CS 6220: Machine Learning Theory

Problem Set 4

Your Name

November 13, 2026

Problem 1

Show that the softmax function is invariant to adding a constant to all its inputs, and use this fact to derive a numerically stable implementation.

Solution. Let $z \in \mathbb{R}^n$ and $c \in \mathbb{R}$. By definition,

$$\text{softmax}(z + c)_i = \frac{e^{z_i + c}}{\sum_j e^{z_j + c}} = \frac{e^c e^{z_i}}{e^c \sum_j e^{z_j}} = \text{softmax}(z)_i.$$

(a) Since the identity holds for any c, choose $c = -\max_i z_i$ so that every exponent $z_i + c \leq 0$, avoiding overflow.

(b) The stabilized computation is

$$\text{softmax}(z)_i = \frac{e^{z_i - \max_j z_j}}{\sum_k e^{z_k - \max_j z_j}}.$$

Problem 2

Prove that the KL divergence $D_{KL}(P\|Q) \geq 0$ for probability distributions P, Q over a finite set, with equality iff $P = Q$.

Proof. By definition,

$$D_{KL}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

By Jensen's inequality applied to the concave function log,

$$-D_{KL}(P\|Q) = \sum_i P(i) \log \frac{Q(i)}{P(i)} \leq \log \sum_i P(i) \frac{Q(i)}{P(i)} = \log \sum_i Q(i) = \log 1 = 0.$$

Hence $D_{KL}(P\|Q) \geq 0$, with equality iff $Q(i)/P(i)$ is constant over the support, i.e. $P = Q$. ■

Problem 3

Compute the gradient of the cross-entropy loss with respect to the logits when combined with a softmax output layer.

Solution. For logits z, softmax output $\hat{y} = \mathrm{softmax}(z)$, and one-hot target y, the loss is $L = -\sum_i y_i \log \hat{y}_i$. Differentiating gives the well-known simplification

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i.$$

Collaborators: discussed Problem 2 with Mara Okafor during office hours.