PTIC — 1º Ano RH — Noite

3º Bimestre — 2026

Gabarito 3 com Resolução

Fundamentos da Inteligência Artificial Generativa

Conteúdos: Métricas de avaliação, falhas de GANs, tokenização, aprendizado com poucos exemplos, espaço latente e modelos multimodais — aprofundamento

---

1. PPL ≈ 1/p = 1/0,20 = 5. Um valor menor de perplexidade indica um modelo melhor: quanto menor a perplexidade, maior é, em média, a probabilidade que o modelo atribui às palavras corretas do texto, ou seja, o modelo está menos 'surpreso' com o texto real e prevê as palavras com mais confiança.

2. O mode collapse ocorre quando o gerador de uma GAN, em vez de aprender a produzir amostras que cubram toda a diversidade dos dados reais, passa a gerar repetidamente um número limitado de saídas muito semelhantes entre si (um único 'modo' da distribuição dos dados), pois essas saídas específicas conseguem enganar o discriminador com sucesso. Isso compromete a qualidade do modelo generativo, pois reduz drasticamente a variedade dos dados sintéticos produzidos, mesmo que, isoladamente, cada amostra gerada pareça realista.

3. A tokenização é o processo de dividir um texto em unidades menores, chamadas tokens, que podem corresponder a palavras inteiras, partes de palavras (subpalavras) ou até caracteres individuais, dependendo do modelo. Os modelos processam o texto em tokens, e não em caracteres isolados, porque tokens carregam mais significado linguístico e reduzem o tamanho da sequência que o modelo precisa processar, tornando o treinamento e a geração de texto mais eficientes computacionalmente. Tokens ≈ 2000/4 = 500 tokens.

4. No zero-shot learning, o modelo realiza uma tarefa sem receber nenhum exemplo prévio dessa tarefa no prompt, apenas a instrução direta — por exemplo: 'Classifique o sentimento da frase a seguir como positivo ou negativo: [frase]'. No few-shot learning, o prompt inclui alguns exemplos resolvidos da tarefa antes de apresentar o caso a ser resolvido — por exemplo: 'Frase: Adorei o produto! Sentimento: positivo. Frase: Não gostei do atendimento. Sentimento: negativo. Frase: [nova frase]. Sentimento:'. Os exemplos fornecidos no few-shot ajudam o modelo a entender melhor o padrão esperado na resposta.

5. O espaço latente é uma representação numérica compacta e abstrata dos dados, em um espaço matemático de dimensão geralmente muito menor do que os dados originais (como uma imagem), na qual características semelhantes dos dados ficam posicionadas de forma próxima. Em modelos generativos, o espaço latente funciona como uma espécie de 'mapa' de onde novos dados podem ser gerados: ao amostrar um ponto desse espaço (aleatoriamente ou de forma controlada) e decodificá-lo de volta para o formato original (como uma imagem), o modelo consegue gerar novos exemplos plausíveis, muitas vezes permitindo até combinar ou interpolar características de diferentes exemplos ao navegar por esse espaço.

6. $h = x \cdot w_1 + b_1 = 1{,}0 \times 0{,}6 + 0{,}1 = 0{,}7$. Saída $= h \cdot w_2 + b_2 = 0{,}7 \times 0{,}8 + (-0{,}05) = 0{,}56 - 0{,}05 = 0{,}51$

7. Tempo total = 50 × 0,2 = 10 segundos. Cada etapa de denoising remove uma pequena parcela do ruído da imagem, refinando-a gradualmente em direção a uma imagem final coerente com o prompt; quanto mais etapas são realizadas, mais gradual e precisa é essa remoção de ruído, o que tende a produzir uma imagem final de melhor qualidade e mais detalhada — mas, como cada etapa consome um tempo de processamento, mais etapas também significam um tempo total de geração maior.

8. Número de pesos = 10 × 5 = 50 pesos. Redes neurais maiores, com mais neurônios e mais camadas, têm um número muito maior de parâmetros (pesos) ajustáveis — e, para que esses parâmetros sejam ajustados de forma confiável durante o treinamento (evitando, por exemplo, overfitting), é necessária uma quantidade proporcionalmente maior de dados de treinamento, além de mais poder computacional (memória e capacidade de processamento) para realizar os cálculos envolvidos em cada etapa do treinamento.

9. A técnica RAG combina duas etapas principais: (1) recuperação (retrieval), na qual o sistema busca, em uma base de dados ou documentos externos e atualizados, as informações mais relevantes relacionadas à pergunta do usuário; (2) geração (generation), na qual o modelo de IA Generativa utiliza essas informações recuperadas como contexto adicional para elaborar sua resposta final. Essa técnica ajuda a reduzir alucinações porque a resposta do modelo passa a ser fundamentada em informações reais e verificáveis, recuperadas de fontes externas confiáveis, em vez de depender exclusivamente do conhecimento que o modelo memorizou (e possivelmente distorceu ou esqueceu) durante seu treinamento original.

10. Modelos de IA Generativa multimodais são aqueles capazes de processar e/ou gerar mais de um tipo de dado simultaneamente — por exemplo, compreender uma imagem e responder perguntas sobre ela em texto, ou gerar uma imagem a partir de uma descrição textual. Um exemplo de aplicação prática é um assistente de IA que recebe a foto de um prato de comida e o áudio de uma pergunta falada ('Quantas calorias tem esse prato?'), processando tanto a imagem quanto o áudio para gerar uma resposta em texto ou voz — algo que um modelo capaz de lidar com apenas um tipo de dado (só texto, por exemplo) não conseguiria fazer sozinho.