Processamento de Linguagem Natural

Este documento simula uma prova baseada nos tópicos do curso Deep Learning for NLP (DL-NLP) e no estilo dos documentos de avaliação fornecidos.

Prova 1 (Estilo Simulado)

Questões

1. Cite três diferenças principais entre Layer Normalization e Batch Normalization em arquiteturas de redes neurais.

2. Sobre as funções de Regularização e Perda (Loss Functions) em Deep Learning, marque as alternativas corretas.

- ☐ a. O princípio MDL (Minimum Description Length) sugere que o melhor modelo é aquele que consegue descrever os dados e, simultaneamente, minimizar a complexidade do próprio modelo, evitando assim o overfitting.
- ☐ b. A Regularização L2 (l2 Regularization) penaliza a complexidade do modelo adicionando a norma dos parâmetros do modelo à função de perda. Isso pode ser interpretado probabilisticamente como a incorporação de um prior Gaussiano sobre os parâmetros.
- ☐ c. A Cross-Entropy (Entropia Cruzada) é a perda primária utilizada para classificadores, e mede a diferença entre duas distribuições de probabilidade (a distribuição prevista (■) e a distribuição real (y)).
- ☐ d. O Dropout é um método de regularização que desliga permanentemente um subconjunto de ativações durante a fase de treinamento, resultando em detectores de características menos redundantes.

3. Você precisa desenvolver um Modelo de Linguagem Neural (NLM) que use um k-grama de palavras para prever a próxima palavra, mas deseja utilizar um modelo mais simples que o NLM clássico (que usa MLP). Proponha um método de word embedding que atenda a essa simplificação. Explique o papel desse método na minimização da tarefa de previsão, incluindo a simplificação na camada de saída (Softmax).

4. Suponha que você está trabalhando com um corpus que contém muitas palavras raras e com variações morfológicas complexas (como em Português). Você está decidindo qual modelo de word embedding usar: um word2vec padrão ou o FastText.

Qual modelo seria a escolha mais robusta para lidar com a natureza morfológica e as palavras Out-Of-Vocabulary (OOV)? Justifique, explicando a técnica de representação subword usada pelo modelo escolhido.

Respostas

1. Três diferenças principais entre Layer Normalization e Batch Normalization

Cálculo da Média e Variância: Batch Normalization agrega a média e o desvio padrão entre exemplos no batch. Layer Normalization padroniza dentro de cada exemplo, ao longo das características.

Dependência do Tamanho do Batch e Uso em Modelos Grandes: BatchNorm depende do tamanho do batch; LayerNorm é independente, sendo comum em Transformers e setups com batches pequenos.

Otimização e Parâmetros: Ambas estabilizam o sinal e permitem LRs maiores; LayerNorm (assim como BN) aplica transformação afim (escala/viés) com parâmetros aprendidos.

2. Sobre as funções de Regularização e Perda (Loss Functions) em Deep Learning, marque as alternativas corretas.

- ☑ a. O princípio MDL (Minimum Description Length) sugere que o melhor modelo é aquele que consegue descrever os dados e, simultaneamente, minimizar a complexidade do próprio modelo, evitando assim o overfitting.
- ☑ b. A Regularização L2 (l2 Regularization) penaliza a complexidade do modelo adicionando a norma dos parâmetros do modelo à função de perda. Isso pode ser interpretado probabilisticamente como a incorporação de um prior Gaussiano sobre os parâmetros.
- ☑ c. A Cross-Entropy (Entropia Cruzada) é a perda primária utilizada para classificadores, e mede a diferença entre duas distribuições de probabilidade (a distribuição prevista (■) e a distribuição real (y)).
- ☐ d. O Dropout é um método de regularização que desliga permanentemente um subconjunto de ativações durante a fase de treinamento, resultando em detectores de características menos redundantes. (Incorreto, pois o Dropout randomicamente zera ativações e incentiva detectores de características redundantes.)

3. Proposta de método de word embedding simplificado (Word2Vec) e explicação da minimização da Softmax.

Método proposto: word2vec.

Simplificação Arquitetural: remove a camada oculta do NLM, tornando-o um modelo log-linear. O contexto (k-grama) é projetado diretamente para o espaço de embeddings.

Minimização da Tarefa de Previsão: em vez da Softmax completa sobre |V|, utiliza-se Negative Sampling ou Noise Contrastive Estimation para resolver tarefas binárias (par válido vs. inválido), reduzindo o custo e mantendo boa qualidade dos vetores.

4. Escolha e Justificativa do Modelo para Morfologia e OOV (Out-Of-Vocabulary).

Escolha: FastText.

Justificativa Subword: representa cada palavra como um saco de n-gramas de caracteres; cada n-grama tem seu embedding e o vetor da palavra é a soma dos embeddings dos seus n-gramas (tipicamente $3 \leq n \leq 6$).

Vantagem: permite vetorizar termos OOV e capturar regularidades morfológicas, crucial em idiomas com flexão rica (ex.: português).
