DS5220 Supervised Machine Learning and Learning Theory: Homework 4

Due 2024/11/26 11:59pm

Instructions The homework must be submitted in **one pdf file**. Everyone needs to make a submission. Please write your nine-digit NUID and your northeastern.edu email. For the file names, please include your NUID. For example, "DS5220 HW4 (xxxxxxxxx)". Please specify the random seed at an appropriate place in your code chunk. It may not be the at the start, but specify it when you are executing codes that involve randomness. There are 100 points in total in this homework.

Problem 1 [20 pts] What is the self-attention mechanism in transformer networks? Describe the computation of a multi-head self-attention with h attention heads, given an input X, their key matrices, query matrices, and value matrices for the attention heads.

Problem 2 [40 pts] Recall that the cross-entropy loss—given an input $x \in \mathbb{R}^d$ and a label $y \in \{1, 2, \ldots, k\}$, let $\ell(f_W(x), y)$ denote the loss of a linear classifier, with parameters given by $W \in \mathbb{R}^d$.

- [10 pts] Can you write down the mathematical definition of $\ell(f_W(x), y)$?
- [10 pts] Next, can you write down the gradient of the loss over W, that is, can you calculate $\nabla \ell(f_W(x), y)$. Make sure to include the intermediate steps in the derivation.
- [10 pts] Based on the above calculation, can you write down the gradient descent algorithm for minimizing ℓ, on a dataset of n samples, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$?
- [10 pts] Finally, can you explain the idea of AdaGrad (adaptive gradient descent)?

Problem 3 [40 pts]

- [10 pts] What are the three types of language models? Can you explain the difference between encoder-only models and decoder-only models?
- [5 pts] What is a recurrent neural network (RNN)? Give one example of an RNN architecture.

- [10 pts] Suppose you are working on a customer review dataset. You would like to build a sentiment classifier. How would you build a simple feed-forward neural networks for this problem? Explain the key steps in the implementation.

- [5 pts] What is the concept of entropy, and how would you measure the entropy of a language model? In addition, can you explain the concept of cross entropy?

- [5 pts] What is the N-gram model, and how can you calculate a bi-gram model from a corpus of text?

- [5 pts] Name at three three different ways of performing fine-tuning, and explain the differences between them.