

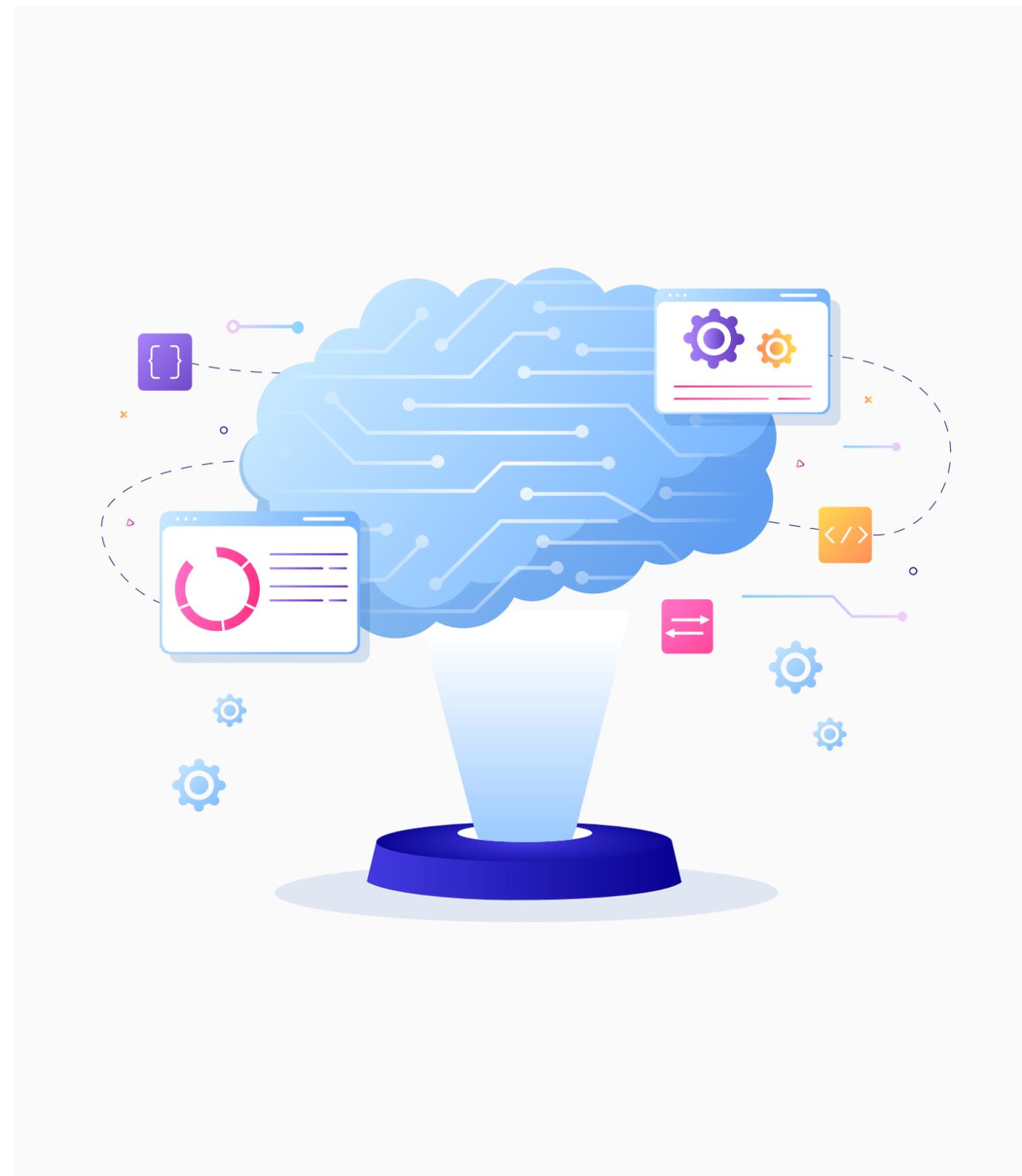
PRINCIPAIS ALGORITMOS DE MACHINE LEARNING

Introdução

Com a evolução da inteligência artificial, pesquisadores da ciência da computação estão a todo momento sugerindo novas abordagens para tornar os computadores ainda mais autônomos e capazes de executar tarefas com precisão.

Por essa razão, tem aumentado a utilização de algoritmos de machine learning para resolver problemas complexos. Entre seus benefícios está a capacidade das máquinas aprenderem sem serem explicitamente programadas para determinada atividade.

Neste e-book, serão apresentadas as 3 principais abordagens para aprendizagem de máquinas: supervisionada, não supervisionada e por reforço. Além disso, será explicado, resumidamente, como funcionam os principais algoritmos de machine learning, mostrando exemplos de casos de uso.



Sumário

04 O que é Machine Learning?

05 Métodos de Machine Learning

07 Quais os principais algoritmos de Machine Learning?

11 Como escolher o algoritmo ideal?

13 Conclusão

O que é Machine Learning?

Tradicionalmente, os programadores sempre desenvolveram sistemas especificando cada passo em código, de forma clara. Para isso, engenheiros de software precisavam conhecer bem os problemas que buscavam resolver e ter uma boa noção de todas as variáveis e possibilidades envolvidas.

Machine learning (ML) é um subcampo da inteligência artificial que visa resolver problemas mais complexos, dos quais os programadores ainda não têm um conhecimento tão amplo.

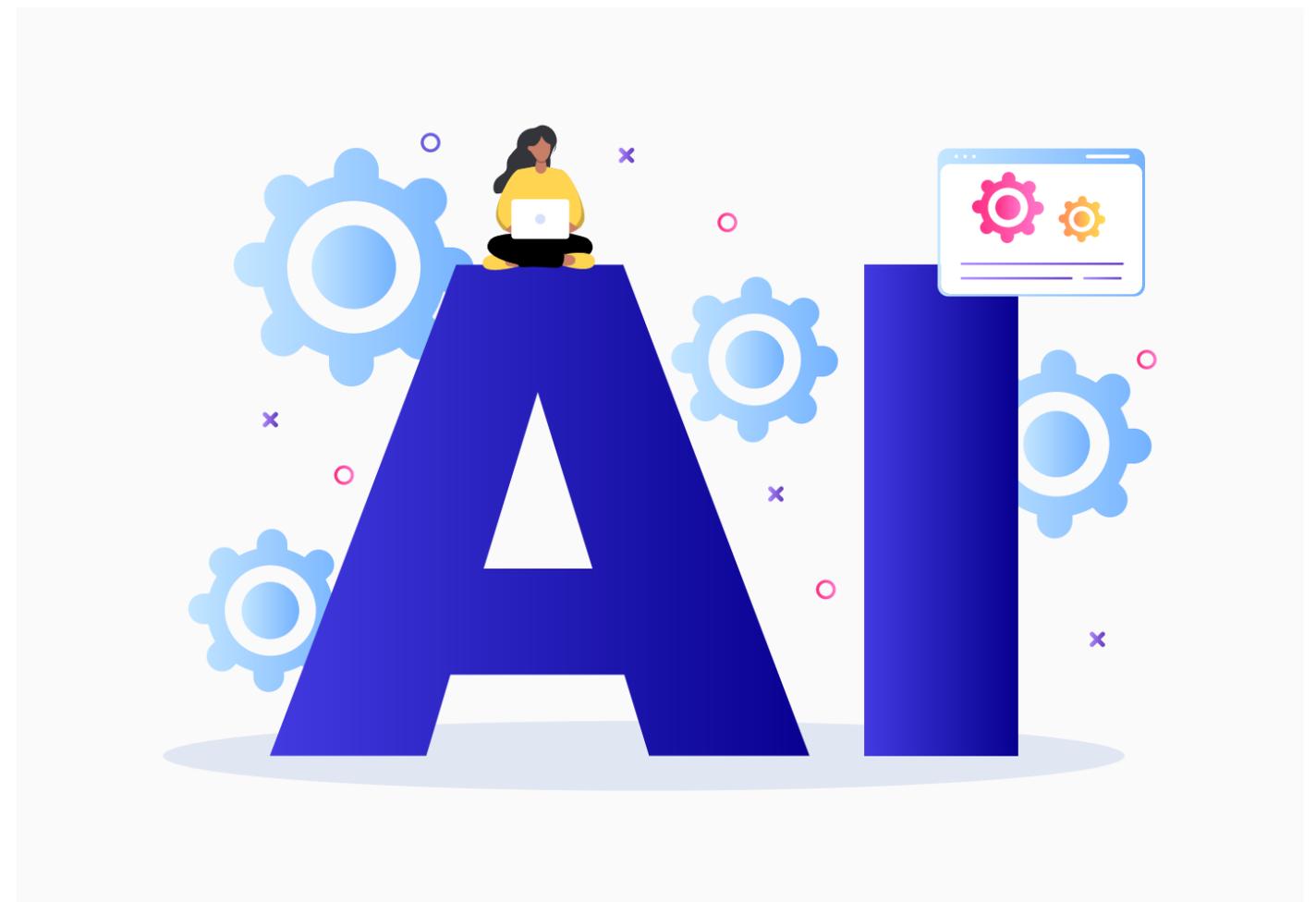
Também se aplica quando é inviável mapear todas as variáveis e possibilidades. Para essa finalidade, os sistemas são estimulados a solucionar problemas de forma autônoma, aprendendo sobre eles e propondo uma saída.

Nesse contexto, os algoritmos analisam dados de entrada, processam e preveem saídas possíveis dentro de um intervalo definido. No processo, eles tentam diferentes abordagens e otimizam sua capacidade de chegar ao resultado.

Os sistemas artificiais no ML passam por duas fases: treino e

execução. Na primeira, são alimentados por um conjunto de entrada para que sejam contextualizados com a situação específica que envolve o problema.

Então, na segunda fase, eles começam a deduzir os caminhos a partir do que aprenderam na etapa anterior. Essas duas fases variam a depender do tipo de algoritmo, como veremos mais adiante neste e-book.



Embora não seja novo, o machine learning é um campo do conhecimento em desenvolvimento, ou seja, ganha novas abordagens e novos modelos frequentemente. Essas evoluções buscam resolver problemas com ainda mais precisão e oferecer opções especiais para que cada ferramenta se adapte ao problema específico.

Métodos de Machine Learning

A seguir, discutiremos os principais métodos do ML e suas características.

APRENDIZAGEM SUPERVISIONADA

A aprendizagem supervisionada é útil nos casos em que uma propriedade (rótulo) está disponível para um determinado conjunto de dados (conjunto de treinamento). Todas as entradas e saídas são conhecidas, mas precisam ser previstas para outras instâncias.

Então, esses dados são passados para o sistema de aprendizagem, que tem como função descobrir caminhos e ajustar seu próprio modelo para chegar aos resultados esperados.

Um exemplo prático seria ensinar uma máquina a reconhecer e

categorizar e-mails, separando os que são relevantes dos que são SPAM.

Na fase de treinamento, são transmitidas para o algoritmo amostras de ambos os casos, visto que os dois blocos de informação são plenamente conhecidos. Já na fase de execução, o sistema terá a capacidade de determinar se um e-mail novo é ou não uma mensagem indesejada.

Ou seja, o software já sabe quais entradas estão associadas com quais saídas, mas precisa aprender um meio de entender essa associação.

Nesse processo, ele tenta identificar padrões e estabelecer previsões que ajudam a otimizar a abordagem. Esse modelo tem esse nome porque é como se o operador humano estivesse sempre dando assistência ao sistema, ensinando-o de uma forma direta.

Os resultados desse tipo de algoritmo são geralmente marcados como classificação e regressão. O primeiro diz respeito a uma forma de mapear elementos iguais em categorias específicas, como no exemplo acima, “spam” e “não spam”. Já o segundo consiste em identificar uma tendência para os dados que permite, inclusive, predizer o futuro com base em dados históricos.

Uma aplicação comum é no setor de segurança, com o uso de ferramentas para identificar comportamentos suspeitos em uma rede e garantir proteção contra ataques.

APRENDIZAGEM NÃO-SUPERVISIONADA

Já na aprendizagem não-supervisionada o desafio é descobrir relações implícitas em um conjunto de dados não rotulados. Dessa forma, o algoritmo fica encarregado de identificar padrões para rotular os dados.

Um exemplo muito comum é a empresa que fornece como base para um sistema de ML um conjunto de dados sobre os clientes e espera que ele identifique possíveis atributos comuns e padrões de comportamento para criar ofertas específicas e segmentadas. No começo, a companhia não conhece bem as informações a ponto de saber as possíveis saídas.

É como um aprendizado sem assistência humana, com o sistema ditando os caminhos a seguir. Nesse modelo de algoritmo, as estratégias comuns são *clustering* e redução de dimensão.

Abordagens como essas podem ser utilizadas em sistemas de recomendação, em que, com base em informações colaborativas ou específicas, um software consegue filtrar conteúdos ideais em meio a

uma miríade deles.

É muito usado em lojas de e-commerce, para recomendar produtos que possam ser interessantes para o cliente que visita o site.

Similarmente, há serviços de streaming que sugerem músicas ou vídeos que devem ser vistos por um usuário. Tanto Netflix como **Amazon** utilizam essa abordagem para criar seus algoritmos de recomendação.

APRENDIZAGEM POR REFORÇO

Nesse método de aprendizagem, que segue o estilo utilizado no início dos estudos sobre Inteligência Artificial, o computador é estimulado a aprender com base em tentativas e erros, otimizando o processo na prática direta. Com essa abordagem, é possível, por exemplo, ensinar um sistema a priorizar hábitos em detrimento de outros, com recompensas proporcionais ao acerto.

A aprendizagem por reforço foi inspirada por psicólogos comportamentalistas, que acreditavam na eficácia de recompensas e punições na educação dos seres humanos. Também lembra o procedimento de adestração de animais domésticos.

É um método, portanto, baseado na construção de experiência. A partir

disso, o algoritmo sabe quais caminhos são melhores que outros e quais processos são mais ágeis.

Exemplos de aplicação são os veículos autônomos e máquinas que jogam xadrez. O sistema aprende com múltiplas tentativas, que envolvem erros como uma jogada ruim ou um choque contra um obstáculo.

Quais os principais algoritmos de Machine Learning?

Existem uma infinidade de algoritmos utilizados em machine learning, cada um com uma finalidade específica. Tais algoritmos podem realizar previsões a partir de amostras ou tomar decisões guiadas unicamente por dados, sem qualquer tipo de programação.

Embora semelhante da estatística computacional, que faz previsões com o uso dos computadores, o aprendizado de máquina é usado em tarefas computacionais onde a programação de algoritmos explícitos é impraticável. A escolha de um algoritmo, entretanto, não é fácil, pois há características que podem inviabilizar a escolha do modelo mais preciso para determinado problema, como o uso de alto poder computacional.

ALGORITMOS PARA APRENDIZAGEM SUPERVISIONADA

O que é: Um algoritmo usa dados de treinamento e feedback de humanos para aprender a relação de dados insumos para um determinado produto (por exemplo, como os insumos “época do ano” e “taxas de juros” prevêm os preços das moradias).

Quando usar: Quando você sabe como classificar os dados de entrada e o tipo de comportamento que deseja prever, mas precisa do algoritmo para calculá-lo em novos dados.

Regressão linear

Método padrão altamente interpretável para modelar a relação passada entre variáveis de entrada independentes e variáveis de saída dependentes (que podem ter um número infinito de valores) para ajudar a prever valores futuros das variáveis de saída.

Exemplo de casos de uso:

- Compreender os impulsionadores das vendas de produtos, como preços da concorrência, distribuição, propaganda, etc;
- Otimizar os pontos de preço e estimar a elasticidade de preços dos produtos.

Regressão logística

Extensão da regressão linear que é usada para tarefas de classificação, o que significa que a variável de saída é binária (por exemplo, apenas preto ou branco) em vez de contínua (por exemplo, uma lista infinita de cores potenciais).

Exemplo de casos de uso:

- Classificar os clientes com base na probabilidade de eles pagarem um empréstimo;
- Prever se uma lesão de pele é benigna ou maligna com base em suas características (tamanho, forma, cor, etc).

Análise discriminante linear/quadrática

Atualiza uma regressão logística para lidar com problemas não lineares – aqueles em que mudanças no valor das variáveis de entrada não resultam em mudanças proporcionais nas variáveis de saída.

Exemplo de casos de uso:

- Prever a rotatividade do cliente;
- Prever a probabilidade de fechamento de um lead de vendas.

Árvore de decisão

Classificação altamente interpretável ou modelo de regressão que divide os valores dos recursos de dados em ramos nos nós de decisão (por exemplo, se um recurso é uma cor, cada cor possível se torna um novo ramo) até que uma saída de decisão final seja feita.

Exemplo de casos de uso:

- Fornecer uma estrutura de decisão para a contratação de novos funcionários;
- Compreender os atributos que tornam um produto mais provável de ser comprado.

Naive Bayes

Técnica de classificação que aplica o teorema de Bayes e permite que a probabilidade de um evento seja calculada com base no conhecimento dos fatores que podem afetar esse evento (por exemplo, se um e-mail contém a palavra “dinheiro”, então a probabilidade de ser spam é alta).

Exemplo de casos de uso:

- Analisar sentimento para avaliar a percepção do produto no mercado;
- Criar classificadores para filtrar e-mails de spam.

Support Vector Machine

Uma técnica que normalmente é usada para classificação, mas que pode ser transformada para também realizar regressões. O algoritmo traça uma divisão ideal entre as classes (a mais ampla possível). Também pode ser rapidamente generalizado para resolver problemas não lineares.

Exemplo de casos de uso:

- Prever quantos pacientes um hospital precisará atender em um determinado período;
- Prever a probabilidade de alguém clicar em um anúncio online.

Random Forest

Modelo de classificação ou regressão que melhora a precisão de uma árvore de decisão simples, gerando várias árvores de decisão e selecionando a maioria dos votos para prever o resultado, que é uma variável contínua (por exemplo, idade) para um problema de regressão

e é uma variável discreta (por exemplo, preto, branco ou vermelho) para classificação.

Exemplo de casos de uso:

- Prever o volume de chamadas em call centers para gerar decisões relacionadas à equipe de atendimento;
- Prever o uso de energia em uma rede de distribuição elétrica.

AdaBoost

Técnica de classificação ou regressão que usa uma infinidade de modelos para chegar a uma decisão, mas os pondera com base em sua precisão na previsão do resultado.

Exemplo de casos de uso:

- Detectar atividades fraudulentas em transações com cartão de crédito. Alcança uma precisão menor do que o aprendizado profundo (deep learning);
- Forma simples e de baixo custo para classificar imagens (por exemplo, reconhecer o uso da terra a partir de imagens de satélite para modelos de mudanças climáticas). Também atinge uma precisão menor do que o aprendizado profundo.

Árvores com aumento de gradiente

Técnica de classificação ou regressão que gera árvores de decisão sequencialmente, onde cada árvore se concentra em corrigir os erros provenientes do modelo de árvore anterior. O resultado final é uma combinação dos resultados de todas as árvores.

Exemplo de casos de uso:

- Previsão de demanda de produtos e níveis de estoque;
- Prever o preço dos carros com base em suas características (por exemplo, idade e quilometragem).

Rede neural simples

Modelo no qual neurônios artificiais (calculadoras baseadas em software) formam três camadas (uma camada de entrada, uma camada oculta onde os cálculos ocorrem e uma camada de saída) que podem ser usados para classificar dados ou encontrar a relação entre variáveis em problemas de regressão.

Exemplo de casos de uso:

- Prever a probabilidade de um paciente ingressar em um programa de saúde;

- Prever se os usuários registrados estarão dispostos ou não a pagar um determinado preço por um produto.

ALGORITMOS PARA APRENDIZAGEM NÃO-SUPERVISIONADA

O que é: Um algoritmo explora dados de entrada sem receber uma variável de saída explícita (por exemplo, explora dados demográficos do cliente para identificar padrões).

Quando usar: Você não sabe como classificar os dados e deseja que o algoritmo encontre padrões e classifique os dados para você.

K-means Clustering (agrupamento)

Coloca os dados em vários grupos (k), cada um contendo dados com características semelhantes (conforme determinado pelo modelo, não antecipadamente por humanos).

Exemplo de casos de uso:

- Segmenta os clientes em grupos por características distintas (por exemplo, faixa etária) para melhor designar campanhas de marketing ou evitar churn.

Modelo de mistura gaussiana

Uma generalização do agrupamento k-means que fornece maior flexibilidade no tamanho e na forma dos grupos (clusters).

Exemplo de casos de uso:

- Segmentar os clientes para atribuir melhor as campanhas de marketing usando características menos distintas do cliente (por exemplo, preferências de produto);
- Segmentar funcionários com base na probabilidade de desgaste.

Agrupamento hierárquico

Divide ou agrega clusters ao longo de uma árvore hierárquica para formar um sistema de classificação.

Exemplo de casos de uso:

- Agrupar os clientes de cartões de fidelidade em grupos progressivamente mais microsegmentados;

- Informar o uso/desenvolvimento do produto agrupando clientes que mencionam palavras-chave em dados de mídia social.

Sistema de recomendação

Usa a previsão do comportamento do cluster para identificar os dados importantes necessários para fazer uma recomendação.

Exemplo de casos de uso:

- Recomendar quais filmes os consumidores devem assistir com base nas preferências de outros clientes com atributos semelhantes;
- Recomendar artigos de notícias que um leitor possa querer ler com base no artigo que está lendo.

Como escolher o algoritmo ideal?

Como são várias as formas de se chegar a um resultado com ML, por existir diversas abordagens, é preciso saber como escolher o algoritmo mais adequado ao problema proposto. Veremos algumas dicas a seguir.

FINALIDADE DO PROBLEMA

É importante conhecer bem a finalidade e o contexto do caso que deve

ser processado pela máquina, com detalhes específicos, pois eles ajudam a selecionar o melhor jeito de tratar os dados.

Caso exista um objetivo bem definido, com saídas conhecidas, o ideal pode ser um método de aprendizagem supervisionada. Dentro dessa abordagem, caso o resultado seja um valor numérico, é possível usar regressão. Caso sejam categorias, classificação.

QUANTIDADE DE DADOS

Outro aspecto relevante é a quantidade de dados que será utilizada para alimentar o software. Redes neurais, por exemplo, geralmente precisam de grandes conjuntos de entrada.

É importante conhecer bem e se ater ao tamanho das bases para estabelecer qual método seguir, pois cada um lida com a quantidade de uma forma diferente. Além disso, caso a base de dados seja muito grande e o algoritmo demasiado complexo, pode ser preciso fazer uma avaliação a respeito da capacidade computacional necessária para a execução do algoritmo.

COMPLEXIDADE DO PROBLEMA E ACURÁCIA

A complexidade do problema também é uma questão importante, pois isso determina o grau de acurácia desejado. Com isso, é possível

selecionar qual algoritmo se encaixa melhor nesse grau de precisão analisando fatores como a periculosidade de um erro e as consequências que eles podem trazer.

TEMPO

E, por fim, o tempo deve ser levado em conta. Existem problemas que vão requerer decisões em tempo real. Outros, no entanto, com um intervalo aceitável. Isso deve ser pensado na hora de escolher o tipo de ML que resolva o caso da melhor forma.



Conclusão

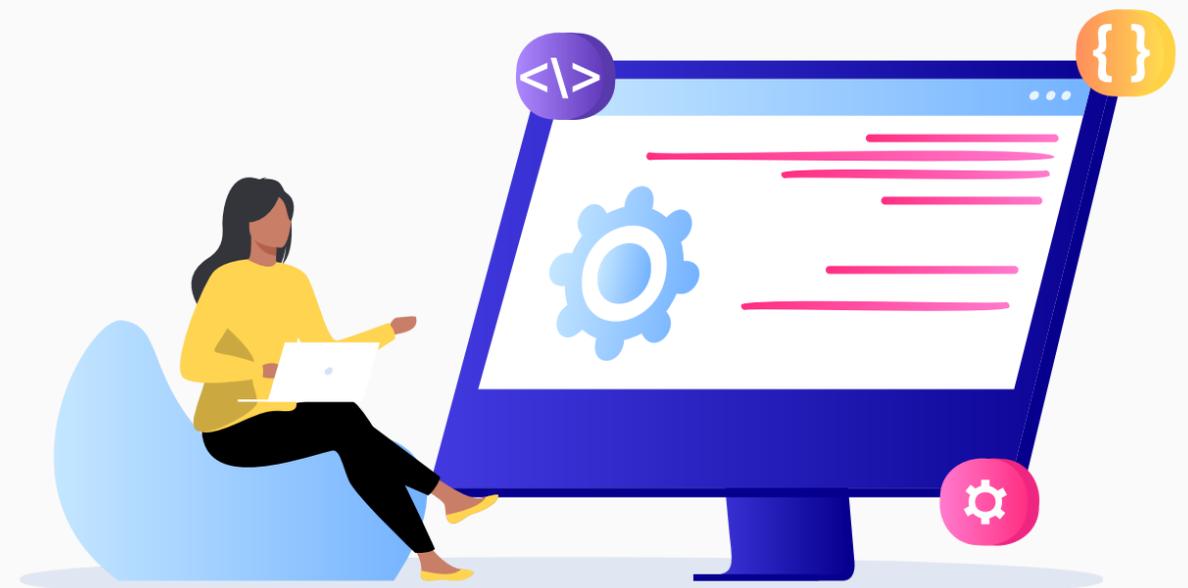
Neste e-book, vimos que as soluções de machine learning são classificadas em três principais grandes categorias e em algumas subcategorias menores. Cada uma tem sua própria característica e natureza, além de terem sido desenvolvidas para resolver problemas específicos. Ou seja, algoritmos de machine learning não são a panaceia para resolver qualquer tipo de problema que envolva dados.

Portanto, os algoritmos de machine learning se adaptam a problemas específicos e contextos especiais. Para selecionar a solução que melhor resolve o seu caso é preciso analisar alguns fatores, como: complexidade, tempo e quantidade de dados.

Diferente do que se divulga na mídia, machine learning está muito mais ligado a operações matemáticas para medir distâncias entre pontos e cálculo de probabilidade do que recriar o cérebro humano artificialmente.

Apesar de muitas atividades estarem sendo automatizadas por sistemas baseados algoritmos, ainda há muito espaço para o ser humano utilizar e explorar a sua criatividade. A maior parte das atividades automatizadas hoje em dia são extremamente repetitivas e é um verdadeiro desperdício da capacidade humana brigar por esse espaço com as máquinas.

Esperamos que este material possa servir como um ponto de partida para ajudar os profissionais a desenvolverem seus projetos de machine learning. Continue aprendendo, pois a jornada é longa!



GOSTOU DO NOSSO CONTEÚDO E
**QUER SABER MAIS SOBRE AS NOSSAS
TECNOLOGIAS E SERVIÇOS?**

Somos especializados em desenvolver projetos para a coleta e visualização de dados de clientes de médio e grande porte.

[ACESSE O NOSSO SITE](#)

Sobre a OpServices

Empresa especializada em **monitoração e visualização de dados**, com quase 20 anos de mercado. É reconhecida por desenvolver soluções e serviços de monitoramento da infraestrutura e da rede de qualquer ambiente de TI.

 marisa

 LEAR

 Grendene

 SantaCruz

 ENGIE

 Zaffari BOURBON

 Unimed
Curitiba

 Klabin

 Unimed
Porto Alegre

 digi+

 FINDES
PELO FUTURO DA INDÚSTRIA

 UNIGEL

Conheça nossas redes sociais:   