Strecht Simulados · Guia de preparação



Guia de preparação EXIN Machine Learning

Tudo o que a prova cobra, organizado por domínio: conceitos-chave, comparações que confundem, pegadinhas reais, gatilhos do enunciado, questões comentadas e um plano para chegar pronto ao exame.

4
capítulos por domínio

28
questões comentadas

65%
nota de corte

6
provas completas no curso

ANTES DE COMEÇAR

Como usar este guia

Este guia foi feito para ser estudado junto com as 6 provas completas do curso EXIN Machine Learning: o guia explica o que cai e onde estão as armadilhas; as provas treinam o ritmo e mostram o que você ainda não domina.

Quando	O que fazer
Semana 1	Faça o teste diagnóstico (página seguinte) e leia os capítulos 1 e 2. No fim da semana, faça a Prova 1 do curso no tempo do exame.
Semana 2	Leia os demais capítulos, começando pelos domínios em que você errou mais no diagnóstico. Faça as provas 2 e 3.
Semana 3	Releia os alertas de pegadinha e as tabelas de comparação. Faça as provas 4 e 5 e revise cada erro pela explicação.
Semana 4	Estude a folha de cola, faça a Prova 6 e refaça as provas abaixo de 80%. Agende o exame quando estiver acima de 80% em todas (a nota de corte é 65%).

Mapa do exame

Cada prova do curso tem 18 questões e 30 minutos, no formato do exame. Referência de aprovação: 65%.

Domínio (capítulo)	Peso	Neste guia
1. Domínio 1: O que é aprendizado de máquina?	20%	18 conceitos · 14 pegadinhas
2. Domínio 2: Codificação para aprendizado de máquina	20%	16 conceitos · 14 pegadinhas
3. Domínio 3: Algoritmos usados no aprendizado de máquina	30%	18 conceitos · 14 pegadinhas
4. Domínio 4: Aprendizado de máquina na prática	30%	18 conceitos · 14 pegadinhas

Como ler os capítulos: os **tópicos-chave** são o que mais cai; a **tabela de memorização** separa o que a prova costuma confundir; os **alertas de prova** mostram as armadilhas das alternativas erradas; os **gatilhos** ligam palavras do enunciado à resposta; e o **checklist** diz se você está pronto para o domínio.

Estude um pouco todo dia, de graça

Instagram @strecht.simulados

Um post de prova por dia: questão comentada, pegadinhas e o "não confunda" de cada certificação.

Seguir no Instagram →

YouTube Strecht Simulados

Shorts diários com pegadinhas de prova e, aos sábados, um simulado comentado.

Inscrever-se no canal →

DIAGNÓSTICO

Você já sabe isso?

Responda às 12 questões sem consultar nada. As respostas, com a explicação e o capítulo para revisar, estão no fim do guia. Comece o estudo pelos capítulos em que errou.

D1 · Domínio 1: O que é aprendizado de máquina?

Na catraca de uma empresa, um sistema de reconhecimento facial identifica os funcionários e acerta cada vez mais à medida que registra novos acessos. O que esse sistema exemplifica?

A. Um sistema baseado em regras fixas
B. Um agente de aprendizagem autônomo
C. Um banco de dados de imagens
D. Uma automação por roteiro programado

D2 · Domínio 1: O que é aprendizado de máquina?

Uma varejista acumula bilhões de registros de vendas e de navegação no site. Segundo o guia, que papel o aprendizado de máquina cumpre diante desse Big Data?

A. Armazenar os registros em um banco relacional mais barato
B. Compactar os arquivos antigos para liberar espaço em disco
C. Aplicar algoritmos aos dados para obter compreensão (insight)
D. Substituir a coleta de dados por regras definidas pela gerência

D3 · Domínio 1: O que é aprendizado de máquina?

Na etapa de revisão, que vem antes da implantação do modelo, quais atividades o guia inclui?

A. Limpeza, integração e transformação dos dados
B. Gráficos para explorar os dados antes do treino
C. Definição do problema e seleção dos dados de origem
D. Revisão por pares e identificação do melhor modelo

D4 · Domínio 2: Codificação para aprendizado de máquina

A documentação oficial do Python informa que a linguagem suporta módulos e pacotes. Que benefício isso traz?

A. Licença paga com suporte do fabricante
B. Modularidade e a reutilização de código
C. Dispensa de qualquer biblioteca externa
D. Execução sem precisar de interpretador

D5 · Domínio 2: Codificação para aprendizado de máquina

O guia pede que o candidato identifique softwares de código aberto e proprietários. O que caracteriza um software de código aberto?

A. Uso liberado só para quem paga uma licença de assinatura anual
B. Funcionamento restrito a computadores com GPU
C. Código-fonte guardado em sigilo pelo fabricante
D. Código-fonte disponível para estudar, alterar e redistribuir

D6 · Domínio 2: Codificação para aprendizado de máquina

Entre os softwares que o guia lista para a codificação do aprendizado de máquina, qual destes é proprietário, e não de código aberto?

A. Scikit-Learn
B. R Studio
C. TensorFlow
D. MATLAB

D7 · Domínio 3: Algoritmos usados no aprendizado de máquina

Em um modelo, cada cliente é descrito por uma lista de números, como visitas, compras e devoluções, e o conjunto de clientes forma uma tabela de linhas e colunas. Que área da matemática dá a base para trabalhar com essas estruturas?

A. Estatística inferencial
B. Probabilidade
C. Álgebra linear
D. Teorema de Bayes

D8 · Domínio 3: Algoritmos usados no aprendizado de máquina

Uma análise mostra que, nos meses em que se vendem mais ventiladores, também se vendem mais sorvetes. Que conclusão é adequada?

A. A venda de sorvetes causa a de ventiladores
B. A venda de ventiladores causa a de sorvetes
C. As duas vendas não têm nenhuma relação estatística
D. Há correlação, o que não prova causa e efeito

D9 · Domínio 3: Algoritmos usados no aprendizado de máquina

No aprendizado por reforço, como o sistema aprende?

A. Por tentativa e erro, com recompensas e punições
B. Por regras escritas por especialistas humanos
C. Por exemplos rotulados com a resposta correta
D. Por grupos formados a partir de dados sem rótulo

D10 · Domínio 4: Aprendizado de máquina na prática

Durante os testes, a equipe combina os resultados de vários modelos em um só, em vez de depender de um único modelo. Que técnica é essa?

A. Revisão por pares do código
B. Conjunto de modelos (ensemble)
C. Ajuste fino das configurações
D. Teste estatístico de hipóteses

D11 · Domínio 4: Aprendizado de máquina na prática

Carla criou um modelo que prioriza chamados de suporte de uma operadora de internet. Nos testes, o modelo encaminhou corretamente 9 em cada 10 chamados urgentes e reduziu a espera média de 40 para 15 minutos. Ela vai apresentar o resultado ao diretor de atendimento, que não é da área técnica. Como Carla deve apresentar os resultados?

A. Destacar o ganho no atendimento em linguagem simples
B. Usar termos técnicos para reforçar a credibilidade
C. Detalhar as fórmulas e os parâmetros internos do algoritmo
D. Mostrar o código-fonte e as bibliotecas usadas

D12 · Domínio 4: Aprendizado de máquina na prática

Dois algoritmos tiveram desempenhos parecidos no compartimento de teste. Para que serve aplicar um teste estatístico nessa comparação?

A. Corrigir os rótulos errados dos dados de teste
B. Converter os resultados em linguagem simples
C. Aumentar a quantidade de dados de treinamento
D. Verificar se a diferença não se deve ao acaso

01

CAPÍTULO 1 · 20% DA PROVA

Domínio 1: O que é aprendizado de máquina?

Aqui eu preciso dominar as definições do item 1 do guia quase palavra por palavra: o que é aprendizado de máquina, as seis aplicações, agente de aprendizagem, aprendizagem profunda, rede neural, sistema baseado em conhecimento e o processo de AM com dados. A prova cobra reconhecimento e compreensão (Bloom 1 e 2), então quase sempre a alternativa certa repete a ideia do guia e as erradas trazem a definição de outro item. Onde mais erro: confundir aprendizado de máquina (aprende com dados) com sistema baseado em conhecimento (aplica regras humanas), trocar recomendações por IA generativa e misturar as etapas do processo (pré-processamento x revisão x implantação). Também caem frases do tipo 'qual contraria o guia', em que preciso achar a única afirmação invertida.

Neste capítulo

18 tópicos-chave · 14 pegadinhas · 8 macetes · 14 gatilhos · 4 questões comentadas

Tópicos-chave

Aprendizado de máquina (AM)

É um subconjunto da inteligência artificial que aplica algoritmos a dados fornecidos para obter uma compreensão (insight). É usado na ciência de dados, inclusive com Big Data. A ideia central é 'aprender com a experiência'.

IA não é conceito novo

Pelo guia, a IA em si não é um conceito novo; o aprendizado de máquina é mais uma etapa na evolução da IA. Qualquer frase dizendo que a IA nasceu com o aprendizado de máquina contraria o guia.

Aprender com a experiência

Significa que o modelo melhora conforme processa novos exemplos de dados. Quem aprende é o próprio sistema, e a experiência vem dos dados. Não é o programador ganhando prática, nem regras novas a cada versão, nem o hardware ficando mais rápido.

Definição de Tom Mitchell

Definição acadêmica de aprendizado iterativo e contínuo (Machine Learning, 1997). Um programa aprende com a experiência E, em relação a uma tarefa T e a uma medida de desempenho P, quando o desempenho em T, medido por P, melhora com E. Mais dados sem melhora em P não é aprendizado.

Cientista de dados

O guia associa o AM à exigência de talento para aprendizado e matemática, 'ou seja, cientista de dados'. Na prova, talento para matemática e aprendizado aponta para esse perfil; o AM é usado na ciência de dados para obter insight de dados e Big Data.

Seis aplicações do AM (item 1.2)

Predição, reconhecimento de objetos, classificação, agrupamento (clustering), recomendações (por exemplo, Netflix e Spotify) e IA generativa (por exemplo, ChatGPT e Copilot). Os exemplos de marca são cobrados exatamente como estão no guia.

Predição

Usar dados históricos para estimar um resultado futuro, como quantos leitos estarão ocupados na próxima semana ou as vendas do próximo trimestre. Pergunta de 'quanto' ou 'quando' no futuro aponta para predição.

Reconhecimento de objetos

Identifica o que aparece em imagens ou vídeos, como placas de trânsito, espécies de plantas ou pedestres. Se a entrada é uma imagem e a saída diz o que há nela, é reconhecimento de objetos.

Agente de aprendizagem

Projetado para realizar uma tarefa específica, usando uma determinada quantidade de dados, de forma autônoma, e melhora com a repetição. Conteúdo indicativo: dados, tarefa única e aprender com a experiência. Exemplos do guia: chatbots, carros sem motorista e reconhecimento facial.

Aprendizagem profunda

Subconjunto do aprendizado de máquina que treina grandes redes neurais profundas para processar e analisar grandes quantidades de dados, obter mais insight e resolver problemas mais complexos. O guia a descreve como técnica universal para resolver um conjunto maior de problemas e como redes neurais combinadas com grandes conjuntos de dados.

Rede neural

Finalidade: entrada > identificar padrões nos dados > resultado (output), e tomada de decisões. É usada particularmente com dados não estruturados ou não rotulados, como imagens e documentos manuscritos. Exemplo do guia: no reconhecimento de voz, a voz capturada é comparada com padrões armazenados para identificar o output (ação, aceitação de comando de voz, conversão de texto em fala).

Sistema baseado em conhecimento

Forma de IA que captura a experiência e o conhecimento humano em uma base de conhecimento e aplica um conjunto de regras para identificar um resultado por meio de um mecanismo de inferência. Não depende de grandes volumes de dados rotulados.

Base de conhecimento x mecanismo de inferência

A base de conhecimento guarda o conhecimento dos especialistas. O mecanismo de inferência é o componente que aplica as regras para chegar ao resultado.

Tecnologias de IA complementares

O AM complementa sistemas baseados em conhecimento. No carro sem motorista do guia, um sistema baseado em conhecimento toma decisões informadas ou identifica a causa provável de uma falha, e uma rede neural faz o reconhecimento de imagem para a navegação com a câmera.

Processo de AM com dados (item 1.7)

Analisar o problema, seleção de dados, pré-processamento, visualização de dados, seleção de um modelo (treinar, testar, repetir) e revisão. Só quando dados e algoritmos apresentam desempenho satisfatório o modelo pode ser implantado.

Subetapas do pré-processamento

Limpeza, integração, transformação, redução e manipulação. Revisão por pares não está aqui: ela pertence à etapa de revisão.

Etapa de revisão

Reúne revisão por pares, aprendizado a partir de vários algoritmos e identificação do melhor modelo de aprendizado de máquina. Vem antes da implantação.

Compartimentos e falta de método único

Pelo guia, não existe um método de fato no AM e o aprendizado por meio da experiência é de vital importância. O teste envolve criar dados de teste corretos e compartimentos para aprender e compartimentos para o que se quer testar.

Macetes para memorizar

💡 Aplicações do item 1.2

PRCARG: Predição, Reconhecimento de objetos, Classificação, Agrupamento, Recomendações (Netflix/Spotify), Generativa (ChatGPT/Copilot).

💡 Agente de aprendizagem

DTA: Dados + Tarefa única + Aprender com a experiência (e sozinho).

Mitchell

T-P-E: Tarefa, Performance (desempenho), Experiência. Aprender = P sobe em T graças a E.

Sistema baseado em conhecimento

A base GUARDA, o mecanismo de inferência APLICA.

Rede neural

Entra, compara, sai: entrada > padrões > resultado.

Pré-processamento

LITRM: Limpeza, Integração, Transformação, Redução, Manipulação.

Processo de AM

Problema > Seleção > Pré-processamento > Visualização > Modelo (treina, testa, repete) > Revisão > Implanta.

Carro sem motorista

Olho = rede neural (câmera); cérebro do mecânico = sistema baseado em conhecimento (falha).

Tabela de memorização: não confunda

Quando é...	...e quando é
Aprendizado de máquina Baseado em dados: aprende padrões e deriva resultados por algoritmos, como uma rede neural.	**Sistema baseado em conhecimento** Baseado em conhecimento humano: base de conhecimento + regras aplicadas por um mecanismo de inferência.
Base de conhecimento Componente que guarda a experiência e o conhecimento dos especialistas.	**Mecanismo de inferência** Componente que aplica o conjunto de regras para identificar o resultado.
Recomendações Sugere itens com base no histórico do usuário; exemplos do guia: Netflix e Spotify.	**IA generativa** Gera conteúdo novo a partir de um pedido; exemplos do guia: ChatGPT e Copilot.
Predição Estima um valor futuro a partir do histórico (ocupação de leitos, vendas do trimestre).	**Agrupamento (clustering)** Organiza dados em grupos por semelhança, sem estimar quanto ou quando.
Rede neural (no carro sem motorista) Reconhecimento de imagem para navegar com a câmera.	**Sistema baseado em conhecimento (no carro sem motorista)** Decisões informadas e identificação da causa provável de uma falha.
Agente de aprendizagem Executa sozinho uma tarefa específica com dados e melhora a cada repetição (ex.: reconhecimento facial na catraca).	**Sistema de regras fixas ou automação por roteiro** Repete passos ou regras escritas previamente e não melhora com novos dados.

Pré-processamento de dados Antes do treino: limpeza, integração, transformação, redução e manipulação.	**Revisão** Depois de treinar e testar: revisão por pares, vários algoritmos e escolha do melhor modelo.

Números que caem

Item	Valor
Peso do domínio no exame	**20% (4 de 20 pontos; 4 questões por prova)**
Peso do item 1.1 (definir o AM)	**5%**
Peso dos itens 1.2 a 1.7	**2,5% cada**
Obra de Tom Mitchell citada no guia	**Machine Learning, 1997**
Aplicações listadas no item 1.2	**6**

Alertas de prova (pegadinhas)

⚠ **Alerta de prova:** Aceitar que 'a IA é um conceito recente, nascido com o aprendizado de máquina'.
Como não cair: A IA veio antes; o AM é mais uma etapa da evolução da IA e um subconjunto dela.

⚠ **Alerta de prova:** Escolher a frase que diz que o AM aplica regras humanas e o sistema baseado em conhecimento aprende com dados.
Como não cair: Os papéis estão invertidos: AM aprende com dados; sistema baseado em conhecimento aplica regras humanas.

⚠ **Alerta de prova:** Achar que a base de conhecimento é quem aplica as regras.
Como não cair: A base guarda; quem aplica as regras é o mecanismo de inferência.

⚠ **Alerta de prova:** Marcar agrupamento ou reconhecimento de objetos para Netflix e Spotify.
Como não cair: No guia, Netflix e Spotify são exemplos de recomendações; ChatGPT e Copilot, de IA generativa.

⚠ **Alerta de prova:** Colocar limpeza, integração e transformação dentro da etapa de revisão.
Como não cair: Essas são subetapas do pré-processamento; a revisão é revisão por pares, vários algoritmos e melhor modelo.

⚠ **Alerta de prova:** Implantar logo após a seleção de dados ou ao terminar o pré-processamento.
Como não cair: Ainda faltam visualização, treino, teste e revisão; implantação só com desempenho satisfatório.

⚠ **Alerta de prova:** Implantar antes do teste para coletar opinião de usuários reais.
Como não cair: O teste sempre vem antes da implantação no processo do guia.

⚠ **Alerta de prova:** Interpretar 'aprender com a experiência' como o programador ganhando prática ou o equipamento ficando mais rápido.
Como não cair: Quem aprende é o modelo, e a experiência são os dados que ele processa.

⚠ **Alerta de prova:** Na definição de Mitchell, aceitar que basta E (dados) aumentar, ou que basta executar T mais rápido que uma pessoa.
Como não cair: Só há aprendizado quando o desempenho em T, medido por P, melhora com a experiência E.

⚠ **Alerta de prova:** Descrever o agente de aprendizagem como executor de muitas tarefas gerais, sem dados, ou como coordenador que não aprende.
Como não cair: Agente = tarefa específica + dados + autonomia + melhora com a repetição.

⚠ **Alerta de prova:** Achar que a aprendizagem profunda funciona melhor com poucos registros em tabelas simples ou é limitada a jogos.
Como não cair: Ela é universal, para problemas complexos, e depende de grandes redes neurais com grandes volumes de dados.

⚠ **Alerta de prova:** Associar redes neurais a tabelas pequenas, relatórios resumidos ou listas de regras.
Como não cair: O guia destaca dados não estruturados ou não rotulados: imagens, manuscritos, voz.

⚠ **Alerta de prova:** No carro sem motorista, pôr o sistema baseado em conhecimento nas imagens ou o mecanismo de inferência na câmera.
Como não cair: Imagem = rede neural; diagnóstico de falha e decisões informadas = sistema baseado em conhecimento.

⚠ **Alerta de prova:** Dizer que o papel do AM diante do Big Data é armazenar ou compactar os registros.
Como não cair: Big Data é matéria-prima; o AM aplica algoritmos para obter compreensão (insight).

Como cai na prova

📝 **Como cai:** Definição direta: 'qual descrição corresponde a...' (agente, rede neural, aprendizagem profunda).
Como resolver: Busque na alternativa as palavras do conteúdo indicativo do guia (tarefa única, entrada > padrões > resultado, grandes redes + muitos dados). Elimine as que descrevem outro item, sobretudo regras de especialistas.

📝 **Como cai:** Cenário curto de negócio pedindo a aplicação (hospital, streaming, câmera).
Como resolver: Identifique a saída pedida: valor futuro = predição; o que há na imagem = reconhecimento de objetos; sugestão ao usuário = recomendações; conteúdo novo = IA generativa.

📝 **Como cai:** Pergunta negativa: 'qual contraria o guia' ou 'qual NÃO é subetapa'.
Como resolver: Três alternativas repetem o guia; ache a única que inverte a ordem ou vem de outra etapa (IA nova; revisão por pares dentro do pré-processamento).

📝 **Como cai:** Em que momento implantar ou o que faz parte da revisão.
Como resolver: Percorra a sequência do item 1.7 e veja o que ainda falta; implantação só depois de testar e revisar com desempenho satisfatório.

📝 **Como cai:** Diferença entre AM e sistema baseado em conhecimento, ou combinação no carro sem motorista.
Como resolver: Dados e algoritmos = AM; conhecimento humano e regras = sistema baseado em conhecimento. Confira se os papéis não estão trocados.

Gatilhos do enunciado

Se o enunciado trouxer...	Pense em
qual frase contraria o guia sobre IA e AM	**a que diz que a IA é nova ou nasceu com o AM**
regras escritas por especialistas, base de conhecimento	**sistema baseado em conhecimento**
componente que aplica as regras	**mecanismo de inferência**
tarefa específica, autônoma, melhora a cada repetição; chatbot, reconhecimento facial	**agente de aprendizagem**
grandes redes neurais + grandes quantidades de dados, problema complexo	**aprendizagem profunda**
entrada comparada com padrões armazenados para gerar resultado; voz, imagem, manuscrito	**rede neural**
Netflix, Spotify	**recomendações**

ChatGPT, Copilot, redigir texto novo	**IA generativa**
estimar quantos/quanto no futuro com base no histórico	**predição**
foto ou câmera que diz o que há na imagem	**reconhecimento de objetos**
tarefa T, desempenho P, experiência E	**definição de Tom Mitchell: P melhora em T por causa de E**
talento para matemática e aprendizado	**cientista de dados**
revisão por pares, vários algoritmos, melhor modelo	**etapa de revisão, antes da implantação**
limpeza, integração, transformação, redução, manipulação	**pré-processamento de dados**

🎯 Se você lembrar de só uma coisa deste capítulo...

- AM é subconjunto da IA; aprendizagem profunda é subconjunto do AM; a IA não é nova.
- Netflix e Spotify = recomendações; ChatGPT e Copilot = IA generativa.
- Base de conhecimento guarda, mecanismo de inferência aplica as regras.
- Revisão por pares é etapa de revisão, não de pré-processamento.
- Implantação só quando dados e algoritmos mostram desempenho satisfatório.

Essenciais para a prova (checklist)

☐ Sei definir AM como aplicação de algoritmos a dados para obter insight, subconjunto da IA.

☐ Sei explicar a definição de Tom Mitchell com T, P e E.

☐ Sei listar as seis aplicações do item 1.2 com os exemplos de marca.

☐ Sei os três elementos do agente de aprendizagem e seus exemplos (chatbots, carros sem motorista, reconhecimento facial).

☐ Sei descrever a aprendizagem profunda como técnica universal com grandes redes neurais e muitos dados.

☐ Sei a sequência da rede neural (entrada > padrões > resultado) e que ela brilha em dados não estruturados.

☐ Sei separar base de conhecimento e mecanismo de inferência.

☐ Sei qual tecnologia faz o quê no carro sem motorista do guia.

☐ Sei a ordem do processo do item 1.7 e as cinco subetapas do pré-processamento.

☐ Sei o que compõe a etapa de revisão e quando o modelo pode ser implantado.

Questões de revisão do capítulo

R1.1 · Domínio 1: O que é aprendizado de máquina?

Qual afirmação diferencia o aprendizado de máquina de um sistema baseado em conhecimento?

A. Os dois dispensam a experiência de especialistas humanos
B. O primeiro aplica regras humanas; o segundo aprende com dados
C. O primeiro aprende com dados; o segundo aplica regras humanas
D. Os dois dependem de grandes volumes de dados rotulados

R1.2 · Domínio 1: O que é aprendizado de máquina?

Como o guia caracteriza a aprendizagem profunda quanto ao alcance dos problemas que ela resolve?

A. Técnica limitada a jogos de tabuleiro com poucas jogadas
B. Técnica restrita a cálculos simples em tabelas pequenas de dados
C. Técnica universal para resolver um conjunto maior de problemas
D. Técnica voltada a regras escritas para um único domínio

R1.3 · Domínio 1: O que é aprendizado de máquina?

Qual sequência resume a finalidade de uma rede neural, segundo o guia?

A. Armazenar a entrada e ordená-la por data de chegada
B. Gerar o resultado e depois buscar dados que o confirmem
C. Receber regras, consultar especialistas e gerar um relatório
D. Receber a entrada, identificar padrões e gerar o resultado

R1.4 · Domínio 1: O que é aprendizado de máquina?

Qual descrição corresponde à função de um agente de aprendizagem no aprendizado de máquina?

A. Coordenar outros sistemas sem aprender com os resultados
B. Realizar sozinho uma tarefa específica e melhorar com a prática
C. Executar muitas tarefas gerais, sem depender de dados de entrada
D. Armazenar o conhecimento de especialistas em regras fixas

Respostas comentadas

R1.1. C) O primeiro aprende com dados; o segundo aplica regras humanas

Por quê: Pelo item 1.6, o sistema baseado em conhecimento captura o conhecimento humano em uma base de conhecimento e aplica regras por um mecanismo de inferência. O aprendizado de máquina é baseado em dados e deriva resultados por meio de algoritmos, como uma rede neural.

✗ A) Incorreta. O sistema baseado em conhecimento existe justamente para capturar a experiência de especialistas humanos.

✗ B) Incorreta. A afirmação inverte os papéis: quem aplica regras humanas é o sistema baseado em conhecimento.

✗ D) Incorreta. O sistema baseado em conhecimento não depende de grandes volumes de dados rotulados; ele usa uma base de conhecimento e regras.

> **Dica de prova:** Dados e algoritmos: aprendizado de máquina. Conhecimento humano e regras: sistema baseado em conhecimento.

R1.2. C) Técnica universal para resolver um conjunto maior de problemas

Por quê: No item 1.4, o conteúdo indicativo traz dois pontos: técnica universal para resolver um conjunto maior de problemas e redes neurais combinadas com grandes conjuntos de dados.

✗ A) Incorreta. O guia não limita a aprendizagem profunda a jogos; ela é descrita como técnica universal.

✗ B) Incorreta. A aprendizagem profunda é indicada para grandes quantidades de dados e problemas complexos, e não para cálculos simples em tabelas pequenas.

✗ D) Incorreta. Regras escritas para um domínio descrevem um sistema baseado em conhecimento, e não a aprendizagem profunda.

> **Dica de prova:** Universal e para problemas mais complexos: assim o guia descreve a aprendizagem profunda.

R1.3. D) Receber a entrada, identificar padrões e gerar o resultado

Por quê: No item 1.5, o conteúdo indicativo traz 'entrada > identificar padrões nos dados > resultado (output)' e a tomada de decisões. A rede neural compara a entrada com padrões aprendidos para determinar o resultado.

✗ A) Incorreta. Ordenar a entrada por data é uma tarefa de armazenamento e não identifica padrões.

✗ B) Incorreta. A ordem está invertida: a rede neural parte da entrada e dos padrões para chegar ao resultado.

✗ C) Incorreta. Receber regras e consultar especialistas descreve um sistema baseado em conhecimento, e não uma rede neural.

> **Dica de prova:** Rede neural: entrada, padrões, resultado, nessa ordem.

R1.4. B) Realizar sozinho uma tarefa específica e melhorar com a prática

Por quê: Pelo item 1.3, cada agente de aprendizagem é projetado para uma tarefa específica, com uma determinada quantidade de dados, executada de forma autônoma. Com a repetição, ele aprende a melhorar. Exemplos: chatbots, carros sem motorista e reconhecimento facial.

✗ A) Incorreta. O que caracteriza o agente é justamente aprender com a repetição da tarefa e melhorar.

✗ C) Incorreta. O agente é projetado para uma tarefa específica e usa uma determinada quantidade de dados; não é um executor de tarefas gerais sem dados.

✗ D) Incorreta. Guardar o conhecimento de especialistas em regras descreve um sistema baseado em conhecimento.

> **Dica de prova:** Agente de aprendizagem: uma tarefa, dados, autonomia e melhora com a repetição.

02

CAPÍTULO 2 · 20% DA PROVA

Domínio 2: Codificação para aprendizado de máquina

Neste domínio 'codificação' quer dizer programação: linguagens e softwares, e não codificação de variáveis categóricas. Eu preciso saber de cor as quatro linguagens orientadas a objetos do guia (Python, R, C++ e Java), o papel das bibliotecas e os cinco softwares do item 2.2 (TensorFlow, R Studio, CUDA, Scikit-Learn e MATLAB), separando código aberto de proprietário. A prova cobra muito a associação software x origem x função: TensorFlow do Google, CUDA da NVIDIA para GPU, R Studio da Posit para R, Scikit-Learn biblioteca de Python, MATLAB proprietário. Onde mais erro: alternativas que trocam a descrição de um software pela de outro e linguagens intrusas como HTML, SQL ou YAML. Também cai o nível esperado: familiaridade, não fluência nem proficiência.

Neste capítulo

16 tópicos-chave · 14 pegadinhas · 7 macetes · 14 gatilhos · 4 questões comentadas

Tópicos-chave

Codificação = programação

No guia, 'codificação para aprendizado de máquina' trata de linguagens de programação e softwares. Não confunda com codificação de variáveis categóricas.

Linguagens orientadas a objetos do item 2.1

Python, R, C++ e Java. Qualquer outra (HTML, SQL, YAML, Markdown, Kotlin, COBOL, Swift) não está na lista do guia.

Orientação a objetos

Paradigma em que o programa é organizado em objetos que reúnem dados e o código que opera sobre eles. Não é consulta a banco (SQL), nem fórmulas de planilha, nem marcação de páginas (HTML/CSS).

Python

O guia o destaca como linguagem muito popular no AM e na ciência de dados. A documentação oficial o descreve como interpretado, orientado a objetos e de alto nível, com semântica dinâmica. Interpretador e biblioteca padrão estão disponíveis sem custo e podem ser distribuídos livremente.

Módulos e pacotes do Python

O suporte a módulos e pacotes estimula a modularidade e a reutilização de código. É isso que permite usar bibliotecas prontas de AM; não dispensa bibliotecas externas.

R

O site oficial descreve o R como linguagem e ambiente para computação estatística e gráficos, software livre sob a licença GNU GPL e ampliável por pacotes. É uma linguagem funcional que também suporta orientação a objetos.

Bibliotecas/modelos

Pelo guia, as bibliotecas são usadas para agrupar funcionalidades em modelos, que podem ser usados com diferentes linguagens, por exemplo Python. Não guardam dados brutos nem apresentam resultados.

NumPy

Descrito pela documentação oficial como o pacote fundamental de computação científica em Python. Serve de base para muitas bibliotecas de AM; não é software proprietário nem ambiente para R.

pandas

Biblioteca de código aberto de Python para análise e manipulação de dados em tabelas. Não acelera GPU e não é ferramenta de painéis como o Power BI.

Nível de domínio esperado

Linguagens: familiaridade com as linguagens comuns e seu uso, sem exigência de fluência. Softwares: não se espera proficiência, mas familiaridade com pelo menos um. Certificação ou experiência comprovada não são exigidas.

Código aberto x proprietário

Código aberto: código-fonte disponível para estudar, alterar e redistribuir nos termos da licença. Proprietário: código-fonte em sigilo e uso condicionado a licença. A diferença é o acesso ao código, não o hardware.

TensorFlow

Construído no Google; API e implementação de referência liberadas como código aberto sob a licença Apache 2.0 em novembro de 2015. O site oficial o define como plataforma de ponta a ponta para criar e implantar modelos de AM, com a API de alto nível Keras.

R Studio

Mantido pela Posit, que o descreve como ambiente de desenvolvimento (IDE) livre e de código aberto para R, com editor de código pensado para ciência de dados. Não é biblioteca de Python nem software do Google.

CUDA

Plataforma de computação paralela e modelo de programação desenvolvidos pela NVIDIA. O CUDA Toolkit é um ambiente para criar aplicativos de alto desempenho acelerados por GPU, com bibliotecas, compilador C/C++ e ferramentas de depuração.

Scikit-Learn

Lema 'Machine Learning in Python': biblioteca de AM de código aberto, utilizável comercialmente sob licença BSD, construída sobre NumPy, SciPy e matplotlib. Áreas: classificação, regressão, agrupamento e pré-processamento.

MATLAB

Entre os softwares do item 2.2, é o exemplo de software proprietário. TensorFlow, Scikit-Learn e R Studio são de código aberto.

Macetes para memorizar

💡 Linguagens do item 2.1

'Python Roda C++ e Java': Python, R, C++, Java. O resto é intruso.

💡 Softwares do item 2.2

TRCSM: TensorFlow, R Studio, CUDA, Scikit-Learn, MATLAB. Só o M é proprietário.

💡 Origem de cada software

TensorFlow = Google; CUDA = NVIDIA; R Studio = Posit; Scikit-Learn = Python/BSD.

💡 CUDA

CUDA = GPU da NVIDIA: computação paralela, com compilador C/C++ no CUDA Toolkit.

💡 Scikit-Learn

Sci-kit = kit de AM pronto em Python: classifica, regride, agrupa e pré-processa.

💡 Nível exigido

Familiar, não fluente: conheço o uso, não preciso ser expert.

💡 Código aberto

Aberto = código à vista; proprietário = código no cofre.

Tabela de memorização: não confunda

Quando é...	...e quando é
TensorFlow Plataforma de ponta a ponta do Google, código aberto, para criar, treinar e implantar modelos.	**CUDA** Plataforma de computação paralela da NVIDIA para acelerar aplicativos com GPU.
R Studio Ambiente de desenvolvimento (IDE) livre para a linguagem R, da Posit.	**Scikit-Learn** Biblioteca de Python com algoritmos prontos de classificação, regressão, agrupamento e pré-processamento.
MATLAB Software proprietário.	**Scikit-Learn / TensorFlow / R Studio** Código aberto (BSD, Apache 2.0 e IDE livre, respectivamente).
NumPy Pacote fundamental de computação científica em Python.	**pandas** Análise e manipulação de tabelas de dados em Python.
Linguagem orientada a objetos (Python, R, C++, Java) Organiza o programa em objetos com dados e código; listadas no item 2.1.	**SQL / HTML / YAML / Markdown** Consulta a banco, marcação de páginas, configuração ou formatação de texto; fora da lista do guia.
Código aberto Código-fonte público para estudar, alterar e redistribuir.	**Proprietário** Código-fonte em sigilo pelo fabricante e uso mediante licença.

Números que caem

Item	Valor
Peso do domínio no exame	**20% (4 de 20 pontos; 4 questões por prova)**
Peso do item 2.1 (linguagens)	**10%**
Peso do item 2.2 (softwares)	**10%**
Linguagens orientadas a objetos listadas	**4 (Python, R, C++, Java)**
Softwares listados no item 2.2	**5 (TensorFlow, R Studio, CUDA, Scikit-Learn, MATLAB)**
Abertura do TensorFlow	**novembro de 2015, licença Apache 2.0**

Alertas de prova (pegadinhas)

⚠ **Alerta de prova:** Aceitar HTML, SQL, YAML ou Markdown como linguagem orientada a objetos do guia.
Como não cair: A lista do item 2.1 tem só quatro: Python, R, C++ e Java.

⚠ **Alerta de prova:** Descrever Python como compilado, de marcação ou de baixo nível.
Como não cair: Python é interpretado, orientado a objetos e de alto nível.

⚠ **Alerta de prova:** Achar que módulos e pacotes do Python dispensam bibliotecas externas ou exigem licença paga.
Como não cair: Eles trazem modularidade e reutilização, e o interpretador é distribuído livremente.

⚠ **Alerta de prova:** Dizer que o R foi criado para páginas web, planilhas ou jogos.
Como não cair: R é para computação estatística e gráficos.

⚠ **Alerta de prova:** Classificar o R como linguagem de consulta, de marcação ou de montagem.
Como não cair: R é funcional, com suporte à orientação a objetos.

⚠ **Alerta de prova:** Chamar o R Studio de biblioteca de Python, plataforma de GPU ou software do Google.
Como não cair: R Studio é IDE livre para R, mantido pela Posit.

⚠ **Alerta de prova:** Marcar Scikit-Learn como software proprietário ou mantido pela NVIDIA.
Como não cair: Scikit-Learn é biblioteca de Python, código aberto, licença BSD.

⚠ **Alerta de prova:** Associar o MATLAB a código aberto ou à comunidade do R.
Como não cair: MATLAB é o software proprietário da lista do guia.

⚠ **Alerta de prova:** Descrever o TensorFlow como ambiente proprietário ou como editor de código R.
Como não cair: TensorFlow é código aberto do Google, plataforma de ponta a ponta para criar e implantar modelos.

⚠ **Alerta de prova:** Escolher pandas, R Studio ou planilha para aproveitar GPUs NVIDIA.
Como não cair: GPU da NVIDIA = CUDA.

⚠ **Alerta de prova:** Pensar que código aberto significa rodar só com GPU ou pagar assinatura anual.
Como não cair: Código aberto fala de acesso ao código-fonte, não de preço por assinatura nem de hardware.

⚠ **Alerta de prova:** Achar que o guia exige fluência, proficiência em todos os softwares, certificação ou experiência comprovada.
Como não cair: Basta familiaridade com as linguagens comuns e com pelo menos um software.

⚠ **Alerta de prova:** Atribuir às bibliotecas a função de guardar dados brutos, definir o problema ou apresentar resultados.
Como não cair: Bibliotecas agrupam funcionalidades em modelos para reutilizar no código.

⚠ **Alerta de prova:** Descrever o NumPy como software proprietário de matrizes ou como ambiente para R.
Como não cair: NumPy é o pacote fundamental de computação científica em Python.

Como cai na prova

📝 **Como cai:** 'Qual NÃO está entre as linguagens do guia' ou 'além de Python e R, quais duas' (escolher 2).
Como resolver: Tenha a lista Python, R, C++ e Java na cabeça; elimine as quatro e o que sobra é a resposta, ou marque exatamente as que faltam (C++ e Java).

📝 **Como cai:** Cenário com necessidade técnica (GPU NVIDIA, IDE para R, biblioteca de Python com algoritmos prontos).
Como resolver: Extraia a palavra-chave (GPU, R, Python + pronto) e associe ao software: CUDA, R Studio, Scikit-Learn.

📝 **Como cai:** Associações software x característica (escolher 2) ou 'qual afirmação NÃO está de acordo'.
Como resolver: Teste cada par contra origem e licença: Google/aberto, NVIDIA/GPU, Posit/IDE R, Python/BSD, MATLAB/proprietário. A descrição de um software colada em outro é o erro.

📝 **Como cai:** 'Como a documentação oficial descreve...' (Python, R, NumPy, TensorFlow).
Como resolver: Procure a frase-assinatura: interpretado/OO/alto nível; estatística e gráficos; computação científica; ponta a ponta para criar e implantar.

📝 **Como cai:** O que o guia espera do candidato em linguagens e softwares.
Como resolver: Marque a opção mais modesta: familiaridade com linguagens comuns e com pelo menos um software.

Gatilhos do enunciado

Se o enunciado trouxer...	Pense em
linguagem muito popular em AM e ciência de dados	**Python**
interpretada, orientada a objetos, alto nível	**Python**
computação estatística e gráficos, licença GNU GPL	**R**
funcional com suporte à orientação a objetos	**R**
GPU NVIDIA, computação paralela, compilador C/C++	**CUDA**
Google, código aberto, Apache 2.0, criar e implantar modelos, Keras	**TensorFlow**
Posit, IDE livre, editor para ciência de dados em R	**R Studio**
'Machine Learning in Python', licença BSD, classificação/regressão/agrupamento prontos	**Scikit-Learn**
software proprietário da lista do guia	**MATLAB**
pacote fundamental de computação científica em Python	**NumPy**
analisar e manipular tabelas de dados em Python	**pandas**
módulos e pacotes	**modularidade e reutilização de código**
nível esperado do candidato	**familiaridade, não fluência nem proficiência**
para que servem as bibliotecas segundo o guia	**agrupar funcionalidades em modelos**

🎯 **Se você lembrar de só uma coisa deste capítulo...**

- Codificação = programação (linguagens e softwares).
- Quatro linguagens: Python, R, C++ e Java; Python é a muito popular.
- MATLAB é o único proprietário entre os softwares do guia.
- CUDA = NVIDIA/GPU; TensorFlow = Google; R Studio = Posit/IDE R; Scikit-Learn = Python/BSD.
- Exige-se familiaridade, não fluência nem proficiência.

Essenciais para a prova (checklist)

- ☐ Sei listar as quatro linguagens orientadas a objetos do item 2.1.
- ☐ Sei explicar o que é orientação a objetos e diferenciá-la de SQL e HTML.
- ☐ Sei descrever Python como interpretado, orientado a objetos e de alto nível.
- ☐ Sei para que o R foi criado e seu paradigma.

- ☐ Sei a função das bibliotecas segundo o guia e citar NumPy e pandas.
- ☐ Sei associar TensorFlow, R Studio, CUDA, Scikit-Learn e MATLAB à origem e à função.
- ☐ Sei separar código aberto de proprietário entre os softwares do guia.
- ☐ Sei o nível de domínio que o guia espera em linguagens e em softwares.

Questões de revisão do capítulo

R2.1 · Domínio 2: Codificação para aprendizado de máquina

Segundo o site oficial do projeto R, para que essa linguagem e ambiente foram criados?

A. Computação estatística e gráficos
B. Edição de planilhas eletrônicas
C. Desenvolvimento de páginas web
D. Programação de jogos para consoles

R2.2 · Domínio 2: Codificação para aprendizado de máquina

Bruno vai treinar uma rede neural grande e quer aproveitar as GPUs NVIDIA do servidor para acelerar o processamento. Que tecnologia citada no guia atende a esse objetivo?

A. Biblioteca pandas
B. Plataforma CUDA
C. Ambiente R Studio
D. Planilha Excel

R2.3 · Domínio 2: Codificação para aprendizado de máquina

Como a documentação oficial do Python descreve a linguagem?

A. Compilada, de marcação e voltada a páginas web
B. Declarativa, de consulta e restrita a bancos
C. Proprietária, de baixo nível e ligada ao hardware
D. Interpretada, orientada a objetos e de alto nível

R2.4 · Domínio 2: Codificação para aprendizado de máquina

Segundo o guia, qual linguagem é muito popular no aprendizado de máquina e na ciência de dados?

A. Python
B. Kotlin
C. COBOL
D. Swift

Respostas comentadas

R2.1. A) Computação estatística e gráficos

Por quê: O site r-project.org descreve o R como 'a language and environment for statistical computing and graphics', disponível como software livre sob a licença GNU GPL e ampliável por pacotes.

✗ B) Incorreta. O R não é um editor de planilhas; ele é uma linguagem de programação para análise de dados.

✗ C) Incorreta. O R não é voltado ao desenvolvimento de páginas web; seu foco é estatística e gráficos.

✗ D) Incorreta. Programar jogos para consoles não é a finalidade do R descrita no site oficial.

> **Dica de prova:** R nasceu para estatística e gráficos, por isso é tão usado com dados.

R2.2. B) Plataforma CUDA

Por quê: O guia cita o CUDA no item 2.2. A NVIDIA descreve o CUDA Toolkit como um ambiente de desenvolvimento para criar aplicativos de alto desempenho acelerados por GPU, com bibliotecas, compilador C/C++ e ferramentas de depuração.

✗ A) Incorreta. O pandas manipula tabelas de dados em Python e não é citado no guia nem acelera GPUs.

✗ C) Incorreta. O R Studio é um ambiente de desenvolvimento para R e não é a tecnologia que acelera o uso de GPUs.

✗ D) Incorreta. Uma planilha eletrônica não aproveita GPUs para treinar redes neurais e não é citada no guia.

> **Dica de prova:** GPU da NVIDIA no enunciado: a resposta tende a ser o CUDA.

R2.3. D) Interpretada, orientada a objetos e de alto nível

Por quê: Segundo python.org, 'Python is an interpreted, object-oriented, high-level programming language with dynamic semantics'. Isso confirma o Python entre as linguagens orientadas a objetos do item 2.1 do guia.

✗ A) Incorreta. O Python não é uma linguagem de marcação para páginas web; ele é interpretado e de propósito geral.

✗ B) Incorreta. Linguagens declarativas de consulta, como SQL, servem a bancos de dados; não é o caso do Python.

✗ C) Incorreta. O interpretador do Python é distribuído livremente, e a linguagem é de alto nível, e não de baixo nível.

> **Dica de prova:** Python: interpretada, orientada a objetos e de alto nível.

R2.4. A) Python

Por quê: O guia lista Python, R, C++ e Java no item 2.1 e destaca o Python como linguagem muito popular no aprendizado de máquina e na ciência de dados.

✗ B) Incorreta. Kotlin é mais conhecida no desenvolvimento de aplicativos e não é a linguagem que o guia destaca para o aprendizado de máquina.

✗ C) Incorreta. COBOL é tradicional em sistemas corporativos legados e não aparece no guia entre as linguagens de aprendizado de máquina.

✗ D) Incorreta. Swift é usada principalmente no desenvolvimento para plataformas da Apple e não é citada no guia.

> **Dica de prova:** Entre as linguagens do guia, a que ele chama de muito popular em aprendizado de máquina é o Python.

03

CAPÍTULO 3 · 30% DA PROVA

Domínio 3: Algoritmos usados no aprendizado de máquina

É o domínio de maior peso junto com o 4 e o que mais aparece em cada prova, então preciso dele afiado. Tenho que dominar três blocos: a matemática do item 3.1 (probabilidade/teorema de Bayes, estatística descritiva e inferencial, álgebra linear), os algoritmos do item 3.2 (regressão linear e polinomial; classificação com K-vizinhos, árvores de decisão e regressão logística; agrupamento com K-means e hierárquico; como desafio extra, impulsionamento, florestas de decisão e conjuntos) e os tipos de aprendizado do item 3.3 (supervisionado, não supervisionado e semi-supervisionado, com o reforço citado como complemento, comum em jogos). A prova não pede fórmulas: pede reconhecer qual algoritmo ou tipo de aprendizado serve a um cenário. Onde mais erro: o 'k' do K-means x o do K-vizinhos, a regressão logística (que classifica) e o semi-supervisionado com a proporção invertida. A pergunta-chave em quase tudo é: os dados têm rótulo e a saída é categoria, número ou grupo?

Neste capítulo

18 tópicos-chave · 14 pegadinhas · 8 macetes · 14 gatilhos · 4 questões comentadas

Tópicos-chave

Por que estudar a matemática

O guia pede compreensão básica da matemática do AM mesmo que o software faça os cálculos automaticamente. Não é essencial saber as fórmulas de cada algoritmo, mas a base ajuda a compreender os algoritmos e sua implementação.

Probabilidade e teorema de Bayes

Probabilidade mede a incerteza de um resultado (ex.: 80% de chance de fraude). O teorema de Bayes calcula uma probabilidade quando outras são conhecidas, como a chance de estar doente dado um exame positivo em doença rara, quando muitos positivos podem ser falsos.

Naïve Bayes

Classificador probabilístico baseado no teorema de Bayes que trata cada palavra como independente das demais, o que o torna simples e rápido. Escolha clássica para classificar e-mails como spam pela frequência das palavras.

Estatística descritiva

Descreve e resume os dados disponíveis: tendência central (média, mediana) e dispersão (desvio-padrão). Desvio-padrão alto indica valores mais espalhados em torno da média; ele não depende do tamanho da turma.

Estatística inferencial

Parte de uma amostra para tirar conclusões sobre a população de onde ela veio, como ouvir 1.000 clientes para concluir sobre 200 mil.

Correlação não é causalidade

Duas variáveis que variam juntas têm relação estatística, mas isso não prova que uma causa a outra; um terceiro fator, como a temperatura, pode influenciar as duas.

Álgebra linear

Um registro pode ser um vetor e uma tabela de registros, uma matriz. A transposição troca linhas por colunas: uma matriz 50 x 4 tem transposta 4 x 50. O guia diz que a álgebra linear dá a base para compreender o AM e implementar algoritmos.

Regressão linear

Algoritmo de regressão que prevê o valor de uma variável a partir de outra ajustando uma reta. O método dos mínimos quadrados encontra a reta de melhor ajuste.

Regressão polinomial

Algoritmo de regressão para relações em curva, como consumo de energia x temperatura, quando uma reta não acompanha os dados.

Regressão logística

Apesar do nome, está entre os algoritmos de classificação do guia. Devolve um valor entre 0 e 1 (probabilidade do evento, ex.: inadimplência sim/não) e mostra como cada variável influencia o resultado.

K-vizinhos mais próximos (k-nearest neighbors)

Classificador supervisionado que usa proximidade: a classe do novo registro é a mais comum entre os k vizinhos mais próximos. Aqui k = número de vizinhos consultados.

Árvore de decisão

Classificador supervisionado com nó raiz (pergunta inicial), ramos (ligações), nós internos de decisão e nós folha, que são os resultados possíveis. É treinada com dados rotulados.

K-means

Algoritmo de agrupamento não supervisionado: começa com K centroides e atribui cada ponto ao grupo mais próximo. K = quantidade de grupos a formar. Indicado para segmentar dados sem rótulo em um número definido de grupos.

Agrupamento hierárquico

Algoritmo de agrupamento não supervisionado que junta ou divide grupos sucessivamente, formando grupos aninhados mostrados em um dendrograma (diagrama em forma de árvore).

Florestas, impulsionamento e conjuntos

Desafio extra citado no guia. Floresta de decisão aleatória: várias árvores treinadas de forma independente; na classificação vence a votação da maioria e na regressão, a média. Impulsionamento (boosting): modelos simples treinados em sequência, cada um corrigindo os erros do anterior.

Aprendizado supervisionado

Aplica um algoritmo a dados rotulados, em que se sabe qual será o resultado (output). Cada exemplo de treino traz a resposta correta. Grandes tipos de problema: classificação (categorias) e regressão (valores).

Aprendizado não supervisionado

Aplica um algoritmo a dados não rotulados para descobrir padrões e grupos escondidos, como agrupamento por semelhança, associação (produtos comprados juntos) e detecção de registros atípicos. K-means e hierárquico são seus algoritmos no guia.

Semi-supervisionado e reforço

Semi-supervisionado: o treinamento começa com uma pequena quantidade de dados rotulados e depois introduz uma quantidade maior sem rótulo; útil porque rotular é caro e demorado. Reforço: o agente age por tentativa e erro, recebe recompensas e punições e aprende as melhores ações; o guia o cita como comum em jogos.

Macetes para memorizar

💡 Classificação no guia

KAL: K-vizinhos, Árvores de decisão, Logística.

💡 Os dois K

K-MEANS = K grupos (centroides); K-VIZINHOS = k vizinhos que votam ('diga-me com quem andas').

💡 Estatística

Descritiva descreve o que tenho; Inferencial infere o que não vi (amostra > população).

💡 Árvore de decisão

Raiz = começo, ramos = caminhos, folhas = resultado final.

💡 Tipos de aprendizado

Rótulo em tudo = supervisionado; rótulo em nada = não supervisionado; pouco rótulo e depois muito sem = semi; prêmio e castigo = reforço.

💡 Floresta x boosting

Floresta vota em paralelo; boosting corrige em fila.

Regressões

Linear = reta; Polinomial = curva; Logística = sim/não (é classificação).

Correlação

Ventilador e sorvete sobem juntos por causa do calor: andar junto não é causar.

Tabela de memorização: não confunda

Quando é...	...e quando é
K do K-means Número de grupos a formar em dados sem rótulo (agrupamento).	**k do K-vizinhos mais próximos** Número de vizinhos consultados para decidir a classe por maioria (classificação).
Estatística descritiva Resume os dados coletados: média, mediana, desvio-padrão.	**Estatística inferencial** Generaliza de uma amostra para a população.
Regressão linear Relação em linha reta; reta ajustada por mínimos quadrados.	**Regressão polinomial** Relação em curva.
Regressão logística Classifica sim/não devolvendo probabilidade entre 0 e 1.	**Regressão polinomial** Estima valores numéricos ao longo de uma curva.
Floresta de decisão aleatória Árvores independentes que votam (ou fazem média).	**Impulsionamento (boosting)** Modelos em sequência, cada um corrigindo os erros do anterior.
Agrupamento hierárquico Não supervisionado; grupos aninhados em dendrograma.	**Árvore de decisão** Supervisionado; classificação por perguntas em nós até as folhas.
Semi-supervisionado Poucos rotulados primeiro, muitos sem rótulo depois.	**Supervisionado** Todos os exemplos de treino rotulados.

Números que caem

Item	Valor
Peso do domínio no exame	**30% (6 de 20 pontos; 6 questões por prova)**
Peso do item 3.1 (matemática)	**8,8%**
Peso do item 3.2 (algoritmos)	**8,8%**
Peso do item 3.3 (tipos de aprendizado)	**13,3%**

Saída da regressão logística	**valor entre 0 e 1 (probabilidade)**
Transposta de uma matriz 50 x 4	**4 x 50**

Alertas de prova (pegadinhas)

⚠ **Alerta de prova:** Descartar a regressão logística em problema sim/não por causa da palavra 'regressão'.
Como não cair: Ela é algoritmo de classificação e devolve a probabilidade (0 a 1) do evento.

⚠ **Alerta de prova:** Achar que o K do K-means é o número de vizinhos, de rótulos ou de árvores.
Como não cair: K-means: K = grupos. Vizinhos é o k do K-vizinhos; árvores combinadas é floresta.

⚠ **Alerta de prova:** Dizer que o K-vizinhos decide pela média do treino inteiro ou pelo centro do grupo.
Como não cair: Ele usa a classe da maioria entre os k vizinhos mais próximos; centroide é coisa do K-means.

⚠ **Alerta de prova:** Responder 'árvore de decisão' quando aparece dendrograma.
Como não cair: Dendrograma = agrupamento hierárquico (não supervisionado); árvore de decisão classifica com rótulo.

⚠ **Alerta de prova:** Marcar K-means ou hierárquico como classificação de categorias já definidas.
Como não cair: São algoritmos de agrupamento; classificação no guia = K-vizinhos, árvores de decisão e regressão logística.

⚠ **Alerta de prova:** Inverter a proporção do semi-supervisionado (muito rotulado, pouco sem rótulo).
Como não cair: Semi = pequena parte rotulada primeiro, grande volume sem rótulo depois.

⚠ **Alerta de prova:** Chamar de não supervisionado um caso com poucos exemplos rotulados por especialistas mais muitos sem rótulo.
Como não cair: Se há qualquer parte rotulada no início seguida de muitos sem rótulo, é semi-supervisionado.

⚠ **Alerta de prova:** Chamar de semi-supervisionado um caso sem nenhuma marcação nos dados.
Como não cair: Nenhum rótulo = não supervisionado.

⚠ **Alerta de prova:** Concluir que 'A causa B' porque as duas variáveis sobem juntas, ou que não há relação estatística alguma.
Como não cair: Há correlação (relação estatística), mas correlação não prova causa e efeito.

⚠ **Alerta de prova:** Interpretar desvio-padrão maior como notas mais altas, turma maior ou menos erros de registro.
Como não cair: Desvio-padrão mede só dispersão: maior = valores mais espalhados.

⚠ **Alerta de prova:** Atribuir à álgebra linear a generalização amostra > população ou o cálculo de chance.
Como não cair: Álgebra linear = vetores e matrizes; amostra > população = inferencial; chance = probabilidade/Bayes.

⚠ **Alerta de prova:** Na transposta de 50 x 4, manter 50 x 4 ou criar 50 x 50 ou 4 x 4.
Como não cair: Transpor inverte as dimensões: 4 linhas e 50 colunas, mantendo todos os valores.

⚠ **Alerta de prova:** Dizer que floresta decide por uma única árvore profunda, ou confundir floresta com boosting.
Como não cair: Floresta usa muitas árvores independentes e votação; boosting é sequência corrigindo erros.

⚠ **Alerta de prova:** Achar que a diferença entre supervisionado e não supervisionado está na linguagem, na quantidade de algoritmos ou no uso de dados de teste.
Como não cair: A diferença é a presença ou não de rótulos nos dados de treino.

Como cai na prova

📝 **Como cai:** Cenário de negócio pedindo o algoritmo (crédito aprovado/recusado, alunos em quatro grupos, consumo em curva).
Como resolver: 1) Os dados têm rótulo? 2) A saída é categoria, número ou grupo? Categoria = KNN/ árvore/logística; número = linear (reta) ou polinomial (curva); grupo sem rótulo = K-means ou hierárquico.

📝 **Como cai:** Cenário pedindo o tipo de aprendizado (raios X marcados, transações sem marcação, 300 rotulados + 40 mil sem rótulo, xadrez contra si mesmo).
Como resolver: Conte os rótulos: todos = supervisionado; nenhum = não supervisionado; poucos primeiro e muitos sem depois = semi; recompensa por ação = reforço.

📝 **Como cai:** Qual ramo da matemática ou estatística (vetores, 80% de chance, amostra, média e desvio).
Como resolver: Estrutura de dados = álgebra linear; incerteza = probabilidade/Bayes; resumir = descritiva; generalizar = inferencial.

📝 **Como cai:** Como o algoritmo funciona por dentro (KNN, folhas, K do K-means, floresta, mínimos quadrados).
Como resolver: Cada distrator é o mecanismo de outro algoritmo (centroide, votação, nós e folhas, reta). Associe o mecanismo ao dono e elimine.

📝 **Como cai:** Escolher DUAS ou 'qual NÃO é' do não supervisionado.
Como resolver: Marque exatamente duas; em 'qual NÃO é', procure a única opção que usa dados com resposta conhecida (rotulados).

Gatilhos do enunciado

Se o enunciado trouxer...	Pense em
sim/não com probabilidade e peso de cada variável	**regressão logística**
jogo, recompensa, punição, tentativa e erro	**aprendizado por reforço**
relação em curva	**regressão polinomial**
reta de melhor ajuste	**regressão linear com mínimos quadrados**
vetor, matriz, linhas e colunas, transposta	**álgebra linear**
amostra para concluir sobre a população	**estatística inferencial**
média, mediana, desvio-padrão para resumir	**estatística descritiva**
chance de A dado B a partir de probabilidades conhecidas; doença rara e exame positivo	**teorema de Bayes**
spam, frequência de palavras	**Naïve Bayes**
dendrograma, grupos aninhados	**agrupamento hierárquico**
dados sem rótulo divididos em N grupos	**K-means (não supervisionado)**
classe da maioria entre os registros mais próximos	**K-vizinhos mais próximos**
poucos rotulados por especialistas e depois milhares sem rótulo	**semi-supervisionado**
modelos em sequência corrigindo erros do anterior	**impulsionamento (boosting)**

🎯 **Se você lembrar de só uma coisa deste capítulo...**

- Regressão logística classifica, apesar do nome.
- K-means: K = grupos; K-vizinhos: k = vizinhos.
- Semi-supervisionado: pouco rotulado primeiro, muito sem rótulo depois.
- Dendrograma é agrupamento hierárquico, não árvore de decisão.
- Correlação não prova causa; desvio-padrão só mede dispersão.

Essenciais para a prova (checklist)

☐ Sei explicar para que servem probabilidade/teorema de Bayes, estatística descritiva, inferencial e álgebra linear.

☐ Sei transpor as dimensões de uma matriz e o que é vetor e matriz em dados.

☐ Sei listar os algoritmos de regressão, classificação e agrupamento do item 3.2.

☐ Sei como o K-vizinhos decide a classe e o que representa o K do K-means.

☐ Sei as partes de uma árvore de decisão e o que é um dendrograma.

☐ Sei diferenciar floresta de decisão, impulsionamento e conjuntos.

☐ Sei identificar supervisionado, não supervisionado, semi-supervisionado e reforço pelo tipo de dado.

☐ Sei que supervisionado resolve classificação e regressão, e não supervisionado resolve agrupamento e associação.

☐ Sei explicar por que correlação não é causalidade.

Questões de revisão do capítulo

R3.1 · Domínio 3: Algoritmos usados no aprendizado de máquina

O K-means e o agrupamento hierárquico são algoritmos usados em que tipo de aprendizado?

A. Supervisionado, com dados rotulados
B. Não supervisionado, com dados sem rótulo
C. Por reforço, com recompensas e punições a cada ação
D. Semi-supervisionado, com rótulos de especialistas

R3.2 · Domínio 3: Algoritmos usados no aprendizado de máquina

Qual é o objetivo típico do aprendizado não supervisionado?

A. Prever um valor já conhecido a partir de exemplos
B. Maximizar a recompensa de um agente em um jogo
C. Descobrir padrões e grupos escondidos nos dados
D. Classificar registros em categorias definidas

R3.3 · Domínio 3: Algoritmos usados no aprendizado de máquina

Uma clínica veterinária treina um modelo com milhares de raios X já marcados por veterinários como 'fratura' ou 'sem fratura'. Que tipo de aprendizado é esse?

A. Aprendizado por reforço
B. Aprendizado semi-supervisionado
C. Aprendizado supervisionado
D. Aprendizado não supervisionado

R3.4 · Domínio 3: Algoritmos usados no aprendizado de máquina

Como o algoritmo K-vizinhos mais próximos decide a classe de um novo registro?

A. Pela classe mais comum entre os registros mais próximos
B. Pela média de todos os registros do conjunto de treino
C. Pelo centro do grupo formado com dados sem rótulo
D. Pela sequência de perguntas de uma árvore de decisão

Respostas comentadas

R3.1. B) Não supervisionado, com dados sem rótulo

Por quê: O guia classifica K-means e hierárquico como algoritmos de agrupamento (clustering). O agrupamento é o exemplo de problema do aprendizado não supervisionado.

✗ A) Incorreta. Algoritmos com dados rotulados, como a regressão logística, são de aprendizado supervisionado.
✗ C) Incorreta. O reforço aprende com recompensas e punições, e não agrupando registros.
✗ D) Incorreta. O semi-supervisionado começa com poucos dados rotulados; K-means e hierárquico agrupam dados sem rótulo.

> **Dica de prova:** Algoritmo de agrupamento = aprendizado não supervisionado.

R3.2. C) Descobrir padrões e grupos escondidos nos dados

Por quê: Sem rótulos, o aprendizado não supervisionado busca padrões e grupos por semelhança, como no agrupamento (clustering). Não há um resultado desejado definido previamente.

✗ A) Incorreta. Prever um valor com exemplos de resposta conhecida é aprendizado supervisionado.
✗ B) Incorreta. Maximizar a recompensa de um agente é o objetivo do aprendizado por reforço.
✗ D) Incorreta. Classificar em categorias definidas exige dados rotulados, o que é aprendizado supervisionado.

> **Dica de prova:** Não supervisionado = descobrir estrutura nos dados, não prever resposta conhecida.

R3.3. C) Aprendizado supervisionado

Por quê: No aprendizado supervisionado, o algoritmo é aplicado a dados rotulados, em que se sabe qual será o resultado. Imagens marcadas por especialistas são um exemplo típico.

✗ A) Incorreta. O aprendizado por reforço aprende por recompensas e punições ao agir, e não com imagens rotuladas.
✗ B) Incorreta. O semi-supervisionado combina poucos dados rotulados com muitos sem rótulo, e aqui todos estão rotulados.
✗ D) Incorreta. O aprendizado não supervisionado trabalha com dados sem rótulo, e aqui cada imagem já está marcada.

> **Dica de prova:** Todos os exemplos com rótulo = supervisionado.

R3.4. A) Pela classe mais comum entre os registros mais próximos

Por quê: O K-vizinhos mais próximos é um classificador supervisionado que usa a proximidade: pontos parecidos ficam perto uns dos outros. A classe do novo registro é decidida pela maioria entre os k vizinhos.

✗ B) Incorreta. O algoritmo usa só os vizinhos mais próximos, e não a média do conjunto inteiro.
✗ C) Incorreta. Usar o centro de grupos sem rótulo descreve o K-means, um algoritmo de agrupamento.
✗ D) Incorreta. A sequência de perguntas em nós descreve uma árvore de decisão, e não o K-vizinhos mais próximos.

> **Dica de prova:** K-vizinhos = 'diga-me com quem andas': vence a classe da maioria dos vizinhos.

04

CAPÍTULO 4 · 30% DA PROVA

Domínio 4: Aprendizado de máquina na prática

Aqui eu aplico o processo de ponta a ponta: identificar o problema e os dados (4.1), preparar os dados (4.2), treinar (4.3), testar (4.4) e avaliar e comunicar os resultados às partes interessadas (4.5). É o domínio dos cenários longos com personagem e, como no exame simulado oficial, é onde ficam as duas questões de cenário que valem 2 pontos cada, então errar aqui custa caro. A prova quer que eu reconheça o nome certo da tarefa: limpeza, validação, transformação, redução, importação, ajuste, conjunto, teste estatístico. Onde mais erro: confundir validação com transformação, ajuste com conjunto, e na comunicação escolher a opção técnica demais em vez da que fala com o público. Sobreajuste (ótimo no treino, fraco no teste) também aparece.

Neste capítulo
18 tópicos-chave · 14 pegadinhas · 6 macetes · 14 gatilhos · 4 questões comentadas

Tópicos-chave

Identificação do problema (4.1)

Bom problema de AM liga uma dor clara do negócio a dados que ajudam a resolvê-la, como prever quais elevadores vão falhar com sensores e histórico de quebras. Problemas de regra fixa ou cálculo exato (aplicar desconto legal, somar vendas, ordenar nomes) não precisam de AM.

Requisitos para a coleta de dados

Os dados precisam representar todos os casos e resultados que o modelo vai encontrar: clientes que cancelaram e que ficaram; folhas doentes e saudáveis em condições variadas. Mais do mesmo tipo de dado não ensina o caso que falta.

Propor a solução com AM

A solução deve usar os dados disponíveis para atacar o problema do cenário, como classificar os alunos atuais pelo risco de desistência a partir do histórico de quem desistiu.

Preparação de dados (4.2)

Conteúdo indicativo: pré-processamento, transformação e importação/carregamento de dados. A orientação cita limpeza, validação e transformação para deixar os dados em formato adequado ao software escolhido.

Limpeza de dados

Trata valores ausentes, preenchendo-os com uma estatística da coluna como média ou mediana (sem inventar valores nem usar o máximo), e remove duplicidades, mantendo um registro de cada.

Validação de dados

Confere se os valores estão dentro do esperado: faixas, formatos e regras dos campos, como idade entre 0 e 120 anos ou CEP com oito dígitos.

Transformação de dados

Deixa os dados em formato adequado para o software escolhido: converter libras em quilos, texto em número, unificar padrões.

Redução (remover colunas irrelevantes)

Retirar colunas sem relação com o problema (cor do carro, nome do porteiro) é uma redução no pré-processamento, que acelera o processamento e evita piorar o modelo.

Importação/carregamento de dados

Tarefa que leva os dados já limpos e transformados para dentro do software em que o modelo será treinado.

Requisitos para o treinamento (4.3)

Conteúdo indicativo: requisitos para o treinamento, configuração de compartimentos de treinamento para dados, seleção de um algoritmo, regras e tipo de aprendizado (supervisionado, não supervisionado, semi-supervisionado). Para começar a treinar bastam dados preparados (e rotulados, no supervisionado), compartimentos e algoritmo escolhido; teste estatístico, relatório e apresentação vêm depois.

Compartimentos de treinamento

Separação dos dados preparados em uma parte para o modelo aprender e outra, que ele não viu, para testá-lo. Não é divisão por linguagem, por pessoa ou cópia de segurança.

Seleção do algoritmo

Depende do tipo de aprendizado e do resultado esperado: dados rotulados com categoria = supervisionado com classificação; valor numérico com rótulo = regressão; sem rótulo = agrupamento.

Atividades de teste (4.4)

Teste, ajuste, conjuntos, teste estatístico e revisão. O guia sugere testar e comparar vários algoritmos para saber se o modelo está funcionando bem.

Ajuste

Quando o modelo fica abaixo do esperado, alteram-se as configurações do algoritmo, treina-se de novo e compara-se o novo resultado no mesmo compartimento de teste.

Conjuntos (ensemble)

Combinação dos resultados de vários modelos em uma resposta, porque juntos costumam acertar mais do que cada um isolado.

Teste estatístico

Verifica se a diferença de desempenho entre modelos é real ou se deve ao acaso. Não corrige rótulos nem cria dados.

Sobreajuste (overfitting)

O modelo vai muito bem no treino e mal no teste com dados que não viu: decorou o treino e precisa de ajuste.

Avaliar e comunicar resultados (4.5)

Informações relevantes às partes interessadas: o que foi aprendido, se foi possível resolver o problema, qual é o próximo passo e o aprendizado com a experiência; depois, tirar conclusões. Considerações: tipo de informação, apresentação, linguagem, uso de termos técnicos e estar preparado para responder a perguntas, adaptando o detalhe a cada público.

Macetes para memorizar

Fluxo do domínio

Problema > Dados > Prepara > Treina > Testa > Conta: itens 4.1 a 4.5 na ordem.

Preparação

Limpa (vazios e duplicados), Valida (regras), Transforma (formato), Carrega (para o software).

Atividades de teste

TACTR: Teste, Ajuste, Conjuntos, Teste estatístico, Revisão.

Partes interessadas

3 perguntas + 1: aprendi o quê? Resolveu? E agora (próximo passo)? + o aprendizado com a experiência.

Sobreajuste

Decorou a prova de casa e foi mal na prova de verdade.

Saída do modelo

Número = regressão; categoria = classificação; grupo sem rótulo = agrupamento.

Tabela de memorização: não confunda

Quando é...	...e quando é
Validação de dados Conferir se os valores respeitam faixas, formatos e regras (idade 0 a 120, CEP com oito dígitos).	**Transformação de dados** Converter para o formato exigido pelo software (libras em quilos, texto em número).
Limpeza de dados Tratar ausentes e duplicidades.	**Importação/carregamento** Levar os dados preparados para dentro do software.
Ajuste Mudar configurações de um modelo e testar de novo.	**Conjunto (ensemble)** Combinar vários modelos em uma resposta.
Teste estatístico Dizer se a diferença entre modelos é real ou acaso.	**Ajuste** Melhorar um modelo alterando suas configurações.
Regressão Resultado é um valor numérico (aluguel em reais).	**Classificação** Resultado é categoria (elogio, reclamação, dúvida).
Comunicação com a diretoria Impacto no negócio, linguagem simples, poucos termos técnicos.	**Comunicação com a equipe técnica** Pode receber mais detalhes técnicos; adapte o nível de detalhe ao público.

Números que caem

Item	Valor
Peso do domínio no exame	**30% (6 de 20 pontos; 4 questões por prova, 2 delas de cenário)**
Questões de cenário	**2 por exame, valendo 2 pontos cada**
Peso dos itens 4.1, 4.2 e 4.5	**7% cada**
Peso dos itens 4.3 e 4.4	**4,5% cada**

Alertas de prova (pegadinhas)

⚠ **Alerta de prova:** Preencher vazios com números sorteados, com o maior valor da coluna, ou treinar com os vazios.
Como não cair: Na limpeza, use um valor coerente da própria coluna, como média ou mediana.

⚠ **Alerta de prova:** Manter registros duplicados 'para ter mais dados' ou excluir os clientes repetidos.
Como não cair: Remova a duplicidade na limpeza, mantendo um registro de cada cliente.

⚠ **Alerta de prova:** Resolver unidades diferentes trocando o algoritmo ou visualizando em painel.
Como não cair: Formato ou unidade incompatível com o software = transformação de dados.

⚠ **Alerta de prova:** Chamar de importação ou visualização a conferência de faixas e formatos.
Como não cair: Conferir regras dos campos é validação.

⚠ **Alerta de prova:** Treinar com todas as colunas e decidir depois, ou duplicar colunas para ter mais variáveis.
Como não cair: Colunas sem relação com o problema saem no pré-processamento (redução).

⚠ **Alerta de prova:** Achar que compartimentos dividem o trabalho da equipe ou guardam cópias de segurança.
Como não cair: Compartimentos = dados para aprender + dados para testar.

⚠ **Alerta de prova:** Exigir relatório aprovado ou teste estatístico antes de começar a treinar.
Como não cair: Para treinar bastam dados preparados (e rotulados) e algoritmo escolhido; o resto vem depois.

⚠ **Alerta de prova:** Escolher classificação porque o imóvel tem bairro, ou agrupamento porque 'não há rótulo'.
Como não cair: O bairro é entrada; o valor conhecido do aluguel é o rótulo e a saída numérica pede regressão.

⚠ **Alerta de prova:** Explicar 98% no treino e 61% no teste por excesso de dados de teste ou falta de rótulos.
Como não cair: Ótimo no treino e fraco no teste = sobreajuste.

⚠ **Alerta de prova:** Chamar de 'conjunto' o ato de mudar configurações e testar de novo.
Como não cair: Isso é ajuste; conjunto é combinar vários modelos.

⚠ **Alerta de prova:** Usar termos técnicos, fórmulas ou código-fonte para impressionar o diretor não técnico.
Como não cair: Mostre o ganho no negócio em linguagem simples e esteja pronto para perguntas.

⚠ **Alerta de prova:** Omitir a meta não atingida ou mostrar só números positivos.
Como não cair: Comunique o que foi aprendido, o resultado obtido e o próximo passo, mesmo abaixo da meta.

⚠ **Alerta de prova:** Coletar mais dados do mesmo tipo (só folhas saudáveis, só clientes ativos) ou trocar dados por opinião.
Como não cair: A coleta precisa cobrir todos os resultados possíveis e condições variadas.

⚠ **Alerta de prova:** Propor solução que não usa os dados do problema (recomendar filmes, gerar materiais, agrupar professores).
Como não cair: Escolha a solução que ataca a dor descrita com os dados disponíveis.

Como cai na prova

📝 **Como cai:** Cenário longo com personagem e um problema de dados ('que tarefa de preparação resolve?').
Como resolver: Sublinhe o sintoma: vazio/duplicado = limpeza; regra de campo = validação; unidade ou formato = transformação; coluna irrelevante = remoção/redução; levar ao software = importação.

📝 **Como cai:** Cenário de teste ('que atividade ele está realizando?' ou 'qual a explicação mais provável?').
Como resolver: Mudou configurações e testou de novo = ajuste; juntou modelos = conjunto; diferença pequena entre modelos = teste estatístico; treino alto e teste baixo = sobreajuste.

📝 **Como cai:** Cenário de negócio pedindo o problema, os dados ou a solução (escola, elevadores, assinaturas).
Como resolver: Identifique a dor, veja que dados com desfecho existem e escolha a opção que usa esses dados para prever ou classificar; descarte soluções desconectadas do problema.

📝 **Como cai:** Cenário de comunicação a diretor ou gerente.
Como resolver: Escolha a alternativa que fala do resultado no negócio em linguagem simples, mostra o que foi aprendido, se resolveu e o próximo passo; elimine fórmulas, código, bibliotecas e omissões.

📝 **Como cai:** Escolha do tipo de algoritmo para o treinamento.
Como resolver: Pergunte: há rótulo? A saída é número ou categoria? Rótulo + número = regressão; rótulo + categoria = classificação supervisionada.

Gatilhos do enunciado

Se o enunciado trouxer...	Pense em
dor clara + histórico com o desfecho (falhou/cancelou/desistiu)	**prever ou classificar o risco com aprendizado supervisionado**

regra fixa, soma, ordenação	**não é problema de AM**
dados só de um tipo de resultado	**requisito de coleta: incluir todos os casos e condições variadas**
campos vazios, registros repetidos	**limpeza de dados**
faixa de valores, quantidade de dígitos, regra do campo	**validação de dados**
unidades diferentes, número gravado como texto, formato exigido pelo software	**transformação de dados**
colunas sem relação com o problema	**remover no pré-processamento (redução)**
levar os dados preparados para o software	**importação/carregamento**
separar parte para aprender e parte para testar	**compartimentos de treinamento**
mudou configurações, retreinou e testou de novo	**ajuste**
combinar vários modelos em uma resposta	**conjunto (ensemble)**
desempenhos parecidos, diferença real ou acaso	**teste estatístico**
acerto alto no treino e baixo no teste	**sobreajuste**
apresentar a público não técnico	**impacto no negócio, linguagem simples, pronto para perguntas**

🎯 **Se você lembrar de só uma coisa deste capítulo...**

- As 2 questões de cenário de 2 pontos ficam neste domínio, como no exame simulado oficial: leia o cenário inteiro.
- Validação confere regras; transformação muda formato; limpeza trata vazios e duplicados.
- Ajuste = mexer em configurações e retestar; conjunto = combinar modelos.
- Ótimo no treino e fraco no teste = sobreajuste.
- Partes interessadas: o que foi aprendido, se resolveu, próximo passo, em linguagem do público.

Essenciais para a prova (checklist)

- ☐ Sei reconhecer um bom problema de AM e descartar regras fixas e cálculos exatos.
- ☐ Sei definir requisitos de coleta que cubram todos os resultados possíveis.
- ☐ Sei diferenciar limpeza, validação, transformação, redução e importação.
- ☐ Sei os requisitos para iniciar o treinamento e para que servem os compartimentos.
- ☐ Sei escolher entre regressão, classificação e agrupamento pela saída e pelos rótulos.
- ☐ Sei as cinco atividades de teste do item 4.4 e diferenciar ajuste, conjunto e teste estatístico.

- ☐ Sei identificar sobreajuste pela diferença entre treino e teste.
- ☐ Sei quais informações levar às partes interessadas e como adaptar a comunicação ao público.

Questões de revisão do capítulo

R4.1 · Domínio 4: Aprendizado de máquina na prática

Sérgio testou uma árvore de decisão que prevê falhas em máquinas e obteve resultados abaixo da meta. Ele muda as configurações do algoritmo, treina o modelo de novo e compara os novos resultados com os anteriores no mesmo compartimento de teste. Ele repete isso até o resultado melhorar. Que atividade do teste Sérgio está realizando?

A. Importação de um novo conjunto de dados
B. Validação dos dados de entrada
C. Ajuste do modelo, seguido de novo teste
D. Comunicação dos resultados às partes

R4.2 · Domínio 4: Aprendizado de máquina na prática

Rafael treinou um modelo para prever atrasos de voos. No compartimento de treinamento, o modelo acertou 98% dos casos. No compartimento de teste, com voos que ele não tinha visto, acertou 61%. Rafael usou os mesmos tipos de dados nos dois compartimentos. Qual é a explicação mais provável para essa diferença?

A. Sobreajuste do modelo aos dados de treino
B. Erro na escolha do tipo de visualização
C. Falta de rótulos nos dados de treinamento
D. Excesso de dados no compartimento de teste

R4.3 · Domínio 4: Aprendizado de máquina na prática

Que requisito é necessário para iniciar o treinamento de um modelo supervisionado?

A. Resultados do teste estatístico do modelo final
B. Dados preparados e rotulados e um algoritmo escolhido
C. Um relatório final aprovado pelas partes interessadas
D. Uma apresentação com os gráficos de resultado

R4.4 · Domínio 4: Aprendizado de máquina na prática

Na preparação dos dados, a equipe confere se as idades estão entre 0 e 120 anos e se cada CEP tem oito dígitos. Que tarefa é essa?

A. Importação de dados
B. Visualização de dados
C. Seleção de algoritmo
D. Validação dos dados

Respostas comentadas

R4.1. C) Ajuste do modelo, seguido de novo teste

Por quê: O item 4.4 inclui o ajuste entre as atividades de teste. Quando o modelo fica abaixo do esperado, alteram-se as configurações e o modelo é treinado e testado de novo.

✗ A) Incorreta. Sérgio usa os mesmos dados de teste; não importou um conjunto novo.
✗ B) Incorreta. A validação confere os dados na preparação, e não as configurações do algoritmo.
✗ D) Incorreta. A comunicação vem depois, quando os resultados já foram avaliados.

> **Dica de prova:** Mexeu nas configurações e testou de novo = ajuste.

R4.2. A) Sobreajuste do modelo aos dados de treino

Por quê: O teste com dados que o modelo não viu mostra se ele funciona bem em casos novos. Quando o acerto no treino é muito maior do que no teste, o modelo decorou o treino (sobreajuste) e precisa de ajuste.

✗ B) Incorreta. A visualização ajuda a analisar os dados, mas não muda o desempenho do modelo nos testes.
✗ C) Incorreta. Sem rótulos, Rafael nem conseguiria medir os 98% de acerto no treino.
✗ D) Incorreta. Ter muitos dados de teste torna a avaliação mais confiável; não explica a queda de desempenho.

> **Dica de prova:** Ótimo no treino e fraco no teste = sobreajuste (overfitting).

R4.3. B) Dados preparados e rotulados e um algoritmo escolhido

Por quê: Entre os requisitos para o treinamento (item 4.3) estão os dados preparados, os compartimentos de treino e a seleção do algoritmo. Relatórios, testes e apresentações vêm depois.

✗ A) Incorreta. O teste estatístico faz parte do teste, que só acontece depois do treinamento.
✗ C) Incorreta. O relatório final é produzido depois do teste e da avaliação dos resultados.
✗ D) Incorreta. A apresentação dos resultados é uma etapa posterior, de comunicação.

> **Dica de prova:** Treinar exige dados prontos e algoritmo escolhido; o resto vem depois.

R4.4. D) Validação dos dados

Por quê: O guia cita limpeza, validação e transformação entre as tarefas de preparação de dados. Validar é garantir que os valores estão dentro do esperado antes de usar os dados.

✗ A) Incorreta. A importação carrega os dados no software, sem checar se os valores são válidos.

✗ B) Incorreta. A visualização mostra os dados em gráficos; não confere regras de cada campo.

✗ C) Incorreta. A seleção de algoritmo acontece no treinamento, e não na conferência dos dados.

> **Dica de prova:** Conferir faixas, formatos e regras dos campos = validação.

DIA DA PROVA

Estratégia de prova

O exame testa os níveis Bloom 1 e 2: lembrar e compreender, sem fórmulas nem cálculo pesado. As alternativas certas repetem quase literalmente o conteúdo indicativo e as orientações do guia oficial, e as erradas costumam ser a definição correta de outro item (sistema baseado em conhecimento no lugar de AM, agrupamento no lugar de classificação, CUDA no lugar de TensorFlow). Aparecem perguntas negativas ('qual NÃO', 'qual contraria o guia') e de 'escolher 2 respostas'. No Domínio 4 ficam os cenários longos com personagem; no exame simulado oficial, as 2 questões de cenário que valem 2 pontos cada são desse domínio.

Como responder

✓ Gerencie o tempo: são 18 questões em 30 minutos, menos de 2 minutos por questão; responda primeiro as diretas e reserve fôlego para os 2 cenários de 2 pontos.

✓ Lembre da conta: aprovação com 13 de 20 pontos (65%); os 2 cenários valem 4 pontos juntos, então leia-os com calma.

✓ Em toda questão de algoritmo ou tipo de aprendizado, faça as duas perguntas: os dados têm rótulo? A saída é categoria, número ou grupo?

✓ Nas questões 'escolha DUAS', marque exatamente duas e confira cada uma isoladamente contra o guia.

✓ Em 'qual NÃO' ou 'qual contraria o guia', procure a única alternativa invertida ou vinda de outro item; as outras três repetem o guia.

✓ Quando a alternativa descreve algo verdadeiro, mas de outro conceito (ex.: mecanismo de inferência para reconhecer imagens), elimine: o erro está no dono da definição.

✓ Nos cenários do Domínio 4, sublinhe o sintoma (vazio, unidade, faixa, treino x teste, público) e só depois leia as alternativas.

✓ Na comunicação com partes interessadas, prefira sempre linguagem simples, resultado no negócio e próximo passo; elimine fórmulas, código e omissões.

Erros que derrubam candidatos

✗ Descartar a regressão logística em problema sim/não por causa do nome 'regressão'.

✗ Confundir o K do K-means (grupos) com o k do K-vizinhos (vizinhos).

✗ Inverter AM (aprende com dados) e sistema baseado em conhecimento (regras humanas).

✗ Trocar recomendações (Netflix, Spotify) por IA generativa (ChatGPT, Copilot).

✗ Achar que o guia exige fluência em linguagens ou proficiência em softwares.

✗ Confundir validação com transformação, e ajuste com conjunto.

✗ Escolher a resposta técnica demais quando o enunciado fala de diretor ou gerente não técnico.

Na semana da prova

☐ Releia as listas do guia de memória: seis aplicações, quatro linguagens, cinco softwares, algoritmos por tipo, cinco subetapas do pré-processamento e cinco atividades de teste.

☐ Refaça simulados completos cronometrando 30 minutos e anote os pares que você ainda confunde.

☐ Revise as pegadinhas e os gatilhos de cada domínio, com atenção especial aos Domínios 3 e 4, que somam 60% do peso.

☐ Treine leitura de cenários longos: identificar o sintoma e o nome da tarefa em menos de 2 minutos.

☐ Na véspera, revise só os macetes e os itens 'lembrar'; durma bem, pois o exame é sem consulta e sem anotações.

REVISÃO FINAL

Folha de cola

Leia na véspera e na manhã da prova: os gatilhos e as pegadinhas mais importantes de cada domínio.

1. Domínio 1: O que é aprendizado de máquina?

‣ qual frase contraria o guia sobre IA e AM → **a que diz que a IA é nova ou nasceu com o AM**

‣ regras escritas por especialistas, base de conhecimento → **sistema baseado em conhecimento**

‣ componente que aplica as regras → **mecanismo de inferência**

‣ tarefa específica, autônoma, melhora a cada repetição; chatbot, reconhecimento facial → **agente de aprendizagem**

‣ grandes redes neurais + grandes quantidades de dados, problema complexo → **aprendizagem profunda**

⚠ A IA veio antes; o AM é mais uma etapa da evolução da IA e um subconjunto dela.

⚠ Os papéis estão invertidos: AM aprende com dados; sistema baseado em conhecimento aplica regras humanas.

⚠ A base guarda; quem aplica as regras é o mecanismo de inferência.

2. Domínio 2: Codificação para aprendizado de máquina

‣ linguagem muito popular em AM e ciência de dados → **Python**

‣ interpretada, orientada a objetos, alto nível → **Python**

‣ computação estatística e gráficos, licença GNU GPL → **R**

‣ funcional com suporte à orientação a objetos → **R**

‣ GPU NVIDIA, computação paralela, compilador C/C++ → **CUDA**

⚠ A lista do item 2.1 tem só quatro: Python, R, C++ e Java.

⚠ Python é interpretado, orientado a objetos e de alto nível.

⚠ Eles trazem modularidade e reutilização, e o interpretador é distribuído livremente.

3. Domínio 3: Algoritmos usados no aprendizado de máquina

‣ sim/não com probabilidade e peso de cada variável → **regressão logística**

‣ jogo, recompensa, punição, tentativa e erro → **aprendizado por reforço**

‣ relação em curva → **regressão polinomial**

‣ reta de melhor ajuste → **regressão linear com mínimos quadrados**

‣ vetor, matriz, linhas e colunas, transposta → **álgebra linear**

⚠ Ela é algoritmo de classificação e devolve a probabilidade (0 a 1) do evento.

⚠ K-means: K = grupos. Vizinhos é o k do K-vizinhos; árvores combinadas é floresta.

⚠ Ele usa a classe da maioria entre os k vizinhos mais próximos; centroide é coisa do K-means.

4. Domínio 4: Aprendizado de máquina na prática

‣ dor clara + histórico com o desfecho (falhou/cancelou/desistiu) → **prever ou classificar o risco com aprendizado supervisionado**

‣ regra fixa, soma, ordenação → **não é problema de AM**

‣ dados só de um tipo de resultado → **requisito de coleta: incluir todos os casos e condições variadas**

‣ campos vazios, registros repetidos → **limpeza de dados**

‣ faixa de valores, quantidade de dígitos, regra do campo → **validação de dados**

⚠ Na limpeza, use um valor coerente da própria coluna, como média ou mediana.

⚠ Remova a duplicidade na limpeza, mantendo um registro de cada cliente.

⚠ Formato ou unidade incompatível com o software = transformação de dados.

GABARITO

Respostas do diagnóstico

D1. B) Um agente de aprendizagem autônomo → revise o capítulo 1

Por quê: Agentes de aprendizagem executam uma tarefa específica com dados, de forma autônoma, e melhoram com a repetição. O guia dá como exemplos chatbots, carros sem motorista e reconhecimento facial.

✗ A) Incorreta. Um sistema de regras fixas não melhora sozinho com novos acessos; ele depende de regras escritas previamente.

✗ C) Incorreta. Um banco de dados guarda as imagens, mas não identifica pessoas nem melhora com a experiência.

✗ D) Incorreta. Um roteiro programado repete os mesmos passos e não aprende com os novos acessos.

> **Dica de prova:** Tarefa única + melhora a cada repetição + autonomia = agente de aprendizagem.

D2. C) Aplicar algoritmos aos dados para obter compreensão (insight) → revise o capítulo 1

Por quê: O aprendizado de máquina aplica algoritmos a dados fornecidos para obter compreensão (insight). Por isso é usado na ciência de dados, inclusive com grandes volumes de dados (Big Data).

✗ A) Incorreta. Guardar os registros é tarefa de armazenamento de dados; não é o que caracteriza o aprendizado de máquina.

✗ B) Incorreta. Compactar arquivos economiza espaço em disco, mas não gera compreensão sobre os dados.

✗ D) Incorreta. O aprendizado de máquina depende dos dados; trocar a coleta por regras da gerência vai na direção contrária.

> **Dica de prova:** Big Data é a matéria-prima; o aprendizado de máquina é o meio de extrair insight dela com algoritmos.

D3. D) Revisão por pares e identificação do melhor modelo → revise o capítulo 1

Por quê: No item 1.7, a revisão reúne a revisão por pares, o aprendizado a partir de vários algoritmos e a identificação do melhor modelo. Depois dela, se o desempenho for satisfatório, o modelo pode ser implantado.

✗ A) Incorreta. Limpeza, integração e transformação são subetapas do pré-processamento de dados.

✗ B) Incorreta. Gráficos para explorar os dados pertencem à visualização de dados, que vem antes do treinamento.

✗ C) Incorreta. Definir o problema e selecionar os dados são as etapas iniciais do processo, e não a revisão.

> **Dica de prova:** Revisão = pares + vários algoritmos + escolha do melhor modelo.

D4. B) Modularidade e a reutilização de código → revise o capítulo 2

Por quê: O site python.org afirma que o Python suporta módulos e pacotes, o que estimula a modularidade e a reutilização de código. É essa característica que permite usar bibliotecas prontas no aprendizado de máquina.

✗ A) Incorreta. O interpretador e a biblioteca padrão do Python estão disponíveis sem custo e podem ser distribuídos livremente.

✗ C) Incorreta. Módulos e pacotes facilitam justamente o uso de bibliotecas, como as de aprendizado de máquina.

✗ D) Incorreta. O Python é uma linguagem interpretada; seus programas são executados por um interpretador.

> **Dica de prova:** Módulos e pacotes no Python = código modular e reutilizável.

D5. D) Código-fonte disponível para estudar, alterar e redistribuir → revise o capítulo 2

Por quê: O item 2.2 separa softwares de código aberto (como TensorFlow, Scikit-Learn e R Studio) dos proprietários (como MATLAB). No código aberto, o código-fonte é público e pode ser estudado, alterado e redistribuído nos termos da licença.

✗ A) Incorreta. Condicionar o uso à compra de licença é típico de software proprietário, e não de código aberto.
✗ B) Incorreta. Ser de código aberto não tem relação com o tipo de hardware em que o software funciona.
✗ C) Incorreta. Manter o código-fonte em sigilo é característica de software proprietário.

> **Dica de prova:** Código aberto fala de acesso ao código-fonte, e não de preço ou de hardware.

D6. D) MATLAB → revise o capítulo 2

Por quê: O item 2.2 pede que o candidato identifique softwares de código aberto e proprietários. Entre os citados, TensorFlow, Scikit-Learn e R Studio são de código aberto, enquanto o MATLAB é proprietário.

✗ A) Incorreta. O Scikit-Learn é uma biblioteca de código aberto, sob licença BSD.
✗ B) Incorreta. O R Studio é descrito pela Posit como ambiente de desenvolvimento livre e de código aberto para R.
✗ C) Incorreta. O TensorFlow foi liberado como código aberto, sob a licença Apache 2.0.

> **Dica de prova:** Dos softwares do guia, o MATLAB é o exemplo clássico de software proprietário.

D7. C) Álgebra linear → revise o capítulo 3

Por quê: Na álgebra linear, um registro pode ser um vetor e uma tabela de registros, uma matriz. O guia afirma que entender os princípios da álgebra linear dá a base para compreender o aprendizado de máquina e a implementação de algoritmos.

✗ A) Incorreta. A estatística inferencial generaliza de uma amostra para a população; não é a base para manipular vetores e matrizes.
✗ B) Incorreta. A probabilidade mede a incerteza de um resultado; a estrutura de vetores e matrizes vem da álgebra linear.
✗ D) Incorreta. O teorema de Bayes calcula probabilidades a partir de outras conhecidas, e não organiza os dados em linhas e colunas.

> **Dica de prova:** Vetor, matriz, linhas e colunas: pense em álgebra linear.

D8. D) Há correlação, o que não prova causa e efeito → revise o capítulo 3

Por quê: Correlação indica que duas variáveis variam juntas, mas não prova que uma causa a outra. Um terceiro fator, como a temperatura, pode influenciar as duas.

✗ A) Incorreta. A correlação não indica a direção de causa; os sorvetes não provam o aumento das vendas de ventiladores.
✗ B) Incorreta. Andar junto não prova que os ventiladores provocam a compra de sorvetes.
✗ C) Incorreta. Existe relação estatística, pois as duas vendas variam juntas; o que não existe é prova de causa.

> **Dica de prova:** Correlação não é causalidade: desconfie da conclusão 'A causa B'.

D9. A) Por tentativa e erro, com recompensas e punições → revise o capítulo 3

Por quê: No aprendizado por reforço, o agente não recebe exemplos rotulados: ele age, recebe recompensas ou punições e aprende a escolher ações que atingem o objetivo. O guia o cita como comum em jogos.

✗ B) Incorreta. Regras escritas por especialistas descrevem um sistema baseado em conhecimento, e não o reforço.
✗ C) Incorreta. Exemplos rotulados com a resposta correta caracterizam o aprendizado supervisionado.
✗ D) Incorreta. Formar grupos a partir de dados sem rótulo é o aprendizado não supervisionado.

> **Dica de prova:** Reforço = agir, receber recompensa ou punição e melhorar.

D10. B) Conjunto de modelos (ensemble) → revise o capítulo 4

Por quê: O guia lista teste, ajuste, conjuntos, teste estatístico e revisão no item 4.4. Nos conjuntos, vários modelos são combinados porque, juntos, costumam acertar mais do que cada um isolado.

✗ A) Incorreta. A revisão por pares é a análise do trabalho por outros profissionais, e não a combinação de modelos.
✗ C) Incorreta. O ajuste altera configurações de um modelo para melhorá-lo, sem combinar vários modelos.
✗ D) Incorreta. O teste estatístico avalia se os resultados são significativos; não junta modelos.

> **Dica de prova:** Vários modelos combinados em uma resposta = conjunto (ensemble).

D11. A) Destacar o ganho no atendimento em linguagem simples → revise o capítulo 4

Por quê: O guia destaca o tipo de informação, a apresentação, a linguagem e o uso de termos técnicos ao compartilhar resultados. Para um diretor não técnico, mostre o efeito no problema de negócio com palavras simples.

✗ B) Incorreta. Excesso de termos técnicos dificulta a compreensão; o guia pede atenção à linguagem.
✗ C) Incorreta. Fórmulas e parâmetros interessam à equipe técnica, e não ajudam o diretor a entender o ganho.
✗ D) Incorreta. Código e bibliotecas não respondem se o problema do atendimento foi resolvido.

> **Dica de prova:** Público não técnico: impacto no negócio, linguagem simples, poucos termos técnicos.

D12. D) Verificar se a diferença não se deve ao acaso → revise o capítulo 4

Por quê: O guia inclui o teste estatístico entre as atividades de teste e sugere testar e comparar vários algoritmos. O teste estatístico ajuda a decidir se um algoritmo é de fato melhor ou se a diferença é casual.

✗ A) Incorreta. Corrigir rótulos é tarefa da preparação de dados, e não do teste estatístico.
✗ B) Incorreta. Traduzir os resultados para linguagem simples é parte da comunicação com as partes interessadas.
✗ C) Incorreta. O teste estatístico avalia os resultados; não cria novos dados de treinamento.

> **Dica de prova:** Diferença pequena entre modelos? O teste estatístico diz se ela é real.

Agora, treine no formato real

Faça as 6 provas completas do curso EXIN Machine Learning: 108 questões inéditas, tempo do exame, explicação de cada alternativa e link da documentação para revisar.

Ver as 6 provas completas e o cupom do mês →

Até o dia da prova, uma revisão por dia no seu feed

Instagram @strecht.simulados

Um post de prova por dia: questão comentada, pegadinhas e o "não confunda" de cada certificação.

Seguir no Instagram →

YouTube Strecht Simulados

Shorts diários com pegadinhas de prova e, aos sábados, um simulado comentado.

Inscrever-se no canal →

Material independente de preparação, sem afiliação ou endosso de EXIN e BCS. As questões são inéditas e não reproduzem a prova real.