, o ponto de inflexão para o desenvolvimento dessa tecnologia aconteceu em 2004, com um concurso patrocinado pela DARPA.
- Alguns dos principais componentes de um carro autônomo incluem câmeras de vídeo, LIDAR (lasers que ajudam a processar o meio ambiente) e sensores (para detecção de outros veículos e obstáculos como meio-fio).
- Em termos de definição do que é "autônomo", há cinco níveis. O quinto é quando o veículo é totalmente autônomo.
- Alguns dos desafios para carros autônomos são infraestrutura (as rodovias existentes não são ideais), legislação, custos e adoção pelo consumidor.
- Os Estados Unidos são considerados líderes globais em IA. Entretanto, isso pode mudar em breve. A China está investindo fortemente em IA e tem grandes vantagens, como enormes quantidades de dados e um número significativo de engenheiros qualificados.
- Um dos receios com relação à IA é que ela leve ao desemprego em massa, seja para empregos de colarinho azul ou colarinho branco. É verdade que a tecnologia já impactou diversas indústrias, como a de manufatura, mas os mercados se mostraram adaptáveis. Se a IA é, de fato, transformadora, pode causar um pouco de perturbação. É por isso que provavelmente haverá uma necessidade de treinamento e requalificação para novas carreiras.
- Os drones causaram grande impacto na guerra. Com a IA, entretanto, está se tornando possível permitir que essa tecnologia tome as decisões no campo de batalha. Há muitas pessoas que veem essa possibilidade como um grande problema. No entanto, Estados Unidos, Rússia e outros países parecem estar focados na busca por armas autônomas.
- Quando se trata de guerra – pelo menos no curto prazo –, a IA pode ter um efeito mais imediato na disseminação de informações falsas. Vimos isso com a interferência dos russos nas eleições presidenciais americanas de 2016.
- Espera-se que a IA ajude muito no processo de descoberta de medicamentos. Grandes operadores farmacêuticos, como Gilead, estão explorando a tecnologia. A IA não só pode processar grandes quantidades de dados, mas também detectar padrões que podem não ser discerníveis para os seres humanos.
- À medida que a IA se tornar mais generalizada, haverá preocupações crescentes sobre privacidade e transparência. Devido a isso, tem havido movimentos no

Congresso, iniciados por cidades e estados, para impor regulamentações. Não está claro o que pode acontecer, mas parece provável que veremos mais restrições. Enquanto isso, algumas empresas estão tentando ser proativas e criando os próprios conselhos de ética.

- Inteligência Geral Artificial ou AGI é quando um sistema tem inteligência humana. Estamos provavelmente muito longe dessa realidade, no entanto. A razão é que será necessário haver outras inovações na IA, como a aprendizagem sem supervisão e a criação de novos hardwares.

---

1 www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html

2 www.cnbc.com/2019/03/08/softbank-ceo-ai-will-completely-change-the-way-humans-live-within-30-years.html

3 www.msn.com/en-us/news/technology/the-best-and-worst-technology-predictions-of-all-time/ss-BBIMwm3#image=5

4 www.recode.net/2017/1/9/14215942/iphone-steve-jobs-apple-ballmer-nokia-anniversary

5 www.politico.eu/article/delf-driving-car-born-1986-ernst-dickmanns-mercedes/

6 www.sec.gov/Archives/edgar/data/1759509/000119312519077391/d633517ds1a.htm

7 www.alliedmarketresearch.com/autonomous-vehicle-market

8 Entrevista realizada pelo autor em maio de 2019 com Scott Painter, CEO e fundador da Fair.

9 www.businessinsider.com/waymo-ceo-john-krafcik-explains-big-challenge-for-self-driving-cars-2019-4

10 https://newsroom.aaa.com/2019/03/americans-fear-self-driving-cars-survey/

11 https://medium.com/waymo/riding-with-waymo-one-today-9ac8164c5c0e

12 Ibid.

13 www.diamandis.com/blog/rise-of-ai-in-china

14 www.theverge.com/2019/3/14/18265230/china-is-about-to-overtake-america-in-ai-research

15 Nova York: Houghton Mifflin Harcourt, 2018.

16 www.nature.com/articles/s41591-018-0335-9

17 www.mercurynews.com/2019/03/18/were-not-prepared-for-the-promise-of-artificial-intelligence-experts-warn/

18 www.washingtonpost.com/news/national/wp/2019/02/17/feature/inside-the-race-to-replace-farmworkers-with-robots/

19 www.cnbc.com/2017/02/17/lawyers-could-be-replaced-by-artificial-intelligence.html

20 www.pewresearch.org/fact-tank/2018/08/07/for-most-us-workers-real-wages-have-barely-budged-for-decades/

21 http://fortune.com/2019/02/08/growing-wealth-inequality-us-study/

22 www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html

23 www.popularmechanics.com/military/aviation/a26871027/air-force-ai-fighter-plane-

skyborg/

24 www.theguardian.com/science/2018/jul/18/thousands-of-scientists-pledge-not-to-help-build-killer-ai-robots

25 www.theguardian.com/science/2019/mar/29/uk-us-russia-opposing-killer-robot-ban-un-ai

26 www.rand.org/news/press/2018/04/24.html

27 www.wsj.com/articles/biogen-shares-drop-28-after-ending-alzheimers-phase-3-trials-11553170765

28 www.fiercebiotech.com/biotech/stealthy-insitro-opens-up-starting-gilead-deal-worth-up-to-1-05b

29 https://khn.org/morning-breakout/ups-and-downs-of-artificial-intelligence-ibm-stops-sales-development-of-watson-for-drug-discovery-hospitals-learn-from-ehrs/

30 www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio

31 www.theverge.com/2019/4/10/18304960/congress-algorithmic-accountability-act-wyden-clarke-booker-bill-introduced-house-senate

32 www.wsj.com/articles/our-software-is-biased-like-we-are-can-new-laws-change-that-11553313609?mod=hp_lead_pos8

33 www.geekwire.com/2019/policing-ai-task-industry-government-customers/

34 www.businessinsider.com/ai-experts-call-on-amazon-not-to-sell-rekognition-software-to-police-2019-4

35 http://fortune.com/2019/02/18/facebook-yann-lecun-lawnmowers-deep-learning/

36 www.theverge.com/2019/4/16/18410501/artificial-intelligence-ai-diversity-report-facial-recognition

37 www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

APÊNDICE A

Recursos de IA

Publicações e blogs sobre IA

- aitrends.com: www.aitrends.com/
- The Berkeley Artificial Intelligence (BAIR): https://bair.berkeley.edu/blog/
- KDnuggets: www.kdnuggets.com/news/index.html • Machine Learning Mastery: https://machinelearningmastery.com/blog/
- MIT Technology Review: www.technologyreview.com/
- ScienceDaily – Seção de IA: www.sciencedaily.com/news/computers_math/artificial_intelligence/

Blogs de empresas sobre IA

- Baidu: http://research.baidu.com/
- DeepMind: https://deepmind.com/blog/
- Facebook: https://research.fb.com/blog/
- Google: https://ai.googleblog.com/
- Microsoft: www.microsoft.com/en-us/research/
- NVIDIA: https://blogs.nvidia.com/blog/category/deep-learning/
- OpenAI: https://openai.com/blog/

Feeds do Twitter dos principais pesquisadores de IA

- Fei-Fei Li: https://twitter.com/drfeifei • Ian Goodfellow: https://twitter.com/goodfellow_ian • Demis Hassabis: https://twitter.com/demishassabis • Yann Lecun: https://twitter.com/ylecun?
- Andrew Ng: https://twitter.com/AndrewYNg

Ferramentas e plataformas de IA de código aberto

- Jupyter Notebook: https://jupyter.org/
- Keras: https://keras.io/
- Python language: www.python.org/
- PyTorch: https://pytorch.org/
- TensorFlow: www.tensorflow.org/

Cursos online

- Coursera: `www.coursera.org/`
- Udacity: `www.udacity.com/`
- Udemy: `www.udemy.com/`

Glossário

Agrupamento (Clustering): forma de aprendizado não supervisionado que utiliza dados não rotulados e executa algoritmos para colocar itens semelhantes em grupos.

Agrupamento k-means: algoritmo eficaz para agrupar dados semelhantes não rotulados.

Análise de sentimentos: extrai dados de mídia social e encontra tendências.

Análise preditiva: envolve o uso de dados para fazer previsões.

Aprendizagem não supervisionada: envolve um modelo de IA que usa dados não rotulados. Em geral, isso significa que será necessário utilizar sistemas de deep learning para detectar padrões.

Aprendizagem por reforço: abordagem para criar um modelo de IA no qual o sistema é recompensado pelas previsões corretas e punido pelas erradas.

Aprendizagem supervisionada: envolve um modelo de IA que usa dados rotulados. É a abordagem mais comum.

Armazenamento: envolve a organização de dados em grupos.

Árvore de decisão: algoritmo de machine learning que é um fluxo de trabalho de caminhos de decisão.

Assistente virtual: dispositivo de IA que ajuda uma pessoa com suas tarefas diárias.

Atuadores: dispositivos eletromecânicos, como motores. Auxiliam no movimento de robôs.

Automação robótica e cognitiva de processos (CRPA – Cognitive Robotic Process Automation): sistema RPA que utiliza as tecnologias de IA.

Banco de dados relacional: banco de dados cujas raízes remontam à década de 1970. Permite a criação de relacionamentos entre tabelas de dados e usa uma linguagem de script chamada SQL.

Big Data: categoria de tecnologia que envolve o processamento de grandes quantidades de dados. É geralmente descrito como tendo os três Vs – volume, variedade e velocidade.

Camadas ocultas: diferentes níveis de análise em um modelo de deep learning.

Chatbot: sistema de IA que se comunica com pessoas.

Classificador Naïve Bayes: método de machine learning que usa o teorema de Bayes para fazer previsões, mas as variáveis são independentes umas das outras.

Cobô: robô que trabalha em conjunto com pessoas.

Comitê de ética: comitê que avalia as questões relacionadas aos projetos de IA.

Correlação de Pearson: mostra a força de uma correlação, de 1 a -1. Quanto mais próximo estiver de 1, mais precisa será a correlação.

Córtex cerebral: parte do cérebro humano que tem mais semelhanças com a IA. Ajuda no pensamento e em outras atividades cognitivas.

Dados categóricos: dados que não têm significado numérico, mas têm significado textual, como a descrição de raça ou sexo.

Dados de teste: permitem a avaliação da precisão de um modelo.

Dados de treinamento: dados usados na criação de um algoritmo de IA.

Dados estruturados: geralmente são armazenados em um banco de dados relacional ou planilha, já que as informações estão em uma estrutura pré-formatada (como números de CPF, endereços e informações sobre pontos de venda).

Dados não estruturados: dados sem formato predefinido, como imagens, vídeos e arquivos de áudio.

Dados ordinais: mistura de dados numéricos e categóricos, como a classificação de produtos da Amazon.com.

Data Lake: permite armazenamento e processamento de grandes quantidades de dados estruturados e não estruturados. Geralmente, há pouca ou nenhuma necessidade de reformatação dos dados.

Deepfake: envolve o uso de modelos de deep learning para criar imagens ou vídeos enganosos ou prejudiciais.

Deep Learning: tipo de IA que usa redes neurais que imitam os processos do cérebro. Grande parte da inovação no campo durante a última década ocorreu em pesquisas dessa área.

Desvio padrão: mede a distância média da média, o que dá uma noção da variação nos dados.

Distribuição normal: gráfico de dados que se parece com um sino e no qual o ponto médio é a média.

Engenharia de recursos: ver extração de recursos.

Estemização: descreve o processo de reduzir uma palavra à sua raiz (ou lema) por meio da remoção de afixos e sufixos.

ETL (Extraction, Transformation and Load – Extração, transformação e carga): forma de integração de dados normalmente usada em um data warehouse.

Explicabilidade: processo de compreensão das causas subjacentes de um modelo de deep learning.

Extração de recursos: descreve o processo de seleção de variáveis para um modelo de IA.

Fadiga de automatização: com a automação robótica de processos, provavelmente haverá menos melhorias à medida que mais tarefas forem automatizadas.

Falso positivo: quando a previsão de um modelo mostra que o resultado é verdadeiro mesmo quando ele não é.

Fonemas: unidade mais básica de som em uma língua.

Função de ativação: usada em modelos de deep learning para ajudar a calcular relações não lineares.

GPUs (Graphics Processing Units – Unidade de processamento gráfico): chips originalmente usados para videogames de alta velocidade devido à capacidade de processar grandes quantidades de dados rapidamente. As GPUs também provaram ser hábeis em lidar com aplicativos de IA.

Hadoop: permite gerenciar o big data e possibilita a criação de sofisticados data warehouses.

Hiperparâmetros: recursos em um modelo que não podem ser aprendidos diretamente do processo de treinamento.

IA: ver Inteligência Artificial.

IA forte: é a verdadeira IA, na qual uma máquina é capaz de apresentar habilidades humanas, como participar de discussões abertas.

IA fraca: situação na qual a IA é utilizada para uma aplicação específica, tal como a Siri, da Apple.

Instância: uma linha de dados.

Inteligência artificial: quando computadores conseguem aprender a partir de experiências, o que costuma envolver processamento de dados usando algoritmos sofisticados. A inteligência artificial é uma categoria ampla que inclui subáreas como machine learning, deep learning e Natural Language Processing (NLP – Processamento de linguagem natural).

Inverno da IA (AI winter): período prologado de tempo, como o que se viu nos anos de 1970 a 1980, quando a indústria de IA ficou sob pressão e experimentou momentos ruins, como cortes de orçamentos.

Jupyter Notebook: aplicativo baseado na web que facilita a codificação em Python e R para criar visualizações e importar sistemas de IA.

K-Nearest Neighbor (k-NN): algoritmo de machine learning que classifica dados com base em semelhanças.

Lematização: processo de NLP que remove afixos ou prefixos para se concentrar em

encontrar palavras-raiz semelhantes.

Lidar (Light Detection and Ranging – Detecção de luz e área): dispositivo – geralmente posicionado no topo de um carro autônomo – que dispara raios laser para medir o ambiente.

Machine Learning: quando um computador pode aprender e melhorar processando dados sem precisar ser explicitamente programado. O machine learning é uma subárea da IA.

Machine learning automatizado (AutoML): ferramenta ou plataforma digital que permite que iniciantes criem os próprios modelos de IA.

Marcação de partes do discurso (POS – Parts Of Speech): no processo de NLP, envolve revisar o texto e associar cada palavra à sua forma gramatical adequada, como substantivos, verbos, advérbios *etc.*

Metadados: são dados sobre dados – ou seja, descrições. Um arquivo de música, por exemplo, pode ter metadados como tamanho, duração, data de upload, comentários, gênero, artista *etc.*

Modelagem de tópicos: no processo de NLP, envolve a busca por padrões e clusters ocultos no texto.

Modelagem por agrupamento: envolve o uso de mais de um modelo para gerar previsões.

Modelo oculto de Markov (HMM – Hidden Markov Model): algoritmo usado para decifrar palavras faladas.

Natural Language Processing (NLP – Processamento de linguagem natural): subárea da IA que lida com a maneira como os computadores compreendem e manipulam a linguagem.

Overfitting: quando um modelo não é preciso, porque os dados não refletem o que está sendo testado ou o foco está nos recursos incorretos.

Problema da dissipação do gradiente: explica como a precisão diminui à medida que o modelo de deep learning cresce.

Python: linguagem de programação que se tornou o padrão no desenvolvimento de modelos de IA.

PyTorch: plataforma desenvolvida pelo Facebook que permite a criação de modelos sofisticados de IA.

R-quadrado: fornece uma maneira de avaliar a precisão de uma regressão. Um R-quadrado varia de 0 a 1 e, quanto mais próximo o modelo estiver de 1, maior a precisão.

RDA (Robotic Desktop Automation – Automação robótica de desktop): o sistema RPA

trabalha em conjunto com um empregado para lidar com trabalhos ou tarefas.

Reconhecimento de entidades nomeadas: No processo NLP, envolve identificar as palavras que representam locais, pessoas e organizações.

Recurso: uma coluna de dados.

Rede adversária generativa (GAN – Generative Adversarial Network): desenvolvida pelo pesquisador de IA Ian Goodfellow, é um modelo de deep learning da próxima geração que ajuda a criar novos tipos de saída, como áudio, texto ou vídeo.

Rede neural: modelo sofisticado de IA que imita o cérebro. Uma rede neural possui diversas camadas que tentam encontrar padrões únicos que envolvem diferentes camadas de análise.

Rede Neural Artificial: (ANN – Artificial Neural Network) – estrutura mais básica para um modelo de deep learning. A ANN inclui diversas camadas ocultas que processam dados por meio de algoritmos sofisticados.

Rede Neural Convolucional: (CNN – Convolutional Neural Network) – modelo de deep learning que passa por diferentes variações – ou convoluções – de análise de dados. As CNNs são frequentemente usadas para aplicações complexas, como reconhecimento facial.

Rede Neural Feed-Forward: modelo de deep learning que processa dados em uma direção linear por meio das camadas ocultas. Não há como repetir o ciclo.

Rede Neural Recorrente: (RNN – Recurrent Neural Network) – modelo de deep learning que processa entradas anteriores ao longo do tempo. Um uso comum é quando uma pessoa digita caracteres em um aplicativo de mensagens e a IA prevê a próxima palavra.

Regressão linear: mostra o relacionamento entre determinadas variáveis, o que pode ajudar com previsões para sistemas de machine learning.

Retropropagação (backpropagation): um grande avanço no deep learning. Permite uma atribuição mais eficiente de pesos nos modelos.

Robotic Process Automation (RPA – Automação robótica de processos): categoria de software que automatiza tarefas rotineiras em uma organização. Costuma ser uma forma inicial de implementação da IA.

Robot Operating System (ROS – Sistema operacional de robôs): sistema middleware de código aberto que gerencia as partes críticas de um robô.

RPA assistida: sistema RPA completamente autônomo no qual um bot é executado em segundo plano.

Sensor: costuma ser uma câmera ou um Lidar, que usa um scanner a laser para criar imagens 3D.

Sigmoide: função de ativação comum para um modelo de deep learning. Seu valor varia de 0 a 1. Quanto mais próximo de 1, maior a precisão.

Sistema especialista: tipo inicial de aplicação de IA que surgiu na década de 1980. Utilizava sistemas lógicos sofisticados para ajudar a entender certas áreas como medicina, finanças e manufatura.

Sistema NoSQL: banco de dados da próxima geração. As informações são baseadas em um modelo de documento que permite maior flexibilidade na análise e no manuseio de dados estruturados e não estruturados.

TensorFlow: plataforma de código aberto, apoiada pelo Google, que permite a criação de modelos sofisticados de IA.

Teorema de Bayes: medida estatística usada no machine learning que ajuda a fornecer uma visão mais precisa das probabilidades.

Teste de Turing: criado por Alan Turing, é uma maneira de determinar se um sistema alcançou a verdadeira IA. O teste envolve uma pessoa que faz perguntas a dois participantes – um humano e um computador. Se não está claro quem é o humano, o Teste de Turing foi superado.

Tipo de dado: tipo de informação que uma variável representa, tal como booleano, inteiro, string ou real.

Tokenização: no processo de NLP, essa é a etapa na qual o texto é analisado e segmentado em várias partes.

Três leis da Robótica: baseadas na ficção científica de Isaac Asimov, essas leis sugerem uma estrutura básica de como os robôs devem interagir com a sociedade.

Verdadeiro positivo: quando um modelo faz uma previsão correta.