

MACHINE LEARNING

(Aprendizado de Máquina)

- Principais conceitos da matéria

O que é a predição(Prediction)?

Predição em Machine Learning é como adivinhar o que vai acontecer no futuro, com base em coisas que aconteceram no passado. É como, quando você prevê o clima olhando para os dados anteriores sobre temperaturas e padrões climáticos. O computador faz algo parecido, usando dados antigos para fazer suposições sobre coisas futuras.

A predição, envolve o uso de algoritmos e modelos para fazer estimativas ou previsões sobre os dados futuros com base em padrões identificados em dados históricos. Em outras palavras, é a capacidade de um sistema de aprendizado de máquina(Machine Learning) de deduzir um valor ou uma classe desconhecida com base em exemplos de dados que foram fornecidos.

O que é acurácia(Accuracy) ?

Acurácia é como se fosse uma nota que o modelo ganha por fazer previsões corretas. Se o modelo acertar muitas vezes, ele recebe uma nota alta (alta acurácia). Se errar com frequência, recebe uma nota baixa (baixa acurácia). Quanto maior a nota (ou seja, quanto maior a acurácia), melhor o modelo é em fazer as coisas certas.

A acurácia é uma medida que indica a proporção de previsões corretas feitas por um modelo em relação ao total de previsões feitas. É a capacidade do do modelo fazer previsões corretas. Quanto maior a acurácia, melhor o modelo está em fazer previsões. Existem outras métricas como precisão, recall e F1-score, que são responsáveis por dar uma visão mais completa do desempenho do modelo na qual os dados são desequilibrados.

Aprendizado de máquina Supervisionado (Supervised Learning)

O Aprendizado Supervisionado é como ensinar uma criança usando exemplos. Imagine que você queira ensinar essa criança a reconhecer frutas. Você mostra maçãs e bananas para ela, dizendo o nome de cada uma. Depois de ver muitos exemplos, a criança aprende a diferenciar as frutas. Agora, quando você mostrar uma fruta nova, a criança provavelmente conseguirá dizer se é uma maçã ou uma banana, porque ela aprendeu com os exemplos anteriores. O Aprendizado Supervisionado funciona de maneira semelhante.

Aprendizado Supervisionado é um tipo de abordagem em que ensinamos modelos usando exemplos de dados que já possuem respostas corretas. É como ensinar um computador a reconhecer algo, mostrando a ele vários exemplos já rotulados.

Nesse processo, você fornece ao modelo um conjunto de dados de treinamento que inclui entradas(**Features** = **X**) e suas saídas(**Target** = **Y**) correspondentes. O modelo aprende a relacionar as entradas às saídas corretas ao identificador padrão nos dados. Uma vez treinado, o modelo pode fazer previsões ou classificações para novos dados, com base no que aprendeu.

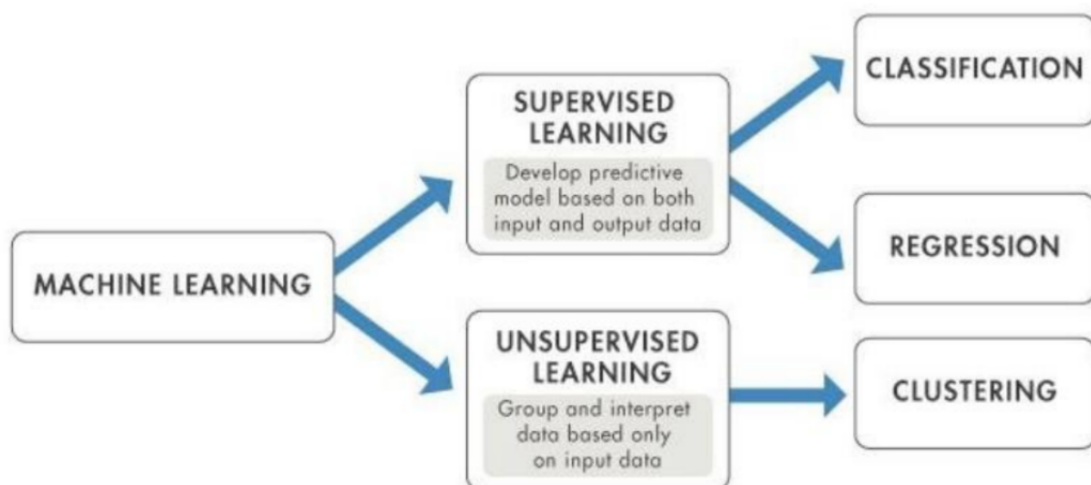
Aprendizado de máquina não supervisionado (Unsupervised Learning)

Aprendizado não supervisionado é como olhar para um monte de coisas misturadas e tentar encontrar padrões nelas sem que ninguém lhe diga o que são essas coisas. É como olhar para um monte de brinquedos e tentar organizá-las de acordo com suas semelhanças, mesmo sem saber o nome de cada brinquedo. O computador faz algo parecido: ele tenta organizar dados sem ter as respostas certas, só procurando coisas parecidas.

Ele é uma abordagem em que o modelo não recebe exemplos rotulados, ou seja, não é dito ao modelo o que as respostas corretas são. Em vez disso, o modelo busca identificar padrões e estruturas nos dados por conta própria. É como dar um monte de peças de quebra-cabeça para o computador e deixá-lo descobrir como elas podem ser agrupadas e organizadas sem saber a imagem final. O modelo procura por semelhanças e diferenças nos dados, tentando criar grupos ou encontrar características importantes.

O aprendizado não supervisionado é muito usado para tarefas como agrupamento (encontrar grupos de dados semelhantes = **Clustering**) e redução de dimensionalidade (ou seja, simplificar dados complexos tornando-os mais fáceis de entender), onde não há respostas corretas conhecidas para comparar.

Para cada grupo de algoritmo (ou seja, os grupos supervisionados e não supervisionados) existem tipos de tarefas que os algoritmos realizam. As principais tarefas são, Classificação, Regressão e Agrupamento.



O que é Classificação(Classification)?

A classificação no aprendizado de máquina é como ensinar um computador a separar coisas em grupos. Imagine que você tem fotos de gatos e cachorros, e você quer ensinar o computador a diferenciá-los. A classificação é como um jogo de adivinhação: você mostra muitas fotos de gatos e cachorros e diz ao computador qual é qual. Com o tempo, o computador aprende padrões, como orelhas pontudas para gatos e focinhos molhados para

cachorros. Depois disso, você pode mostrar uma foto nova, e o computador tenta adivinhar se é um gato ou um cachorro com base no que aprendeu. Classificação é como ensinar o computador a ser um bom detetive de gatos e cachorros.

A classificação em machine learning é um tipo de tarefa na qual um algoritmo de aprendizado automático é treinado para atribuir categorias ou rótulos a dados com base em padrões identificados nos dados de treinamento. Essa tarefa é usada para prever a qual classe ou categoria um novo conjunto de dados pertence com base em suas características. É normalmente usada em problemas como detecção de spam de e-mail, diagnóstico médico, reconhecimento de imagem (por exemplo, identificação de objetos em uma foto) e muito mais. Os algoritmos de classificação procuram aprender a função que mapeia as entradas para as saídas desejadas com base nos exemplos de treinamento fornecidos.

Resumindo, nós queremos encontrar uma classe: é ou não é um spam, é um pato ou uma galinha, o paciente está doente ou não está doente. Quando tentamos classificar e separar o que é um cachorro de um gato, por exemplo, estamos falando de classificação.

O que é regressão(Regression)?

Imagine que você quer prever algo, como quanto um carro usado pode custar. A regressão é como uma fórmula mágica que olha para coisas que podem afetar o preço, como a idade do carro, a quilometragem, a marca, etc. Ela encontra uma maneira de combinar essas coisas e dar a você um palpite aproximado do preço do carro. É como quando você vê muitos exemplos de carros diferentes e aprende a adivinhar os preços dos outros carros. A regressão faz algo parecido, mas com números!

A regressão em aprendizado de máquina é uma técnica que envolve a previsão ou estimativa de um valor numérico com base em variáveis independentes ou recursos. Ela é usada quando se deseja entender a relação entre variáveis e prever um valor contínuo, como prever preços de imóveis, temperatura, receita de vendas, entre outros.

Resumindo, quando tentamos prever um valor específico, com um dado numérico, é regressão.

O que é agrupamento(Clustering)?

O que é modelo de predição?

histogramas

Claro! Vou explicar de forma simples:

Imagine que você tem um monte de notas de testes de alunos e você quer entender quais notas são mais comuns e quais são mais raras. Um histograma é como um gráfico que mostra isso de maneira fácil de entender. Ele divide as notas em grupos, chamados "intervalos", como 0-10, 11-20, etc. Então, para cada intervalo, ele mostra quantos alunos tiraram notas dentro desse intervalo. O histograma é como um mapa que ajuda você a ver quais notas são mais populares e quais são menos populares.

o que são rótulos?

Claro! Rótulos em aprendizado de máquina são como etiquetas que dizem o que algo é. Quando estamos ensinando um computador a reconhecer coisas, como gatos e cachorros em fotos, os rótulos são as respostas certas que dizem "isso é um gato" ou "isso é um cachorro". Os rótulos ajudam o computador a aprender a associar características (como formas e cores) aos objetos corretos. É como mostrar um monte de fotos para o computador e dizer: "Isso é um gato" e "Isso é um cachorro". Com o tempo, o computador aprende a identificar esses objetos sozinho, usando os rótulos como exemplos.

sacionalidade

clusterim

variância: q

— REGRESSÃO

Aprendizado Supervisionado

— CLASSIFICAÇÃO

Features = ENTRADA (X) - Variáveis independentes

Target = SAÍDA (Y) – variável dependente

Overfitting e Underfitting

overfitting: sua função é excelente na base de treino, contudo não tem um desempenho tão bom na base de testes. Isso acontece pois, de certa forma, a sua função “aprendeu o ruído”, ou seja, o modelo se ajusta a erros e fatores aleatórios (exceções ou outliers), que não necessariamente condizem com a realidade dos dados apresentados.

Para exemplificar esse problema, vamos analisar um dataset com duas classificações (azul e vermelho). Na imagem abaixo, o modelo ideal é representado pela linha preta, que separa os pontos vermelhos dos azuis. Observe que esse modelo erra algumas vezes. No entanto, esses erros ocorrem por causa de ruído. A linha verde, que não comete nenhum erro, é muito mais complexa: ela se ajustou ao ruído. Se obtivéssemos pontos novos (que não estavam no dataset), o modelo verde provavelmente cometeria erros nesses pontos.

Em contrapartida, se o seu modelo for muito simples, você pode acabar com um caso de *underfitting*, ou seja, sua função não captura corretamente a complexidade dos dados apresentados. Uma linha reta, por exemplo, nem sempre consegue separar um conjunto de pontos em duas categorias.