APRENDIZADO DE MÁQUINA EM AÇÃO

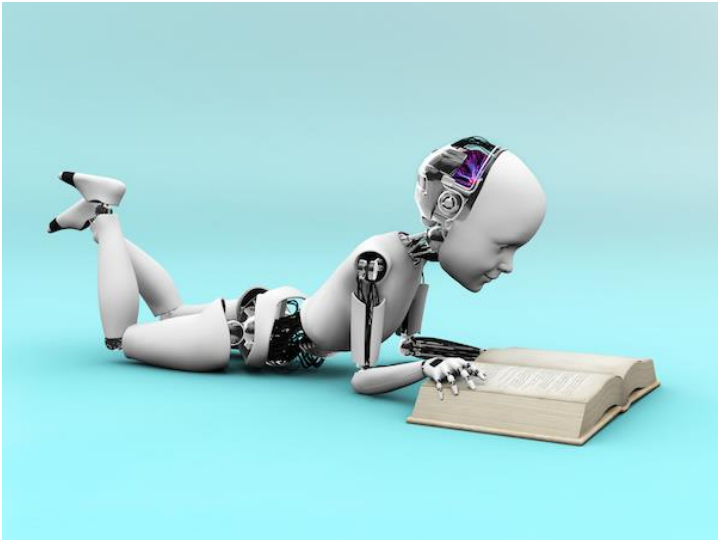
UM MANUAL PARA LEIGOS

ALAN T. NORMAN

ÍNDICE

POR QUE ESCREVI ESSE LIVRO

Bem-vindo ao mundo do aprendizado de máquina!



A inteligência artificial está preparada para mudar o curso da história humana, talvez mais que qualquer tecnologia já criada. Grande parte dessa revolução é o aprendizado de máquina.

Aprendizado de máquina é a ciência de ensinar computadores a fazer previsões baseadas em dados. Basicamente, aprendizado de máquina envolve dar a um computador um conjunto de dados e pedir a ele que faça uma previsão. No início, o computador fará muitas previsões incorretas. No entanto, após fazer milhares de previsões, o computador reconfigurará seu algoritmo, aprimorando suas previsões.

Esse tipo de previsão computacional era impossível. Os computadores simplesmente não conseguiam armazenar muitos dados ou processá-los suficientemente rápido para aprenderem de forma efetiva. Atualmente, a cada ano, a inteligência dos computadores aumenta em uma velocidade muito alta. Avanços no armazenamento de dados e no poder de processamento estão impulsionando essa tendência em direção a máquinas cada vez mais inteligentes. Como resultado, os computadores estão fazendo coisas que seriam impensáveis há apenas uma ou duas décadas.

O aprendizado de máquina já está afetando nossa vida diária. A Amazon usa o aprendizado de máquina para prever quais produtos você vai querer comprar. O Gmail o usa para filtrar mensagens de spam da sua caixa de entrada. Suas recomendações de filmes na Netflix são baseadas em algoritmos de aprendizado de máquina.

Contudo, os impactos do aprendizado de máquina não param por aqui. Os algoritmos de aprendizado de máquina estão fazendo previsões para todo o tipo de atividade, da agricultura à área de saúde. Além disso, seus impactos serão sentidos de novas formas e em novos setores a cada ano. À medida que novas aplicações de aprendizado de máquina surgem, vamos gradualmente aceitando-as como parte da vida normal. No entanto, esta nova dependência de máquinas inteligentes é um ponto de transição na história da tecnologia e uma tendência que se acelera a cada dia.

No futuro, o aprendizado de máquina e a inteligência artificial serão responsáveis pela automação de muitas tarefas feitas hoje por humanos. Carros auto-dirigíveis dependem do aprendizado de máquina para o reconhecimento de imagem, e eles progressivamente se tornarão parte do transporte, assim como caminhões auto-conduzidos e outros veículos para o transporte de mercadorias. Atualmente, grande parte da agricultura e manufatura é automatizada, de forma que o aprendizado de máquina está provendo os alimentos que consumimos e os bens que utilizamos. A tendência para a automação está só acelerando. Outras aplicações de aprendizado de máquina podem fundamentalmente mudar tarefas feitas por humanos no dia-a-dia, na medida em que as máquinas se tornam mais aptas para controlar processos e concluir trabalhos de conhecimento.

Como o aprendizado de máquina terá um impacto tão profundo na vida diária, é importante que todos tenham acesso à informação sobre como isso funciona. É por isso que escrevi esse livro. O cenário atual de informações sobre aprendizado de máquina está dividido.

Primeiramente, há explicações para o público geral, que dificultam os conceitos. Estes "explicadores" fazem o aprendizado de máquina parecer como algo que somente especialistas pudessem entender.

Em segundo lugar, há documentos técnicos escritos por especialistas para especialistas. Eles excluem o público

geral, com jargões e complexidades. Obviamente, escrever e executar um algoritmo de aprendizado de máquina é um feito enormemente técnico, e essas explicações técnicas são importantes. No entanto, há um buraco na atual literatura a respeito de aprendizado de máquina.

E quanto aos leigos que realmente querem entender essa revolução tecnológica, não necessariamente sabendo escrever códigos, mas sim compreender as mudanças que estão ocorrendo à sua volta? A compreensão dos princípios básicos sobre aprendizado de máquina não deveria estar limitada a uma elite tecnológica. Essas mudanças afetarão a todos nós. Elas têm consequências éticas, e é importante que o público saiba sobre todos os benefícios e desvantagens do aprendizado de máquina.

É por isso que escrevi esse livro. Se isso soa interessante para você, eu espero que aproveite.

ESSE LIVRO NÃO É SOBRE ALGORITMOS DE CODIFICAÇÃO DE APRENDIZADO DE MÁQUINA

Caso essa declaração na introdução não tenha sido suficientemente clara: este não é um livro sobre codificação. Não é para cientistas da computação aprenderem como criar algoritmos de aprendizado de máquina.

Para começar, não sou nem um pouco qualificado para escrever um livro sobre isso. Pessoas passam anos

aprendendo as complexidades da escrita de algoritmos e redes de treinamento. Existem programas inteiros de PhD que exploram os meandros desta área, desenhando em álgebra linear e análise preditiva. Se você mergulhar fundo nos detalhes do aprendizado de máquina e amar isso o suficiente para obter um PhD, você poderá facilmente sair ganhando entre 300 e 600 mil dólares, trabalhando para uma grande empresa de tecnologia. É dessa forma que essas atividades são tão raras e valiosas.

Eu não tenho essas qualificações, e não vejo mal algum nisso. Se você chegou até esse livro, você é um iniciante interessado em aprendizado de máquina. Provavelmente, você não é um técnico, ou se é, está buscando um livro sobre seus fundamentos, para iniciar com os conceitos básicos. Como um escritor da área de tecnologia, estou constantemente aprendendo sobre tecnologias. Sou um estudante de aprendizado de máquina e lembro-me como é ser um iniciante. Posso ajudar a explicar os conceitos básicos, de uma forma fácil de entender. Uma vez que tiver lido esse livro, você terá uma sólida compreensão sobre os princípios fundamentais que facilitarão seu acesso a um livro mais avançado, caso queira aprender mais.

Dito isso, caso sinta que já entende os princípios básicos ou realmente queira um livro que o ensine os detalhes práticos sobre como escrever e treinar um algoritmo de aprendizado de máquina, então provavelmente esse livro não é para você.

UM MANUAL PARA LEIGOS

O real objetivo desse livro é ser uma introdução fácil de ler sobre aprendizado de máquina. Meu objetivo é escrever um livro que qualquer um possa ler, mantendo-o, ao mesmo tempo, fiel aos princípios sobre aprendizado de máquina, sem inferiorizar seus conceitos. Estou certo da inteligência de meus leitores, e não acho que um livro para iniciantes tenha que necessariamente sacrificar complexidades e nuances. Sendo assim, este não é um livro grande e nem tampouco abrangente. Aqueles interessados no tema vão querer se aprofundar através de outros livros e pesquisas.

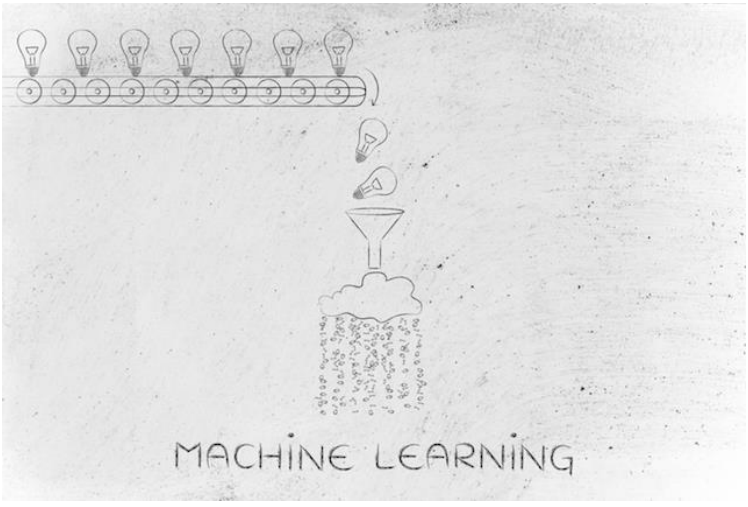
Neste livro, veremos os conceitos básicos e tipos de aprendizado de máquina. Investigaremos como eles funcionam. Então, exploraremos questões sobre conjunto de dados, e escrita e treinamento de algoritmos. Por fim, veremos casos de aplicação do aprendizado de máquina no mundo real, e áreas onde ele deverá ser usado no futuro.

Mais uma vez, bem-vindo ao aprendizado de máquina. Vamos mergulhar nisso...

Capítulo 1. O que é aprendizado de máquina?

O objetivo desse primeiro capítulo é estabelecer a base do resto que você lerá nesse livro. Aqui, identificaremos os conceitos básicos que exploraremos mais detalhadamente em capítulos futuros. Este livro constrói-se por si próprio, e neste capítulo estão os fundamentos.

Dito isso, o ponto lógico para começar é definindo o que queremos dizer quando falamos sobre aprendizado de máquina.



Minha definição simples é a seguinte: aprendizado de máquina permite que um computador aprenda pela experiência.

Isso deve até soar trivial, mas a falha na definição pode provocar implicações profundas. Antes do aprendizado de máquina, os computadores não eram capazes de melhorar através da experiência. Ao invés disso, o que quer que o código dissesse é o que o computador fazia.

Aprendizado de máquina, em sua explicação mais simples, envolve permitir que um computador varie suas respostas, dando um retorno com respostas boas ou ruins. Isso significa que os algoritmos de aprendizado de máquina são fundamentalmente diferentes dos programas de computadores que surgiram antes deles. Entender a diferença entre programação explícita e treino de algoritmos é o primeiro passo para vermos como o aprendizado de máquina muda fundamentalmente a ciência da computação.

PROGRAMAÇÃO EXPLÍCITA VS. TREINAMENTO DE ALGORITMOS

Com algumas exceções recentes, praticamente qualquer software que você já tenha usado na vida foi certamente programado. Isso significa que algum humano escreveu um conjunto de regras para o computador seguir. Tudo, desde o sistema operacional do seu computador até a internet ou aplicativos do seu telefone têm códigos escritos por um humano. Sem humanos para dar a um computador um conjunto de regras para ele funcionar, este não seria capaz de fazer nada.

A programação explícita é ótima. Ela é o fundamento para tudo o que fazemos atualmente com computadores. É ideal para quando você precisa de um computador para gerenciar dados, calcular um valor, ou manter um registro de relacionamentos para você. A programação explícita é muito poderosa, mas ela tem um gargalo: o humano.

Isso passa a ser um problema quando queremos fazer coisas complexas com um computador, como pedir para que reconheça a foto de um gato. Se usássemos a programação explícita para ensinar um computador o quê deve observar em um gato, passaríamos anos escrevendo códigos para cada contingência. E se você não puder ver todas as quatro patas na foto? E se o gato tiver uma cor diferente? O computador seria capaz de distinguir um gato preto em um pano de fundo preto ou um gato branco na neve?

Todas essas coisas são claras para os humanos. Nossos cérebros reconhecem coisas rápido e facilmente em muitos contextos. Computadores não são tão bons nisso, e seriam necessárias milhares de linhas de código explícito para ensinar um computador a identificar um gato. Na verdade, talvez nem seja possível explicitamente programar um computador para identificar gatos com 100% de precisão, porque o contexto pode sempre mudar e confundir seu código.

É aí que os algoritmos entram em cena. Com a programação explícita, estávamos tentando dizer ao computador o que é um gato e dar subsídios para cada

contingência em nosso código. Em contraste, algoritmos de aprendizado de máquina permitem ao computador descobrir o que é um gato.

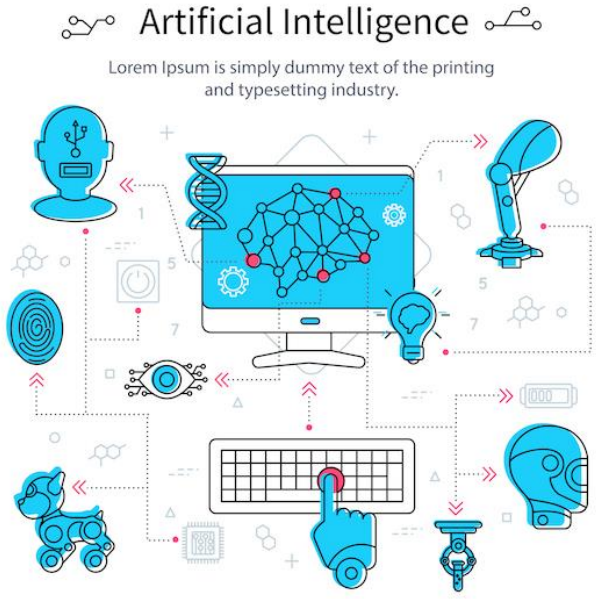
Para começar, o algoritmo deve conter algumas características principais. Por exemplo, podemos dizer ao computador para que procure quatro patas e um rabo. Então, alimentamos o algoritmo com muitas imagens. Algumas das imagens são de gatos, mas outras podem ser de cães, árvores ou imagens aleatórias. Quando o algoritmo faz uma suposição, reforçamos suposições corretas e damos uma resposta negativa para as incorretas.

Com o tempo, o computador usará o algoritmo para construir seu próprio modelo do que procurar para identificar um gato. Os componentes no modelo do computador podem ser coisas em que nem pensamos no início. Com mais reforços e milhares de imagens, o algoritmo se tornará gradualmente melhor na identificação de gatos. Talvez nunca alcance 100% de precisão, mas será preciso o suficiente para substituir uma imagem de gato rotulada por humanos e mais eficiente.

Algoritmos são instruções, porém não são regras explícitas. São uma nova forma de dizer a um computador como abordar uma tarefa. Eles apresentam ciclos de respostas que são corrigidas automaticamente em um processo de centenas ou milhares de tentativas em uma tarefa.

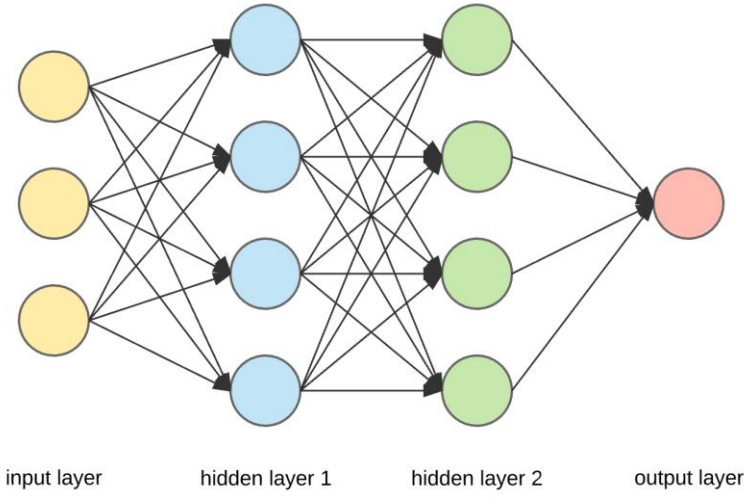
DEFINIÇÕES: INTELIGÊNCIA ARTIFICIAL VS. APRENDIZADO DE MÁQUINA VS. REDES NEURAIS

Este livro é sobre aprendizado de máquina, mas esse termo se enquadra em um contexto mais amplo. Como o aprendizado de máquina está crescendo em popularidade, este assunto tem recebido muita cobertura jornalística. Nesses artigos, jornalistas frequentemente usam os termos inteligência artificial, aprendizado de máquina e redes neurais intercambiáveis. No entanto, existem pequenas variações entre os três termos.



Inteligência artificial (IA) é o mais velho e mais amplo dos três termos. Criada em meados do século 20, a inteligência artificial refere-se a qualquer momento em que uma máquina observa e responde a seu ambiente. Ela está em contraste com a inteligência natural dos humanos e animais. Ao longo do tempo, no entanto, seu escopo tem mudado. Por exemplo, o reconhecimento de caracteres costumava ser um grande desafio para a IA. Hoje, trata-se de uma rotina e não é mais considerado parte da IA. À medida que descobrimos novos usos para a IA, estes são integrados à base de referência do que é normal, e o escopo da IA se amplia para qualquer nova função que surja.

Aprendizado de máquina é um subconjunto específico da IA. Nós já gastamos algum tempo o definindo neste capítulo, mas ele refere-se a dar a uma máquina um ciclo de respostas que a permite aprender pela experiência. Como termo, o aprendizado de máquina existe desde os anos 80. Só recentemente, nos últimos 10 a 15 anos obtivemos o poder de processamento e armazenamento de dados para realmente começar a implementar o aprendizado de máquinas em escala.



Redes neurais são um subconjunto do aprendizado de máquina e são a maior tendência do setor no momento. Um rede neural consiste em muitos nós que trabalham juntos para produzir uma resposta. Cada um dos diminutos nós tem uma função específica. Por exemplo, quando olhamos para uma imagem, os nós de baixo nível podem identificar cores ou linhas específicas. Nós posteriores podem agrupar as linhas em formas, medir distâncias ou procurar a intensidade da cor. Cada um desses nós é então avaliado por seu impacto na resposta final. No início, a rede neural cometerá muitos erros, porém, ao longo de muitas tentativas, atualizará a avaliação de cada nó, aprimorando na identificação da resposta correta.

Agora, quando você ler um artigo sobre IA, aprendizado de máquina ou redes neurais, entenderá a diferença. O fundamental é entender que eles são subconjuntos. Redes neurais são só um tipo de aprendizado de

máquina que, por sua vez, é apenas parte da inteligência artificial.

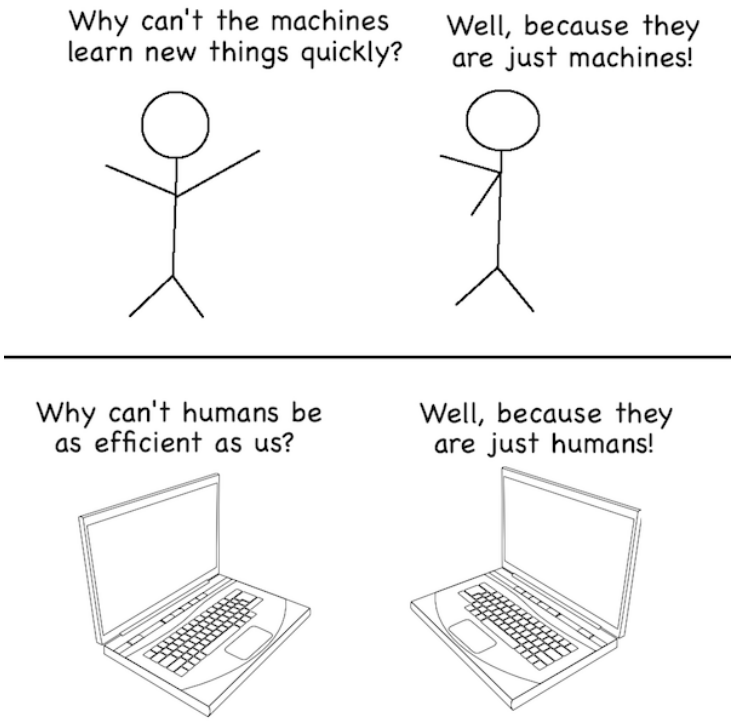
CONCEITOS BÁSICOS

O aprendizado de máquina pode ser implementado em muitos casos. Contanto que haja dados significativos para analisar, o AM pode ajudar a entendê-los. Dessa forma, cada projeto de aprendizado de máquina é diferente. No entanto, existem cinco aspectos principais para qualquer aplicação de aprendizado de máquina:

1. O PROBLEMA

O aprendizado de máquina é útil sempre que você precisar reconhecer padrões e prever comportamentos baseados em dados históricos. Reconhecer padrões pode significar qualquer coisa, desde o reconhecimento de caracteres até previsão de manutenção ou recomendação de produtos a clientes com base nas últimas compras.

Contudo, o computador não compreende inerentemente os dados ou o problema. Ao invés disso, um cientista de dados precisa ensinar o computador o que deve procurar para poder dar a resposta apropriada. Se o cientista de dados não definir bem o problema, mesmo o melhor algoritmo treinado no maior conjunto de dados não produzirá os resultados que você quer.



É claro que o aprendizado de máquina ainda não é adequado para raciocínio simbólico em alto nível. Por exemplo, um algoritmo pode ser capaz de identificar um cesto, ovos coloridos e um campo, porém, não seria capaz de dizer que aquilo se trata de uma caça a ovos de páscoa, como a maioria dos humanos reconheceria.

Tipicamente, projetos de aprendizado de máquina têm um problema bem reduzido e específico para o qual buscam encontrar uma resposta. Um problema diferente precisará de uma nova abordagem e, possivelmente, de um algoritmo distinto.

2. Os Dados

O Aprendizado de máquina é possível em escala por causa da quantidade de dados que começamos a coletar ao longo dos últimos anos. Essa grande revolução de dados é a chave que desbloqueou o treinamento de algoritmos complexos. Os dados estão no centro do ajuste de um algoritmo de aprendizado de máquina, para que este dê a resposta correta.

Como os dados são muito importantes para o AM, os resultados são um reflexo direto de suas entradas. Caso haja um erro nos dados, o algoritmo de AM aprenderá a ser incorreto. Por exemplo, previsões de candidatos a contratação, recomendações de sentenças de tribunais e diagnósticos médicos estão todos usando o aprendizado de máquina, e têm algum nível de cultura, gênero, raça, educação ou outros erros dentro do conjunto de dados que os treinam.

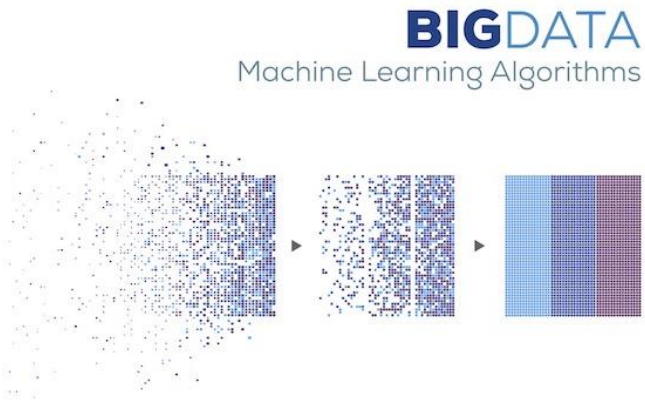
Os erros estendem-se para além do prejuízo no conjunto de dados. Às vezes, dados enganam um algoritmo de outras maneiras. Considere o caso de um modelo de aprendizado de máquina militar treinado para procurar tanques camuflados em uma floresta. Os cientistas de dados treinaram o algoritmo em um conjunto de imagens, algumas das quais tinham tanques entre árvores e outras somente de árvores. Após o treinamento, o modelo pontuou com precisão quase perfeita nos testes de dados feitos pelos cientistas. No entanto, quando o modelo entrou em produção, ele não funcionou na identificação dos tanques. Verificou-se, no conjunto de dados de treinamento, que as fotos de tanques tinham sido tiradas em um dia ensolarado, enquanto as fotos da floresta eram de um dia nublado. O algoritmo tinha aprendido a distinguir dias ensolarados de nublados, não tanques!

Nenhum conjunto de dados é perfeito, mas podemos tomar precauções para fazê-lo ter menos erros. As principais precauções advém da estatística. Quando possível, os dados devem ser uma amostra aleatória da

população alvo. O tamanho da amostra deve ser grande o suficiente para que você possa tirar conclusões significativas dos resultados com um alto nível de confiança. Os dados devem ser rotulados e limpos com precisão, no caso de pontos de dados inválidos ou isolados que possam enganar o algoritmo.

Teremos um capítulo inteiro sobre dados, onde exploraremos essas questões com maior profundidade.

3. Os Algoritmos



Algoritmos são o principal componente sobre o qual as pessoas pensam quando fazem referência ao aprendizado de máquina. Este é o código que diz ao computador o que procurar e como ajustar sua ponderação de possíveis respostas com base nas respostas que recebe.

Atualmente, existem muitos algoritmos de aprendizado de máquina bem estabelecidos. Muitos deles vêm pré-

carregados em bibliotecas de codificação de ciência de dados populares. Criar um modelo básico de aprendizado de máquina é tão simples quanto testar vários algoritmos pré-criados para ver qual se adapta melhor aos dados. Cada modelo tem suas próprias forças, fraquezas, arquitetura e abordagem única de ponderação de resultados.

Se você é um programador lendo esse livro e está pensando em entrar no aprendizado de máquina, não cometa o erro de escrever algoritmos a partir do zero. Eventualmente, sim, qualquer bom especialista em aprendizado de máquina precisará saber como escrever um algoritmo. No entanto, os algoritmos disponíveis estão se tornando padrões da indústria e funcionam em mais de 80% dos casos.

Escrever um algoritmo a partir do zero requer habilidades matemáticas, teoria e codificação. Teremos também um capítulo inteiro sobre algoritmos e como eles funcionam. Basta dizer que os algoritmos são essenciais para um modelo funcional de aprendizado de máquina.

4. O Treinamento

Treinar um algoritmo em um conjunto de dados é onde a mágica acontece no aprendizado de máquina. É a parte em que a máquina realmente aprende. É também a parte onde a aprendizagem de máquinas pode se tornar intensiva em recursos. Se você está tentando fazer algo complexo ou treinar um algoritmo em um

grande conjunto de dados, pode levar tempo e um poder de computação significativo para obter os resultados desejados.

O treinamento geralmente também vem com retornos decrescentes. Para uma determinada tarefa com uma resposta sim / não, você provavelmente pode obter uma precisão de 80% com uma pequena quantidade de treinamento. Chegar a 90% demoraria muito mais tempo. 95% ainda mais, e cada porcentagem adicional de precisão do modelo que você deseja, mais treinamento (e dados de treinamento) você precisará. Esse ajuste de algoritmo para precisão é uma parte importante do trabalho de um cientista de dados.

Normalmente, o treinamento de aprendizado de máquina é estático, o que significa que você não pode treinar o modelo em tempo real. Isso significa que o modelo está em treinamento ou em produção. Com mais uso na produção, o modelo não melhora. Se você quiser melhorar o modelo, terá que treiná-lo novamente de forma separada.

No entanto, é possível treinar dinamicamente um modelo. Essas aplicações são muito mais difíceis e caras de implementar. Eles também exigem que você monitore constantemente os dados em tempo real que o algoritmo está recebendo. A vantagem, claro, é que o modelo permanece responsivo aos dados recebidos e não fica desatualizado com o tempo.

Outro desafio é que durante a fase de treinamento, o algoritmo procura a correlação e não a causalidade. Um ótimo exemplo disso é o detector de camuflagem de tanque militar que mencionei acima. O algoritmo descobriu que os dias nublados estavam relacionados com a obtenção do resultado certo. O treinamento ensina o algoritmo a procurar o resultado certo, mesmo às custas dos motivos certos. Isto é legal quando a aprendizagem da máquina aponta uma variável que se correlaciona para corrigir resultados que não tínhamos pensado anteriormente em procurar. É problemático quando essa correlação acaba sendo um falso positivo de algum tipo.

Posteriormente, teremos também neste livro um capítulo completo sobre treinamento de algoritmo. Este capítulo é apenas um esboço dos conceitos básicos para que possamos começar.

5. Os Resultados

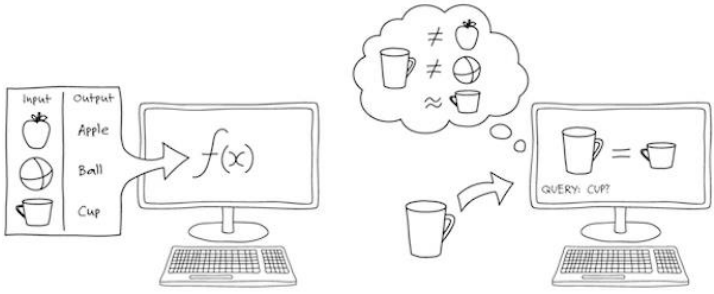
A etapa final, muitas vezes negligenciada, da aprendizagem da máquina é apresentar os resultados. O objetivo do aprendizado de máquina é produzir dados úteis para seres humanos. Um cientista de dados tem muito trabalho a ser feito para explicar o contexto, o problema e a solução de uma aplicação de aprendizagem de máquina. Além de responder como e por que o modelo funciona, os cientistas de dados também precisam apresentar os resultados de uma forma que seja acessível ao público final.

No caso do filtro de spam do Gmail, isso significa demonstrar o valor de redução de spam do filtro de aprendizado de máquina e construir uma integração para o modelo na plataforma do Gmail. Para recomendações de produtos da Amazon, isso significa testar os resultados do modelo no mundo real.

Frequentemente, o ato de preparar e usar os resultados revelará algo que estava faltando no modelo original. Assim, os projetos de aprendizado de máquina costumam ser repetitivos, adicionando mais funcionalidade e combinando vários modelos ao longo do tempo para atender às necessidades dos seres humanos no mundo real.

APRENDIZADO SUPERVISIONADO VS. NÃO SUPERVISIONADO

O aprendizado de máquina pode ser supervisionado, não supervisionado ou semi-supervisionado. As várias categorias dependem do tipo de dados e de seus objetivos sobre o que fazer com esses dados.



Supervised Machine Learning

The computer is given examples of inputs and typical outputs which it uses to develop and refine an algorithm. The algorithm is applied to new data and the outcome is used for further refinement.
E.g. Training a computer to recognize and classify similar objects based on shape.

Unsupervised Machine Learning

Unsupervised machine learning is similar to learning without a teacher. The computer learns by exploring the data and finding structure and data patterns on its own.
E.g. Learning to spot patterns in customer data based on purchasing behaviour.

APRENDIZADO SUPERVISIONADO

O aprendizado supervisionado é a abordagem mais comumente usada e bem compreendida para o aprendizado de máquina. Envolve uma entrada e saída para cada dado em seu conjunto de dados. Por exemplo, uma entrada pode ser uma imagem e a saída pode ser a resposta para "isto é um gato?"

Com o aprendizado supervisionado, o algoritmo precisa de um conjunto de dados de treinamento rotulado com as respostas corretas para que possa aprender. Esses rótulos atuam como um professor supervisionando o aprendizado. À medida que o algoritmo faz suposições sobre se há ou não um gato na imagem, o retorno do professor (os rótulos) ajudará o modelo a se ajustar. O modelo para de aprender quando atinge um nível aceitável de precisão ou fica sem dados de treinamento rotulados.

O aprendizado supervisionado é ótimo para tarefas em que o modelo precisa prever resultados. Esses problemas de previsão podem envolver o uso de estatísticas para adivinhar um valor (por exemplo, 20 kg, $ 1.498, 0,08 cm) ou categorizar dados com base em determinadas classificações (por exemplo, "gato", "verde", "feliz").

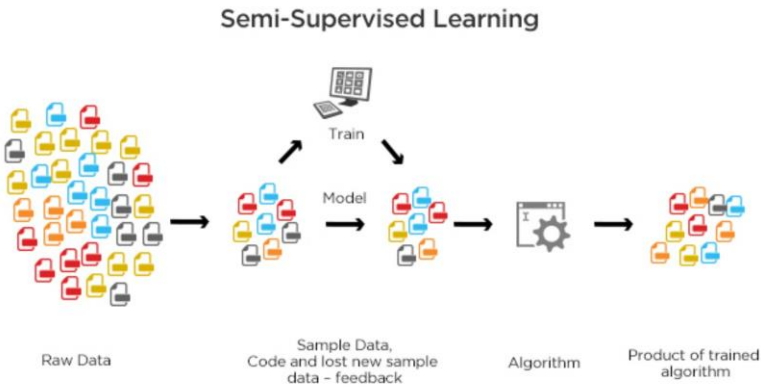
APRENDIZADO NÃO SUPERVISIONADO

Usamos o termo aprendizagem não supervisionada quando o conjunto de dados de treinamento não tem rótulos com uma resposta correta. Em vez disso, permitimos que o algoritmo tire suas próprias conclusões comparando os dados consigo mesmo. O objetivo é descobrir algo sobre a estrutura básica ou distribuição do conjunto de dados.

O aprendizado não supervisionado pode ser usado para problemas de agrupamento, onde os dados devem ser organizados em grupos semelhantes. Também podemos

usá-lo para problemas de associação, a fim de descobrir quais variáveis se correlacionam entre si.

APRENDIZADO SEMI-SUPERVISIONADO



Em muitos casos, apenas parte do conjunto de dados é rotulada, e é aí que entra o aprendizado semi-estruturado. Quando a maioria do conjunto de dados não está rotulada, geralmente devido ao custo de contratação de pessoas para rotular os dados, ainda podemos usar uma combinação de técnicas supervisionadas e não supervisionadas para tirar conclusões dos dados.

O aprendizado não supervisionado pode nos ajudar com a estrutura e distribuição do conjunto de dados. Então, podemos usar os poucos rótulos que temos como dados de treinamento supervisionado. Se usarmos esses dados no restante do conjunto de dados, poderemos potencialmente usar os resultados como dados de treinamento para um novo modelo.

QUAIS PROBLEMAS O APRENDIZADO DE MÁQUINA PODE SOLUCIONAR?

Vamos dar uma olhada em alguns exemplos de problemas que o aprendizado de máquina pode resolver:

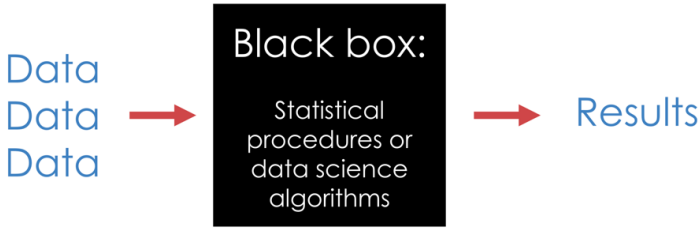
- Clientes que compraram x, provavelmente comprarão y
- Detecção de fraude baseada em dados históricos
- Previsão de ações e negociação automatizada
- Identificação de doenças em imagens médicas
- Reconhecimento de fala para controles de voz
- Previsão de classificações de degustação de vinhos com base em vinhedos e dados climáticos
- Previsão de gostos de música ou programas de TV (Spotify, Netflix)
- Química combinatória para criar novos produtos farmacêuticos
- Diagnóstico de manutenção de aeronaves
- Determinar emoções e escalar incidentes em chamadas de suporte ao cliente
- Carros com direção autônoma (com reconhecimento de objetos na estrada)
- Reconhecimento facial
- Marketing micro-direcionado e publicidade com base em dados demográficos
- Previsão do tempo baseada em padrões anteriores

Basicamente, qualquer aplicação que envolva classificação, previsão ou detecção de anomalias com

base em um grande conjunto de dados é um uso potencial para a aprendizagem de máquinas. O aprendizado de máquina está entrando rapidamente em todos os aspectos de nossas vidas e, nos próximos anos, será uma tecnologia fundamental na sociedade, de algumas formas como a Internet hoje.

A CAIXA PRETA: O QUE NÃO SABEMOS SOBRE APRENDIZADO DE MÁQUINA

Se você ler sobre aprendizado de máquina, especialmente redes neurais e aprendizagem profunda, provavelmente ouvirá referências ao aprendizado de máquina como sendo um modelo de "caixa preta". Quando falamos de caixas pretas, queremos dizer que o funcionamento interno do modelo não é exatamente claro. Por exemplo, o cérebro humano é um tomador de decisões do tipo caixa preta (pelo menos neste momento da história). Sabemos que certas partes do cérebro são responsáveis por certas funções vitais. No entanto, nós realmente não entendemos como o cérebro processa entradas e envia sinais para criar pensamentos e ações (saídas).



Complexidade semelhante se aplica a alguns algoritmos de aprendizado de máquina, especialmente aqueles que envolvem várias camadas de nós neurais ou relacionamentos complexos entre muitas variáveis. Pode ser difícil explicar, de uma maneira humana, o que o algoritmo está fazendo e por que isso funciona.

É claro que essa terminologia de caixa preta é um pouco inadequada no aprendizado de máquina. Podemos, de fato, entender a arquitetura, padrões e pesos dos diferentes nós em um algoritmo. Dessa maneira, podemos olhar para dentro da caixa preta. Porém, o que encontramos lá pode não fazer nenhum sentido racional para nós como humanos.

Nem mesmo os maiores especialistas do mundo podem explicar por que um modelo de aprendizado de máquina ponderou e combinou vários fatores da maneira que fez e, em muitos aspectos, é altamente dependente do conjunto de dados sobre o qual o modelo foi treinado. É possível que um algoritmo treinado em um conjunto de dados de treinamento

diferente possa criar um modelo completamente diferente que ainda gere resultados semelhantes.

Para esclarecer, é útil pensar em algoritmos de aprendizado de máquinas (em cenários de aprendizado supervisionado) como busca de uma função tal que f (entrada) = saída. Quando usamos o aprendizado de máquina para modelar essa função, a função geralmente é confusa, complexa e podemos não compreender totalmente todas suas propriedades relevantes. O AM nos permite dizer exatamente o que a função é, mas podemos não ser capazes de compreender o que a função faz ou por que o faz.

Nesse sentido, os modelos de AM podem ter problemas de caixa preta muito complexos de serem compreendidos. Mas todo o campo do AM não é necessariamente uma caixa preta.

Ainda assim, o fato de que às vezes não conseguimos entender e explicar os resultados do aprendizado de máquina é preocupante. À medida que a adoção dessa tecnologia está crescendo, o AM está entrando em partes de nossas vidas que têm consequências profundas e duradouras. Quando uma caixa preta prevê planos de tratamento para doenças, aciona o piloto automático de um avião ou determina sentenças de prisão, queremos ter certeza de que entendemos como essas decisões estão sendo tomadas? Ou confiamos nas máquinas e nos cientistas por trás dos algoritmos para cuidar de nossos interesses?

Este é um debate contínuo no centro da revolução do aprendizado de máquina. Por um lado, confiar nos algoritmos e modelos pode levar a salvar vidas, maior prosperidade e realizações científicas. No entanto, a implicação em transparência é real. Não seremos capazes de dizer definitivamente por que nossas previsões estão corretas, apenas que o algoritmo acredita que há 97,2% de chance de que estejam.

Não tenho uma resposta que possa encerrar bem este debate. Em vez disso, você terá que formar suas próprias opiniões com base nas vantagens e desvantagens que você vê no aprendizado de máquina ao longo deste livro e outras leituras. Se você estiver interessado neste problema, recomendo o artigo "O segredo obscuro no coração da IA" da MIT Technology Review (disponível online) para começar a aprender mais.

INDO MAIS A FUNDO

Esperamos que este capítulo tenha dado uma visão geral e fácil de digerir sobre como tudo se encaixa e o que esperar de cada capítulo componente. Nos capítulos a seguir, vamos nos aprofundar nos detalhes do aprendizado de máquina.

CAPÍTULO 2. LIMPEZA, ROTULAGEM E CURADORIA DE CONJUNTO DE DADOS

Depois que um cientista de dados define um problema que gostaria de resolver, a primeira etapa em qualquer aventura de aprendizado de máquina é encontrar um conjunto de dados com o qual trabalhar. Isso pode ser mais difícil do que parece no início. Embora estejamos certamente vivendo na era do *Big Data* (grandes dados), encontrar dados limpos e bem rotulados para o aprendizado supervisionado com as variáveis necessárias pode ser um desafio.

Escolher o conjunto de dados certo e ter dados suficientes para treinamento é fundamental para o sucesso de um projeto de aprendizado de máquina. Dados distorcidos ou incompletos podem levar à criação de um modelo de aprendizado de máquina tendencioso ou totalmente inútil.

A boa notícia é que há muitos dados em potencial por aí. Normalmente, quando um cientista de dados trabalha em um ambiente corporativo, a empresa já terá alguns dados específicos para analisar. Estes dados corporativos também podem precisar ser vinculados a dados de fontes públicas.

Por exemplo, as imagens de satélite do Landsat são atualizadas diariamente no Amazon Web Services e você pode rastrear a construção ou o desmatamento com um algoritmo de aprendizado de máquina. O mapeamento de código aberto do OpenStreetMap pode

formar a base de um problema de mapeamento do cliente. As informações do Censo dos EUA podem fornecer informações demográficas sobre uma área. Você pode encontrar genomas humanos sequenciados e disponíveis para investigar a variação genética. O Deutsche Bank divulga dados do mercado financeiro em tempo real que permitiriam um projeto de aprendizado de máquina sobre as tendências do mercado.

Não há falta de projetos potenciais. Mas antes de usar todos esses dados, os cientistas de dados precisam ter certeza de que atendem a alguns critérios.

LIMPANDO O CONJUNTO DE DADOS

Isso é bastante simples, mas a falha em remover valores ruins terá impacto no desempenho do modelo. O primeiro passo para limpar um conjunto de dados é remover quaisquer registros que estejam faltando variáveis-chave. Então, métodos estatísticos simples ajudam os pesquisadores a identificar e remover dados discrepantes. Outras informações que os cientistas de dados freqüentemente removem são, a qualquer momento, múltiplas colunas que são altamente correlacionadas. Eles também procuram por variáveis em que todo o conjunto de dados mostra uma variação próxima de zero.

Essa limpeza de dados pode frequentemente reduzir um grande conjunto de dados em uma fração de seu tamanho original e pode ser usada para o aprendizado de máquina.

NECESSIDADE DE CONJUNTOS DE DADOS MUITO GRANDES PARA AM

Alguns algoritmos simples podem aprender em um pequeno conjunto de dados. No entanto, se você tiver um problema complexo que deseja resolver com o aprendizado de máquina, precisará de um grande conjunto de dados de treinamento. Há alguns motivos para esta situação.

Pequenos conjuntos de dados podem funcionar com sucesso para aprendizado de máquina quando você usa um modelo de baixa complexidade. Porém, quanto mais nuances você quiser que seus resultados tenham, maior a probabilidade de que você ajuste o modelo aos dados. O sobre-ajuste ("*overfitting*", em inglês) ocorre quando o modelo faz suposições amplas com base em dados limitados. É chamado de sobre-ajuste porque o modelo irá inclinar em direção a pontos de dados altos, baixos ou remotos/discrepantes. A resposta correta pode estar em algum lugar mais próximo do meio, mas desde que o conjunto de dados foi limitado, o modelo captará a mensagem e os dados de treinamento com ruídos. Em essência, o modelo aprendeu os dados de treinamento muito bem e não conseguiu obter uma visão geral.

Com mais dados, o modelo pode obter médias mais precisas e começar a classificar através do ruído. Isso faz sentido intuitivamente, mas como os cientistas de dados decidem quantos dados são suficientes?

Bem, essa resposta é parte estatística e parte recursos de computação disponíveis. Isso também depende da complexidade do algoritmo.

Curvas de aprendizado

Quando os cientistas de dados têm muitos dados, eles usam algo chamado curva de aprendizado para representar graficamente a precisão da previsão versus o tamanho do conjunto de treinamento. Por exemplo, o algoritmo pode atingir 80% de precisão após 100 amostras de treinamento e 90% após 200 amostras. Os cientistas de dados podem continuar seguindo essa curva para ver onde a precisão atinge o máximo e quantas amostras de treinamento eles precisarão para chegar lá.

Validação cruzada

Outra consideração para saber se você tem dados suficientes é a validação cruzada. Além dos dados de treinamento, os cientistas de dados reservam parte do conjunto de dados original para testar se o algoritmo é bem-sucedido. Por exemplo, um esquema comum é a validação cruzada de 10 grupos (em inglês, "*10-fold cross-validation*"). O conjunto de dados original é dividido em 10 grupos iguais. Um grupo é separado e os cientistas de dados treinam o modelo usando os nove grupos restantes. Então, quando o treinamento do modelo é concluído, eles executam o modelo nos dados que separaram para testar a precisão do seu desempenho.

A validação cruzada leva mais tempo porque você precisa treinar os modelos e depois executá-los, geralmente comparando vários algoritmos para ver qual tem o melhor desempenho. Porém, o tempo extra vale a pena. A validação cruzada é essencial para construir um modelo de aprendizado de máquina bem-sucedido, pois permite que os pesquisadores identifiquem e corrijam erros no início do processo.

NECESSIDADE DE SER BEM ROTULADO

Para o aprendizado não supervisionado, tudo que você precisa é de um conjunto de dados grande e bom. A partir daí, você pode tirar algumas conclusões sobre tendências ou agrupamentos nos dados. No entanto, as aplicações de aprendizagem não supervisionadas são limitadas nos tipos de conclusões que podem tirar. Para a maioria das aplicações de aprendizado de máquina em que você queira usar variáveis de entrada para prever um resultado, precisará realizar o aprendizado supervisionado.

O aprendizado supervisionado requer um conjunto de dados rotulado com as respostas corretas. Uma maneira simples de pensar sobre isso é que o algoritmo fará uma suposição e, em seguida, usará o rótulo para verificar sua resposta. Se acertar a resposta, o algoritmo sabe como aumentar o peso que atribui aos fatores que contribuem para a resposta certa. Se obtiver a resposta incorreta, o algoritmo diminuirá ou ajustará o peso que atribui aos fatores que produzem a resposta errada.

Obviamente, o desafio é que a maioria dos dados não é rotulada. Empresas e governos coletam uma enorme quantidade de dados todos os anos, mas esses dados não vêm convenientemente com as respostas. (Se assim fosse, não haveria muito uso para aprendizado de máquina ou estatísticas preditivas!) Antes de podermos treinar um algoritmo de aprendizado supervisionado, precisamos adicionar rótulos aos dados brutos para torná-los úteis.

Por exemplo, um algoritmo pode estar funcionando em visão computacional e precisamos dele para identificar corretamente placas de sinalização pare. Podemos ter um monte de imagens, mas precisamos verificar e rotular se há ou não uma placa de pare em cada uma das imagens.

A rotulagem de dados pode ser uma das partes mais caras e demoradas do treinamento de um algoritmo de aprendizado de máquina. Também há um risco de uma rotulagem pobre ou imprecisa introduzir erros no conjunto de dados de treinamento e comprometer todo o projeto.

Se os dados ainda não tiverem rótulos, geralmente há duas maneiras de adicionar esses rótulos.

DADOS ROTULADOS POR HUMANOS

Frequentemente, usamos o aprendizado de máquina para ensinar os computadores a realizar tarefas nas quais nós, humanos, somos intuitivamente bons. A placa de pare é um bom exemplo. Quando vemos uma forma

vermelha octogonal com PARE, sabemos o que estamos olhando. Nossos cérebros são ótimos para entender o contexto. Mesmo que não possamos ver a placa inteira, se ela estiver pichada ou em um ângulo estranho, ainda podemos identificar uma placa de pare quando vemos uma. Máquinas não conseguem fazer isso intuitivamente.

Sendo assim, com frequência, a melhor forma de rotular conjuntos de dados é quando essa tarefa é feita por humanos. Cientistas de dados empregam pessoas reais para examinar conjuntos de dados e fazer o trabalho que, eventualmente, o computador aprenderá a fazer. Pode ser identificar placas de pare em fotos, estimar distâncias, ler palavras, reconhecer expressões faciais, interpretar mapas ou até mesmo fazer julgamentos estéticos ou éticos. Dizem que a rotulagem de dados pode ser o novo trabalho de "colarinho azul" da era da Inteligência Artificial. A demanda por rotuladores será muito grande, pois cada nova aplicação de AM requer um conjunto de dados de treinamento.

Rotuladores humanos são ótimos nessas tarefas. No entanto, comparados aos computadores, eles são lentos. Pagar pessoas reais para rotular os dados também é caro, além de proibido em alguns casos. Como já comentamos anteriormente, os humanos também cometem erros. Se um rotulador ou grupo de rotuladores cometem erros, então esses erros provavelmente aparecerão no modelo final.

Uma consideração adicional é que às vezes os humanos não são tão bons em rotular. Eles podem julgar erroneamente ou tirar conclusões precipitadas. Como humanos, temos excesso de confiança em nossas próprias opiniões, às vezes às custas da verdade objetiva. Quando implantamos o aprendizado de máquina em casos com maiores nuances, todas essas são considerações que devemos levar em conta.

Dito tudo isso, os humanos ainda são os melhores rotuladores de dados que temos. Entretanto, atualmente existem tentativas de fazer com que os computadores também participem da parte de rotulagem do aprendizado de máquina.

DADOS SINTÉTICOS

Os dados sintéticos são um campo em expansão no aprendizado de máquina. A ideia básica é usar um computador para gerar do zero conjuntos de dados rotulados.

Vejamos, por exemplo, nosso problema com a placa de pare. Poderíamos modelar uma placa de pare em um ambiente 3D CGI ("*computer-generated imagery*", em português: imagens geradas por computador). Então, conseguiríamos compilar imagens dessa placa de pare em diferentes panos de fundo, ângulos e condições de iluminação. O conjunto de dados resultante teria uma grande quantidade de variações que seríamos capazes de controlar. Ele já estaria rotulado com base na

condição de a placa de pare aparecer ou não na imagem compilada.

Essa abordagem é interessante porque nos permite criar conjuntos de dados complexos muito rapidamente. Os dados sintéticos já vêm pré-rotulados e formatados para serem introduzidos em um algoritmo. Também sabemos que os rótulos são objetivamente corretos. Podemos medir variáveis diversas no conjunto de dados sintético com alta precisão.

Mas é claro que também existem desvantagens. O maior desafio é a transferência de domínio. Essas compilações de imagens e outros tipos de dados sintéticos precisam ser fiéis ao mundo real. No fim das contas, o objetivo é que o modelo de aprendizado de máquina funcione no mundo real. O receio é de treinarmos a máquina com dados gerados por computador e o modelo passar a ser bom em reconhecer placas de pare computadorizadas, mas não as reais. Resolver esses problemas de fidelidade e transferência de domínio é um grande desafio para os defensores dos dados sintéticos.

Os dados sintéticos também podem não ser necessariamente mais baratos do que os dados rotulados por humanos. A criação de um conjunto de dados sintéticos requer um alto nível de especialização. Pagar esses especialistas envolveria um investimento inicial significativo. Essa abordagem provavelmente só faz sentido quando você precisa de milhares de pontos de dados, já que um evento de geração de dados

sintéticos pode ser feito muito mais facilmente em escala do que com dados rotulados por humanos.

Por fim, os dados sintéticos não podem ajudar com rótulos que são inerentemente baseados em humanos, como estética ou ética. A conclusão é que provavelmente acabaremos com uma combinação de dados rotulados sintéticos e humanos para o aprendizado supervisionado.

Capítulo 3. Escolhendo ou Escrevendo um Algoritmo de AM

Este capítulo pode rapidamente ficar muito bagunçado e confuso. Isso ocorre porque os algoritmos de aprendizado de máquina dependem de estatísticas e matemática complexas para direcionar seus resultados. Para realmente entender os algoritmos de AM, você teria que estudar aprendizado supervisionado e não supervisionado, análise de dados topológicos, métodos de otimização, estratégias de redução de dimensionalidade, geometria diferencial computacional e equações diferenciais. Porém, como este é um livro para iniciantes, e eu não sou de forma alguma um especialista em algoritmos de AM, evitarei a matemática e farei o meu melhor para explicá-los de forma simples.

Existem programas inteiros de doutorado sobre algoritmos de aprendizado de máquina. Você poderia passar anos se tornando um especialista neste campo. Portanto, não há como explicar tudo em um capítulo de livro. Dito isto, se o conteúdo deste capítulo lhe interessar, a busca por um doutorado em aprendizado de máquinas pode valer muito a pena. As empresas de tecnologia estão recrutando doutores com PhD e oferecendo a eles salários de 300 a 600 mil dólares por ano para escrever algoritmos para as melhores e mais novas aplicações de AM.

Não tenho doutorado em aprendizado de máquina e, se você está lendo este livro, provavelmente é um iniciante nesses conceitos. Então, vamos dar uma olhada nas funções mais básicas de um algoritmo de AM, sem entrar em matemática.

CONCEITOS BÁSICOS

Já abordamos os fundamentos de como funciona o aprendizado de máquina. Agora, vamos nos aprofundar no que exatamente um algoritmo faz com os dados. Cada algoritmo é diferente, mas existem algumas semelhanças entre eles:

- Entradas - todos os algoritmos precisam de algum tipo de dado de entrada. Em aplicações de ciência de dados, isso pode ser apenas uma única variável. Mais provavelmente, no entanto, o modelo estará aprendendo a relação entre dezenas, centenas ou mesmo milhares de variáveis a qualquer momento.

  Para aplicações mais complexas, como visão computacional, precisamos de meios para transformar informações visuais em variáveis que o computador possa entender. Existem diferentes abordagens, dependendo do contexto e do problema que você está tentando resolver. Nem precisamos dizer que até inserir dados em um algoritmo pode ser complicado, antes mesmo que a máquina faça qualquer aprendizado.

A escolha ou criação de um algoritmo depende muito dos dados que você tem para alimentá-lo e do contexto.

- Vetores de saída - no final de qualquer projeto de aprendizado de máquina, você precisa de algum tipo de saída. No entanto, nem sempre está claro exatamente quais dados você precisa para satisfazer seu projeto. Escolher vetores de saída pode ser mais complicado do que parece à primeira vista.

  Claro que para muitos projetos a saída será óbvia, dependendo de seus objetivos. Porém, à medida que o aprendizado de máquina penetra em áreas com mais nuances e ambíguas, escolher e coordenar as saídas pode ser uma tarefa em si. Você não pode escolher o algoritmo certo para o seu projeto se não tiver uma ideia clara do resultado esperado.

- Ajuste - algoritmos de aprendizado de máquina usam ciclos de resposta para ajustar um modelo aos dados. Isso pode acontecer de diferentes maneiras. Às vezes, um algoritmo tentará uma combinação aleatória de fatores até que um comece a funcionar, e essa combinação receberá um peso maior em testes de treinamento futuros. Outras vezes, o algoritmo possui um método integrado para localizar e ajustar uma tendência

nos dados que se ajusta gradualmente ao longo do tempo.

É aqui que os cientistas de dados devem ser cuidadosos. Às vezes, um algoritmo aprende a ajustar seus dados de treinamento muito bem. Ou seja, o modelo se tornou muito específico para os dados em que foi treinado e não prevê mais tendências gerais ou classificações no mundo real. Em essência, o algoritmo aprendeu seus dados de treinamento muito bem. Isso é chamado de "*overfitting*" (sobre-ajuste) e é um conceito importante para entender no aprendizado de máquina. Quando cientistas de dados treinam modelos, eles precisam se certificar de que seus modelos percorram uma linha tênue entre fazer previsões específicas e serem precisos em geral.

Os cientistas de dados passam muito tempo pensando e ajustando seus algoritmos para mitigar o sobre-ajuste. No entanto, eles também testam vários algoritmos ao mesmo tempo, lado a lado, para ver quais funcionam melhor após o treinamento.

Uma parte importante da escolha ou escrita de um algoritmo é entender como o algoritmo se ajusta ao longo do tempo em resposta aos dados de treinamento. Esses ciclos de respostas são frequentemente onde a matemática complexa

entra em jogo para ajudar o algoritmo a decidir quais fatores contribuíram para seu sucesso e, portanto, devem ser mais ponderados. Eles também ajudam o algoritmo a determinar quanto aumentar ou diminuir o peso de um fator contribuinte.

TIPOS DE ALGORITMOS POPULARES

Ok, então cobrimos uma visão geral de como funciona um algoritmo. Vejamos alguns dos mais populares para obter detalhes mais específicos sobre como cada um funciona.

REGRESSÃO LINEAR

Este é um algoritmo simples que se baseia em conceitos ensinados na maioria das aulas de Estatística 1. A regressão linear é o desafio de ajustar uma linha reta a um conjunto de pontos. Esta linha tenta prever a tendência geral para um conjunto de dados e você pode usá-la para fazer uma previsão de probabilidade de novos pontos de dados.

Existem várias abordagens para a regressão linear, mas cada uma é essencialmente focada em encontrar a equação de uma linha reta que se ajusta aos dados de treinamento. Conforme você adiciona mais dados de treinamento, a linha se ajusta para minimizar a distância de todos os pontos de dados. Assim, a regressão linear funciona melhor em conjuntos de dados muito grandes.

Este é um tipo de algoritmo bastante simples, mas uma das principais máximas do aprendizado de máquina é não usar um algoritmo complexo onde um simples funciona muito bem.

REGRESSÃO LOGÍSTICA

Se a regressão linear era uma linha reta em um plano 2D, a regressão logística é sua irmã mais velha que usa linhas curvas em uma área multidimensional. É muito mais poderosa do que a regressão linear, mas também é mais complexa.

A regressão logística pode lidar com mais de uma variável explicativa. É um algoritmo de classificação e suas saídas são binárias (uma escala de 0 a 1). Como resultado, ele modela a probabilidade (por exemplo, “0,887” ou “0,051”) de que a entrada é parte de uma determinada classificação. Se você aplicá-lo a várias classificações, obterá a probabilidade do ponto de dados pertencer a cada classe. O mapeamento dessas probabilidades dá a você uma curva multiplanar não linear conhecida como "sigmoide". A regressão logística é o algoritmo mais simples para aplicações não lineares.

ÁRVORES DECISÓRIAS

Se você já viu um fluxograma, entenderá a ideia básica por trás de uma árvore de decisão. A árvore estabelece um conjunto de critérios; se o primeiro critério for um “sim”, o algoritmo se move ao longo da árvore para a direção sim. Se for um "não", o algoritmo se move na outra direção. Os algoritmos da árvore de decisão

ajustam os critérios e as respostas possíveis até que forneçam uma boa resposta de forma consistente.

No aprendizado de máquina moderno, é raro ver uma única árvore de decisão. Em vez disso, muitas vezes são incorporadas a outras árvores simultaneamente para construir algoritmos de tomada de decisão eficientes.

Floresta Aleatória

A floresta aleatória é um tipo de algoritmo que combina várias árvores de decisão. Ela apresenta o conceito de "aprendiz fraco" (em inglês, *weak learner*) ao algoritmo. Basicamente, um aprendiz fraco é um preditor que se sai mal sozinho, mas quando usado em conjunto com outros aprendizes fracos, o conhecimento de muitos produz um bom resultado.

Árvores de decisão implementadas aleatoriamente são os aprendizes fracos em uma floresta aleatória. Cada árvore de decisão aprende como parte da implementação do algoritmo. No entanto, um preditor forte e abrangente também está aprendendo como combinar os resultados das várias árvores.

Método K-Means de Clusterização

Este é um algoritmo de aprendizado não supervisionado que tenta agrupar os dados em k número de clusters (agrupamentos). Embora não seja supervisionado, o cientista de dados precisa fornecer orientações desde o início. Eles definirão imagens ou pontos de dados que devem ser o centro de cada cluster. Em outras palavras,

pontos de dados que são arquetípicos do que o cluster representa. Durante o treinamento, todas as imagens ou pontos de dados são associados ao cluster de que estão mais próximos. Eventualmente, esses pontos de dados convergem com seus clusters apropriados.

Existem outros métodos mais rápidos ou mais otimizados para o agrupamento não supervisionado. No entanto, K-means continua popular porque é bem estabelecido, documentado e geralmente eficaz.

K Vizinhos Mais Próximos

K-Vizinhos Mais Próximos (KNN) é um algoritmo de classificação. Ele compartilha algumas semelhanças com o agrupamento K-Means, mas é fundamentalmente diferente porque é um algoritmo de aprendizagem supervisionado, enquanto o K-Means não é supervisionado. Daí a ligeira diferença na terminologia de agrupamento para classificação. O KNN é treinado usando dados rotulados para que possa rotular dados futuros. Já o K-Means é capaz apenas de tentar agrupar pontos de dados.

O KNN compara novos pontos de dados com os pontos de dados existentes do conjunto de dados de treinamento rotulado. Em seguida, procura os "vizinhos mais próximos" desses novos dados e associa esses rótulos.

Análise de Componentes Principais

A Análise de Componentes Principais (PCA) reduz um conjunto de dados às suas tendências principais. É um algoritmo não supervisionado que você usaria em um conjunto de dados muito grande para entender os dados em termos mais simples. Ela reduz as dimensões de seus dados. Mas ela também se concentra na grande variação entre as dimensões (ou componentes principais) para que você não perca o comportamento do conjunto de dados original.

O QUE É NECESSÁRIO PARA ESCREVER UM NOVO ALGORITMO

Abordamos alguns dos principais algoritmos e ainda há vários outros que constituem a base da teoria do aprendizado de máquina. Além desses algoritmos básicos, no entanto, é raro alguém inventar algo realmente novo. Normalmente, novos algoritmos são melhorias em teorias existentes. Ou eles personalizam um algoritmo para uso em um novo cenário.

Parte do motivo pelo qual novos algoritmos raramente são inventados é porque isso é muito difícil. A criação de um algoritmo requer um grande domínio de matemática complexa. Também requer provas e testes muito amplos. Além disso, os algoritmos mais simples e óbvios já foram inventados.

Mas isso não é tudo. Bons algoritmos são eficazes e eficientes, uma combinação complicada de definir. O aprendizado de máquina é tanto um problema

computacional com milhares de pontos de dados, quanto é um problema de matemática. Algoritmos de depuração também podem ser muito difíceis, pois não é simples identificar onde as coisas deram errado.

Sempre que possível, um projeto de aprendizado de máquina deve utilizar os algoritmos existentes, testados e revisados. Codificar seus próprios algoritmos do zero ou remendar uma abordagem híbrida é desaprovado porque pode introduzir erros e resultados lentos.

Às vezes, os desenvolvedores e cientistas de dados precisarão ajustar ou implementar um algoritmo existente em um novo contexto. Ou talvez um algoritmo existente não seja rápido o suficiente para a aplicação desejada. No entanto, a maioria dos aplicativos de aprendizado de máquina pode usar algoritmos e bibliotecas existentes com eficácia, sem a necessidade de codificar do zero.

Capítulo 4. Treinar e implementar um algoritmo

Este é o passo onde acontece o verdadeiro aprendizado da máquina. Depois de preparar o conjunto de dados, os cientistas de dados selecionam vários algoritmos semelhantes que eles acham que podem funcionar para realizar a tarefa em questão. Agora, o desafio é treinar esses algoritmos no conjunto de dados e comparar os resultados.

Muitas vezes, pode ser difícil dizer qual algoritmo funcionará melhor para uma aplicação de aprendizado de máquina antes de iniciar. Por esse motivo, a prática recomendada é treinar vários algoritmos primeiro, selecionar um ou alguns que tenham o melhor desempenho e, em seguida, ajustar esses algoritmos até obter um modelo que funcione melhor para suas necessidades.

Quando dizemos "melhor", isso pode significar várias coisas. Obviamente, queremos que o modelo faça previsões precisas, portanto a precisão é um componente importante. No entanto, se o modelo exigir muitos recursos ou tempo para obter esses resultados, pode fazer mais sentido escolher um algoritmo mais simples. Obteremos resultados um pouco menos precisos, mas eles virão muito mais rapidamente.

Programação envolvida

O aprendizado de máquina está situado na interseção entre a estatística, o cálculo e a ciência da computação. Como estamos lidando com máquinas, naturalmente precisaremos escrever instruções de aprendizado de máquina em uma linguagem de programação. Com o aumento do interesse no AM, ele rapidamente tem se tornando uma área com enorme crescimento para novos desenvolvedores de software. Habilidades em aprendizado de máquina são altamente valiosas

Até agora, não falamos sobre as linguagens de programação e abordagens que os desenvolvedores usam para codificar e criar suas aplicações de aprendizado de máquina. Esta seção será apenas uma breve visão geral de seus principais atores.

Python é de longe a linguagem mais popular para a criação de aplicações de AM. É também a linguagem preferida em pesquisas de desenvolvedores sobre aprendizado de máquina. Uma grande parte do sucesso do Python é sua simplicidade em comparação com outras linguagens de programação. Além disso, a biblioteca de código aberto do Google para algoritmos de AM, chamada TensorFlow, é baseada em Python. São grandes os recursos e a comunidade para aplicações de aprendizado de máquina desenvolvidos em Python.

As linguagens Java e C ou C ++ seguem a Python com uma ampla margem de popularidade. São linguagens mais antigas e permitem uma otimização de nível inferior para o ambiente onde o algoritmo será executado. Java e C e C ++ são usadas em muitas

aplicações, não apenas no aprendizado de máquina. Isso significa que há muitos desenvolvedores por aí que entendem essas linguagens. Existem algumas bibliotecas de AM para essas linguagens, porém nada na escala do TensorFlow.

R é outra linguagem de programação que freqüentemente entra na conversa sobre aprendizado de máquina. É uma linguagem especializada projetada para aplicações de ciência de dados. Apesar do fato de a R certamente ter seu lugar no aprendizado de máquina, é raro acontecer de um projeto a escolher como sua linguagem principal ou preferida. Ao invés disso, trata-se mais de uma linguagem complementar às listadas acima.

Claro que é possível escrever códigos de AM em muitas linguagens diferentes. Existem outras linguagens que se especializam em certas áreas de estatística, ciência de dados ou modelagem. Julia, Scala, Ruby, Octave, MATLAB e SAS são opções que ocasionalmente surgem em projetos de aprendizado de máquina. No entanto, essas linguagens são as exceções e não a regra.

ESTÁTICA VS. DINÂMICA

Depois de escolher uma linguagem de programação e instalar uma biblioteca para ajudá-lo a implementar os algoritmos que deseja executar, você está pronto para começar a treinar seus algoritmos.

Existem dois tipos de treinamento de aprendizado de máquina. O primeiro é o estático, que recebe treinamento offline e depois é concluído até o momento em que os cientistas de dados iniciam uma nova sessão de treinamento. O segundo é o dinâmico, onde o modelo continua aprendendo na produção, indefinidamente.

Os modelos estáticos são muito mais fáceis de construir. Eles também são mais fáceis de testar a precisão e tendem a apresentar menos problemas na implantação. Se seus dados não mudam com o tempo, ou mudam muito lentamente, um modelo estático é o melhor caminho a seguir, já que é mais barato e fácil de manter.

Modelos dinâmicos são muito mais trabalhosos para serem implementados. Eles também exigem monitoramento constante dos dados recebidos para garantir que não distorçam o modelo de maneira inadequada. Uma vez que os modelos dinâmicos se adaptam aos dados variáveis, eles são muito melhores em prever coisas como mercados ou clima, onde os padrões estão constantemente em fluxo.

AJUSTE E ENGENHARIA DE ATRIBUTOS

O trabalho de um cientista de dados não consiste apenas em escolher um punhado de algoritmos e deixá-los rodar. Para obter o desempenho ideal, a pessoa que programa o algoritmo deve definir os parâmetros de entrada que entrarão no algoritmo. Como os problemas de aprendizado de máquina costumam ser complexos,

pode ser difícil decidir quais parâmetros são relevantes e quantos incluir.

Tentar diferentes combinações de parâmetros e refinar o melhor mix é conhecido como ajuste de algoritmo. Não há uma resposta absolutamente correta aqui. Em vez disso, cada trabalho de ajuste é uma questão de corresponder o algoritmo ao contexto em que está sendo implantado.

Outro conceito relacionado ao ajuste é a engenharia de atributos (em inglês, *feature engineering*). Às vezes, como no caso do reconhecimento de imagem, alimentar um computador com um fluxo de dados não é suficiente para que ele dê sentido ao que está vendo. Embora a aprendizagem profunda (em inglês, *deep learning*) e as redes neurais tenham progredido com os computadores, aprendendo com imagens, a engenharia de atributos é uma maneira prática de dizer a um computador o que procurar. Você pode criar um atributo que ajude um computador a identificar uma linha reta ou a borda de um objeto. Uma vez que codificamos esse atributo manualmente, este tecnicamente não é um aprendizado de máquina, porém agora a máquina sabe o que procurar.

Os atributos de engenharia podem aumentar drasticamente o desempenho.

DESCARTANDO UM ALGORITMO

Se tudo correr bem, o resultado é um modelo que aprendeu a fazer com precisão previsões, agrupamentos ou classificações em seus dados.

No entanto, o lado negro do aprendizado de máquina são os algoritmos que não funcionam. Atualmente, há muito tempo e dinheiro sendo investidos em aplicações de AM. Infelizmente, muitas dessas aplicações não serão bem-sucedidas.

Talvez os algoritmos tenham sido mal escolhidos ou implementados. Mais provavelmente, o projeto não tem o suficiente ou o tipo certo de dados para ter sucesso. É subnotificada a frequência com que os projetos de aprendizado de máquina falham.

O frustrante é que pode ser difícil dizer por que seu projeto está falhando. Você poderia ter toneladas de dados e testar e ajustar muitos algoritmos sem sucesso. Isso é especialmente verdadeiro com problemas complexos ou algoritmos que implementam redes neurais de várias camadas ou florestas aleatórias. É difícil dizer onde as coisas deram errado. Às vezes, os cientistas de dados investem muito tempo em um projeto, apenas para descobrir que precisam jogar tudo fora e começar de novo com mais dados novos ou diferentes.

Esta pode parecer uma seção estranha de ser incluída em um livro tão otimista sobre aprendizado de máquina. Entretanto, acho importante destacar o fato de que ainda há muito que não sabemos sobre como

criar e usar projetos de AM. Projetos falham o tempo todo e conserta-los é difícil. Essa é uma realidade importante do aprendizado de máquina. É fundamental reconhecer que só porque um modelo de AM produz uma resposta, não significa que ele está sempre certo ou incontestável.

Devemos respeitar e admirar o AM como uma ferramenta. Mas no final, é apenas isso: uma ferramenta.

CAPÍTULO 5. APLICAÇÕES DO APRENDIZADO DE MÁQUINA NO MUNDO REAL

Agora que você tem um entendimento básico de como o aprendizado de máquina funciona, é interessante dar uma olhada nos exemplos do dia a dia em que você pode nem o ter reconhecido.

TRANSPORTES

Ao abrir o Google Maps para obter rotas, você está usando um modelo dinâmico de AM. Ele usa dados anônimos de telefones celulares de motoristas em sua área para obter os tempos de trajetos para vários caminhos. O modelo também integra dados do Waze sobre bloqueios de estradas, acidentes e outros relatórios de usuários. Com a união destes dados, o

modelo prevê a rota mais rápida e o tempo estimado de chegada com base em informações em tempo real.

Os aplicativos Uber e Lyft baseiam-se nesses dados com seus próprios algoritmos de aprendizado de máquina que direcionam o cálculo dinâmico de preços e tarifas. Eles também informam quanto tempo você espera por um motorista e o tempo provável para chegar ao seu destino, até mesmo contando com a possibilidade de pegar e levar outras pessoas no caso de opções de compartilhamento de carona do Uber Pool ou do Lyft Line.

Esses mesmos cálculos de rota, logística e chegada também se aplicam a caminhões de longa distância, fretes e até mesmo à navegação aérea. Os modelos ajudam a prever a maneira mais rápida e segura de transportar mercadorias e pessoas, maximizando a eficiência.

Recomendações de Produtos

Basicamente, sempre que uma empresa faz uma recomendação on-line, você pode presumir que um algoritmo de aprendizado de máquina ajudou a fazer essa previsão. A Amazon sabe em quais produtos você pode estar interessado, com base no que você viu e comprou antes. A Netflix sabe quais filmes você gostaria, porque aprende com todos os filmes que você assistiu antes.

Isso vai mais além do que apenas dar recomendações personalizadas, mas também se aplica à publicidade. O Facebook conhece muitos dados pessoais seus, e eles os usam para personalizar os anúncios que mostram a você. O mesmo pode ser dito para o YouTube, Twitter, Instagram e todas as outras mídias sociais.

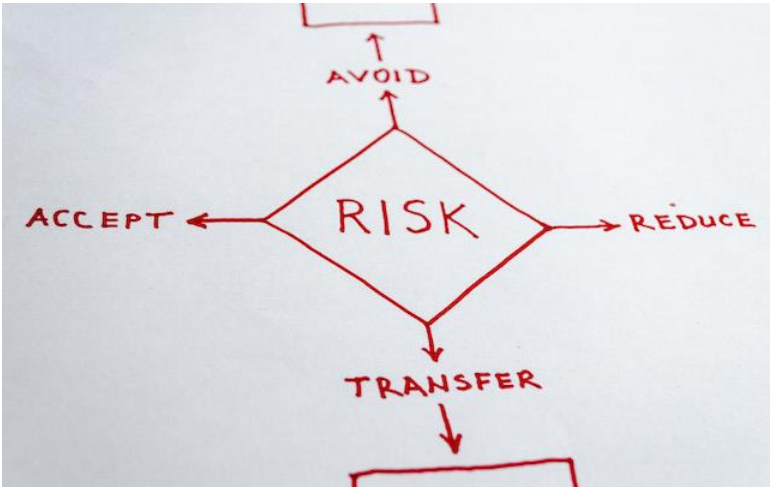
Além disso, o Google usa suas informações pessoais para personalizar os resultados que você recebe ao realizar uma pesquisa. Por exemplo, é mais provável que lhe recomende empresas locais de sua cidade ou artigos de sites ou escritores que você já visitou. Semelhante à mídia social, o Google também está personalizando seus anúncios para você. Não acredita em mim? Faça uma pesquisa no Google em seu

navegador e, em seguida, faça a mesma pesquisa em uma janela anônima do navegador (elimine os cookies e as informações de login). Para a maioria das pesquisas, especialmente tópicos que você pesquisou antes, você verá que está obtendo resultados diferentes.

Até mesmo o aprendizado de máquina presencial mudará a forma como compramos produtos. Os principais varejistas estão procurando aplicações de visão computacional que identifiquem o que você já tem em sua cesta e possam fazer recomendações. Outros sistemas estão utilizando o reconhecimento facial para identificar quando os clientes estão perdidos ou confusos e, assim, podem solicitar a um funcionário para que ajude. Esses sistemas ainda estão em sua infância, mas representam as maneiras como o aprendizado de máquina está se integrando a todos os aspectos da vida, incluindo as interações entre humanos.

FINANÇAS

Todos os grandes bancos estão usando o aprendizado de máquina para ajudar a simplificar suas operações. Em tecnologia regulatória, os algoritmos de AM podem ajudar os bancos a identificar se seus processos e documentação estão em conformidade com os padrões governamentais. Outros algoritmos de aprendizado de máquina preveem tendências de mercado ou dão ideias de investimento.



Para solicitações de empréstimo ou linhas de crédito, o AM pode ajudar os bancos a prever o risco de emprestar a um determinado cliente. Esses modelos podem sugerir termos e taxas individualizadas para o requerente. No setor bancário, o reconhecimento de caracteres com a tecnologia de AM torna possível depositar um cheque usando a câmera de seu smartphone. O aprendizado de máquina também pode detectar e impedir que transações fraudulentas sejam compensadas em sua conta.

Assistentes de Voz, Casas Inteligentes e Carros

As assistentes de voz Siri e Alexa contam com o aprendizado de máquina para entender e responder à fala humana. A Inteligência Artificial (IA) de

conversação é o que há de mais moderno em AM e treinamento de rede neural. Temos sido muito bons no reconhecimento de fala e respondendo a perguntas básicas como "Como vai ser o tempo hoje?" O próximo desafio é obter uma IA de conversação que possa falar sobre música, literatura, eventos atuais ou outras ideias complexas.

O papel da voz só continuará a se expandir nos próximos anos, conforme passarmos a contar cada vez mais com nossos assistentes pessoais. Isso é especialmente poderoso quando combinado com o movimento em direção a casas inteligentes e veículos autônomos. É possível imaginar um futuro onde você possa controlar todos os aspectos da sua casa e transporte intuitivamente, falando com um assistente de voz. Por sua vez, cada um desses sistemas - como termostatos inteligentes, sistemas de segurança

inteligentes e carros autônomos - usa seus próprios algoritmos de aprendizado de máquina para realizar as tarefas que exigimos deles.

Conclusão

Claro, existem muitos outros casos de uso para aprendizado de máquina em saúde, manufatura, agricultura e em todos os outros lugares de nossas vidas. O AM é útil em qualquer lugar onde haja dados e precisarmos de ajuda para entender, prever ou usar esses dados.

O aprendizado de máquina é poderoso e continuará ganhando destaque em nossa vida diária. Como tal, é importante que todos tenham uma compreensão básica de como funciona, suas possíveis falhas e enormes oportunidades. Esperançosamente, este guia rápido para iniciantes forneceu uma base sólida para o leigo interessado no básico.

Dito isto, há muito mais sobre aprendizado de máquina e que não é abordado neste livro! Existem ótimos recursos disponíveis online e impressos para expandir ainda mais seu conhecimento sobre essa importante tecnologia. Espero que este seja apenas o começo de sua jornada no aprendizado de máquina.

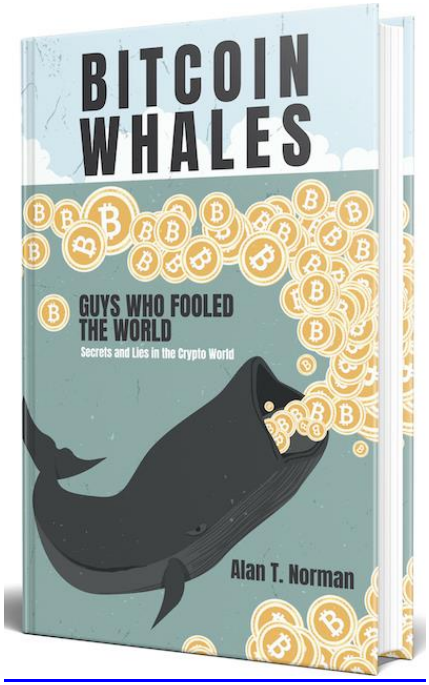
Obrigado pela leitura.

SOBRE O AUTOR

Alan T. Norman é um orgulhoso, experiente, e ético hacker da cidade de São Francisco. Após receber seu diploma de Bacharel em Ciências pela Universidade de Stanford, Alan agora trabalha para uma empresa de médio porte de Tecnologia Informacional no coração de São Francisco. Ele aspira um dia poder trabalhar para o governo dos Estados Unidos como hacker de segurança, mas também ama ensinar pessoas sobre o futuro da tecnologia. Além disso Alan acredita firmemente que o futuro dependerá bastante de "geeks" (nerds) de computador, tanto para a segurança quanto para o sucesso de empresas e futuros ramos parecidos. E em seu tempo livre ele adora destrinchar e analisar tudo sobre basquete.

LIVRO BÔNUS BALEIAS DE BITCOINS

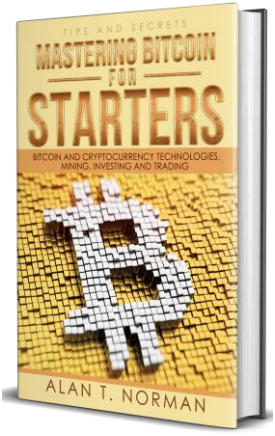
ENCONTRE O LINK PARA O LIVRO BÔNUS ABAIXO



Link do Livro

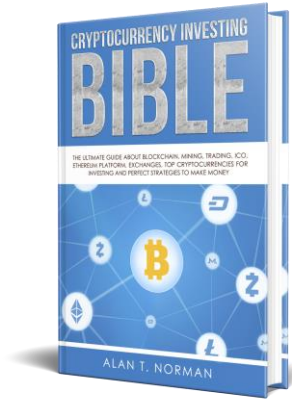
OUTROS LIVROS DO ALAN T. NORMAN:

Dominando Bitcoins para Iniciantes *(http://amzn.to/2AwSNy0)*



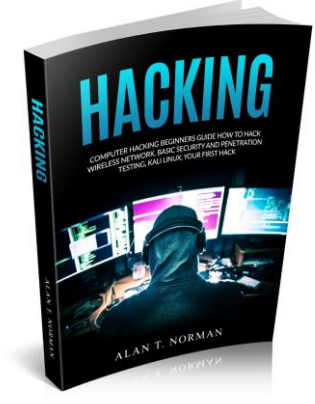
Bíblia do investimento em criptomoeda

(http://amzn.to/2zzB8IP)

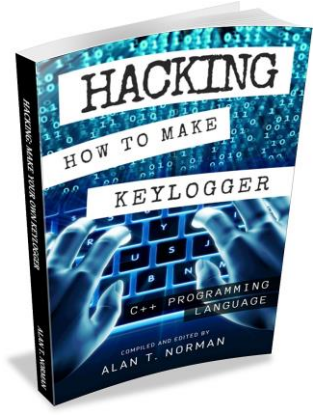


[Tecnologia Blockchain Explicada](http://mybook.to/BlockchainExplained)
http://mybook.to/BlockchainExplained

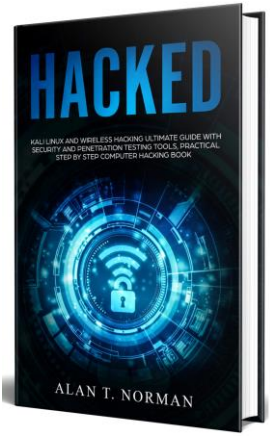
Hacking: Guia para iniciantes em pirataria informática *(www.amazon.com/dp/B01N4FFHMW)*



Hacking: Como fazer seu próprio Keylogger em linguagem de programação C++



HACKED: Guia definitivo de Kali Linux e Wireless Hacking *(https://www.amazon.com/dp/B0791WSRNZ)*



UMA ÚLTIMA COISA...

VOCÊ GOSTOU DO LIVRO?

CASO SIM, CONTE-ME DEIXANDO UM COMENTÁRIO NA AMAZON! As resenhas são a força vital de autores independentes. Eu apreciaria até mesmo algumas palavras e avaliação, caso você tenha tempo para isso.

SE VOCÊ NÃO GOSTOU DESTE LIVRO, POR FAVOR, ME CONTE! Envie-me um email para alannormanit@gmail.com e diga-me o que não gostou! Talvez eu possa mudar isso. No mundo de hoje, um livro não precisa ficar estagnado, ele pode melhorar com o tempo e o feedback de leitores como você. Você pode impactar este livro e agradeço seus comentários. Ajude a tornar este livro melhor para todos!