Machine Learning: Perguntas e Respostas

Q1- Qual é o compromisso (trade-off) entre viés e variância?

Essencialmente, se tornar o modelo mais complexo e adicionar mais variáveis, perderá erro de viés (bias), mas ganhará erro de variância. Para obter a redução ideal de erro, é necessário equilibrar os dois. O erro de alto viés causa subajuste (underfitting) e a alta variância causa sobreajuste (overfitting).

Q2- Qual é a diferença entre aprendizagem supervisionada e não supervisionada?

A aprendizagem supervisionada requer o treino com dados rotulados. Por exemplo, para fazer classificação, é necessário rotular os dados primeiro. A aprendizagem não supervisionada, pelo contrário, não requer a rotulagem explícita dos dados.

Q3- Como o KNN é diferente do agrupamento k-means?

A diferença crítica é que o KNN precisa de pontos rotulados e é, portanto, aprendizagem supervisionada, enquanto o k-means não precisa — sendo aprendizagem não supervisionada.

Q4- Explique como funciona uma curva ROC.

A curva ROC é uma representação gráfica do contraste entre as taxas de verdadeiros positivos (TPR) e as taxas de falsos positivos (FPR) em vários limiares. É frequentemente usada como uma métrica para o equilíbrio entre a sensibilidade do modelo (verdadeiros positivos) versus a probabilidade de disparar um alarme falso (falsos positivos).

Q5- Defina precisão e recall.

O Recall é conhecido como a taxa de verdadeiros positivos: a quantidade de positivos que o modelo identifica em comparação com o número real de positivos nos dados. A Precisão é o valor preditivo positivo: a medida de quão exatos são os positivos que o modelo afirma em comparação com o total de positivos que ele previu.

Q6- O que é o Teorema de Bayes? Como é útil no contexto de machine learning?

O Teorema de Bayes fornece a probabilidade posterior de um evento acontecer dado um conhecimento prévio. Matematicamente, é expresso como: $P(A|B) = P(B|A) \cdot P(A) / P(B)$. É a base de algoritmos como o Naive Bayes.

Q7- Por que o "Naive" Bayes é ingénuo?

É considerado ingénuo porque assume que todos os recursos (features) no conjunto de dados são mutuamente independentes, uma condição que raramente ou nunca acontece na realidade.

Q8- Explique a diferença entre regularização L1 e L2.

A regularização L2 tende a espalhar o erro entre todos os termos, enquanto a L1 é mais binária/esparsa, atribuindo pesos zero ou um a muitas variáveis. A L2 corresponde a um prior Gaussiano e a L1 a um prior de Laplace.

Q10- Qual é a diferença entre erro de Tipo I e Tipo II?

O erro de Tipo I é um falso positivo: afirmar que algo aconteceu quando não aconteceu. O erro de Tipo II é um falso negativo: afirmar que nada está a acontecer quando, na verdade, algo está.

Q12- Qual é a diferença entre probabilidade (probability) e verosimilhança (likelihood)?

A probabilidade refere-se à chance de sucesso em contextos discretos. A verosimilhança é a probabilidade condicional em contextos contínuos, ou seja, a chance de sucesso dadas certas variáveis de entrada.

Q13- O que é Deep Learning e como se diferencia de outros algoritmos?

É um subconjunto do machine learning focado em redes neurais. Utiliza retropropagação (backpropagation) e princípios da neurociência para modelar grandes conjuntos de dados não estruturados ou semiestruturados.

Q14- Qual a diferença entre um modelo generativo e um discriminativo?

Um modelo generativo aprende as categorias dos dados, enquanto um modelo discriminativo aprende a distinção entre as diferentes categorias. Modelos discriminativos geralmente superam os generativos em tarefas de classificação.

Q15- Que técnica de validação cruzada usaria num conjunto de dados de séries temporais?

Não se deve usar k-fold aleatório. Deve-se usar uma técnica de "janela rolante": treinar em dados passados e testar em dados futuros sucessivamente (ex: treinar em [Jan], testar em [Fev]; treinar em [Jan, Fev], testar em [Mar]).

Q16- Como é feita a poda (pruning) numa árvore de decisão?

A poda ocorre quando se removem ramos que têm baixo poder preditivo. Isto ajuda a reduzir a complexidade do modelo e a evitar o sobreajuste (overfitting).

Q17- O que é mais importante: acurácia ou performance do modelo?

A performance é mais importante. A acurácia é apenas a proporção de previsões corretas, mas existem modelos com alta acurácia que têm baixo poder preditivo real (ex: num conjunto de dados desequilibrado).

Q18- O que é o F1 score? Como o usaria?

É uma média ponderada da precisão e do recall. É usado em tarefas de classificação onde os verdadeiros negativos não são tão importantes, sendo que resultados próximos de 1 são os ideais.

Q19- Descreva algumas medidas de performance.

F1 Score, Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE), Área sob a curva (AUC), Matriz de Confusão e Log Loss.

Q20- Como lidaria com um conjunto de dados desequilibrado?

Pode-se coletar mais dados, mudar a métrica para ROC/AUC, fazer reamostragem (oversampling/undersampling), usar SMOTE para gerar amostras sintéticas ou usar algoritmos penalizados.

Q21- Dê um exemplo onde técnicas de ensemble seriam úteis.

São úteis para reduzir o sobreajuste e tornar o modelo mais robusto. Exemplos comuns incluem Random Forest (Bagging) e XGBoost (Boosting).

Q22- Como evitar o sobreajuste (overfitting)?

1. Manter o modelo simples (reduzir variáveis).
2. Usar técnicas de validação cruzada.
3. Usar técnicas de regularização como LASSO (L1) ou Ridge (L2).

Q23- Que abordagens de avaliação usaria para medir a eficácia de um modelo?

Usaria validação cruzada (cross-validation) combinada com métricas específicas como F1 Score, acurácia e matriz de confusão.

Q25- O que é o "kernel trick" e por que é útil?

O truque do kernel permite projetar dados num espaço de maior dimensão onde classes que não eram linearmente separáveis passam a ser, sem a necessidade de calcular explicitamente as coordenadas nesse novo espaço.

Q26- Como lida com dados ausentes ou corrompidos?

Pode-se eliminar as linhas/colunas afetadas ou usar técnicas de imputação para substituir os valores por medidas como a média, mediana ou moda.

Q29- Quais são as diferenças entre uma lista ligada e um array?

Um array é uma coleção ordenada onde cada elemento tem o mesmo tamanho. Uma lista ligada é uma série de objetos com ponteiros que indicam o próximo elemento, sendo mais fácil de redimensionar e reorganizar.

Q30- Descreva uma tabela hash.

É uma estrutura de dados que produz um array associativo, onde uma chave é mapeada para valores específicos através de uma função hash. É muito utilizada para indexação em bases de dados.

Q31- Que bibliotecas de visualização de dados utiliza?

Matplotlib e Seaborn são as mais comuns em Python.

Q32- Como implementaria um sistema de recomendação?

Poderia ser feito associando palavras-chave aos produtos. Quando um utilizador vê um item, o sistema recomenda outros itens que partilham as mesmas palavras ou categorias.

Q33- Como as suas competências de ML podem gerar receita?

Através da análise de sentimento em redes sociais, análise de clusters para campanhas de marketing direcionadas, otimização de logística e previsão de vendas para gestão de stocks.

Q35- Quais foram os últimos artigos de ML que leu?

(Resposta baseada no texto): Um artigo sobre um novo algoritmo de regressão diferenciável usado para prever preços de imóveis e produzir mapas de contorno geográficos.

Q37- Quais são os seus casos de uso favoritos para modelos de ML?

Previsão de vendas, análise de churn (cancelamento de clientes) e sistemas de recomendação.

Q38- Como reduzir a dimensão num conjunto de dados gigante (1M linhas, 1000 colunas)?

Usar PCA em variáveis numéricas, amostragem de colunas, ou modelos que suportem aprendizagem online (online learning) como SGD (Gradiente Descendente Estocástico).

Q39- A rotação é necessária no PCA? Porquê?

A rotação (como varimax) maximiza a diferença entre as variâncias capturadas pelos componentes, tornando-os mais fáceis de interpretar.

Q40- Se faltarem valores espalhados em 1 desvio padrão da mediana, que percentagem de dados não é afetada?

Assumindo uma distribuição normal, cerca de 68% dos dados estão dentro de 1 desvio padrão, o que significa que cerca de 32% dos dados não estariam nesse intervalo.

Q41- O que é a verosimilhança (likelihood)?

É a probabilidade de classificar uma observação como 1 na presença de outra variável.

Q42- Pode uma regressão superar uma árvore de decisão em séries temporais?

Sim. Se os dados tiverem uma relação linear forte, a regressão linear terá melhor performance que uma árvore de decisão, que é mais adequada para relações não lineares.

Q43- Que algoritmo de ML pode salvar uma empresa de entrega de comida?

Problemas de otimização de rotas e previsão de tempo de entrega, pois envolvem padrões complexos e grandes volumes de dados que equações matemáticas simples não resolvem.

Q44- Como baixar a variância de um algoritmo?

Usando regularização para penalizar a complexidade ou selecionando apenas as variáveis mais importantes (top features).

Q45- O que usar para baixo viés e alta variância?

Algoritmos de Bagging (como Random Forest) ou Boosting (como GBM).

Q46- Deve remover variáveis correlacionadas antes do PCA?

Sim, porque variáveis altamente correlacionadas podem inflar artificialmente a variância explicada por um componente específico.

Q47- Por que combinar 5 modelos GBM pode não melhorar a acurácia?

Ensembles funcionam melhor quando os modelos são independentes. Se os 5 modelos GBM estiverem altamente correlacionados, eles estarão a cometer os mesmos erros.

Q48- Explique o algoritmo KNN.

O KNN (K-Nearest Neighbors) classifica uma nova observação com base na maioria das etiquetas dos seus k vizinhos mais próximos no espaço de dados.

Q49- Como o TPR (True Positive Rate) e o Recall estão relacionados?

Eles são exatamente a mesma coisa. A fórmula para ambos é $TP / (TP + FN)$.

Q50- Liste as métricas da matriz de confusão.

Acurácia, Precisão, Recall (Sensibilidade), Especificidade e F1 Score.

Q51- Se eu remover o intercepto de um modelo de regressão múltipla, o R^2 pode aumentar de 0.3 para 0.8?

Sim, isso pode acontecer. Quando você força a linha de regressão a passar pela origem (removendo o intercepto), a fórmula padrão do R^2 é alterada. Isso pode inflar o valor estatístico, mas não significa necessariamente que o modelo tenha uma capacidade preditiva melhor na prática.

Q52- Como verificar multicolinearidade sem perder informação?

Você pode criar uma matriz de correlação para identificar variáveis altamente relacionadas (ex: correlação acima de 0.75). Para tratar o problema sem excluir variáveis e perder informação, a **Regressão Ridge** é ideal, pois ela lida com a multicolinearidade penalizando os coeficientes.

Q53- Quando a regressão Ridge é favorável em relação à Lasso?

- **Lasso (L1)**: É melhor quando você tem poucas variáveis com efeitos grandes; ela realiza seleção de variáveis ao zerar coeficientes inúteis.
- **Ridge (L2)**: É preferível quando você tem muitas variáveis com efeitos pequenos ou quando as variáveis são altamente correlacionadas entre si.

Q54- Correlação implica causalidade (Ex: Piratas vs. Temperatura)?

Não. A correlação indica apenas que duas variáveis variam juntas. No exemplo, o aumento da temperatura global e a redução dos piratas podem ser coincidência ou causados por uma terceira variável oculta (conhecida como variável de confusão).

Q55- Como você seleciona variáveis importantes?

Os métodos incluem:

- Remoção de variáveis com alta correlação.
- Uso de p-values em regressão linear.
- Seleção Forward, Backward ou Stepwise.
- Gráficos de importância de variáveis (Variable Importance) em Random Forest ou XGBoost.
- Regressão Lasso.

Q56- Qual a diferença entre covariância e correlação?

A correlação é a versão padronizada da covariância. Enquanto a covariância depende da escala das variáveis, a correlação coloca o valor em um intervalo fixo entre -1 e 1, permitindo comparar variáveis com unidades diferentes.

Q57- É possível capturar a correlação entre uma variável contínua e uma categórica?

Sim, é possível utilizar a técnica **ANCOVA** (Análise de Covariância) para medir essa associação.

Q58- Qual a diferença entre Random Forest e Gradient Boosting (GBM)?

- **Random Forest**: Usa Bagging (treina árvores em paralelo) e foca na redução da variância.
- **GBM**: Usa Boosting (treina árvores sequencialmente, onde a próxima corrige o erro da anterior) e foca na redução do viés e da variância.

Q59- Como ocorre a divisão (split) em uma árvore de decisão?

O algoritmo busca a característica que melhor separa os dados nos "nós filhos". Isso é medido através do **Índice Gini** ou da **Entropia**. O objetivo é tornar os grupos resultantes o mais homogêneos (puros) possível.

Q60- Erro de treino 0.00 e erro de validação 34.23 em Random Forest. O que aconteceu?

O modelo sofreu um **overfitting** (sobreajuste) severo. Isso geralmente acontece quando o número de árvores é muito maior do que o necessário ou a profundidade é excessiva. É preciso ajustar os hiperparâmetros.

Q61- O que fazer quando o número de variáveis (p) é maior que o de observações (n)?

O método tradicional de Mínimos Quadrados (OLS) não funciona (matriz não invertível). Você deve usar regressão penalizada (**Lasso ou Ridge**) ou técnicas de redução de dimensionalidade como PCA.

Q62- One-hot encoding vs. Label encoding?

- **One-hot**: Cria uma nova coluna para cada categoria (aumenta a dimensão).
- **Label**: Atribui um número (0, 1, 2...) a cada categoria. O Label encoding pode confundir o modelo fazendo-o pensar que existe uma ordem numérica onde não há.

Q63- Como lidar com variáveis que têm mais de 30% de valores ausentes?

Não as descarte de imediato. Verifique se o fato de o dado estar ausente possui relação com a variável alvo. Você pode imputar os valores ou criar uma categoria específica para representar o "valor ausente".

Q64- Qual algoritmo para "Pessoas que compraram isso, também compraram..."?

Filtragem Colaborativa (Collaborative Filtering). Ele analisa o histórico de comportamento e as preferências de usuários semelhantes para fazer recomendações.

Q65- Erro de Tipo I e Tipo II no contexto de hipóteses?

- **Tipo I**: Rejeitar a hipótese nula quando ela é verdadeira (Falso Positivo).
- **Tipo II**: Aceitar a hipótese nula quando ela é falsa (Falso Negativo).

Q66- Acurácia de validação alta, mas de teste baixa (mesmo com modelo simples).

Isso pode ser um problema de amostragem. Use **Amostragem Estratificada** em vez de aleatória para garantir que as classes do alvo estejam representadas na mesma proporção tanto no treino quanto no teste.

Q67- Critérios para avaliar um modelo de regressão?

Use o **R^2 Ajustado**, pois ele penaliza a adição de variáveis que não melhoram o modelo. Também verifique a estatística de Tolerância ($1/VIF$) para identificar multicolinearidade.

Q68- Por que não usar a distância de Manhattan no K-means?

A distância de Manhattan mede apenas caminhos horizontais e verticais. A distância Euclidiana é preferida pois permite calcular a distância direta "em linha reta" em qualquer direção no espaço.

Q69- Como explicar Machine Learning para uma criança de 5 anos?

É como um bebê aprendendo a andar: ele tenta, cai e sente dor. O cérebro dele entende que cair é ruim (erro) e, na próxima vez, ele ajusta o equilíbrio ou busca apoio para não cair de novo. Ele aprendeu com a experiência.

Q70- Como avaliar um modelo de regressão logística?

Utilize a curva **AUC-ROC** e a matriz de confusão. Além disso, o **AIC** (Critério de Informação de Akaike) é útil para comparar modelos; quanto menor o AIC, melhor o ajuste.

Q71- Como decidir qual algoritmo usar?

Depende da natureza do dado: Linearidade (Regressão Linear), Imagens/Áudio (Deep Learning), Interações não lineares (Boosting), ou necessidade de explicação simples (Árvores de Decisão).

Q72- Variável categórica pode ser tratada como contínua?

Apenas se ela for de natureza **ordinal** (onde a ordem numérica importa, como níveis de escolaridade).

Q73- Quando a regularização se torna necessária?

Quando o modelo apresenta overfitting ou sobreeajuste, o que é percebido quando o erro de treino é muito baixo, mas o erro de teste é alto.

Q74- OLS vs. Máxima Verossimilhança?

O OLS busca minimizar a distância quadrada entre os pontos. A Máxima Verossimilhança (Maximum Likelihood) busca os parâmetros que tornam os dados observados o mais prováveis possível de ocorrer.

Q78- Técnicas de redução de dimensionalidade?

- **PCA**: Não supervisionado, ideal para dados numéricos.
- **LDA**: Supervisionado, foca na separação de classes.
- **t-SNE**: Técnica não linear, excelente para visualização de dados complexos em 2D ou 3D.

Q79- Qual distribuição de frequência se espera para regressão linear?

Espera-se uma **Distribuição Normal** (Gaussiana) para os resíduos do modelo.

Q80- Pressupostos da Regressão Logística?

Diferente da linear, ela não exige homocedasticidade, não requer relação linear e os resíduos não precisam ser distribuídos normalmente.

Q81- Tamanho da amostra para Regressão Logística?

Uma regra prática é: $N = (10 \times \text{número de variáveis}) / \text{probabilidade da classe menos frequente}$.

Q82- O que é Clustering?

É o processo de agrupar objetos de forma que os itens dentro de um grupo (cluster) sejam mais parecidos entre si do que com itens de outros grupos.

Q83- Tipos de Clustering?

- **Hard**: O ponto pertence estritamente a um grupo.
- **Soft**: O ponto tem uma probabilidade de pertencer a vários grupos.
- Métodos: Centróide (K-means), Hierárquico, Densidade (DBSCAN).

Q84- Como funciona o K-means?

Ele escolhe centros aleatórios, associa cada ponto ao centro mais próximo, recalcula a média do grupo para mover o centro e repete o processo até que os grupos não mudem mais.

Q87- Número ideal de clusters no Hierárquico?

Deve-se observar o **Dendrograma** e identificar a maior distância vertical que não é cortada por nenhuma linha horizontal de outros clusters.

Q88- K-means vs. Hierárquico?

O K-means é mais rápido e escala melhor para grandes volumes de dados, mas o Hierárquico fornece uma visualização melhor de como os dados se relacionam em diferentes níveis.

Q90- Aplicações de Clustering?

Detecção de anomalias, sistemas de recomendação, segmentação de clientes e análise de redes sociais.

Q91- Relação entre Clustering e Regressão?

Você pode usar clusters como variáveis de entrada para uma regressão ou criar um modelo de regressão específico para cada grupo identificado pelo clustering.

Q94- Como otimizar o K-means?

Utilizar o "método do cotovelo" (Elbow Method) para achar o número ideal de clusters e testar diferentes inicializações de centróides.

Q95- Quais algoritmos precisam de escala (scaling)?

LDA e PCA, pois são baseados em distâncias ou variâncias.

Q96- Quais algoritmos NÃO precisam de escala?

Algoritmos baseados em árvores, como Árvores de Decisão e Random Forest.

Q97- Z-score vs. Min-Max?

- **Z-score (Padronização)**: Melhor para PCA e Análise de Clusters.
- **Min-Max (Normalização)**: Melhor para Redes Neurais e Processamento de Imagens.

Q98- Qual transformação usar antes da regressão linear?

A transformação **Box-Cox** é frequentemente usada para estabilizar a variância e tornar os dados mais parecidos com uma distribuição normal.

Q99- Pressupostos da Regressão Linear?

Homocedasticidade (variância constante), resíduos com distribuição normal, relação linear entre X e Y, e observações independentes.

Q100- O que é necessário para calcular o tamanho da amostra?

Tamanho da população, nível de confiança desejado, margem de erro aceitável e a variabilidade (desvio padrão) dos dados.

Q101- Como lidar com matrizes esparsas em sistemas de recomendação?

Redução de dimensionalidade com svd e fatoração de matrizes,uso de matrizes esparsas do numpy,embeddings(transformam IDs esparsos como o ID de um produto em vetores densos),usar estratégia baseada em conteúdo e metadados. Lidar com Cold Start recomendando itens populares ou usando metadados para recomendar.