Prova 1 - 2025.2

Processamento de Linguagem Natural

Observação: prova idêntica ao período de 2025.1.

1) Se a perda no treinamento aumenta com o número de épocas, qual das seguintes opções poderia ser um possível problema no processo de aprendizagem?

a. A regularização está muito baixa e o modelo está sobreajustando (overfitting).

a. A regularização está muito alta e o modelo está subajustando (underfitting).

a. A taxa de aprendizagem (step size) está muito alta.

a. A taxa de aprendizagem (step size) está muito pequena.

2) Vetores de palavras densos aprendidos por meio de word2vec têm muitas vantagens em relação ao uso de vetores de palavras únicas. Qual dos itens a seguir NÃO é uma vantagem que os vetores densos têm sobre os vetores esparsos? Por quê?

a. Modelos que usam vetores de palavras densos generalizam melhor para palavras invisíveis do que aqueles que usam vetores esparsos.

a. Modelos que usam vetores de palavras densos generalizam melhor para palavras raras do que aqueles que usam vetores esparsos.

a. Vetores de palavras densos codificam semelhança entre palavras, enquanto vetores esparsos não.

a. Vetores de palavras densos são mais fáceis de incluir como recursos no aprendizado de máquina sistêmico do que vetores esparsos.

3) Um modelo de rede neural multicamadas (MLP) treinado usando gradiente descendente estocástico no mesmo conjunto de dados, com diferentes inicializações para seus parâmetros, tem garantia de aprender os mesmos parâmetros ao final do treinamento? Justifique sua resposta.

4) Explique o conceito de amostragem negativa (negative sampling) no algoritmo skip-gram. Descreva o motivo pelo qual o skip-gram com amostragem negativa é mais rápido de treinar do que o modelo skip-gram original.

5) Redes neurais têm seus parâmetros inicializados de maneira aleatória. Descreva uma maneira (mesmo que improvável) pela qual uma rede neural poderia ser inicializada de forma inadequada e qual efeito isso poderia ter.

6) Um certo site de notícias permite que os leitores postem comentários anônimos em qualquer artigo publicado no site. O editor o contratou para escrever um software que identifica “comentários bons”. O editor planeja deletar automaticamente os comentários ruins. O editor insiste que você deve alcançar 90% de precisão na tarefa de identificar comentários bons. O que isso significa? Justifique brevemente.

a a. Seja gentil com os escritores. Tudo bem se acidentalmente deletarmos alguns dos comentários bons, mas não mais do que 10% deles.

a b. Seja gentil com os leitores. Tudo bem se alguns dos comentários que aparecem em nosso site forem ruins, mas não mais do que 10% deles.

a c. Seja preciso. Tudo bem cometer alguns erros ao classificar erroneamente alguns comentários, mas não mais do que 10% deles.

7) Você decide tratar o problema do editor como um problema de classificação supervisionada. Seu primeiro passo é coletar alguns dados. Que conjunto de dados você prepararia? Dê detalhes.