Sistemas de aprendizado de máquina

Introdução aos sistemas de aprendizado de máquina, seus métodos de aprendizado de acordo com o sinal, e as categorias de problemas de aprendizado de máquina em função da saída.

Profa. Daisy Albuquerque

Propósito

Compreender o conceito de aprendizado de máquina, identificando os seus métodos – tais como supervisionado, não supervisionado, semissupervisionado e aprendizado por reforço – e discutindo as categorias de problemas de aprendizado de máquina em função da saída.

Objetivos

- Identificar conceitos de aprendizado de máquina
- Reconhecer os métodos de aprendizado de máquina de acordo com o sinal
- Distinguir as categorias de problemas de aprendizado de máquina em função da saída

Introdução

Aprendizado de máquina, ou *machine learning*, é uma área da inteligência artificial relacionada à busca de um conjunto de regras e procedimentos para permitir que as máquinas possam agir e tomar decisões baseadas em dados, ao invés de serem programadas para realizar determinada tarefa (SAMUEL, 1959).

Quando as máquinas analisam um grande volume de informações, elas são capazes de identificar padrões e de tomar decisões com o auxílio de modelos. Sendo assim, as máquinas se tornam aptas a fazer predições por meio do processamento de dados.

Veremos os principais conceitos e as definições do aprendizado de máquina, quem é adepto, quais os seus desafios, bem como suas diferenças em relação à inteligência artificial. Ao longo do texto, usaremos os termos aprendizado de máquina e *machine learning* indistintamente, assim como as respectivas siglas AM e ML.

Para entender melhor o que é aprendizado de máquina, vamos identificar seus métodos e suas categorias de problemas.

Neste vídeo, exploramos o intrigante mundo do aprendizado de máquina. Descubra como máquinas aprendem, reconhecem padrões e fazem predições usando dados. Identifique conceitos, métodos e categorias de problemas neste fascinante campo da inteligência artificial.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Inteligência artificial e suas linhas de pensamento

Machine learning e inteligência artificial são sinônimos?

Inteligência Artificial versus Machine Learning

Assista ao vídeo a seguir e acompanhe um resumo das principais diferenças entre IA e ML.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

A inteligência artificial (IA) é uma área recente da ciência da computação criada com o propósito de desenvolver sistemas que simulem a capacidade humana na percepção de um problema, identificando seus componentes, resolvendo problemas, propondo soluções e, até mesmo, tomando decisões.

Saiba mais

As ideias relacionadas com inteligência artificial são bem anteriores ao surgimento da tecnologia que tornou isso possível. Na antiguidade, a ideia de uma inteligência não humana que pudesse executar atividades por si própria já era idealizada. Sendo assim, podemos afirmar que a origem da inteligência artificial pode ter vindo dos filósofos gregos, como Aristóteles. Aristóteles – professor de Alexandre, o Grande, rei da Macedônia – pensava em como substituir os escravos por objetos, como uma vassoura de limpeza. O filósofo imaginava a substituição da mão de obra humana e escrava por um elemento que pudesse ter vontade própria e ser capaz de estabelecer um sistema básico de arrumação. Dessa forma, não precisaria usar seres humanos como mão de obra escrava. Podemos concluir que ele inventou a robótica em 300 a.C.

O ser humano sempre idealizou uma máquina que agisse e pensasse como humano, e estudos de várias áreas começaram a ir por esse caminho especificamente durante a Segunda Guerra Mundial.

1943

Primeiro estudo sobre redes neurais
Warren McCulloch e Walter Pitts apresentaram o artigo *A Logical Calculus of Ideas Immanent in Nervous Activity,* que tratava pela primeira vez de redes neurais, isto é, estruturas de raciocínio artificiais em forma de modelo matemático que imitam o nosso sistema nervoso.

1950

Programação de máquina para jogar xadrez
Outro artigo importante da época é o de Claude Shannon sobre como programar uma máquina para jogar xadrez com cálculos de posição simples, porém eficientes.

1956

Uso do termo "inteligência artificial"
Na chamada Conferência de Dartmouth, foi realizado o encontro que reuniu Nathan Rochester, da IBM; Claude Shannon, autor do artigo sobre o jogo de xadrez; Marvin, do SNARC e John McCarthy, professor da Universidade Stanford, que, nessa conferência, começou a usar o termo inteligência artificial.

Na época, já existiam várias teorias de complexidade, simulação de linguagem, redes neurais e máquinas de aprendizagem.

> McCarthy resolveu dar o nome de inteligência artificial para esses sistemas de imaginação humana que usam a ciência da computação.

E assim nasceu a ciência e engenharia de produzir máquinas inteligentes; porém, apenas recentemente começou a fazer parte do nosso cotidiano.

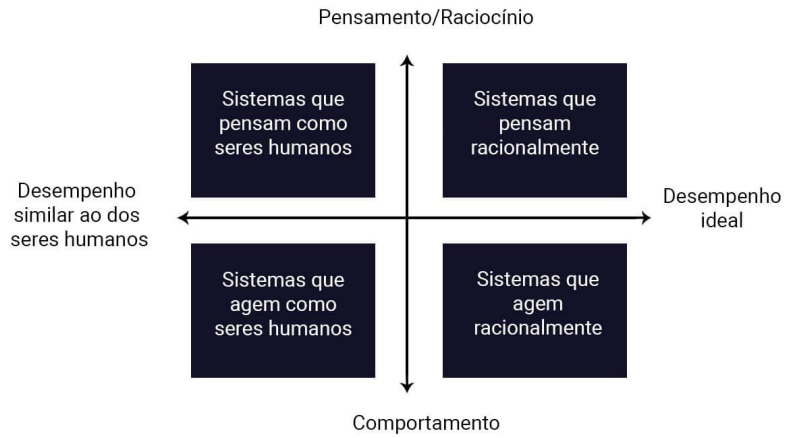
Linhas de Pensamento da Inteligência Artificial

Neste vídeo, mergulhamos nas origens e linhas de pensamento da inteligência artificial explorando as quatro fascinantes linhas de pensamento que moldam a IA moderna!

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

A inteligência artificial pode ser apresentada em quatro linhas de pensamento (RUSSEL; NORVIG, 2009).



Linhas de pensamento da inteligência artificial

Nos dois quadrantes superiores, encontramos os sistemas que pensam como seres humanos e os que pensam racionalmente, relacionados com os processos de pensamento e raciocínio. Já na parte inferior, estão os sistemas que agem como seres humanos e aqueles que agem racionalmente, referentes ao comportamento humano.

Os sistemas do lado esquerdo da figura medem o sucesso em termos de desempenho, como acontece com os seres humanos; os do lado direito visam o desempenho ideal de inteligência, chamado de racionalidade.

Vamos então conhecer um pouco de cada uma dessas linhas de pensamento:

Sistemas que pensam como seres humanos

Um exemplo são as redes neurais artificiais. Esses sistemas automatizam atividades como tomada de decisões, resolução de problemas e aprendizagem. Para isso, é necessário saber como os humanos pensam. Nesse caso, a modelagem cognitiva é trabalhada com modelos computacionais de inteligência artificial e técnicas experimentais da Psicologia para construir teorias precisas e verificáveis a respeito dos processos de funcionamento da mente humana. Trabalham com o aprendizado por observação, realizando investigações experimentais de seres humanos ou animais.

Sistemas que pensam racionalmente

São os sistemas inteligentes. Eles tentam simular o pensamento lógico racional dos humanos, isto é, são capazes de fazer as máquinas entender e raciocinar.

Sistemas que agem como seres humanos

Um exemplo são os robôs, máquinas que executam tarefas de um jeito semelhante ao das pessoas. O Teste de Turing, criado em 1950 por Alan Turing, fez com que 30% dos humanos consultados acreditassem que a máquina que participava do teste, ao se fazer passar por humano, era, de fato, um ser humano.

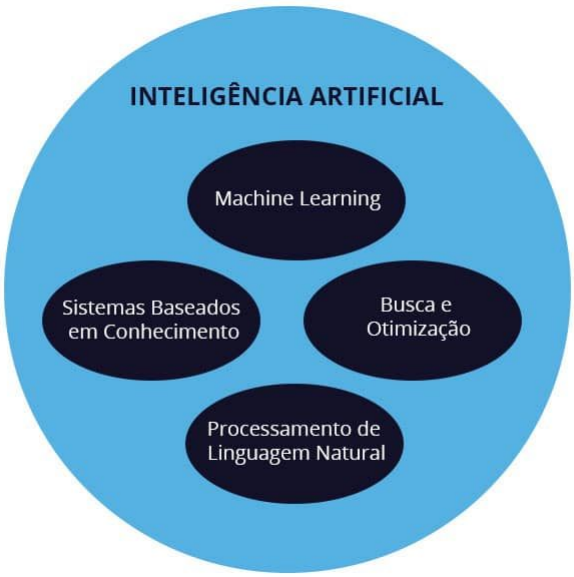
Sistemas que agem racionalmente

São os agentes inteligentes. Idealmente, aqueles que tentam imitar de forma racional o comportamento humano. Agir racionalmente significa agir de modo a alcançar seus objetivos, dadas as suas crenças. Nesse caso, a IA é vista como o estudo e a construção de agentes racionais.

A inteligência artificial abrange uma enorme variedade de áreas, desde as de uso geral, como aprendizado e percepção, até aquelas que envolvem tarefas específicas, como jogos e diagnóstico de doenças.

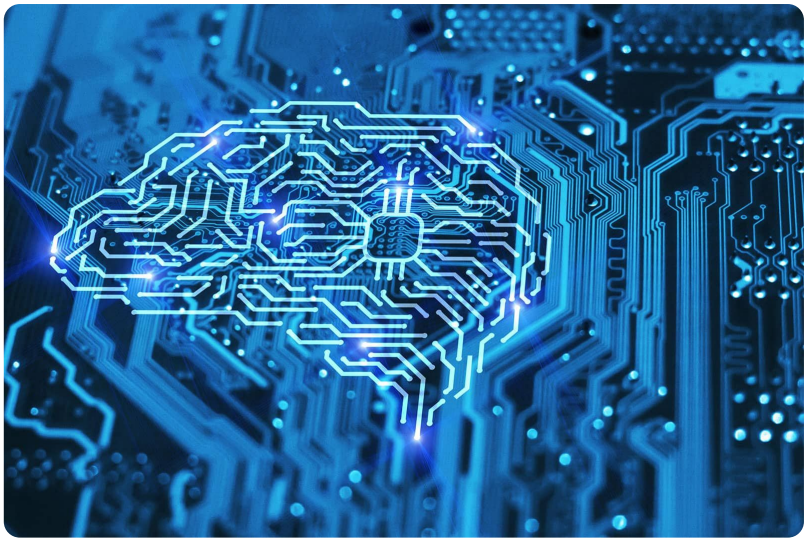
Machine learning

Dentro da IA, podemos destacar vários segmentos, como: o processamento de linguagem natural; os sistemas baseados em conhecimento; a busca e otimização com os algoritmos genéticos e o aprendizado de máquina. *Machine learning*, portanto, não é sinônimo de inteligência artificial, mas sim uma de suas áreas ou componentes.



Áreas da Inteligência Artificial

Conceitos e definições de machine learning



Decifrando Machine Learning: Da Teoria à Prática

Neste vídeo, desvendamos os conceitos fundamentais do Machine Learning (ML). Viaje desde a visão pioneira de Arthur Lee Samuel em 1959 até as aplicações práticas, como tradução automática e reconhecimento facial. Descubra como ML revoluciona tarefas complexas, libertando-se das limitações da programação tradicional.

> Conteúdo interativo
>
> Acesse a versão digital para assistir ao vídeo.

Machine Learning (ML) é uma evolução da teoria do aprendizado computacional em IA.

Arthur Lee Samuel, cientista da computação estadunidense e um dos pioneiros no campo dos jogos de computador com aprendizado de máquina e inteligência artificial, publicou em 1959 o artigo *Some Studies in Machine Learning Using the Game of Checkers*, conceituando o aprendizado de máquina como campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados (SAMUEL, 1959).

Em 1991, Sholom M. Weiss e Casimir A. Kulikowski, autores do livro *Computer Systems that Learn*, definiram um sistema de aprendizado como um programa de computador que toma decisões baseadas na experiência contida em exemplos solucionados com sucesso.

Antes do *machine learning*, os algoritmos de computação necessitavam de instruções na sua totalidade.

Exemplo

Se você quisesse criar um programa que joga o jogo da velha, seria necessário mapear todas as possíveis jogadas e codificá-las como instruções dentro do programa. Veja a seguir instruções para o programa jogo da velha: InstruçãoAçãoInstrução 1Se o oponente ocupar duas casas seguidas, ocupe a terceira.Instrução 2Se não, e havendo algum movimento que cria duas linhas com duas casas ocupadas, faça-o.Instrução 3Se não, e caso o espaço do centro esteja vazio, ocupe-o.Instrução 4Se não, e caso o oponente tenha preenchido uma quina, preencha a quina contrária.Instrução 5Se não, e caso haja uma quina vazia, preencha-a.Instrução 6Se nenhuma dessas condições acontecer, pode preencher qualquer espaço vazio. Segundo Pedro Domingos, no seu livro O algoritmo mestre, esse algoritmo não perde o jogo da velha, isso porque suas regras são simples.

Agora imagine um jogo de xadrez, em que cada jogada tem, em média, 35 movimentos possíveis. Mapear todas essas instruções em um algoritmo, cobrindo o passo a passo para todas as possibilidades de jogada é inviável.

> Em situações ainda mais complexas, como processamento da linguagem natural, ou reconhecimento facial, é mais difícil ainda mapear todas as possíveis regras.

Os programas de computação baseados no aprendizado de máquina são algoritmos capazes de fazer previsões ou tomadas de decisão a partir de grandes volumes de dados, o *big data*, sem qualquer tipo de programação prévia.

Apesar da semelhança com a estatística computacional, que também faz previsões com o uso de computadores, os algoritmos de aprendizado de máquina são empregados em tarefas computacionais nas quais a criação e a programação de algoritmos explícitos são impraticáveis.

Quem está usando machine learning

A inteligência artificial cria máquinas inteligentes com o poder de reconhecer objetos, vozes, faces, além de racionalmente solucionar problemas, com a capacidade de planejamento e de manipular e mover objetos.

Essa funcionalidade, possível por meio do aprendizado de máquina, utiliza uma programação prévia, uma apresentação de certas características para as máquinas e a inserção de um grande volume de informações relacionadas ao mundo.

As máquinas iniciam suas funções ao serem programadas e treinadas a partir do acesso a imagens, definições e características de objetos, categorias, propriedades e relação entre todos eles para implementar a engenharia do conhecimento.

> Mas como funciona na prática?

Para compreendermos o aprendizado de máquina, vamos ver como funciona um tradutor automático de documentos. Para traduzir um documento, teremos que entender o texto completamente antes de traduzi-lo. Usaremos um conjunto de regras elaboradas por um linguista computacional bem versado nas duas línguas

que gostaríamos de traduzir. Entretanto, isso seria um desafio muito complexo, uma vez que um texto nem sempre está gramaticalmente correto e o contexto utilizado pode gerar confusões de entendimento.

Poderíamos simplesmente lançar mão de exemplos de documentos semelhantes traduzidos, de modo que a máquina aprendesse a tradução entre as duas línguas. Em outras palavras, poderíamos usar exemplos de traduções para aprender como traduzir. Essa abordagem de aprendizado de máquina mostrou-se bem-sucedida e é muito utilizada, uma vez que a própria internet é uma ótima base de dados.

Outro exemplo de aplicação de *machine learning* é o reconhecimento facial. Muitas aplicações de segurança, como o controle de acesso, usam esse recurso como um de seus componentes. Mediante uma foto ou gravação de vídeo de uma pessoa, é possível reconhecê-la. Nesse caso, o sistema classifica o rosto a partir das categorias existentes, como os nomes das pessoas cadastradas; caso não o reconheça, decide que é um rosto desconhecido.

Cada vez mais o aprendizado de máquina tem sido usado para automatizar o processo de criação de um bom mecanismo de pesquisa, como o Google. Outro exemplo é a filtragem colaborativa – técnica utilizada pelos **sistemas de recomendação**.

O *machine learning* pode ser aplicado em diversas áreas, como:

Sistemas de recomendação

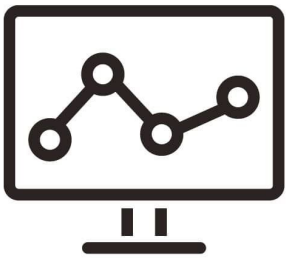
Um sistema de recomendação combina técnicas computacionais para selecionar itens personalizados com base nos interesses dos usuários e conforme o contexto no qual estão inseridos. Lojas online, como a Amazon, ou sites de streaming, como Netflix, utilizam esse sistema para atrair usuários a comprar produtos adicionais ou assistir a mais filmes.

Pesquisa na internet

Coleta e análise de dados

Rastreamento de mensagens de spam

Organização e classificação de informações



Busca de fraudes na internet

Desafios do aprendizado de máquina

Neste vídeo, exploramos os desafios cruciais enfrentados no Aprendizado de Máquina. Da qualidade dos dados à ética, descubra como a tecnologia lida com dilemas complexos. Desvende as barreiras da generalização e a necessidade de profissionais especializados, enquanto mergulhamos nas intricadas questões éticas que moldam o futuro do ML.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Como vimos, o AM é, em termos gerais, uma tecnologia desenvolvida com o objetivo de criar programas a partir dos dados. Antes dele, desenvolvíamos softwares programando manualmente cada etapa. No caso do aprendizado de máquina, é realizado um treinamento com dados para o algoritmo com objetivo de criar um outro programa, ou melhor, um modelo a ser seguido durante a operação.

> Atualmente, utilizamos essa nova tecnologia para a tomada de decisão mais assertiva, no caso de recursos escassos e para suprir a falta de especialistas, dentre outras situações.

Devido à evolução tecnológica – Hadoop, computação distribuída, execução em memória, computação em nuvem etc. –, os algoritmos de *machine learning*, com suas técnicas de predição, estão cada vez mais

acurados e precisos. Porém, é necessário tomar cuidado com a qualidade e as maneiras como a análise dos dados tem sido realizada. Colocam-se, então, alguns desafios.

Desafio 1: Qualidade dos dados

Sabemos que o aprendizado de máquina depende dos dados, pois a partir deles é criado o modelo de aprendizado. Sendo assim, um dos desafios está na qualidade desses dados – os insumos do aprendizado de máquina. Esse é um grande desafio, pois dificilmente uma empresa possui uma base de dados perfeita que possa servir de base para os algoritmos de *machine learning*.

O aprendizado de máquina utiliza grandes bases de dados, como o *big data*, com potencial para leitura de dados não estruturados. Porém, isso significa a capacidade de armazenar e processar dados diversos, como imagens, textos, voz e vídeos.

O grande desafio está na organização, ou melhor, na padronização desses dados armazenados. Isso leva tempo e é nesse ponto que muitos projetos de ML acabam nascendo já mortos.

Desafio 2: Generalização

Esse é outro grande desafio. Os algoritmos de *machine learning* devem ser generalizáveis para que tenham resultados em aplicações futuras. Dificilmente uma empresa detém todas as informações necessárias para realizar previsões tão acuradas. Nesse caso, o algoritmo de ML será capaz de generalizar a predição de interesse. Porém, é comum acontecer o overfitting, que ocorre quando o modelo de aprendizado criado é tão específico que apenas prevê, com bons resultados, ao usar a base em que foi treinado. Essa circunstância gera resultados ruins quando ele é colocado em produção.

Desafio 3: Profissionais especializados

No AM, é importante montar um time multidisciplinar, com profissionais que entendam de computação, outros com expertise em estatística, programadores, cientistas de dados, além daqueles que conhecem o negócio.

O grande desafio está na contratação de profissionais com conhecimento técnico, interessados e esforçados, com ótima capacidade de comunicação, que busquem a resolução do problema, mantendo a base bem-feita. É uma forma muito mais produtiva para interpretar e transferir a mensagem contida nos números.

Desafio 4: Ética

Os parâmetros dos algoritmos de ML nem sempre incluem ética. Um algoritmo pode criar um orçamento nacional com o objetivo de “maximizar o PIB/produtividade no trabalho/expectativa de vida”, mas sem limitações éticas programadas no modelo; pode eliminar orçamentos para escolas, hospitais psiquiátricos e o meio ambiente, porque não aumentam diretamente o PIB.

A inclusão da ética no processamento da máquina se torna um grande desafio. Nesse caso, as questões éticas precisam ser incorporadas e monitoradas ao longo do tempo. Por exemplo, as opiniões sobre questões como os direitos LGBT, matrimônio inter-racial ou entre pessoas de diferentes grupos econômicos podem mudar significativamente de uma geração para outra.
Se as questões morais variam entre grupos do mesmo país, imagine entre países. Por exemplo, na China, usar o reconhecimento facial para vigilância em massa se tornou a norma. Mas será que outros países veem esse assunto da mesma forma?

Desafio 5: Evitar o envenenamento dos dados

Os resultados do aprendizado de máquina dependem muito dos dados de referência, que formam a base do conhecimento. Contudo, os dados podem ficar distorcidos, por acidente ou por conduta duvidosa; este último caso é geralmente chamado de “envenenamento”.

Se os dados utilizados como amostra de treinamento para um algoritmo de contratação forem obtidos de uma empresa com práticas racistas, eles também serão.

Comentário

A Microsoft ensinou um chatbot a se comunicar no Twitter, permitindo que qualquer pessoa conversasse com ele. Tiveram que cancelar o projeto em menos de 24 horas porque internautas “gentis” ensinaram o bot a falar palavrões e a recitar trechos de Mein Kampf (em português, Minha Luta), livro escrito por Adolf Hitler.

Verificando o aprendizado

Questão 1

A inteligência artificial é uma área da ciência da computação criada com o propósito de desenvolver sistemas que simulem a capacidade humana na percepção de um problema, identificando seus componentes e, com isso, resolvendo situações, propondo soluções e, até mesmo, tomando decisões. Nesse contexto, a IA pode ser apresentada em linhas de pensamento. Dentre as opções, assinale a que contém essas linhas de pensamento:

A

Sistemas que imitam os seres humanos; sistemas que substituem os seres humanos.

B

Sistemas inteligentes; sistemas semi-inteligentes; sistemas não inteligentes.

C

Sistemas que pensam e agem como seres humanos.

D

Sistemas que pensam como seres humanos; sistemas que pensam racionalmente; sistemas que agem como seres humanos; sistemas que agem racionalmente.

E

Sistemas que seguem as instruções programadas; sistemas que criam as instruções; sistemas que reconhecem imagens; sistemas que otimizam a programação.

A alternativa D está correta.

Segundo Russel e Norvig, a inteligência artificial pode ser dividida em quatro linhas de pensamento: sistemas que pensam como seres humanos; sistemas que pensam racionalmente; sistemas que agem como seres humanos; e sistemas que agem racionalmente.

Questão 2

A partir das definições de inteligência artificial e *machine learning*, podemos concluir que:

A

Machine learning é sinônimo de inteligência artificial.

B

Inteligência artificial não tem nada a ver com *machine learning*.

C

Inteligência artificial é uma ciência que se utiliza de quatro linhas de pensamento, uma das quais é o *machine learning*.

D

Machine learning é o termo inglês para inteligência artificial.

E

Inteligência artificial é uma ciência composta por várias áreas, e uma delas é o *machine learning*.

A alternativa E está correta.

A inteligência artificial contempla várias áreas e o *machine learning* é apenas uma delas, com o objetivo principal de ensinar a máquina a realizar atividades à semelhança dos seres humanos.

Aprendizado de máquina

> [...] A definição formal de aprendizado de máquina registra que um programa de computador “aprende” a respeito de uma tarefa T, por meio de uma experiência E, em relação a uma métrica de desempenho P, se o seu desempenho na tarefa T, quando medido pela métrica P, aumenta com a experiência E.
>
> —
>
> MITCHELL,1997

Comparação dos métodos de Aprendizado de Máquina

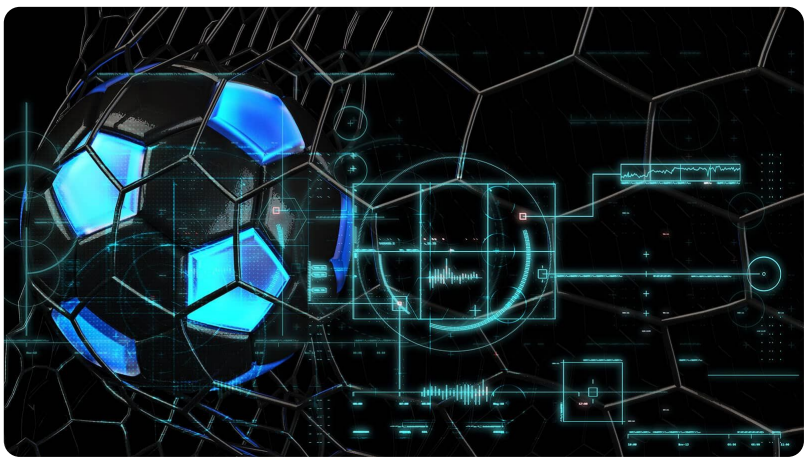
Assista ao vídeo a seguir e acompanhe um comparativo dos principais métodos de ML.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Ou seja, para realizar uma **tarefa T**, um computador aprenderá a partir de uma **experiência E**, procurando melhorar uma **métrica de desempenho P** (performance).

Os algoritmos de AM atingem um determinado objetivo, aprendendo a partir de um grande volume de dados, que são as suas experiências.



1

Tarefa T

Prever o resultado de um jogo de futebol.

2 Experiência E
Informações sobre partidas de futebol, como formação tática, composição do time, técnico, resultados etc.

3 Métrica de desempenho P
Percentual de acertos do resultado do jogo.

Outro exemplo do uso do aprendizado de máquina é o diagnóstico de exames. Suponhamos que agora a **tarefa T** seja determinar, com base em imagens de exames, se um tumor é maligno ou benigno. Aqui, a base de dados para compor a **experiência E** serão as imagens de tumores obtidas de exames médicos, como raios X, ultrassom, tomografia, ressonância magnética, entre outros. E a **métrica de desempenho P** será o percentual de diagnósticos corretos.

Resumindo:

1 Tarefa T
Diagnosticar se o tumor é maligno ou benigno.

2 Experiência E
Imagens de tumores obtidas por exames médicos (raio X, ultrassom, tomografia, ressonância magnética etc.).

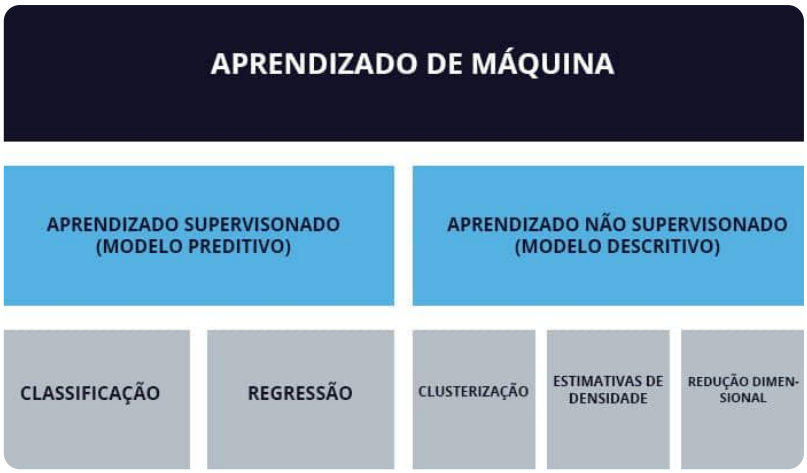
3 Métrica de desempenho P
Percentual de diagnósticos corretos.

> Conclui-se que, quanto mais dados são apresentados, quanto mais experiências são fornecidas, melhores serão os resultados, ou seja, melhor será a performance do algoritmo.

Os algoritmos de aprendizado de máquina podem ser categorizados conforme o tipo de aprendizagem. Existem diferentes tipos de cenários em que o AM pode ser implementado e em cada um desses cenários pode-se aplicar um método para o conjunto de algoritmos.

O aprendizado de máquina se utiliza de programas de computador que aprendem com seus erros e fazem previsões sobre dados a partir de abordagens de aprendizagem, como: supervisionada, não supervisionada, semissupervisonada e por reforço. Assim, acabam por tomar decisões e gerar resultados confiáveis e repetíveis.

Os principais métodos de aprendizado de máquina em que estão inseridos os diferentes cenários de algoritmos estão descritos na figura abaixo.



Métodos de aprendizado de máquina

Aprendizado supervisionado

O aprendizado de máquina supervisionado é um método cuja forma de aprender é a partir de experiências predefinidas, de dados rotulados.

Desvendando o Aprendizado Supervisionado: Da Teoria à Prática

Neste vídeo, mergulhamos no universo do Aprendizado Supervisionado. Descubra como os algoritmos aprendem com dados rotulados, criando modelos para identificar e classificar novas entradas. Explore as fases de treinamento e teste, e entenda como esse método diferencia classificação e regressão.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Os algoritmos de AM supervisionado utilizam uma série de experiências ou exemplos (entradas) rotulados com suas respectivas classes (saídas), para aprender padrões dentre esses dados e, assim, formular um modelo de conhecimento para classificar novas entradas.

Podemos dividir o AM supervisionado em duas fases.

Primeira fase

É a formulação do modelo, chamada de treinamento.

Segunda fase

É o teste, quando o modelo é usado para identificar corretamente os dados de fora do conjunto de treinamento e não rotulados.

Para entendermos melhor, imaginemos uma base de dados composta por um conjunto de imagens de veículos identificadas por carros, motos, caminhões e ônibus. Esses dados foram mapeados cada um de acordo com o tipo de veículo que representa:

Moto

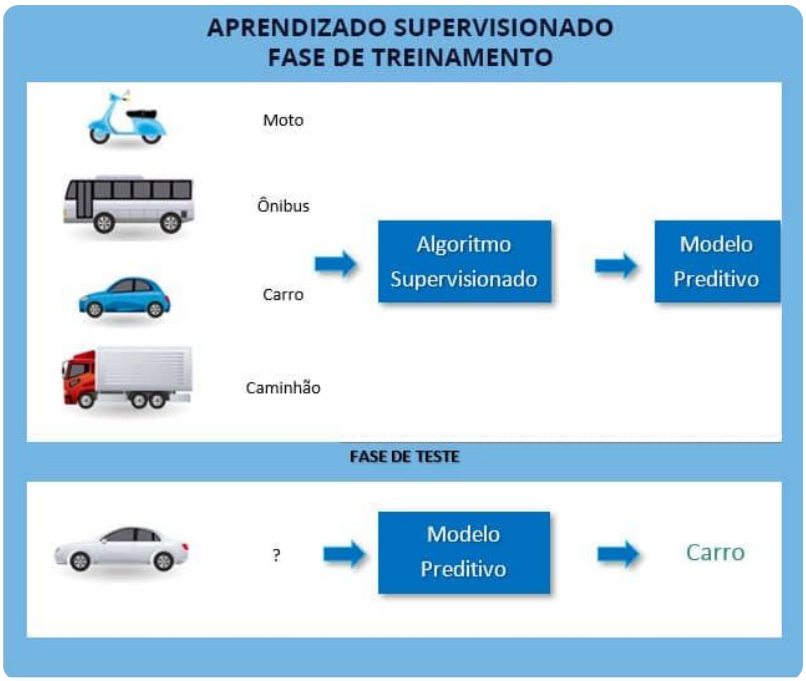
Ônibus

Carro

Caminhão

Nesse caso, as imagens de veículos são as entradas e os tipos de veículos são as saídas. Cada imagem foi rotulada pelo seu tipo e essa base de dados rotulada será o modelo de conhecimento gerado pela fase de treinamento do algoritmo de AM.

Na fase de teste, o modelo é aplicado pelo AM para identificar cada nova entrada de acordo com o seu tipo correspondente:



Fases do aprendizado de máquina supervisionado

Nesse tipo de aprendizado, pode-se estimar números reais ou valores dentro de um conjunto finito. A estimativa dos rótulos se dá de duas formas:

Classificação

Quando se estima com base em um conjunto finito de rótulos, que pode ser utilizado para classificar produtos defeituosos, clientes inadimplentes etc.

Regressão

Quando se estimam valores reais, por exemplo, o número de vendas, quantidade de matéria-prima necessária para um determinado período, número de produtos com defeito etc.

> Dessa maneira, se o rótulo é um número real, a tarefa chama-se regressão. Se o rótulo vem de um conjunto finito e não ordenado, então a tarefa chama-se classificação.

Para entender melhor, imagine uma base de dados de clientes inadimplentes, na qual diversas características – como limite de compra, quantidade de pedidos, valor médio de cada pedido etc. – podem levar à inadimplência. Os algoritmos de classificação são capazes de analisar as características de cada cliente e predizer as chances de um determinado cliente, ainda não rotulado, ser inadimplente.

Já os algoritmos de regressão podem estimar o número de vendas, a quantidade de matéria-prima necessária para um determinado período, o número de produtos com defeito. Como exemplo, pense em uma base de dados de imóveis para locação, em que diversas características – como a quantidade de quartos e de banheiros, se tem garagem, tamanho do imóvel, valor do aluguel etc. – levam ao valor de locação do imóvel. Os algoritmos de regressão são capazes de analisar as características de cada imóvel e predizer o valor do aluguel correspondente.

Como exemplo, pense em uma base de dados de imóveis para locação, em que diversas características – como a quantidade de quartos e de banheiros, se tem garagem, tamanho do imóvel, valor do aluguel etc. – levam ao valor de locação do imóvel. Os algoritmos de regressão são capazes de analisar as características de cada imóvel e predizer o valor do aluguel correspondente.

> Sendo assim, esse tipo de aprendizado é capaz de tomar decisões precisas quando recebe novos dados não rotulados a partir de um treinamento com dados que têm rótulos conhecidos.

Aprendizado de máquina não supervisionado

Ao contrário do aprendizado supervisionado que acabamos de discutir, no aprendizado não supervisionado não existem experiências predefinidas a serem utilizadas como referência para aprender.

Desvendando o Aprendizado Não Supervisionado: Da Teoria à Prática

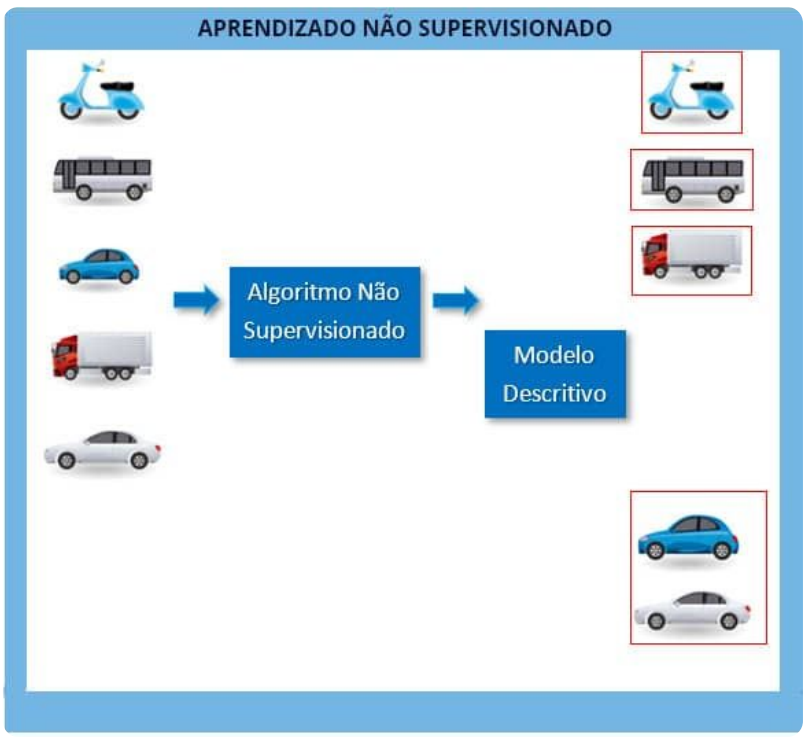
Neste vídeo, adentramos no Aprendizado de Máquina Não Supervisionado. Sem rótulos predefinidos, os algoritmos revelam padrões em dados não rotulados, formando agrupamentos e descobrindo características únicas. Entenda como essa abordagem redefine a análise de dados e a tomada de decisões.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Os algoritmos de AM não supervisionados utilizam conjunto de dados de entrada não rotulados (sem um resultado conhecido) e, apenas com base em suas características, aprendem a detectar agrupamentos implícitos ou características especialmente úteis para a sua categorização.

Vamos retornar para a base de dados composta por um conjunto de imagens de veículos, porém agora ela não conta com o rótulo que identifica o tipo de veículo. Sendo assim, o algoritmo de AM não supervisionado atuará agrupando as imagens dos veículos de acordo com as suas semelhanças, gerando *clusters*:



Aprendizado não supervisionado

Para entender melhor, imagine a base de dados de clientes, porém sem ter explicitamente a informação de que um cliente é inadimplente. Nesse caso, um algoritmo de aprendizado não supervisionado vai criar agrupamentos a partir das características dos clientes e, muito provavelmente, os inadimplentes estarão no mesmo grupo.

Além disso, como cada agrupamento vai conter clientes com características semelhantes, especialistas podem definir um nível de risco para cada grupo criado. Em vez de definidos apenas como inadimplentes ou não, os clientes podem ser agrupados de acordo com o nível de risco de se tornarem inadimplentes.

Outro exemplo seria o caso de um comércio que desejasse conhecer o perfil dos seus consumidores. Pode haver um perfil de consumidor que sempre compra vinho e queijo ou que compra carne e carvão ou ainda leite em pó e fralda. Se colocarmos esses produtos em prateleiras distantes, é possível que ocorra aumento de vendas, pois aumentará o tempo e o percurso do cliente no comércio.

No entanto, não estamos registrando para cada compra o perfil ao qual o consumidor pertence. Sequer sabemos quantos perfis de consumidores existem. Nesse caso, o computador terá que descobrir esses perfis por meio de semelhanças entre os dados.

Atenção

Uma opção seria observar, nos registros de compras, se existem padrões repetidos, os quais permitiriam a inferência de um grupo ou perfil de consumidor. Outra opção seria ver diretamente quais produtos são frequentemente comprados juntos e, então, aprender uma regra associativa entre eles. Dessa maneira, é possível aplicar técnicas que utilizam regras de associação e clusterização. Essas técnicas podem ser aplicadas na fase de pré-processamento ou mineração de dados, para se encontrar anomalias (os outliers), realizar redução de dimensionalidade em features etc.

Aprendizado semissupervisionado

O aprendizado de máquina semissupervisionado é um método que abrange ambos os métodos: supervisionado e não supervisionado.

> Esse aprendizado utiliza tanto o conjunto de dados rotulados como também o conjunto de dados não rotulados.

Desvendando o Aprendizado Semissupervisionado: Da Teoria à Prática

Neste vídeo, exploramos o Aprendizado Semissupervisionado, uma fusão estratégica do supervisionado e não supervisionado. Descubra como algoritmos combinam dados rotulados e não rotulados, otimizando a precisão e a eficiência em tarefas de classificação e clusterização.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

No caso de dados rotulados, pode-se utilizar aprendizado supervisionado para induzir classificadores a partir desses dados. Ao contrário, quando os rótulos dos dados não estão definidos, pode-se utilizar aprendizado não supervisionado com o objetivo de encontrar *clusters*, ou seja, agrupamento entre os dados por semelhança.

Atenção

Já o aprendizado semissupervisionado consiste em utilizar algoritmos que aprendem a partir de exemplos rotulados e não rotulados. Isso se deve ao fato de existirem em abundância dados não rotulados e os rotulados geralmente serem mais escassos.

Além disso, a rotulação de dados pode ser custosa, como nos casos de indexação de vídeo, categorização de textos e diagnósticos médicos, entre outros.

Outra motivação vem do fato de que, geralmente, os algoritmos induzidos exclusivamente a partir de um pequeno conjunto de dados rotulados não apresentam boa precisão. Sendo assim, o aprendizado semissupervisionado se utiliza dos dados rotulados para obter informações sobre o problema e utilizá-las no processo de aprendizado a partir dos dados não rotulados.

Aprendizado semissupervisionado pode ser utilizado tanto em tarefas de classificação como em tarefas de *clustering*.

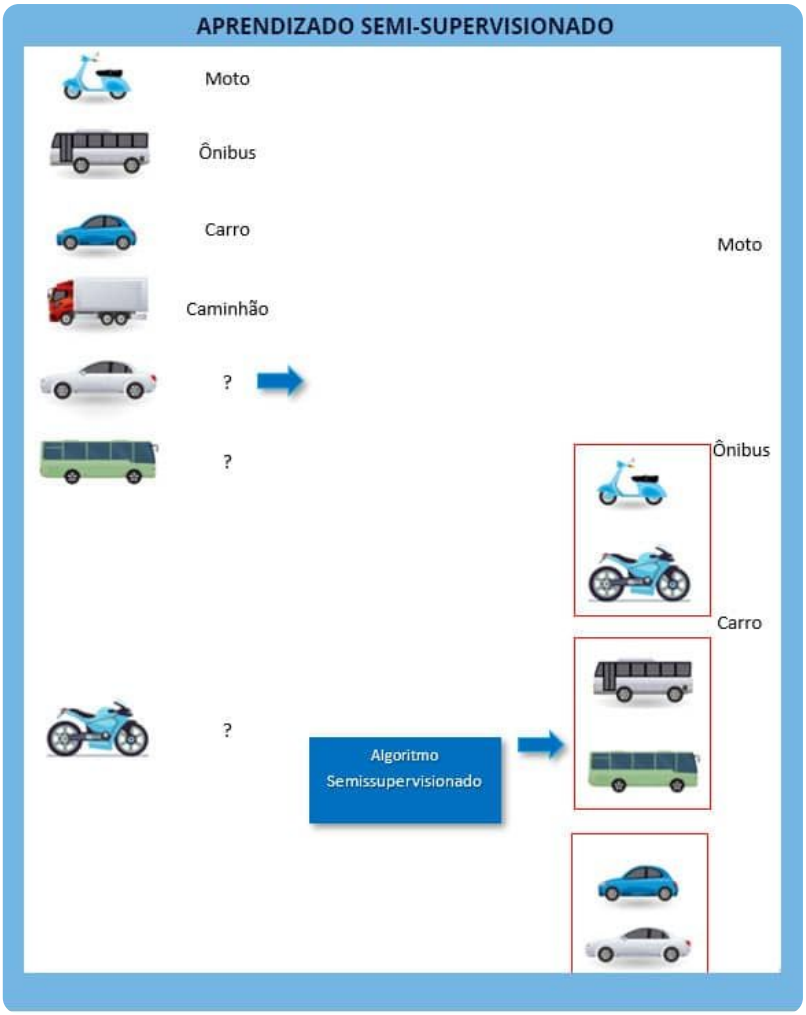
Algoritmos de classificação semissupervisionados

Rotulam, com uma certa margem de segurança, alguns dos dados no conjunto de dados não rotulados, os quais são posteriormente utilizados durante a fase de treinamento do classificador, gerando, assim, uma classificação mais precisa.

Algoritmos de clusterização semissupervisionados

Os dados rotulados são utilizados no processo de criação dos *clusters*, servindo como conhecimento expresso na forma de restrições e, dessa forma, resultando em melhores *clusters*.

Voltando à base de dados de imagens de veículos, suponha que, dentre as imagens, algumas sejam rotuladas com o tipo de veículo e outras estejam sem rótulo. O algoritmo de aprendizado semissupervisionado usaria as imagens rotuladas para fazer inferências sobre qual o tipo de veículo das imagens não rotuladas:



Aprendizado semissupervisionado

Aprendizado por reforço

O entendimento do conceito de aprendizado por reforço fica mais fácil com a ajuda da Psicologia, cuja corrente denominada behaviorismo define o comportamento humano como resultado das influências do ambiente. Sendo assim, **podemos entender que o comportamento pode ser moldado de acordo com estímulos e respostas**.

Os principais expoentes do behaviorismo são os psicólogos **Ivan Pavlov e Burrhus Frederic Skinner**.

Ivan Pavlov e Burrhus Frederic Skinner

Ivan PavlovIvan Pavlov, por sua vez, desenvolveu a teoria do comportamento em resposta aos estímulos do ambiente. Burrhus Frederic SkinnerSkinner, dentre outros experimentos, usou a ideia de aprendizado por reforço com recompensas e punições no treinamento de pombos, com o objetivo de conduzir mísseis na Segunda Guerra Mundial.

> [...] De acordo com Pavlov, o requisito fundamental é que qualquer estímulo externo seja o sinal (estímulo neutro) de um reflexo condicionado e se sobreponha à ação de um estímulo absoluto.
>
> —
>
> "LA ROSA, 2003"

Desvendando o Aprendizado por Reforço na Máquina: Explorando o Behaviorismo na Tecnologia

Neste vídeo, exploramos o Aprendizado por Reforço, conectando Psicologia e Inteligência Artificial. Como Pavlov e Skinner influenciam a máquina a agir através de recompensas e punições? Entenda esse paradigma, das estratégias de carros autônomos ao desafio de equilibrar exploration e exploitation.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Uma situação que ilustra bem esse tipo de comportamento é o adestramento de um cão; por exemplo, ensiná-lo a sentar a partir de um comando. Primeiramente, o animal não executará o comando requerido, e o tutor responde a isso com um “reforço negativo” (punição). Quando o cão se aproxima do que deveria fazer, é a vez de oferecer-lhe “reforços positivos” como sinal de aprovação ou incentivo. Se o cão de fato se sentar após o comando, a ele será dada uma recompensa – um biscoitinho, por exemplo. Com várias repetições desse mesmo experimento, com o tempo, o cão passa a associar a relação de “causa-efeito” entre o comando e a recompensa a ser recebida e assim “aprende” a obedecer a esse comando.

O famoso experimento do “cão de Pavlov” mostra esse paradigma de aprendizagem. Ivan Pavlov apresentou a ideia do “reflexo condicionado”, baseado no seguinte experimento:

Estímulo neutro

Apresentando um pedaço de carne a um cão, o animal salivava, desejando o alimento.

Estímulo incondicionado

Em vez de apresentar apenas a carne, Pavlov soava uma campainha sempre que isso acontecia.

Condicionamento

Com a repetição, o cão passava a associar os dois "estímulos" (carne e campainha).

Estímulo condicionado

Portanto, salivando assim que ouvia a campainha.

> Aprendizado por reforço é uma área do AM que avalia como a máquina deve agir em determinados ambientes de acordo com as recompensas ou punições recebidas e acumuladas com o tempo.

É um método de aprendizado diferente dos anteriores, pois não existem conjunto de treinamento, dados rotulados e dados não rotulados.

⚠

Atenção

A máquina aprende executando ações e avaliando recompensas. Primeiramente ela observa o ambiente e seu conjunto de cenários futuros possíveis. Com base nisso, escolhe uma ação a ser executada e recebe a recompensa associada a ela. O processo se repete até que a máquina seja capaz de escolher a melhor ação para cada um dos cenários possíveis. E isso vai depender das circunstâncias nas quais a ação será executada. Ou seja, uma recompensa ou punição é dada à máquina, dependendo da decisão tomada.

Esse processo funciona com o tempo e, com a repetição dos experimentos, a máquina vai associando as ações que geram maior recompensa para cada situação que o ambiente apresenta, passando a evitar as ações que geram punição ou recompensa menor.

Nesse tipo de aprendizado, muda-se um pouco o método com relação aos demais. Geralmente é aplicado quando se conhecem as regras, mas não se sabe a melhor sequência de ações a executar. As regras são iterativamente aprendidas, como num jogo de xadrez ou num videogame, onde o fundamental são as ações tomadas pelo jogador ou pela máquina.

Exemplo

Outro exemplo do uso desse método são os carros autônomos, que tomam decisões dependendo do cenário ao redor, recebendo recompensas negativas quando colidem com o ambiente ou com outros veículos; com repetidas etapas, aos poucos, eles aprendem a contornar os obstáculos.

> E se o objetivo for testar novas combinações de ações ótimas em uma sequência de estados não realizados anteriormente em busca de uma recompensa maior?

Vamos voltar ao adestramento de cães; suponha que, depois de tê-lo ensinado a sentar com sucesso, o tutor queira fazer um teste de obediência – manter o cão sentado enquanto anda para trás, até chegar a uma distância de cinco metros dele. Caso o cão se sente, ele ganha o biscoitinho (assim como antes); caso ele não se levante até o tutor ficar a cinco metros dele, o cão recebe uma recompensa maior (um pedaço de carne, por exemplo).

Essa situação ilustra um *trade-off*: o animal pode escolher por "testar" novas combinações de ações ótimas em uma sequência de "estados" não realizada anteriormente, em busca de uma recompensa maior, mas não imediata (o chamado *exploration*); ou simplesmente se ater à recompensa que ele pode obter por meio de uma ação já conhecida (o chamado *exploitation*).

> Fazer a máquina ser capaz de encontrar o meio termo ótimo entre *exploration* e *exploitation* é um dos principais desafios do aprendizado por reforço, e é bastante pertinente para aplicações mais complicadas, como ensinar uma máquina a jogar xadrez, por exemplo.

Uma estratégia vencedora frequentemente envolve abrir mão de uma vantagem imediata, ou até mesmo o sacrifício de peças, visando o sucesso a longo prazo – um bom jogador deve ser capaz de levar em consideração as consequências de sua jogada várias rodadas adiante, sabendo que a resposta do oponente também estará visando um benefício futuro, e assim por diante.

Verificando o aprendizado

Questão 1

Identifique, dentre as opções a seguir, que método do aprendizado de máquina não se baseia em experiências predefinidas a serem utilizadas como referência para aprender:

A

Aprendizado supervisionado.

B

Aprendizado não supervisionado.

C

Aprendizado semissupervisionado.

D

Aprendizado por reforço.

E

Aprendizado max-supervisionado.

A alternativa B está correta.

O aprendizado não supervisionado é um método de AM que se baseia em dados não rotulados, buscando a observação de semelhança entre os dados.

Questão 2

Identifique, dentre as opções a seguir, que método do AM avalia como a máquina deve agir em determinados ambientes de acordo com as recompensas ou punições recebidas e acumuladas com o tempo:

A

Aprendizado supervisionado.

B

Aprendizado não supervisionado.

C

Aprendizado semissupervisionado.

D

Aprendizado por reforço.

E

Aprendizado max-supervisionado.

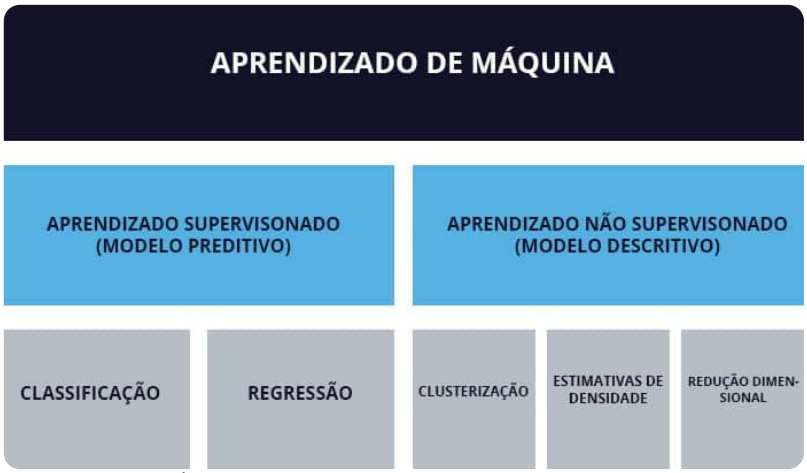
A alternativa D está correta.

O aprendizado por reforço é um método de AM que não tem por base um conjunto de treinamento, dados rotulados e dados não rotulados. Nesse método, o aprendizado é realizado mediante a execução de ações e a avaliação de recompensas.

Modelos analíticos de aprendizado de máquina

O aprendizado de máquina automatiza a criação de modelos analíticos utilizados para predição, simulação e análise de dados. Dessa forma, os sistemas são capazes de aprender com o conjunto de dados, identificar padrões, além de auxiliar na tomada de decisão.

O AM pode ser realizado de duas maneiras, na predição ou no auxílio da descrição dos dados:



Aprendizado de máquina – modelo preditivo e descritivo

A predição dos dados ou análise preditiva é aplicada para situações em que se deseja prever algum comportamento ou resultado. A utilidade da análise preditiva está em verificar tendências de consumo, flutuações econômicas, por exemplo. Dessa maneira, é realizada uma análise de dados ao longo do tempo em busca de padrões comportamentais, variações, para prever o comportamento futuro, dadas as condições atuais.

> A análise de dados é feita mediante um treinamento com os dados históricos e rotulados, e o resultado desse treinamento é o modelo preditivo a ser empregado nos dados não rotulados.

A descrição dos dados ou a análise descritiva consiste em detalhar todos os acontecimentos passados, com foco em parâmetros e referências que possam subsidiar a tomada de decisão. Ou seja, a análise descritiva compreende os dados históricos, para gerar a fotografia do presente como auxílio à tomada de decisão. Sendo assim, o modelo descritivo é gerado com base na análise dos dados históricos e no cruzamento de informações, e o resultado é um panorama preciso para o momento com explicações sobre o que está acontecendo com base em dados existentes.

Modelo preditivo

É usado pelo aprendizado supervisionado com foco em predizer os rótulos desconhecidos. Na prática, o aprendizado supervisionado utiliza os dados para aprender como resolver determinado problema, que pode ser de classificação ou de regressão.

Classificação

Os problemas de classificação são aqueles que buscam encontrar uma classe, dentro das possibilidades existentes. A ideia é prever os resultados em uma saída discreta, a partir do mapeamento de variáveis de entrada em categorias distintas.

Exemplo

Por exemplo, uma classe pode identificar se determinado aluno foi aprovado ou reprovado, um paciente pode saber se um exame oferece diagnóstico para determinada doença ou não, dentre outras situações.

Por exemplo, imagine uma base de dados composta de seis registros para prever se uma pessoa irá ou não jogar tênis com base nas condições climáticas.

Tempo	Temperatura	Humidade	Vento	Jogar Tênis (sim ou não)
Ensolarado	**Quente**	**Alta**	**Fraco**	**Não**
Ensolarado	Quente	Alta	Forte	Não
Nublado	Quente	Alta	Fraco	Sim
Chuvoso	Moderada	Alta	Fraco	Sim
Ensolarado	Moderada	Normal	Normal	Sim
Chuvoso	Quente	Alta	Forte	Não
Chuvoso	Quente	Normal	Fraco	**?**

Base de dados – jogar ou não jogar tênis

Para prever se uma pessoa deve sair de casa ou não, para jogar tênis, temos que considerar algumas variáveis, como: aspecto do céu (tempo), temperatura do dia, umidade e força do vento.

Sendo assim, um dia chuvoso (tempo chuvoso), mas com temperatura quente, pode inviabilizar a saída de casa, pois há chance de começar a chover durante o jogo. Por outro lado, se o tempo está chuvoso, mas com temperatura moderada, resulta em “sim” para ir jogar tênis, pois o clima está agradável.

Nesse caso, o problema de classificação, ou seja, a identificação da classe – jogar tênis (sim ou não) – será resolvido pelo modelo preditivo gerado de acordo com o conjunto de dados rotulados.

O vídeo a seguir ilustra outro exemplo simples como o apresentado acima, com emprego de um algoritmo clássico de classificação, o "Naive Bayes". Para isso, é usado um conhecido software de Machine Learning com licença livre, denominado WEKA (do inglês Waikato Environment for Knowledge Analysis), disponível no site da Universidade de Waikato, Nova Zelândia, e descrito em (Witten et. al., 2011).

Demonstração prática do problema de classificação

Assista ao vídeo a seguir e observe um caso prático de problema de classificação.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Regressão

Os problemas de regressão são aqueles que precisam prever um valor numérico específico. Nesse caso, a ideia é prever os resultados em uma saída contínua, o que significa mapear variáveis de entrada para uma função contínua.

Nesse tipo de problema, o modelo preditivo apresenta um valor como resposta.

Exemplo

Por exemplo, o valor numérico do preço de um produto, o peso ou a altura de uma pessoa, a metragem de uma casa, dentre outras situações.

Imagine uma base de dados de casas do mercado imobiliário. Um problema de regressão é tentar prever o preço de venda de determinada casa, levando em consideração o modelo preditivo gerado a partir das características presentes na base de dados.

Nesse caso, estamos prevendo um valor – o preço de venda da casa – em função das características da casa, o que é uma saída contínua. O preço da casa (variável dependente) é calculado a partir das características dos imóveis na base de dados (variáveis independentes), como: tamanho da casa, tamanho do lote, quantidade de quartos, quartos com granito, banheiros reformados e preço de venda do imóvel.

A tabela abaixo 2 ilustra o conjunto de dados com as características dos imóveis-base e as características da casa cujo preço de venda se deseja saber.

Tamanho da casa (m^2)	Tamanho do lote (m^2)	Qtd. quartos	Granito (sim ou não)	Banheiro reformado (sim ou não)	Preço de venda
353	**919**	**6**	**0**	**0**	**R$ 205.000,00**
325	1.006	5	1	1	R$ 224.900,00
403	1.015	5	0	1	R$ 197.900,00
240	1.415	4	1	0	R$ 189.900,00
220	960	4	0	1	R$ 195.000,00
353	1.999	6	1	1	R$ 325.000,00

Tamanho da casa (m^2)	Tamanho do lote (m^2)	Qtd. quartos	Granito (sim ou não)	Banheiro reformado (sim ou não)	Preço de venda
298	936	5	0	1	R$ 230.000,00
320	967	5	1	1	**?**

Base de dados de imóveis

O modelo preditivo gerado vai sugerir o preço da casa de acordo com o conjunto de dados. Mas talvez a presença de granito nos quartos não seja relevante para o preço; ou o tamanho da casa influencie negativamente no preço do imóvel; ou ainda o número de quartos e a reforma do banheiro sejam mais importantes que o tamanho do lote, por terem pesos maiores do que o relativo a este.

Atenção

Os problemas do modelo preditivo de classificação são diferentes dos problemas do modelo preditivo de regressão. A classificação é a tarefa de prever um rótulo de classe discreto, e a regressão é a tarefa de prever um valor contínuo.

O que define o tipo de problema não é a distinção entre a previsão de um número e a de uma letra ou de uma palavra. É possível prever valores numéricos nos casos de problemas de classificação, mas a previsão será uma categoria. Isso quer dizer que é possível encontrar valores como 0 e 1 representando classes.

Por exemplo, o valor 0 pode indicar a reprovação de um aluno e o valor 1, a aprovação. Isso significa que o número 0 é algo diferente do seu valor, que pode ser substituído por qualquer letra, palavra ou até mesmo outro número, sem prejudicar o entendimento das previsões.

Demonstração prática do problema de Regressão

Assista ao vídeo a seguir e observe um caso prático de problema de regressão.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Modelo descritivo

> O modelo descritivo é usado pelo aprendizado não supervisionado para dados não rotulados, com foco na identificação e na indicação de similaridades no conjunto de dados.

O AM não supervisionado descobre padrões previamente desconhecidos em dados. Como não se sabe quais devem ser os resultados, ainda não há como determinar a precisão deles, tornando o aprendizado de máquina não supervisionado mais aplicável aos problemas do mundo real do que os outros tipos de aprendizado de máquina.

Comentário

Na prática, o aprendizado não supervisionado nos permite abordar problemas com praticamente nenhum conhecimento sobre os resultados; isso significa que não fazem uso de atributos de saída para aprender como resolver determinado problema.

Os problemas de aprendizado não supervisionado podem se apresentar nas seguintes categorias: *clustering*, estimativas de densidade e redução dimensional.

Clustering

Uma das principais categorias utilizadas pelo aprendizado de máquina não supervisionado é a clusterização (*clustering*), que divide automaticamente o conjunto de dados em grupos, de acordo com a similaridade.

> O objetivo dessa técnica é agrupar os dados utilizando medidas de similaridade e incluí-los em grupos, ou seja, em *clusters*.

No caso de alguns dados serem largamente distintos dos demais grupos, podemos assumir que eles são anomalias. A detecção de anomalias é útil para reconhecer transações fraudulentas, descobrir peças defeituosas ou identificar um caso discrepante ocasionado por erro humano na entrada de dados.

Por exemplo, imagine uma base de dados históricos de uma concessionária da BMW, na qual estão armazenadas todas as informações de vendas passadas, bem como informações a respeito de cada pessoa que comprou uma BMW, olhou uma BMW ou apenas procurou algo no salão de exposição da BMW. A concessionária acompanhou quantas pessoas passaram pela concessionária e pelo salão de exibição, quais carros elas olharam e com que frequência compraram.

Um problema de aprendizado não supervisionado para esse caso é encontrar padrões nos dados, gerando agrupamento baseado em comportamentos semelhantes entre os clientes. A partir da base de dados, podemos identificar cinco *clusters*:

Cluster 0 - Sonhadores

- Andam pela concessionária olhando os carros estacionados.
- Apresentam baixa taxa de quem entra na concessionária.
- Não compram nada.

Cluster 1 - Amantes do modelo M5

- Tendem a ir diretamente em direção ao modelo M5.
- Ignoram os carros dos modelos 3-series e Z4.
- Apresentam baixa taxa de compra – somente 52%.

Cluster 2 - Jogados fora

- Não são estatisticamente relevantes.
- Não se pode tirar nenhuma conclusão sobre seu comportamento.

Cluster 3 - Bebês da BMW

- Sempre acabam comprando um carro.
- Sempre acabam financiando.
- Andam pelo estacionamento olhando os carros.
- Usam a pesquisa do computador disponível na concessionária.
- Tendem a comprar os modelos M5 ou Z4, mas nunca o modelo 3-series.

Cluster 4 - Começando com a BMW

- Sempre olham o modelo 3-series; nunca olham para o M5, que é o mais caro.
- Entram diretamente no salão de exibição e preferem não andar pelo estacionamento.
- Tendem a ignorar os terminais de pesquisa por computador.
- 50% chegam ao estágio de financiamento e somente 32% acabam finalizando a transação.

Ao analisar as características de cada *cluster*, a concessionária pode ter condições de identificar padrões em seus clientes e, assim, propor melhorias.

Exemplo

Por exemplo, o cluster 1 – Amantes do M5 – identifica um potencial problema, pois apresenta uma baixa taxa de compra; sendo assim, a concessionária poderia adotar, como possível melhoria, o envio de mais vendedores para a seção do M5. Outra análise pode vir do cluster 4 – Começando com a BMW. A concessionária poderia concluir que os clientes que esperam comprar seu primeiro BMW sabem exatamente o tipo de carro desejado, e esperam qualificar-se para o financiamento de modo a ter condições financeiras de comprá-lo; sendo assim, a concessionária poderia aumentar suas vendas, relaxando seus padrões de financiamento ou reduzindo os preços do 3-series.

Demonstração prática do problema de Clustering

Assista ao vídeo a seguir e observe um caso prático de problema de Clustering.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Estimativas de densidade

Esse método tem por base a função densidade de probabilidade – conceito fundamental em estatística. Nele se identificam as regiões densas, permitindo a formação de grupos com diferentes formatos.

A estimativa de densidade se apoia no conceito de **ε-vizinhança**, um espaço bidimensional.

Podemos então definir esse método como uma técnica não paramétrica para estimação de curvas de densidades, em que cada observação é ponderada pela distância em relação a um valor central, o núcleo.

O método usa o valor de densidade para agrupar aqueles pontos que têm valores similares. Dessa maneira, é possível identificar vizinhanças densas nas quais a maioria dos pontos está contida.

Por exemplo, imagine um conjunto de dados com os resultados de uma análise química de vinhos cultivados na região da Itália. Cada instância do conjunto de dados corresponde a pontos em um espaço n-dimensional, conforme a figura abaixo:

Ɛ vizinhança

A ε-vizinhança de um ponto p é o conjunto de pontos contidos em um círculo de raio ε, centrado em p.

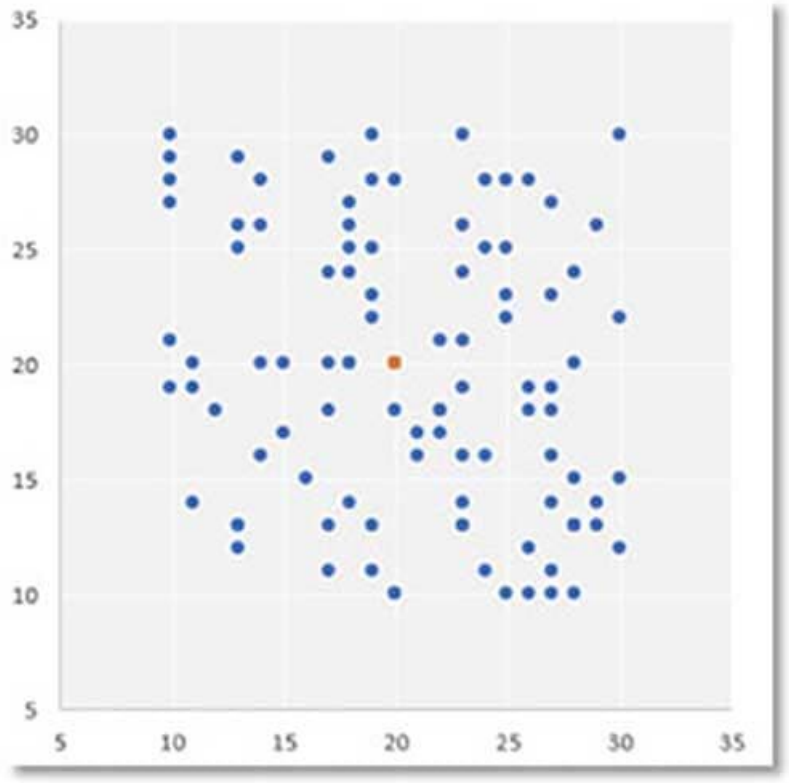


Gráfico de amostra de vinho

Nesse exemplo, temos 100 vinhos cultivados na região da Itália (pontos azuis) e distribuídos em um espaço bidimensional, de acordo com os resultados da análise química. O novo vinho está localizado no espaço pelo ponto vermelho, e será o ponto p (p = (20,20)).

Supomos que o ε é igual a 5; nesse caso, com um raio de 5, encontraremos 19 pontos no espaço de 100, conforme se pode ver na próxima figura:

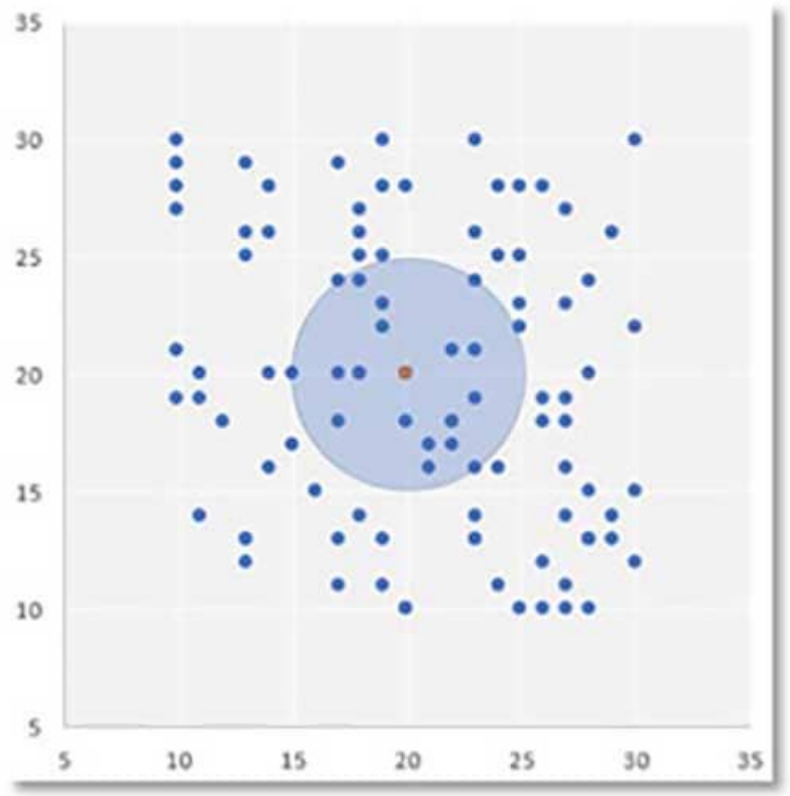


Gráfico de amostra de vinho (ε = 5)

Se diminuirmos o valor de ε, menor será o número de vizinhos de p. Se o ε for igual a 2,5 significa que teremos uma área do círculo menor e com apenas quatro vizinhos, como na figura abaixo:

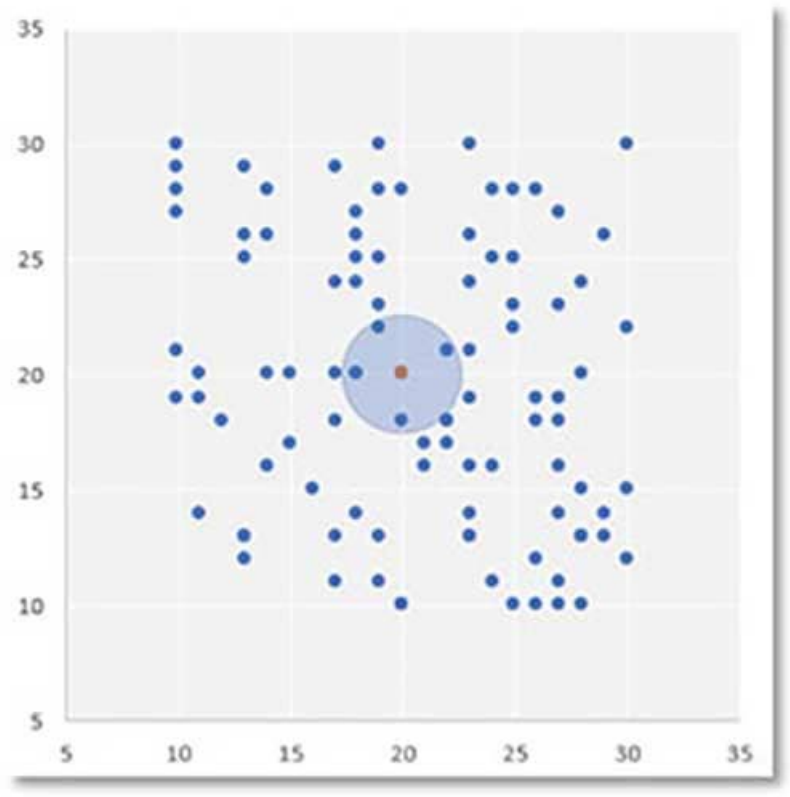


Gráfico de amostra de vinho (ε = 0,25)

A densidade é calculada pela divisão da massa pelo volume (Densidade = Massa / Volume). Sendo assim, para o caso de ε ser igual a 5, temos o volume como a área do círculo ($A=\pi r^2$), $\pi(5)^2$. Dessa forma, Massa/Volume será igual a 19 pontos/$\pi(5)^2$.

Demonstração prática do problema de Estimativas de densidade

Assista ao vídeo a seguir e observe um caso prático de problema de Estimativas de densidade.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Redução dimensional

> O método de redução dimensional se aplica quando temos um conjunto de dados com uma quantidade de variáveis bem grande em um problema.

Às vezes, há muitas variáveis altamente correlacionadas, ou seja, elas são redundantes ou são variáveis que não apresentam informações úteis ao problema. Essa situação faz o modelo selecionado apresentar muitos parâmetros, e pode acabar causando *overfitting*, significando que o modelo tem um desempenho excelente no treino, porém quando o utilizamos em dados de teste, o resultado é ruim.

Curiosidade

Podemos entender que, nesse caso, o modelo aprendeu tão bem as relações existentes no treino, que acabou apenas decorando o que deveria ser feito. Ao receber as informações novas de teste, o modelo tenta aplicar as mesmas regras decoradas. Acontece que, com dados diferentes, essa regra não tem validade e o desempenho é afetado.

A redução da dimensionalidade é o processo de diminuir o número de variáveis que serão inseridas em um modelo para treino. Para fazer isso, precisamos identificar quais são as variáveis principais, ou seja, as mais importantes.

> Uma das técnicas mais populares de redução de dimensionalidade é a *principal components analysis* (PCA) ou análise de componentes principais.

A PCA consiste em um procedimento algébrico que converte as variáveis originais (que são tipicamente correlacionadas) em um conjunto de variáveis não correlacionadas (linearmente), que se designam por componentes principais (PC) ou variáveis latentes.

Atenção

A PCA fornece o mapeamento de um espaço com N dimensões (em que N é o número de variáveis originais) para um espaço com M dimensões (onde M é tipicamente muito menor do que N).

Essa técnica não é paramétrica; é, sim, linear, simples e rápida, e tem por base a variância dos dados. Sendo assim, ela tenta criar uma representação dos dados, com uma dimensão menor, porém mantendo a variância entre eles.

> Outra técnica é a *curvilinear distance analysis* (CDA), poderosa para reduzir dimensionalidade, por ser não linear.

Essa técnica é bem complicada de implementar e seu custo computacional é bem elevado, se comparado à PCA, pois se baseia em distância de grafos.

Resumindo

Em resumo, podemos concluir que a redução de dimensionalidade é capaz de simplificar modelos, reduzir o tempo de treino e o overfitting.

Por exemplo, imagine uma base de dados de jogadores de basquete. Nosso objetivo é determinar a qualidade de um jogador. Dentre todas as variáveis da base, existem a média da quantidade de dias que o jogador treina por semana e a média de treinos por mês. Quando analisamos os resultados, percebemos que as variáveis são correlacionadas, ou seja, na maioria dos casos, quando um jogador treina muitas vezes por semana, acaba realizando muitos treinos por mês.

Isso mostra que não é necessário utilizar as duas variáveis para treinar o modelo. A premissa é que as variáveis redundantes podem ser removidas sem perda de informação/valor para o modelo.

Demonstração prática do problema de Redução dimensional

Assista ao vídeo a seguir e observe um caso prático de problema de Redução dimensional.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

O aprendizado de máquina supervisionado utiliza os dados rotulados para aprender como resolver determinado problema. Identifique as categorias que resolvem esse problema de AM:

A

Clustering e regressão.

B

Redução dimensional e estimativas de densidade.

C

Classificação e regressão.

D

Associação e regressão.

E

Classificação e clusterização.

A alternativa C está correta.

O aprendizado de máquina supervisionado gera, a partir de dados rotulados, o modelo preditivo usado para predizer os rótulos desconhecidos. Para resolver esse tipo de problema, o aprendizado se utiliza de duas categorias: a classificação e a regressão.

Questão 2

O aprendizado de máquina não supervisionado utiliza os dados não rotulados com foco em identificar e indicar similaridades no conjunto de dados. Identifique as categorias que resolvem problemas de aprendizado não supervisionado:

A

Clustering e regressão.

B

Redução dimensional e estimativas de densidade.

C

Classificação e regressão.

D

Associação e regressão.

E

Classificação e clusterização.

A alternativa B está correta.

O aprendizado não supervisionado aborda problemas com praticamente nenhum conhecimento sobre os resultados, ou seja, não faz uso de atributos de saída para aprender como resolver a situação. Esse problema de aprendizado não supervisionado apresenta as seguintes categorias: *clustering*, estimativas de densidade e redução dimensional.

Considerações finais

Nas últimas décadas, o aprendizado de máquina tornou-se um dos pilares da Tecnologia da Informação. Com a crescente quantidade de dados disponíveis, a análise inteligente se tornará cada vez mais difundida como um ingrediente necessário para o progresso tecnológico.

Iniciamos conceituando aprendizado de máquina. Após passar pelos conceitos básicos, seus desafios e quem o está utilizando, identificamos os métodos de aprendizado, suas diferenças e apresentamos exemplos para cada método.

De acordo com a base de dados, a forma como estão organizados, se rotulados ou não, torna-se possível utilizar classificação ou regressão, ou até clusterização, para resolver determinados problemas de aprendizado.

Finalizamos abordando os problemas de AM nos métodos de aprendizado supervisionado e aprendizado não supervisionado.

Podcast

Agora, o especialista encerra o tema respondendo a alguns questionamentos sobre *Machine Learning*.

Conteúdo interativo

Acesse a versão digital para ouvir o áudio.

Explore+

Veja como o canal GCFAprendeLivre e o Canal Marcelo Tas abordam o tema Aprendizado de Máquina de uma forma clara e descontraída.

Referências

FITCH, F. B. **Warren S. McCulloch and Walter Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity**. Bulletin of Mathematical Biophysics, v. 5 (1943), pp. 115-133. *In*: Journal of Symbolic Logic, 1944, v. 9 (2), pp. 49-50. Cambridge University Press [online]. Publicado em: 12 mar. 2014.

LA ROSA, J. **Psicologia e educação**: o significado do aprender. Porto Alegre: PUCRS, 2003.

RUSSEL, S.; NORVIG, P. **Inteligência artificial**: uma abordagem moderna. 3. ed. São Paulo: LTC, 2009.

SAMUEL, A. L. (1959). **Some studies in machine learning using the game of checkers**. *In*: IBM Journal of Research and Development, v. 3 (3), jul. 1959.

WEISS, S. M.; KULIKOWSKI, C. A. **Computer systems that learn**: classification and prediction methods from statistics, neural nets, machine learning and expert systems. San Mateo: Kaufmann, 1991.

Witten, Ian H.; Frank, Eibe; Hall, Mark A.; Pal, Christopher J. (2011). **"Data Mining: Practical machine learning tools and techniques, 3rd Edition"**. Morgan Kaufmann, San Francisco (CA).