Estratégia Concursos

Aula 00

BACEN (Analista - Área 1 - Tecnologia da Informação) Passo Estratégico de Conhecimentos Específicos

Autor:
Fernando Pedrosa Lopes

11 de Janeiro de 2025

MACHINE LEARNING

Sumário

CONTEÚDO

Aprendizado de Máquina. Conceitos fundamentais. Aplicações. Viés e variância. Underfitting. Overfitting. Modelos de aprendizagem de máquina. Aprendizagem supervisionada. Aprendizagem não supervisionada. Aprendizagem por reforço.

ANÁLISE ESTATÍSTICA

Inicialmente, convém destacar o percentual de incidência do assunto, dentro da disciplina **Banco de Dados e Business Intelligence** em concursos/cargos similares. Quanto maior o percentual de cobrança de um dado assunto, maior sua importância.

Obs.: *um mesmo assunto pode ser classificado em mais de um tópico devido à multidisciplinaridade de conteúdo*.

Assunto	Relevância na disciplina em concursos similares
SQL	**21.6 %**
BI (Business Intelligence)	**9.0 %**
DW - Data Warehouse	**7.2 %**
SQL Server	**7.2 %**
Oracle	**6.3 %**
Banco de Dados Multidimensionais	**5.4 %**
Data Mining	**5.4 %**
Administração de banco de dados	**3.6 %**
Banco de Dados	**2.7 %**
Formas normais	**2.7 %**
ETL (Extract Transform Load)	**2.7 %**
Banco de Dados Relacionais	**2.7 %**
Arquitetura de Banco de Dados	**1.8 %**
SGBD - Sistema de Gerenciamento de Banco de Dados	**1.8 %**
OLAP (On-line Analytical Processing)	**1.8 %**
Segurança	**1.8 %**
MS-Access	**1.8 %**
Modelo relacional	**1.8 %**
Metadados e Metainformação	**1.8 %**
Álgebra relacional	**0.9 %**
Banco de Dados Paralelos e Distribuídos	**0.9 %**
Gerência de Transações	**0.9 %**
Modelagem de dados	**0.9 %**
Gatilhos (Triggers)	**0.9 %**
DER - Diagrama de Entidade e Relacionamento	**0.9 %**
Visão (View)	**0.9 %**
Banco de Dados Textuais	**0.9 %**
Índices	**0.9 %**
PostgreSQL	**0.9 %**
MySQL	**0.9 %**
Big Data	**0.9 %**

GLOSSÁRIO DE TERMOS

Faremos uma lista de termos que são relevantes ao entendimento do assunto desta aula. Caso tenha alguma dúvida durante a leitura, esta seção pode lhe ajudar a esclarecer.

Aprendizado de máquina: É um subcampo da Inteligência Artificial que se concentra em desenvolver algoritmos capazes de aprender a partir de dados e realizar tarefas sem serem explicitamente programados para isso.

Viés: Refere-se à tendência de um modelo de aprendizado de máquina em fazer previsões incorretas devido a suposições erradas sobre os dados. Um modelo com alto viés é geralmente simples demais para aprender a complexidade dos dados.

Variância: Refere-se à tendência de um modelo de aprendizado de máquina em ser muito sensível aos dados de treinamento específicos, o que pode levar a previsões incorretas em novos dados. Um modelo com alta variância é geralmente muito complexo e superajustado aos dados de treinamento.

Underfitting: É um problema em que um modelo de aprendizado de máquina não consegue capturar bem a complexidade dos dados de treinamento e, como resultado, também não é capaz de fazer previsões precisas em novos dados. É geralmente causado por modelos muito simples com alto viés.

Overfitting: É um problema em que um modelo de aprendizado de máquina se ajusta demasiadamente aos dados de treinamento, capturando o "ruído" dos dados e não sendo capaz de fazer previsões precisas em novos dados. É geralmente causado por modelos muito complexos com alta variância.

Aprendizagem supervisionada: É uma técnica de aprendizado de máquina em que um modelo é treinado em um conjunto de dados rotulados, em que o objetivo é aprender a mapear entradas para saídas conhecidas e, posteriormente, fazer previsões precisas em novos dados não vistos.

Aprendizagem não supervisionada: É uma técnica de aprendizado de máquina em que um modelo é treinado em um conjunto de dados não rotulados, em que o objetivo é encontrar padrões e estruturas nos dados que possam ser úteis para análise e previsão.

Aprendizagem por reforço: É uma técnica de aprendizado de máquina em que um agente aprende a tomar decisões em um ambiente através da interação com o mesmo. O objetivo é maximizar uma recompensa ou minimizar uma penalidade ao realizar ações em um ambiente.

Paradigma simbólico: um paradigma de machine learning que utiliza representações simbólicas para modelar o conhecimento e raciocínio, como lógica e linguagem natural.

Árvores de decisão: um método de machine learning baseado no paradigma simbólico que usa uma árvore para representar possíveis decisões e suas consequências em um processo de tomada de decisão.

Rede Semântica: um método de machine learning baseado no paradigma simbólico que representa o conhecimento em uma rede de conceitos interconectados para inferir novas informações.

Paradigma estatístico: um paradigma de machine learning que se baseia em modelos probabilísticos e estatísticos para fazer previsões e inferências a partir de dados.

Rede Bayesiana: um método de machine learning baseado no paradigma estatístico que usa grafos acíclicos direcionados para representar a dependência entre variáveis aleatórias.

Paradigma baseado em exemplos: um paradigma de machine learning que se baseia em exemplos conhecidos para inferir padrões e fazer previsões sobre novos exemplos.

k-NN: um método de machine learning baseado no paradigma baseado em exemplos que classifica novos exemplos com base na classe dos k exemplos mais próximos no espaço de características.

Paradigma evolutivo: um paradigma de machine learning que se baseia em algoritmos genéticos para buscar soluções ótimas em um espaço de busca.

Algoritmo genético: um método de otimização baseado no paradigma evolutivo que simula o processo de seleção natural para encontrar soluções ótimas em um espaço de busca.

ROTEIRO DE REVISÃO

A ideia desta seção é apresentar um roteiro para que você realize uma revisão completa do assunto e, ao mesmo tempo, destacar aspectos do conteúdo que merecem atenção.

Introdução

Aprendizado de máquina (ou machine learning em inglês) é um subcampo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem que um computador aprenda a realizar tarefas específicas sem **ser explicitamente programado para isso**.

Em outras palavras, em vez de um programador escrever um código que especifica como um computador deve realizar uma tarefa, o aprendizado de máquina permite que o computador aprenda a executar a tarefa a partir de dados.

Isso é feito através do treinamento de um modelo de aprendizado de máquina em um conjunto de dados de treinamento, que consiste em exemplos de entrada e saída. O modelo usa esses exemplos para identificar padrões nos dados e, em seguida, usa esses padrões para fazer previsões ou tomar decisões sobre novos dados de entrada.

Aplicações

O aprendizado de máquina é usado em uma ampla gama de aplicações, desde reconhecimento de fala e imagem até detecção de fraudes e recomendações personalizadas. Vejamos:

Reconhecimento de voz: muitos assistentes virtuais, como a Siri e a Alexa, usam aprendizado de máquina para entender a fala dos usuários e fornecer respostas relevantes.

Reconhecimento de imagem: os aplicativos de reconhecimento facial usam o aprendizado de máquina para identificar rostos em imagens e vídeos.

Detecção de fraude: o aprendizado de máquina é usado em bancos e outras instituições financeiras para detectar atividades suspeitas e prevenir fraudes.

Medicina: a análise de dados médicos pode ajudar médicos e pesquisadores a desenvolver tratamentos mais eficazes para doenças, usando algoritmos de aprendizado de máquina para identificar padrões nos dados.

Previsão de demanda: empresas de varejo usam o aprendizado de máquina para prever a demanda de produtos e otimizar a gestão de estoques.

Carros autônomos: o aprendizado de máquina é uma das tecnologias-chave por trás dos carros autônomos, permitindo que eles "aprendam" a reconhecer objetos e tomar decisões com base em dados do ambiente.

Esses são apenas alguns exemplos, mas o aprendizado de máquina é usado em uma variedade de outras aplicações, como reconhecimento de fala, tradução automática, recomendações personalizadas, detecção de spam, entre outros.

Viés e Variância

No aprendizado de máquina, o objetivo é construir modelos que sejam precisos em prever a saída para novos dados de entrada. No entanto, quando se trabalha com modelos de aprendizado de máquina, é importante considerar dois conceitos importantes: **viés e variância**.

O **viés** se refere à tendência do modelo de aprendizado de máquina de fazer suposições simplistas sobre a relação entre os dados de entrada e a saída esperada.

Por exemplo, um modelo de regressão linear que assume que a relação entre as variáveis de entrada e de saída é linear pode ter um viés elevado se a relação real for mais complexa do que isso. Um modelo com um viés elevado tende a subestimar a complexidade dos dados e, portanto, pode ter um desempenho insuficiente na previsão de novos dados de entrada.

Por outro lado, a **variância** se refere à sensibilidade do modelo a pequenas variações nos dados de treinamento. Um modelo com alta variância pode se ajustar excessivamente (ou "memorizar") os dados de treinamento e, portanto, pode ter um desempenho insuficiente na previsão de novos dados de entrada. Em outras palavras, um modelo com alta variância pode ser muito complexo e, portanto, pode ser sensível a ruídos nos dados de treinamento.

O objetivo é encontrar um equilíbrio entre o viés e a variância para criar um modelo de aprendizado de máquina que seja preciso e generalize bem para novos dados de entrada. Isso é conhecido como o trade-off viés-variância. A escolha do algoritmo de aprendizado de máquina e a configuração dos seus parâmetros afetam o viés e a variância do modelo, e é importante ajustar esses parâmetros para encontrar o equilíbrio certo.

Underfitting e Overfitting

Underfitting e overfitting são problemas comuns no aprendizado de máquina, que afetam o desempenho do modelo na previsão de novos dados.

Underfitting

O **underfitting** ocorre quando um modelo de aprendizado de máquina não é capaz de capturar a complexidade dos dados de treinamento, resultando em um desempenho insuficiente na previsão de novos dados. Isso geralmente ocorre quando o modelo é **muito simples** para a tarefa em questão ou quando o conjunto de treinamento é muito pequeno. Um exemplo de underfitting seria um modelo de regressão linear que é incapaz de capturar a relação não-linear entre as variáveis de entrada e saída. Nesse caso, o modelo teria um desempenho insuficiente na previsão de novos dados de entrada.

Um exemplo de underfitting seria o treinamento de um modelo de regressão linear para prever o preço de casas com base em uma única variável de entrada, como a área da casa. Nesse caso, o modelo seria muito simples para capturar a complexidade dos dados, uma vez que há várias outras variáveis, como localização, número de quartos, número de banheiros etc., que podem influenciar no preço da casa.

Overfitting

Já o **overfitting** ocorre quando um modelo de aprendizado de máquina se ajusta demais aos dados de treinamento, resultando em um desempenho insuficiente na previsão de novos dados. Isso geralmente ocorre quando o modelo é muito complexo ou quando o conjunto de treinamento é muito pequeno em relação à complexidade do modelo.

Um exemplo de overfitting seria o treinamento de um modelo de árvore de decisão para prever se um paciente tem uma doença cardíaca com base em várias variáveis de entrada, como idade, histórico familiar, pressão arterial, níveis de colesterol etc.

Se o modelo de árvore de decisão for muito complexo ou o conjunto de dados de treinamento for muito pequeno, o modelo pode ajustar-se perfeitamente aos dados de treinamento. Isso significa que o modelo seria capaz de prever corretamente a presença ou ausência da doença para cada paciente nos dados de treinamento.

No entanto, quando o modelo é aplicado a novos dados de entrada, ele pode ter um desempenho insuficiente na previsão da doença cardíaca. Isso ocorre porque o modelo pode ter "decorado" as características específicas dos pacientes no conjunto de dados de treinamento, em vez de aprender a relação subjacente entre as variáveis de entrada e a saída.

Esse é um exemplo clássico de overfitting, onde o modelo se ajusta demais aos dados de treinamento e não consegue generalizar bem para novos dados de entrada. Para evitar o overfitting, é importante ajustar adequadamente os parâmetros do modelo e utilizar técnicas de validação cruzada e conjunto de dados de teste para avaliar o desempenho do modelo na previsão de novos dados.

Para evitar o underfitting e o overfitting, é importante escolher um modelo apropriado para a tarefa em questão e ajustar seus parâmetros adequadamente. É importante também ter um conjunto de dados de treinamento suficientemente grande e representativo e validar o modelo em um conjunto de dados de teste para avaliar seu desempenho na previsão de novos dados.

Modelos (modos) de Aprendizagem de Máquina

Modelos de aprendizado de máquina são algoritmos treinados em dados para realizar uma tarefa específica, como classificação, regressão, clusterização etc. Esses modelos aprendem

a partir dos dados de entrada, identificando padrões e relacionamentos entre as variáveis, e usam essas informações para fazer previsões ou tomar decisões sobre novos dados de entrada.

De forma geral, podemos falar em três tipos de modelos de aprendizagem de máquina: aprendizagem supervisionada, aprendizagem não supervisionada e por reforço.

Aprendizagem Supervisionada

Aprendizagem supervisionada é um tipo de aprendizado de máquina em que o modelo é treinado usando um conjunto de dados rotulados. Isso significa que o conjunto de dados de treinamento inclui pares de entrada e saída correspondentes, onde a saída é **conhecida** e fornecida ao modelo durante o treinamento.

O objetivo da aprendizagem supervisionada é ensinar o modelo a prever a saída correta para novas entradas que ele nunca viu antes. Para isso, o modelo é ajustado aos dados de treinamento, tentando encontrar a melhor função que relaciona as entradas aos rótulos de saída.

Por exemplo, um modelo de aprendizagem supervisionada poderia ser treinado para classificar e-mails como spam ou não spam. O conjunto de dados de treinamento incluiria vários e-mails, alguns marcados como spam e outros não. O modelo aprenderia a partir desses exemplos, tentando identificar padrões e características que distinguem os e-mails de spam dos e-mails legítimos.

Em seguida, o modelo seria testado em um conjunto de dados de teste separado, contendo novos e-mails que nunca foram vistos antes. O modelo tentaria prever se cada e-mail é spam ou não com base nas informações que ele aprendeu durante o treinamento.

A aprendizagem supervisionada é amplamente utilizada em várias aplicações de aprendizado de máquina, como reconhecimento de voz, classificação de imagens, detecção de fraudes, entre outras. Ela é particularmente útil quando se tem uma grande quantidade de dados rotulados disponíveis para treinar o modelo e prever a saída para novos dados de entrada.

Algoritmo: Random Forest

O Random Forest é um algoritmo de aprendizagem supervisionada que usa múltiplas árvores de decisão para classificar ou prever valores. Cada árvore é construída a partir de uma amostra aleatória dos dados de treinamento e é baseada em um subconjunto aleatório das variáveis explicativas.

O algoritmo começa dividindo aleatoriamente os dados de treinamento em várias árvores de decisão e cada árvore é treinada em um subconjunto diferente dos dados. Durante a construção de cada árvore, um subconjunto aleatório de variáveis explicativas é selecionado

para cada divisão de nó. Isso garante que cada árvore tenha uma perspectiva única sobre os dados e ajuda a evitar o overfitting.

Uma vez que todas as árvores são construídas, o algoritmo faz uma previsão combinando as previsões de cada árvore, usando votação majoritária para classificação ou média para regressão.

Por exemplo, suponha que temos um conjunto de dados fictício com informações sobre animais e queremos prever se um animal é um cachorro, um gato ou um pássaro com base em suas características. O conjunto de dados possui as seguintes variáveis explicativas: idade, peso, altura, comprimento da cauda e cor do pelo.

Usando o algoritmo Random Forest, podemos treinar várias árvores de decisão em subconjuntos aleatórios do conjunto de dados e variáveis explicativas. Em seguida, podemos usar a votação majoritária das previsões das árvores para determinar a classe final do animal.

Neste cenário, suponha que temos uma nova observação de um animal com 2 anos de idade, pesando 4 kg, com altura de 30 cm, comprimento de cauda de 10 cm e pelo marrom. Podemos usar o modelo Random Forest treinado para prever a classe do animal, que seria "cachorro" se a maioria das árvores classificou como cachorro, "gato" se a maioria classificou como gato ou "pássaro" se a maioria classificou como pássaro.

Algoritmo: Support Vector Machines

O algoritmo Support Vector Machines (SVM) é um modelo de aprendizagem de máquina supervisionado utilizado para resolver problemas de classificação e regressão. O SVM funciona encontrando o hiperplano que maximiza a margem entre as classes, ou seja, a distância entre o hiperplano e as observações mais próximas de cada classe.

No contexto da classificação, o SVM é capaz de lidar com dados linearmente separáveis e não separáveis, utilizando diferentes tipos de funções de kernel. Algumas das funções de kernel comuns são a linear, a polinomial e a RBF (Radial Basis Function).

A seguir, um exemplo (em Python) de como utilizar o algoritmo SVM para classificar dados fictícios:

Suponha que você queira classificar frutas em duas classes: maçãs e laranjas, com base em duas características: a cor e o tamanho. Para isso, você coletou dados de várias frutas e criou um conjunto de dados com as seguintes informações:

Fruta	Cor (RGB)	Tamanho (cm)	Classe
1	(255,0,0)	7	Maçã
2	(255,50,0)	6	Maçã

3	(255,100,0)	5	Maçã
4	(255,150,0)	4	Maçã
5	(255,200,0)	3	Laranja
6	(255,250,0)	2	Laranja
7	(255,255,0)	1	Laranja

O próximo passo seria treinar o modelo SVM com esse conjunto de dados. Primeiro, é necessário separar os dados em dois conjuntos: um conjunto de treinamento e um conjunto de teste. Para este exemplo, vamos separar aleatoriamente 70% dos dados para treinamento e 30% para teste:

```
from sklearn.model_selection import train_test_split



X = df.iloc[:, :-1].values

y = df.iloc[:, -1].values



X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=0)
```

A seguir, criamos o modelo SVM e ajustamos os dados de treinamento:

```
from sklearn.svm import SVC



classifier = SVC(kernel='linear', random_state=0)

classifier.fit(X_train, y_train)
```

Por fim, podemos fazer previsões utilizando o modelo treinado e avaliar sua acurácia:

```
y_pred = classifier.predict(X_test)



from sklearn.metrics import accuracy_score



print(accuracy_score(y_test, y_pred))
```

ara este exemplo, o resultado da acurácia foi de 100%, o que indica que o modelo SVM foi capaz de classificar as frutas corretamente com base nas características de cor e tamanho.

Aprendizagem Não Supervisionada

A aprendizagem não supervisionada é um tipo de aprendizado de máquina em que o modelo é treinado em um conjunto de dados sem rótulos ou sem nenhuma informação prévia sobre a saída esperada. O objetivo é encontrar padrões, estruturas ou relações nos dados sem o auxílio de informações externas.

Diferente da aprendizagem supervisionada, onde o modelo recebe pares de entrada e saída, na aprendizagem não supervisionada, o modelo recebe apenas as entradas e deve descobrir padrões ou estruturas que existem no conjunto de dados.

Por exemplo, um modelo de clusterização é um tipo comum de aprendizagem não supervisionada. Ele agrupa pontos de dados em clusters baseados em suas similaridades. O modelo analisa as características dos pontos de dados e encontra grupos naturais de pontos que compartilham características semelhantes.

Outro exemplo é a redução de dimensionalidade, que é usada para encontrar representações mais simples dos dados de entrada, eliminando características menos relevantes ou redundantes. O modelo reduz a dimensionalidade do conjunto de dados, criando uma representação mais compacta dos dados de entrada.

A aprendizagem não supervisionada é amplamente utilizada em várias aplicações de aprendizado de máquina, como análise de dados, detecção de anomalias, segmentação de mercado, entre outras. Ela é particularmente útil quando não há informações rotuladas disponíveis ou quando não se sabe o que procurar nos dados.

Aprendizagem por Reforço

A aprendizagem por reforço é um tipo de aprendizado de máquina em que um agente aprende a realizar uma tarefa através de tentativa e erro, interagindo com um ambiente dinâmico e recebendo feedback através de **recompensas ou punições**.

O agente recebe um estado do ambiente como entrada e, em seguida, executa uma ação que afeta o ambiente. O ambiente responde ao agente com um novo estado e uma recompensa que pode ser positiva ou negativa, dependendo do desempenho do agente. O objetivo do agente é maximizar a recompensa ao longo do tempo, aprendendo a tomar decisões que levam a ações mais bem-sucedidas.

Por exemplo, um agente de aprendizado por reforço pode ser treinado para jogar um jogo de tabuleiro. O agente recebe o estado atual do tabuleiro como entrada e decide a próxima

jogada com base em suas experiências anteriores. O ambiente fornece uma recompensa positiva se o agente vencer o jogo e uma recompensa negativa se ele perder.

Ao longo do tempo, o agente aprende a jogar melhor, levando em consideração as recompensas e punições recebidas. Ele ajusta suas estratégias de acordo com o desempenho passado, tentando maximizar a recompensa total acumulada.

A aprendizagem por reforço é amplamente utilizada em várias aplicações de aprendizado de máquina, como jogos, robótica, navegação, entre outras. Ela é particularmente útil quando não há dados rotulados disponíveis ou quando é difícil especificar uma função de custo precisa que guie o aprendizado do agente.

Tabela Comparativa - Modelos de Aprendizagem

Segue uma tabela comparativa entre aprendizagem supervisionada, não supervisionada e por reforço:

	Aprendizagem Supervisionada	Aprendizagem Não Supervisionada	Aprendizagem por Reforço
Exemplos	**Entradas e Saídas Rotuladas**	**Apenas Entradas**	**Estado e Recompensas**
Objetivo Geral	**Predição de Saídas**	**Descoberta de Padrões**	**Maximização de Recompensas**
Algoritmos Comuns	**Árvores de Decisão, Regressão Linear, Redes Neurais**	**Clusterização, Redução de Dimensionalidade**	**Q-Learning, SARSA, Deep Reinforcement Learning**
Tipo de Feedback	**Feedback Supervisionado**	**Sem Feedback Direto**	**Feedback por Recompensas**
Exemplos de Aplicação	**Reconhecimento de Imagem, Previsão de Preços, Classificação de Emails**	**Agrupamento de Dados, Detecção de Anomalias, Segmentação de Mercado**	**Jogos, Navegação Robótica, Controle de Processos**

Paradigmas de Aprendizagem de Máquina

Os paradigmas de machine learning são abordagens teóricas e práticas que guiam o desenvolvimento de algoritmos de aprendizado de máquina. Os quatro principais paradigmas são os paradigmas: simbólico, estatístico, baseado em exemplos e evolutivo.

Paradigma Simbólico

O paradigma simbólico, também conhecido como conhecimento baseado em regras, tem como base a utilização de regras lógicas e simbólicas para modelar o conhecimento e raciocínio. Esse paradigma é bastante utilizado em sistemas especialistas, que são sistemas que possuem um conhecimento específico em determinado domínio.

Em um sistema baseado em regras, as regras são escritas manualmente por especialistas do domínio e podem ser utilizadas para tomar decisões, fazer previsões e gerar recomendações. Essas regras podem ser organizadas em uma hierarquia de conceitos e podem ser refinadas à medida que o sistema recebe novas informações ou feedback dos usuários.

Um exemplo de aplicação do paradigma simbólico é um sistema de diagnóstico médico. O sistema pode ser construído com base em um conjunto de regras que descrevem as relações entre sintomas e doenças. Por exemplo, se o paciente apresenta febre, dor de cabeça e vômito, o sistema pode concluir que ele tem meningite. As regras podem ser refinadas com base em dados de pacientes e feedback de especialistas médicos.

Outro exemplo é um sistema de recomendação de filmes. Nesse caso, o sistema pode ser construído com base em regras que descrevem as preferências dos usuários em relação a gêneros de filmes, atores e diretores. O sistema pode utilizar essas regras para recomendar filmes que sejam mais adequados aos gostos dos usuários.

Uma das vantagens do paradigma simbólico é que ele permite uma explicabilidade e interpretabilidade dos resultados gerados pelo sistema. No entanto, a construção de sistemas baseados em regras pode ser trabalhosa e requer a participação de especialistas do domínio na definição das regras. Além disso, esses sistemas podem não ser tão eficazes em lidar com dados não estruturados ou situações em que as regras não são bem definidas.

Técnica: Árvores de Decisão

No paradigma simbólico, a árvore de decisão é um método utilizado para classificação e previsão de valores a partir de regras definidas de maneira explícita. Essas regras são representadas em forma de uma árvore, em que cada nó interno representa uma condição, cada aresta representa um resultado possível para essa condição, e cada folha representa o resultado final da árvore.

O método de árvore de decisão funciona criando uma série de perguntas com base nos atributos dos dados de treinamento e dividindo os dados de acordo com as respostas a essas perguntas. Isso é feito até que todos os dados em cada subconjunto sejam da mesma classe ou tenham um valor numérico previsto, o que resulta em uma árvore de decisão completa.

Por exemplo, suponha que temos um conjunto de dados de animais e queremos criar uma árvore de decisão para classificá-los como mamíferos ou aves. Podemos começar com uma pergunta como "o animal tem penas?" Se a resposta for sim, podemos classificá-lo como ave e parar a árvore. Caso contrário, podemos fazer outra pergunta, como "o animal é quadrúpede?" e assim por diante, até chegarmos a uma classificação final.

As árvores de decisão podem ser usadas em muitas outras áreas, como previsão de preços de imóveis, diagnóstico médico, previsão de falhas em equipamentos, entre outros. A principal vantagem deste método é que a árvore gerada é fácil de interpretar e entender, permitindo que sejam tomadas decisões com base nos resultados obtidos.

Técnica: Rede Semântica

No paradigma simbólico, um dos métodos utilizados em machine learning é a Rede Semântica, que consiste em representar o conhecimento de um domínio através de um grafo, onde os nós são conceitos e as arestas são as relações entre eles.

Essa técnica é útil em situações em que é necessário modelar um conhecimento hierárquico, como por exemplo, em sistemas de recomendação ou na resolução de problemas de raciocínio lógico.

Um exemplo de aplicação de rede semântica é em um sistema de recomendação de filmes. Nesse caso, pode-se construir uma rede em que os nós representam os filmes e as arestas indicam as relações entre eles, como por exemplo, filmes do mesmo gênero, com os mesmos atores, diretor ou roteirista. Dessa forma, quando um usuário assiste a um filme, é possível recomendar outros filmes que estejam relacionados na rede semântica.

Outro exemplo é na resolução de problemas de raciocínio lógico. Por exemplo, em um problema em que é necessário determinar qual objeto pertence a qual pessoa, pode-se construir uma rede em que os nós representam as pessoas e os objetos, e as arestas indicam as possíveis relações entre eles. Com base nas informações fornecidas no problema, é possível inferir as relações entre os nós e encontrar a solução para o problema.

Paradigma Estatístico

O paradigma estatístico do aprendizado de máquina é baseado em métodos estatísticos e probabilísticos para construir modelos de aprendizado a partir de dados. A ideia principal é usar os dados disponíveis para estimar as probabilidades das diferentes classes ou valores das variáveis, e então usar essas estimativas para fazer previsões em novos dados. Em outras palavras, o aprendizado é visto como uma tarefa de inferência estatística.

Existem diversas técnicas de aprendizado de máquina baseadas no paradigma estatístico, tais como:

- **Regressão**: é um método usado para prever valores contínuos a partir de dados históricos. Por exemplo, pode-se usar regressão para prever o preço de uma casa baseado em informações como o número de quartos, a área construída e a localização.
- **Classificação**: é um método usado para classificar objetos em diferentes categorias a partir de dados históricos. Por exemplo, pode-se usar classificação para identificar se um email é spam ou não, baseado no conteúdo do email.
- **Análise de componentes principais (PCA)**: é um método usado para reduzir a dimensão dos dados, preservando ao mesmo tempo a maior parte da informação relevante. Isso é útil

quando se tem muitas variáveis que estão altamente correlacionadas entre si. PCA pode ser usado para análise de dados em diversas áreas, como biologia, finanças e engenharia.

- **Modelos de mistura**: é um método usado para modelar dados que são gerados por uma combinação de diferentes distribuições estatísticas. Por exemplo, pode-se usar modelos de mistura para modelar dados de sensoriamento remoto que contenham diferentes tipos de terrenos.
- **Redes Bayesianas**: é um modelo probabilístico que permite representar e atualizar incertezas em sistemas complexos. Esses modelos são usados em diversas áreas, como diagnóstico médico, análise de riscos, engenharia e finanças.

O paradigma estatístico é amplamente utilizado em diversas áreas, como finanças, marketing, ciência de dados e bioinformática. Alguns exemplos de aplicações práticas incluem a previsão de preços de ações, a análise de dados de exames médicos para identificar doenças, a previsão de padrões climáticos e a análise de dados de tráfego para melhorar o planejamento urbano.

Método: Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é um método de redução de dimensionalidade utilizado para encontrar um conjunto de variáveis não correlacionadas que expliquem a maior variação em um conjunto de dados. É uma técnica muito utilizada em aprendizado de máquina e análise de dados para facilitar a visualização e o processamento de dados com muitas variáveis.

O processo de PCA envolve a decomposição da matriz de covariância dos dados originais em seus componentes principais, que são as novas variáveis não correlacionadas que explicam a maior variação nos dados. Esses componentes principais são ordenados em ordem decrescente de variação explicada e podem ser selecionados para criar um subconjunto menor e mais representativo das variáveis originais.

Por exemplo, suponha que temos um conjunto de dados com quatro variáveis (altura, peso, idade e gênero) para cada uma das 1000 pessoas. Podemos aplicar a PCA para reduzir a dimensionalidade do conjunto de dados para duas novas variáveis, que seriam combinações lineares das quatro variáveis originais. Essas novas variáveis seriam escolhidas para maximizar a variação explicada nos dados originais. Depois de reduzir a dimensionalidade do conjunto de dados, podemos visualizá-lo em um gráfico de dispersão bidimensional com os novos eixos principais representando as novas variáveis não correlacionadas.

Outro exemplo seria a análise de dados genômicos. Os dados genômicos contêm informações sobre milhares de genes e seus níveis de expressão em diferentes amostras biológicas. A PCA pode ser usada para reduzir a dimensionalidade desses dados e identificar os principais padrões de variação na expressão gênica. Isso pode ser útil para identificar genes que estão envolvidos em processos biológicos específicos ou que são importantes para distinguir diferentes tipos de amostras.

Paradigma Baseado em Exemplos

O paradigma baseado em exemplos, também conhecido como aprendizado baseado em instâncias, é uma abordagem de aprendizado de máquina que utiliza exemplos de treinamento para fazer previsões ou classificações em novos dados. Em vez de aprender um modelo geral, como no paradigma simbólico, o aprendizado baseado em exemplos se concentra em encontrar padrões específicos nos dados de treinamento para fazer previsões em novos dados.

O método mais comum de aprendizado baseado em exemplos é o **k-Nearest Neighbors (k-NN)**, que é utilizado para classificação e regressão. No k-NN, o algoritmo utiliza os k exemplos mais próximos do novo exemplo para prever sua classe ou valor de saída. A distância entre os exemplos é geralmente calculada utilizando a distância euclidiana.

Um exemplo de aplicação do k-NN é a classificação de imagens digitais de dígitos manuscritos. Nesse caso, cada imagem é representada por um conjunto de pixels e uma classe correspondente que indica qual dígito foi escrito. Ao receber uma nova imagem, o algoritmo calcula as distâncias entre seus pixels e os das imagens de treinamento e seleciona as k imagens mais próximas. A classe mais frequente entre essas k imagens é escolhida como a classe prevista para a nova imagem.

Outro exemplo é a previsão do preço de uma casa com base em suas características, como número de quartos, banheiros e tamanho do terreno. Nesse caso, o conjunto de treinamento consiste em informações de casas vendidas anteriormente e seus respectivos preços. Ao receber as características de uma nova casa, o algoritmo calcula as distâncias entre seus atributos e os das casas do conjunto de treinamento e seleciona as k casas mais próximas. O valor médio dos preços dessas k casas é usado como o valor previsto para a nova casa.

Paradigma Evolutivo

O paradigma evolutivo de machine learning é baseado na ideia de que os algoritmos de aprendizagem podem ser inspirados por processos biológicos de evolução e seleção natural. Isso significa que o processo de aprendizagem se assemelha a uma população de indivíduos, que passam por processos de seleção, mutação e recombinação genética para melhorar seu desempenho em uma tarefa específica.

Um exemplo de algoritmo de aprendizagem evolutiva é o algoritmo genético. Nesse método, uma população inicial de indivíduos é criada com um conjunto aleatório de características. Cada indivíduo é avaliado em relação ao seu desempenho na tarefa em questão e aqueles que tiverem o melhor desempenho são selecionados para se reproduzir, gerando uma nova geração de indivíduos com características semelhantes às dos pais, mas com pequenas mutações.

Outro exemplo de algoritmo de aprendizagem evolutiva é a programação genética, que é usada para criar programas de computador. Nesse método, uma população de programas é criada com estruturas e regras aleatórias. Cada programa é avaliado em relação ao seu desempenho em uma tarefa específica e aqueles que tiverem o melhor desempenho são

selecionados para se reproduzir, gerando uma nova geração de programas com pequenas modificações.

O paradigma evolutivo de machine learning tem sido usado em diversas aplicações, como em otimização de sistemas de controle, em projetos de robótica, em jogos e em análise de dados. Ele tem se mostrado especialmente útil em problemas de otimização complexos, em que as soluções analíticas não são possíveis ou eficientes.

Técnica: Algoritmo Genético

Um exemplo de algoritmo genético pode ser aplicado para encontrar a melhor solução para um problema de otimização, como por exemplo a programação de uma malha de produção de uma fábrica.

Suponha que temos uma fábrica que precisa produzir três tipos de produtos (A, B e C) com determinadas especificações. Cada produto deve passar por uma sequência de processos de produção (P1, P2 e P3) em uma determinada ordem e com diferentes tempos de produção. O objetivo é encontrar a melhor programação de produção, que minimize o tempo total de produção e atenda às especificações de cada produto.

Para resolver esse problema com um algoritmo genético, podemos representar cada solução (programação de produção) como um cromossomo, que é uma cadeia de genes (neste caso, a ordem de produção de cada produto). Podemos codificar cada gene com um número inteiro que representa o tipo de produto e a ordem de produção.

O algoritmo começa com uma população inicial de soluções aleatórias, representadas por cromossomos. Em cada geração, os cromossomos são avaliados de acordo com um critério de aptidão (neste caso, o tempo total de produção), e os mais aptos são selecionados para a reprodução. A reprodução é feita por meio de cruzamento (combinação de dois cromossomos para gerar um novo) e mutação (alteração aleatória de um gene em um cromossomo).

O processo de seleção, cruzamento e mutação é repetido até que uma solução satisfatória seja encontrada, ou até que um número máximo de gerações seja alcançado. A solução final é a programação de produção que minimiza o tempo total de produção.

Esse é apenas um exemplo simples de como um algoritmo genético pode ser aplicado para resolver problemas de otimização em um contexto de produção. Existem muitas outras aplicações possíveis em diferentes áreas, como engenharia, finanças, medicina, entre outras.

Tabela Comparativa - Paradigmas de Machine Learning

Segue uma tabela comparativa entre os diferentes tipos de paradigmas de Machine Learning:

Paradigma	Principais características	Exemplos de algoritmos

Simbólico	**Utiliza representações explícitas de conhecimento em forma de regras, árvores, grafos, etc.**	**Árvore de Decisão, Rede Semântica**
Estatístico	**Baseado em modelos probabilísticos e teoria estatística. Encontra padrões em dados através de inferências estatísticas.**	**Regressão Linear, Análise Discriminante, Análise de Componentes Principais**
Baseado em exemplos	**Aprendizado por exemplos. Baseia-se na semelhança entre os dados. Não necessita de modelo explícito.**	**k-NN, SVM**
Evolutivo	**Inspirado em processos biológicos. Usa mecanismos de seleção, cruzamento e mutação para encontrar soluções ótimas.**	**Algoritmo Genético, Programação Genética**

APOSTA ESTRATÉGICA

A ideia desta seção é apresentar os pontos do conteúdo que mais possuem chances de serem cobrados em prova, considerando o histórico de questões da banca em provas de nível semelhante à nossa, bem como as inovações no conteúdo, na legislação e nos entendimentos doutrinários e jurisprudenciais[1].

Overfitting e underfitting são dois problemas comuns em modelos de machine learning. O overfitting ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, capturando não apenas o padrão subjacente, mas também o ruído presente nos dados. Isso pode resultar em um desempenho ruim do modelo ao lidar com novos dados, pois ele não consegue generalizar adequadamente.

Por outro lado, o underfitting acontece quando um modelo é muito simples para capturar a complexidade dos dados. Isso ocorre quando o modelo não consegue ajustar-se bem nem aos dados de treinamento, indicando que não está capturando o padrão subjacente. Como resultado, o modelo tem um desempenho fraco tanto nos dados de treinamento quanto nos dados de teste, apresentando uma alta taxa de erro. Encontrar o equilíbrio entre esses dois extremos é essencial para construir modelos de machine learning com bom desempenho e capacidade de generalização.

[1] Vale deixar claro que nem sempre será possível realizar uma aposta estratégica para um determinado assunto, considerando que às vezes não é viável identificar os pontos mais prováveis de serem cobrados a partir de critérios objetivos ou minimamente razoáveis.

QUESTÕES ESTRATÉGICAS

Nesta seção, apresentamos e comentamos uma amostra de questões objetivas selecionadas estrategicamente: são questões com nível de dificuldade semelhante ao que você deve esperar para a sua prova e que, em conjunto, abordam os principais pontos do assunto.

A ideia, aqui, não é que você fixe o conteúdo por meio de uma bateria extensa de questões, mas que você faça uma boa revisão global do assunto a partir de, relativamente, poucas questões.

1. (CESPE / TCE-MG – 2018) Em machine learning, a categoria de aprendizagem por reforço identifica as tarefas em que:

a) um software interage com um ambiente dinâmico, como, por exemplo, veículos autônomos.
b) as etiquetas de classificação não sejam fornecidas ao algoritmo, de modo a deixá-lo livre para entender as entradas recebidas.
c) o aprendizado pode ser um objetivo em si mesmo ou um meio para se atingir um fim.
d) o objetivo seja aprender um conjunto de regras generalistas para converter as entradas em saídas predefinidas.
e) são apresentados ao computador exemplos de entradas e saídas desejadas, fornecidas por um orientador.

Comentários:

(a) Correto! Veículos autônomos seguem uma rota de trânsito em um ambiente dinâmico até o seu destino final. O carro precisa escolher o melhor caminho, evitando colisões e infrações de trânsito, sendo penalizado quando não cumpre seu papel corretamente e recompensado quando o faz. Essa é uma utilização típica de aprendizado por reforço; (b) Errado, esse seria o comportamento típico de uma tarefa não supervisionada; (c) Errado, esse seria o comportamento típico do aprendizado não supervisionado, em que o aprendizado pode ser um objetivo em si mesmo (descobrir novos padrões nos dados) ou um meio para atingir um fim; (d) Errado, esse seria o comportamento típico do aprendizado supervisionado, em que se busca aprender uma regra geral que mapeia entradas de dados em saídas de dados; (e) Errado, esse seria o comportamento típico do aprendizado supervisionado, em que são apresentadas ao computador exemplos de entradas de dados e suas respectivas saídas desejadas por um supervisor/professor/orientador.

Gabarito: A

---

2. (CESPE / ANP – 2022) O algoritmo random forest é um algoritmo de aprendizado de máquina supervisionado em que se agrupam os resultados de várias árvores de decisão de cada nó para se obter uma conclusão própria e aumentar a precisão do modelo, não sendo o referido algoritmo adequado para grandes conjuntos de dados.

Comentários:

O algoritmo Random Forest é um algoritmo de aprendizado de máquina? Sim! Ele é supervisionado? Sim, trata-se de um algoritmo de classificação. Ele agrupa os resultados de várias árvores de decisão de cada nó para se obter uma conclusão própria e aumentar a precisão do modelo? Sim, por meio da utilização de diversas árvores aleatórias. Não é adequado para grandes conjuntos de dados? Opa, ele é ótimo para lidar com grandes conjuntos de dados.

Gabarito: E

---

3. (CESPE / ANP – 2022) As aplicações em inteligência artificial são definidas como uma subárea da área de aprendizagem de máquina (machine learning).

Comentários:

Opa! É o inverso: as aplicações de aprendizagem de máquina (machine learning) são definidas como uma subárea da área de inteligência artificial.

Gabarito: E

---

4. (CESPE / ANP – 2022) A técnica de redução de dimensionalidade (PCA) permite transformar dados que inicialmente pertencem a um espaço de dimensão n em um espaço de dimensão m, em que m < n, sendo utilizada, por exemplo, para reduzir a dimensionalidade de certo conjunto de dados através do descarte de características não úteis e que ainda permita realizar o reconhecimento de padrões.

Comentários:

Perfeito! A redução de dimensionalidade busca justamente reduzir as dimensões eliminando características que não serão úteis para o reconhecimento de padrões. Por meio da seleção de atributos, é possível eliminar atributos irrelevantes ou redundantes.

Gabarito: C

---

5. (CESPE / TCE-MG – 2018) Em machine learning, a categoria de aprendizagem por reforço identifica as tarefas em que:

a) um software interage com um ambiente dinâmico, como, por exemplo, veículos autônomos.
b) as etiquetas de classificação não sejam fornecidas ao algoritmo, de modo a deixá-lo livre

para entender as entradas recebidas.
c) o aprendizado pode ser um objetivo em si mesmo ou um meio para se atingir um fim.
d) o objetivo seja aprender um conjunto de regras generalistas para converter as entradas em saídas predefinidas.
e) são apresentados ao computador exemplos de entradas e saídas desejadas, fornecidas por um orientador.

Comentários:

(a) Correto! Veículos autônomos seguem uma rota de trânsito em um ambiente dinâmico até o seu destino final. O carro precisa escolher o melhor caminho, evitando colisões e infrações de trânsito, sendo penalizado quando não cumpre seu papel corretamente e recompensado quando o faz. Essa é uma utilização típica de aprendizado por reforço; (b) Errado, esse seria o comportamento típico de uma tarefa não supervisionada; (c) Errado, esse seria o comportamento típico do aprendizado não supervisionado, em que o aprendizado pode ser um objetivo em si mesmo (descobrir novos padrões nos dados) ou um meio para atingir um fim; (d) Errado, esse seria o comportamento típico do aprendizado supervisionado, em que se busca aprender uma regra geral que mapeia entradas de dados em saídas de dados; (e) Errado, esse seria o comportamento típico do aprendizado supervisionado, em que são apresentadas ao computador exemplos de entradas de dados e suas respectivas saídas desejadas por um supervisor/professor/orientador.

Gabarito: A

QUESTIONÁRIO DE REVISÃO E APERFEIÇOAMENTO

A ideia do questionário é elevar o nível da sua compreensão no assunto e, ao mesmo tempo, proporcionar uma outra forma de revisão de pontos importantes do conteúdo, a partir de perguntas que exigem respostas subjetivas.

São questões um pouco mais desafiadoras, porque a redação de seu enunciado não ajuda na sua resolução, como ocorre nas clássicas questões objetivas.

O objetivo é que você realize uma auto explicação mental de alguns pontos do conteúdo, para consolidar melhor o que aprendeu ;)

Além disso, as questões objetivas, em regra, abordam pontos isolados de um dado assunto. Assim, ao resolver várias questões objetivas, o candidato acaba memorizando pontos isolados do conteúdo, mas muitas vezes acaba não entendendo como esses pontos se conectam.

Assim, no questionário, buscaremos trazer também situações que ajudem você a conectar melhor os diversos pontos do conteúdo, na medida do possível.

É importante frisar que não estamos adentrando em um nível de profundidade maior que o exigido na sua prova, mas apenas permitindo que você compreenda melhor o assunto de modo a facilitar a resolução de questões objetivas típicas de concursos, ok?

Nosso compromisso é proporcionar a você uma revisão de alto nível!

Vamos ao nosso questionário:

Perguntas

1. O que é aprendizado de máquina?

2. O que é viés no aprendizado de máquina?

3. O que é variância no aprendizado de máquina?

4. O que é underfitting?

5. O que é overfitting?

6. O que é aprendizagem supervisionada?

7. O que é aprendizagem não supervisionada?

8. O que é aprendizagem por reforço?

9. O que é o paradigma simbólico em machine learning?

10. Como funciona o método de Árvore de Decisão no contexto do paradigma simbólico?

11. O que é Rede Semântica no contexto do paradigma simbólico?

12. Qual é o princípio básico do paradigma estatístico em machine learning?

13. O que é uma Rede Bayesiana?

14. O que é o paradigma baseado em exemplos em machine learning?

15. O que é o k-NN e como ele funciona?

16. O que é o paradigma evolutivo em machine learning?

17. O que é um algoritmo genético?

18. O que é seleção de recursos em machine learning?

19. O que é o bias-variance trade-off em machine learning?

20. Quais são os principais desafios enfrentados em problemas de aprendizado de máquina?

21. Quais são as etapas básicas para construir um modelo de machine learning?

22. Quais são os principais benefícios do uso de técnicas de aprendizado de máquina?

Perguntas e Respostas

1. O que é aprendizado de máquina?

Resposta: Aprendizado de máquina é um subcampo da inteligência artificial que se preocupa com o desenvolvimento de algoritmos e modelos que permitem que um sistema automatizado aprenda com os dados.

2. O que é viés no aprendizado de máquina?

Resposta: O viés no aprendizado de máquina refere-se à tendência de um modelo para aprender mais ou menos informações do que o necessário, o que pode levar a resultados imprecisos.

3. O que é variância no aprendizado de máquina?

Resposta: A variância no aprendizado de máquina refere-se à sensibilidade de um modelo a variações nos dados de treinamento, o que pode levar a uma tendência de superajuste.

4. O que é underfitting?

Resposta: Underfitting é um problema no aprendizado de máquina em que um modelo é muito simples para aprender as informações necessárias dos dados, o que resulta em resultados imprecisos.

5. O que é overfitting?

Resposta: Overfitting é um problema no aprendizado de máquina em que um modelo é muito complexo para aprender apenas os padrões nos dados de treinamento, o que pode levar a uma tendência de não generalizar bem em novos dados.

6. O que é aprendizagem supervisionada?

Resposta: Aprendizagem supervisionada é um tipo de aprendizado de máquina em que um modelo é treinado em dados rotulados, o que permite que o modelo aprenda a prever os rótulos para novos dados.

7. O que é aprendizagem não supervisionada?

Resposta: Aprendizagem não supervisionada é um tipo de aprendizado de máquina em que um modelo é treinado em dados não rotulados, o que permite que o modelo encontre padrões e estruturas nos dados.

8. O que é aprendizagem por reforço?

Resposta: Aprendizagem por reforço é um tipo de aprendizado de máquina em que um modelo aprende através da tentativa e erro em um ambiente, recebendo recompensas ou punições com base em suas ações.

9. O que é o paradigma simbólico em machine learning?

Resposta: O paradigma simbólico em machine learning é baseado na lógica e na manipulação de símbolos e regras para realizar tarefas de aprendizado. Ele usa representações explícitas de conceitos e conhecimento para inferir novas informações e tomar decisões.

10. Como funciona o método de Árvore de Decisão no contexto do paradigma simbólico?

Resposta: O método de Árvore de Decisão no paradigma simbólico constrói uma árvore de regras de decisão baseada em uma série de perguntas binárias sobre características dos dados. Cada nó da árvore representa uma questão e cada ramo representa uma possível resposta. A partir das respostas dos ramos, é possível chegar a uma decisão final.

11. O que é Rede Semântica no contexto do paradigma simbólico?

Resposta: Rede Semântica é uma técnica no paradigma simbólico que usa grafos para representar conhecimento e relacionamentos entre conceitos. Cada nó no grafo representa um conceito e as arestas representam as relações entre eles.

12. Qual é o princípio básico do paradigma estatístico em machine learning?

Resposta: O princípio básico do paradigma estatístico em machine learning é a utilização de modelos estatísticos para descrever a relação entre os dados de entrada e de saída. Esses modelos são ajustados aos dados de treinamento e usados para fazer previsões em novos dados.

13. O que é uma Rede Bayesiana?

Resposta: Uma Rede Bayesiana é um modelo gráfico probabilístico que usa a teoria de probabilidade bayesiana para representar e raciocinar sobre incertezas e causalidades entre

variáveis. Ela é composta de nós e arestas, onde cada nó representa uma variável e as arestas representam a dependência probabilística entre elas.

14. O que é o paradigma baseado em exemplos em machine learning?

Resposta: O paradigma baseado em exemplos em machine learning é baseado na ideia de que o conhecimento é adquirido por meio de exemplos específicos em vez de regras gerais. Ele usa um conjunto de exemplos de treinamento para construir um modelo que possa generalizar para novos exemplos.

15. O que é o k-NN e como ele funciona?

Resposta: O k-NN é um algoritmo de aprendizado baseado em exemplos que classifica novos exemplos com base na maioria das classes k exemplos mais próximos em um conjunto de treinamento. Ele funciona medindo a distância entre o novo exemplo e os exemplos de treinamento e selecionando os k exemplos mais próximos.

16. O que é o paradigma evolutivo em machine learning?

Resposta: O paradigma evolutivo em machine learning é baseado em algoritmos de otimização inspirados na teoria da evolução darwiniana. Eles geram uma população de soluções candidatas e aplicam operadores genéticos, como mutação e recombinação, para criar novas soluções a partir da população existente.

17. O que é um algoritmo genético?

Resposta: Um algoritmo genético é uma técnica de otimização inspirada na evolução biológica que busca encontrar soluções para problemas complexos. Ele trabalha com uma população de soluções potenciais que evoluem ao longo do tempo por meio de seleção natural, cruzamento e mutação.

18. O que é seleção de recursos em machine learning?

Resposta: A seleção de recursos é o processo de identificar e selecionar as variáveis mais relevantes e informativas para o modelo de machine learning. Isso ajuda a reduzir a dimensionalidade dos dados e melhorar a eficiência e a precisão do modelo.

19. O que é o bias-variance trade-off em machine learning?

Resposta: O bias-variance trade-off refere-se à relação entre o viés (bias) e a variância de um modelo de machine learning. Um modelo com alto viés tende a ser simplificado e subestimar a complexidade dos dados, enquanto um modelo com alta variância se ajusta demais aos dados de treinamento. Encontrar um equilíbrio entre viés e variância é fundamental para obter um modelo com bom desempenho.

20. Quais são os principais desafios enfrentados em problemas de aprendizado de máquina?

Resposta: Alguns desafios comuns em problemas de aprendizado de máquina incluem a disponibilidade e qualidade dos dados, seleção adequada de recursos, tratamento de dados ausentes ou desbalanceados, lidar com o trade-off entre viés e variância, e evitar overfitting ou underfitting do modelo.

21. Quais são as etapas básicas para construir um modelo de machine learning?

Resposta: As etapas básicas para construir um modelo de machine learning incluem a coleta e preparação dos dados, seleção e engenharia de recursos, divisão dos dados em conjuntos de treinamento e teste, escolha e treinamento do modelo, ajuste de hiperparâmetros e avaliação do desempenho do modelo.

22. Quais são os principais benefícios do uso de técnicas de aprendizado de máquina?

Resposta: As técnicas de aprendizado de máquina oferecem a capacidade de automatizar a tomada de decisões, descobrir padrões e insights ocultos em grandes volumes de dados, melhorar a precisão e eficiência em várias tarefas, como classificação, regressão, agrupamento e recomendação, além de permitir a personalização e adaptação de sistemas.

LISTA DE QUESTÕES ESTRATÉGICAS

1. (CESPE / TCE-MG – 2018) Em machine learning, a categoria de aprendizagem por reforço identifica as tarefas em que:

a) um software interage com um ambiente dinâmico, como, por exemplo, veículos autônomos.
b) as etiquetas de classificação não sejam fornecidas ao algoritmo, de modo a deixá-lo livre para entender as entradas recebidas.
c) o aprendizado pode ser um objetivo em si mesmo ou um meio para se atingir um fim.
d) o objetivo seja aprender um conjunto de regras generalistas para converter as entradas em saídas predefinidas.
e) são apresentados ao computador exemplos de entradas e saídas desejadas, fornecidas por um orientador.

---

2. (CESPE / ANP – 2022) O algoritmo random forest é um algoritmo de aprendizado de máquina supervisionado em que se agrupam os resultados de várias árvores de decisão de cada nó para se obter uma conclusão própria e aumentar a precisão do modelo, não sendo o referido algoritmo adequado para grandes conjuntos de dados.

---

3. (CESPE / ANP – 2022) As aplicações em inteligência artificial são definidas como uma subárea da área de aprendizagem de máquina (machine learning).

---

4. (CESPE / ANP – 2022) A técnica de redução de dimensionalidade (PCA) permite transformar dados que inicialmente pertencem a um espaço de dimensão n em um espaço de dimensão m, em que m < n, sendo utilizada, por exemplo, para reduzir a dimensionalidade de certo conjunto de dados através do descarte de características não úteis e que ainda permita realizar o reconhecimento de padrões.

---

5. (CESPE / TCE-MG – 2018) Em machine learning, a categoria de aprendizagem por reforço identifica as tarefas em que:

a) um software interage com um ambiente dinâmico, como, por exemplo, veículos autônomos.
b) as etiquetas de classificação não sejam fornecidas ao algoritmo, de modo a deixá-lo livre para entender as entradas recebidas.
c) o aprendizado pode ser um objetivo em si mesmo ou um meio para se atingir um fim.
d) o objetivo seja aprender um conjunto de regras generalistas para converter as entradas em saídas predefinidas.
e) são apresentados ao computador exemplos de entradas e saídas desejadas, fornecidas por um orientador.

Gabaritos

1. A
2. E
3. E
4. C
5. A

Questões Adicionais

As questões apresentadas a seguir integram o Banco de Questões do Passo Estratégico. Recomenda-se utilizá-las como um recurso complementar para a prática e consolidação dos conhecimentos adquiridos no material teórico, de acordo com o estilo adotado pela banca organizadora.

Bom estudo!

1. Overfitting ocorre frequentemente quando um modelo de aprendizado de máquina é treinado por muito tempo em um conjunto de dados limitado, capturando não apenas os padrões relevantes, mas também o ruído, o que resulta em um desempenho fraco em novos dados.

2. Árvores de decisão são modelos de aprendizado de máquina que são inerentemente resistentes a dados ausentes e podem lidar automaticamente com a ausência de valores sem ajustes adicionais.

3. Um dos principais desafios ao utilizar o algoritmo k-NN é determinar o valor de k, já que valores muito altos de k podem aumentar a sensibilidade do modelo a outliers, enquanto valores muito baixos podem levar a um underfitting severo.

4. Em uma árvore de decisão, a métrica de entropia é utilizada para medir a pureza de um nó, e o ganho de informação é calculado com base na redução da entropia após a divisão dos dados.

5. Random Forest é um exemplo de algoritmo que combina várias árvores de decisão para melhorar o desempenho e reduzir a variância do modelo.

6. Em machine learning, a regularização é uma técnica utilizada para aumentar a complexidade do modelo e garantir que ele capture todas as variações nos dados de treinamento.

7. O paradigma simbólico é uma abordagem de aprendizado de máquina que se baseia principalmente em algoritmos estatísticos.

8. As redes Bayesianas são limitadas a modelos de aprendizado supervisionado e não podem ser utilizadas para problemas não supervisionados.

9. A técnica de Análise de Componentes Principais (PCA) é frequentemente usada para aumentar a dimensionalidade dos dados, criando novos atributos baseados nas combinações lineares das variáveis originais, o que melhora a complexidade do modelo.

10. Overfitting é uma situação onde o modelo de aprendizado de máquina se ajusta tão bem aos dados de treinamento que seu desempenho em dados não vistos anteriormente é significativamente reduzido.

11. O algoritmo k-NN classifica uma instância de teste atribuindo-lhe a classe da maioria dos k vizinhos mais próximos no conjunto de treinamento.

12. O algoritmo k-NN (k-Nearest Neighbors) é conhecido por sua rapidez e eficiência em lidar com grandes volumes de dados, sendo ideal para aplicações em tempo real.

13. Redes Bayesianas são modelos gráficos que representam relações probabilísticas entre variáveis através de um grafo direcionado acíclico.

14. A validação cruzada é uma técnica que envolve dividir o conjunto de dados em treinamento e teste uma única vez.

15. A regularização é uma técnica utilizada para aumentar a complexidade dos modelos de aprendizado de máquina, promovendo um ajuste mais detalhado aos dados de treinamento e melhorando a precisão do modelo em dados futuros.

16. A validação cruzada é uma técnica utilizada para avaliar o desempenho de um modelo de aprendizado de máquina, dividindo os dados em múltiplos subconjuntos e treinando o modelo diversas vezes para garantir uma estimativa robusta.

17. Underfitting é um problema que ocorre quando o modelo é excessivamente complexo, capturando ruído nos dados de treinamento, o que resulta em uma falta de capacidade de generalização para novos dados.

18. A técnica de redução de dimensionalidade PCA visa aumentar a dimensionalidade dos dados para melhorar a qualidade das informações.

19. A técnica de validação cruzada é essencialmente utilizada para melhorar o desempenho do modelo em dados de treinamento, garantindo que o modelo aprenda melhor os padrões existentes no conjunto de dados original.

20. Em uma árvore de decisão, o ganho de informação é uma métrica crucial que ajuda a determinar o melhor atributo para dividir os dados em cada nó, maximizando a pureza das subdivisões.

GABARITOS E COMENTÁRIOS

1. Overfitting ocorre frequentemente quando um modelo de aprendizado de máquina é treinado por muito tempo em um conjunto de dados limitado, capturando não apenas os padrões relevantes, mas também o ruído, o que resulta em um desempenho fraco em novos dados.

Gabarito: C

Comentários: Certo. Overfitting ocorre quando o modelo se ajusta tão bem aos dados de treinamento que começa a capturar o ruído, prejudicando sua capacidade de generalização para novos dados.

2. Árvores de decisão são modelos de aprendizado de máquina que são inerentemente resistentes a dados ausentes e podem lidar automaticamente com a ausência de valores sem ajustes adicionais.

Gabarito: E

Comentários: Errado. Árvores de decisão podem ser afetadas por dados ausentes e geralmente requerem técnicas específicas para lidar com a ausência de valores, como imputação ou o uso de critérios de divisão modificados.

3. Um dos principais desafios ao utilizar o algoritmo k-NN é determinar o valor de k, já que valores muito altos de k podem aumentar a sensibilidade do modelo a outliers, enquanto valores muito baixos podem levar a um underfitting severo.

Gabarito: E

Comentários: Errado. Valores muito altos de k tendem a suavizar o modelo e podem levar ao underfitting, enquanto valores muito baixos de k podem resultar em um modelo mais sensível a outliers e potencialmente a overfitting, não o contrário.

4. Em uma árvore de decisão, a métrica de entropia é utilizada para medir a pureza de um nó, e o ganho de informação é calculado com base na redução da entropia após a divisão dos dados.

Gabarito: C

Comentários: Certo. A entropia mede a impureza de um nó, e o ganho de informação é calculado para avaliar o quão eficaz uma divisão dos dados é em termos de reduzir essa impureza.

5. Random Forest é um exemplo de algoritmo que combina várias árvores de decisão para melhorar o desempenho e reduzir a variância do modelo.

Gabarito: C

Comentários: Correta. Random Forest é uma técnica de ensemble que combina múltiplas árvores de decisão para obter um modelo mais robusto e geralmente com menor variância.

6. Em machine learning, a regularização é uma técnica utilizada para aumentar a complexidade do modelo e garantir que ele capture todas as variações nos dados de treinamento.

Gabarito: E

Comentários: Errado. A regularização é usada para reduzir a complexidade do modelo, penalizando coeficientes muito grandes e ajudando a prevenir o overfitting.

7. O paradigma simbólico é uma abordagem de aprendizado de máquina que se baseia principalmente em algoritmos estatísticos.

Gabarito: E

Comentários: Errada. O paradigma simbólico se concentra em representar o conhecimento de forma simbólica, como regras lógicas e representações explícitas de conhecimento.

8. As redes Bayesianas são limitadas a modelos de aprendizado supervisionado e não podem ser utilizadas para problemas não supervisionados.

Gabarito: E

Comentários: Errado. Redes Bayesianas podem ser utilizadas tanto para aprendizado supervisionado quanto não supervisionado, modelando relações probabilísticas entre variáveis.

9. A técnica de Análise de Componentes Principais (PCA) é frequentemente usada para aumentar a dimensionalidade dos dados, criando novos atributos baseados nas combinações lineares das variáveis originais, o que melhora a complexidade do modelo.

Gabarito: E

Comentários: PCA é uma técnica usada para reduzir a dimensionalidade dos dados, não para aumentá-la. Ela cria novos atributos, chamados componentes principais, que capturam a maior parte da variância dos dados originais.

10. Overfitting é uma situação onde o modelo de aprendizado de máquina se ajusta tão bem aos dados de treinamento que seu desempenho em dados não vistos anteriormente é significativamente reduzido.

Gabarito: C

Comentários: Correto. Overfitting ocorre quando o modelo captura tanto os padrões reais quanto o ruído dos dados de treinamento, resultando em uma capacidade de generalização ruim.

11. O algoritmo k-NN classifica uma instância de teste atribuindo-lhe a classe da maioria dos k vizinhos mais próximos no conjunto de treinamento.

Gabarito: C

Comentários: Correta. O algoritmo k-NN atribui a classe de uma instância de teste baseando-se na maioria dos k vizinhos mais próximos no conjunto de treinamento.

12. O algoritmo k-NN (k-Nearest Neighbors) é conhecido por sua rapidez e eficiência em lidar com grandes volumes de dados, sendo ideal para aplicações em tempo real.

Gabarito: E

Comentários: Errado. O k-NN pode ser computacionalmente caro, especialmente com grandes volumes de dados, pois a classificação de uma nova instância requer o cálculo das distâncias para todas as instâncias de treinamento.

13. Redes Bayesianas são modelos gráficos que representam relações probabilísticas entre variáveis através de um grafo direcionado acíclico.

Gabarito: C

Comentários: Correta. Redes Bayesianas de fato representam relações probabilísticas entre variáveis usando grafos direcionados acíclicos.

14. A validação cruzada é uma técnica que envolve dividir o conjunto de dados em treinamento e teste uma única vez.

Gabarito: E

Comentários: Errada. A validação cruzada envolve dividir o conjunto de dados em treinamento e teste várias vezes para avaliar a capacidade de generalização do modelo.

15. A regularização é uma técnica utilizada para aumentar a complexidade dos modelos de aprendizado de máquina, promovendo um ajuste mais detalhado aos dados de treinamento e melhorando a precisão do modelo em dados futuros.

Gabarito: E

Comentários: Errado. A regularização é usada para reduzir a complexidade do modelo, penalizando coeficientes elevados, o que ajuda a evitar o overfitting e melhora a generalização do modelo para novos dados.

16. A validação cruzada é uma técnica utilizada para avaliar o desempenho de um modelo de aprendizado de máquina, dividindo os dados em múltiplos subconjuntos e treinando o modelo diversas vezes para garantir uma estimativa robusta.

Gabarito: C

Comentários: Correto. A validação cruzada envolve dividir os dados em múltiplos subconjuntos, treinando e testando o modelo várias vezes para obter uma estimativa mais confiável de seu desempenho.

17. Underfitting é um problema que ocorre quando o modelo é excessivamente complexo, capturando ruído nos dados de treinamento, o que resulta em uma falta de capacidade de generalização para novos dados.

Gabarito: E

Comentários: Errado. Underfitting ocorre quando o modelo é muito simples para capturar os padrões nos dados de treinamento, não quando é excessivamente complexo.

18. A técnica de redução de dimensionalidade PCA visa aumentar a dimensionalidade dos dados para melhorar a qualidade das informações.

Gabarito: E

Comentários: Errada. A técnica PCA busca reduzir a dimensionalidade dos dados mantendo a maior parte da variância e preservando as informações mais importantes.

19. A técnica de validação cruzada é essencialmente utilizada para melhorar o desempenho do modelo em dados de treinamento, garantindo que o modelo aprenda melhor os padrões existentes no conjunto de dados original.

Gabarito: E

Comentários: Errado. A validação cruzada é usada para avaliar o desempenho do modelo em dados não vistos durante o treinamento, ajudando a garantir que o modelo possa generalizar bem para novos dados, não apenas para melhorar o desempenho em dados de treinamento.

20. Em uma árvore de decisão, o ganho de informação é uma métrica crucial que ajuda a determinar o melhor atributo para dividir os dados em cada nó, maximizando a pureza das subdivisões.

Gabarito: C

Comentários: Correto. O ganho de informação mede a redução na entropia ou impureza ao dividir os dados, ajudando a selecionar o melhor atributo para cada divisão.

1.C	2.E	3.E	4.C	5.C
6.E	7.E	8.E	9.E	10.C
11.C	12.E	13.C	14.E	15.E
16.C	17.E	18.E	19.E	20.C

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.