

Aprendizado de Máquina (Machine Learning, ML) é um ramo da inteligência artificial que permite que computadores "aprendam" padrões e façam previsões ou decisões sem serem explicitamente programados para isso. Em vez de seguir regras fixas definidas por um programador, os algoritmos de ML analisam dados, identificam padrões e ajustam seus modelos para melhorar a precisão das previsões ao longo do tempo

Inteligência Artificial (IA): Área da computação que desenvolve sistemas capazes de realizar tarefas que normalmente exigiriam inteligência humana, como reconhecimento de padrões, tomada de decisão e aprendizado a partir de dados.

Aprendizado de Máquina (ML): Subárea da IA que cria algoritmos capazes de aprender padrões a partir de dados e fazer previsões ou decisões sem programação explícita.

Aprendizado Supervisionado: O modelo aprende com dados rotulados, ou seja, recebe entradas e saídas corretas para identificar padrões. Usado em classificação e regressão

Aprendizado Não Supervisionado: O modelo trabalha com dados sem rótulos e encontra padrões ocultos sozinho.

Aprendizado por Reforço: Um agente aprende interagindo com um ambiente, recebendo recompensas ou penalidades para otimizar suas ações. Aplicado em jogos, robótica e controle

Deep Learning (DL): Subcampo do aprendizado de máquina que usa redes neurais profundas para modelar padrões complexos em grandes quantidades de dados. É amplamente utilizado em visão computacional, processamento de linguagem natural e outras áreas avançadas.

Redes Neurais Recorrentes (RNNs): Tipo de rede neural projetada para lidar com dados sequenciais, como texto e séries temporais, pois mantém uma memória do que já foi processado.

Redes Neurais Convolucionais (CNNs): Redes especializadas no processamento de dados espaciais, como imagens. Utilizam filtros para extrair características relevantes, permitindo identificar padrões como bordas e texturas. Aplicadas em reconhecimento facial, diagnóstico médico e visão computacional.

Visão Computacional: Área da IA que permite que máquinas "vejam" e interpretem imagens ou vídeos, usada em reconhecimento facial, carros autônomos e diagnóstico médico.

YOLO (You Only Look Once): Algoritmo de detecção de objetos em tempo real que identifica e localiza múltiplos objetos em uma imagem de forma rápida e eficiente.

CNNs (Redes Neurais Convolucionais): Redes especializadas no processamento de imagens, extraindo características como bordas e texturas para tarefas como reconhecimento de padrões.

Robótica Inteligente: Robôs que usam IA para tomar decisões e se adaptar ao ambiente, aplicados em indústrias, saúde e veículos autônomos.

Processamento de Linguagem Natural (PLN): Área que ensina máquinas a entender e gerar linguagem humana, usada em chatbots, assistentes virtuais e tradução automática.

A maior diferença entre machine learning supervisionado e não supervisionado é o tipo de dados usado. O aprendizado supervisionado usa dados de treinamento rotulados, ao contrário do aprendizado não supervisionado.

Com o aprendizado supervisionado, o algoritmo usa um conjunto de dados de amostra para treinar e fazer previsões, ajustando-se iterativamente para minimizar erros. Esses conjuntos de dados são rotulados para fornecer contexto, fornecendo os valores de saída desejados para permitir que um modelo dê uma resposta "correta".

Por outro lado, os algoritmos de aprendizado não supervisionados trabalham de forma independente para aprender a estrutura inerente dos dados, sem orientações ou instruções específicas. Basta fornecer dados de entrada não rotulados e deixar o algoritmo identificar padrões que ocorrem naturalmente no conjunto de dados.

Supervisionados:

Regressão Linear: Modelo que ajusta uma linha reta aos dados para prever valores contínuos com base em relações lineares.

Regressão Logística: Usa uma função sigmoide para classificar dados em categorias binárias ou multiclases.

Redes Neurais Artificiais (RNA): Estruturas compostas por camadas de neurônios que ajustam pesos para reconhecer padrões complexos.

Naïve Bayes: Classificador baseado no Teorema de Bayes, assume independência entre as variáveis e é eficiente para texto e probabilidades.

SVM (Support Vector Machine): Separa dados em classes usando hiperplanos, ótimo para problemas de classificação.

Random Forest: Conjunto de várias árvores de decisão que votam para melhorar precisão e reduzir overfitting.

KNN (K-Nearest Neighbors): Classifica um dado com base nos "K" vizinhos mais próximos em um espaço de características.

Árvore de Decisão: Estrutura hierárquica que divide dados com base em perguntas de "sim" ou "não", facilitando interpretações.

Não Supervisionados:

t-SNE: Técnica de redução de dimensionalidade que preserva a estrutura dos dados ao projetá-los em um espaço menor para visualização.

UMAP: Similar ao t-SNE, mas mais eficiente e escalável para reduzir dimensões mantendo a estrutura global dos dados.

PCA (Análise de Componentes Principais): Transforma dados em novas variáveis (componentes principais) para reduzir dimensionalidade mantendo máxima variação.

Hierarchical Clustering: Agrupa dados formando uma árvore de similaridade, podendo ser divisivo ou aglomerativo.

K-Means: Algoritmo que divide dados em "K" grupos baseando-se na proximidade a centróides ajustados iterativamente.

DBSCAN: Algoritmo de clustering baseado em densidade, identificando grupos e outliers sem precisar definir o número de clusters.

SOM (Self-Organizing Map): Rede neural não supervisionada que organiza dados em um mapa bidimensional preservando relações topológicas.

Variável Independente: É a entrada do modelo, ou seja, a característica usada para prever um resultado. No ML, são os atributos (features) dos dados, como idade, altura ou temperatura.

Variável Dependente: É a saída que queremos prever, ou seja, o valor que depende das variáveis independentes. Em um modelo de regressão, pode ser o preço de um imóvel baseado em tamanho e localização.

<https://medium.com/data-hackers/principais-m%C3%A9tricas-de-classifica%C3%A7%C3%A3o-de-modelos-em-machine-learning-94eeb4b40ea9>

Em Machine Learning, temos os métodos supervisionados e não supervisionados. No aprendizado supervisionado, há duas categorias principais: regressão e classificação. Dentro da classificação, temos a binária e a multiclasse. Já no aprendizado não supervisionado, um dos principais métodos é o clustering.

O aprendizado supervisionado utiliza dados rotulados, ou seja, cada exemplo no conjunto de dados possui um rótulo associado. Um exemplo clássico é o dataset **Iris**, que contém medições de flores associadas às suas respectivas espécies. Outro exemplo são classificações médicas, como prever se um paciente tem ou não uma doença com base em exames. Modelos de classificação também são usados para identificar e-mails como **spam** ou **não spam**.

A **regressão** é um tipo de análise onde temos uma ou mais variáveis independentes usadas para prever uma variável dependente. Por exemplo, ao analisar dados como **IMC e glicose**, podemos prever a probabilidade de um paciente desenvolver **diabetes**.

Na **classificação binária**, o objetivo é prever entre duas classes, como determinar se uma pessoa está **doente ou saudável** com base em dados médicos. Já na **classificação multiclasse**, há mais de duas categorias possíveis, como identificar se uma imagem contém um **gato, um cachorro ou um pássaro**.

No aprendizado **não supervisionado**, os dados não possuem rótulos e são usados para identificar padrões. Técnicas como **clustering** agrupam dados por similaridade sem um rótulo pré-definido. Isso é útil para **segmentação de clientes, detecção de anomalias e recomendações**. Outra técnica comum é a **redução de dimensionalidade**, usada para simplificar dados complexos sem perder informações importantes.

Maca e laranja, tamanho da casa e o preto e o preto

Acurácia

Beleza! Mas por que isso é bom para a acurácia?

Imagine que de 1000 registros que a base de pacientes possua, 900 deles são não diabéticos e 100 são portadores da doença. O modelo aprenderá com muito mais dados de uma classe do que de outra, assim, tenderá a classificar os pacientes como não diabéticos e terá mais dificuldade para acertar quando receber informações de uma pessoa que de fato tenha a doença. Assim, se de 100 registros para os dados de teste, 90 forem da classe sem diabetes e 10 com diabetes, o modelo terá uma certa facilidade para acertar grande parte das análises (90%), gerando uma acurácia alta, que certamente será **enganosa**. Para bases desbalanceadas, há outras métricas que podem ser exploradas para machine learning, por exemplo a **Precisão e Revocação**.

Precisão

Perceba que a métrica leva em consideração os FP, mas os FN ficam de fora. Isso pode ser ruim dependendo do tipo de problema que está sendo analisado. Neste caso, como é uma situação que envolve uma doença, os FN ficando de fora significa que há pacientes que realmente são diabéticos e que não estão fazendo parte da avaliação da métrica. Imagine que você possua uma doença e, depois de todos os exames, você é liberado para casa como se não tivesse nada. Situação **BEM PERIGOSA**

Revocação

Por levar em conta os FP, seria uma boa métrica para se considerar em problemas que envolvem a área de saúde, pois a métrica prefere classificar como doente quem não possui a doença, do que classificar como saudável e deixar o paciente ir embora, sendo que possui a doença. É bem “menos pior” receber um diagnóstico de diabetes e depois perceber que era um engano, do que achar que está tudo bem quando na verdade se está com a doença, não é mesmo??

