MCB128

Quiz b4

April 3, 2026

Name: _____ **Score:** _____

---

1. **About learning with a probabilistic model.**

   (a) (**40 pts**) Show that optimizing the closeness between the data distribution P_{data} and model distributions P_θ using the KL divergence

   $$\min_\theta D_{KL}(P_{data}||P_\theta) = \min_\theta E_{x\sim P_{data}}\left[\log \frac{P_{data}(x)}{P_\theta(x)}\right]$$

   is equivalent to optimizing the maximum likelihood estimation (MLE)

   $$\max_\theta E_{x\sim P_{data}}\left[\log P_\theta(x)\right].$$

   (b) (**10 pts**) The KL divergence is not symmetric. What would go wrong if you try to optimize the reverse KL divergence?

   (c) (**5 pts**) Which other name have we used in class to refer to the MLE quantity above (more precisely to -MLE)?

   (d) (**5 pts**) What would go wrong if the model gives probability zero to any of the examples in the training set?

[Notation: $E_{x\sim P_{data}}[f(x)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i), \quad \text{for } \{x_i\}_{i=1}^{N} \in \text{ training set.}$]

(a)

$$\begin{aligned}\min_\theta D_{KL}(P_{data}||P_\theta) &= \min_\theta E_{x\sim P_{data}}\left[\log \frac{P_{data}(x)}{P_\theta(x)}\right] \\ &= \min_\theta E_{x\sim P_{data}}\left[\log P_{data}(x) - \log P_\theta(x)\right] \\ &= \min_\theta E_{x\sim P_{data}}\left[-\log P_\theta(x)\right] \\ &= \max_\theta E_{x\sim P_{data}}\left[\log P_\theta(x)\right]\end{aligned}$$

(b)

$$D_{KL}(P_\theta||P_{data}) = E_{x\sim P_\theta}\left[\log P_\theta(x) - \log P_{data}(x)\right]$$

Several things would go wrong

i. We don't have a ready sample of P_θ

ii. Even if we did created one, we would have to get a different sample each time we changed θ since the distribution would change.

iii. There are two terms to calculate to do the optimization, and the contribution of the data could go away.

(c) CrossEntropy(P_{data}, P_θ)

(d) if $P_\theta(x_i) = 0$ then $\log P_\theta(x_i) = -\infty$, and one cannot optimize the MLE (or cross entropy) $E_{x\sim P_{data}}\left[\log P_\theta(x)\right] = -\infty$

2. (**5 pts each/40 pts total**) **About generative versus discriminative models**.

Which of these tasks is **G**enerative (no labels needed) versus **D**iscriminative (external labels needed to determine the task) :

(a) To predict whether two genes are expressed together in a cell D

(b) To predict missing (masked) amino acids in a protein G

(c) To predict a new RNA homolog for an RNA family G

(d) To predict the gene-expression profile of "new" sampled cells G

(e) To predict the binding affinity of an RNA to a protein D

(f) To predict cellular responses to drug-induced perturbations D

(g) To predict the future state of a cell G

(h) To predict the type of a cell from its expression profile D