Homework 3, Part 2

CS 685, Fall 2025

[Update 11/10: added AI disclosure as explicit question]

This is Part 2 of 2 for HW3. Part 1 is a Colab notebook; see the schedule page.

For these questions, submit all answers as a PDF to Gradescope. Feel free to use latex, colab, ms word, handwriting, whatever lets you author your own PDF that is easy for us to read and understand. If you handwrite some answers, write very neatly if you want to receive full credit.

1 KL divergence

We discussed KL divergence in the context of DPO. It's a way to define a distance of dissimilarity between two distributions, P and Q:

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

where x is summed over the space of all its possible values. In lecture we did a derivation and talked about it as a measure that compares the expected Q surprisal (cross-entropy), compared to the baseline expected P surprisal (entropy), when the expectations based on the P distribution. (The "surprisal" form a probability p is just its negative log-prob, $-\log p = \log(1/p)$; surprisals are always non-negative.)

1.1

If the distributions are the same, what KL divergence do you get? Show why.

1.2

KL divergence is not symmetric. Construct an example to show this. We suggest using a discrete distribution over just two or three outcomes, to keep things simple.

1.3

The definition of KL divergence typically has an additional property to handle zero probabilities in P. If for any x, $P(x) = 0$, then it's defined that $P(x) \log P(x) = 0$.

Does KL divergence make sense if there are zero probabilities in Q? (That is, if there are some x where $Q(x) = 0$.) Show why or why not.

2 Decoding

In nucleus and top-k sampling, for each you can decrease the filtering threshold: shrinking the p for top-p, or shrinking the k toward top-k. As you decrease them, what decoding method does it converge to?

3 Finding an annotation error

Here, you'll do some manual data analysis of a very widely used annotated dataset GLUE, which includes several English-language sentence understanding tasks (mostly semantics, and some sentiment and syntax).

Take a look at any of GLUE's per-task datasets, which can be accessed from either: https://gluebenchmark.com/tasks (one task per row) or https://huggingface.co/datasets/nyu-mll/glue (the "subset" dropdown).

Choose one of these datasets and skim the paper or documentation about it so you know what the labels are supposed to mean. Report which dataset/task you're analyzing. Within labeled examples, find one example where you disagree with the label. Show it and describe what you think the correct label should be, and why. (This can be very short, like a sentence or two.) Find and show two examples where you agree with the dataset's given label, to contrast with it and to help make sure you're finding what ought to be considered an error.

4 AI Disclosure (applies to both Part 1 and Part 2 of HW3)

Did you use any AI assistance to complete this homework? If so, please also specify what AI systems or models you used.

If you answered "yes" above:

If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which problem is associated with which prompt.

Free response: For each problem for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good answer, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to get the answer or check your own answer?