INFR10086 Machine Learning (MLG) Semester 2, 2024/5

Practice Exam

1. In this question, we will look at the hinge loss for binary classification. Recall that a linear classifier has the form

$$f(x) = \begin{cases} +1 & \text{if } w^\top x \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{1}$$

The hinge loss for binary classification with linear classifier is defined as

$$L_{\text{hinge}}(x, y; w) = \max(1 - yw^\top x, 0), \tag{2}$$

where $x \in \mathbb{R}^d$ and $y \in \{+1, -1\}$.

(a) Show that the hinge loss is an upper bound of the zero-one loss

$$L_{01}(x, y; w) = \mathbb{1}_{yw^\top x < 0}. \tag{3}$$

In other words, show that

$$L_{01}(x, y; w) \leq L_{\text{hinge}}(x, y; w) \tag{4}$$

for all $x \in \mathbb{R}^d$, $y \in \{+1, -1\}$, and $w \in \mathbb{R}^d$.

(b) In the following three steps, we will look at the convexity of hinge loss.

(i) Show that

$$\max(a + b, c + d) \leq \max(a, c) + \max(b, d) \tag{5}$$

for any $a, b, c, d \in \mathbb{R}$.

(ii) Let

$$h(x) = \max(f(x), g(x)) \tag{6}$$

for any two convex functions f and g. Use (b) and show that h is convex in x.

(iii) Use (c) and show that the hinge loss L_{hinge} is convex in w for any $x \in \mathbb{R}^d$ and $y \in \{+1, -1\}$.

(c) If we happen to find a linear classifier that achieves a hinge loss of 0 on a data set, what does that tell us about the optimal value of log loss on that particular data set?

2. In this question, we are going to implement a layer called layer normalization in a neural network library. Formally, layer normalization is a function

$$f(x) = \begin{bmatrix} \frac{x_1-\mu}{\sigma} \\ \frac{x_2-\mu}{\sigma} \\ \vdots \\ \frac{x_d-\mu}{\sigma} \end{bmatrix} \tag{7}$$

where

$$\mu = \frac{1}{d}\sum_{i=1}^{d} x_i \qquad \sigma^2 = \frac{1}{d}\sum_{i=1}^{d}(x_i-\mu)^2 \tag{8}$$

(a) Show that

$$\sigma^2 = \frac{1}{d}\sum_{i=1}^{d} x_i^2 - \mu^2. \tag{9}$$

(b) The forward function is as defined, and is straightforward to implement. The backward function (as part of the backpropagation) is more involved. Given the forward computation, the backward computation can be worked out using the total derivative

$$\frac{\partial L}{\partial x_j} = \sum_{i=1}^{d} \frac{\partial L}{\partial f_i}\frac{\partial f_i}{\partial x_j} + \frac{\partial L}{\partial \mu}\frac{\partial \mu}{\partial x_j} + \frac{\partial L}{\partial \sigma}\frac{\partial \sigma}{\partial x_j}, \tag{10}$$

where f_i is a shorthand for the i-th coordinate of $f(x)$ and L is the loss function. Note that $\partial L/\partial f_i$ will be given during backpropagation. Our goal is the derive the rest of the terms.

i. Show that

$$\frac{\partial \mu}{\partial x_j} = \frac{1}{d}. \tag{11}$$

ii. Show that

$$\frac{\partial f_i}{\partial x_j} = \frac{1}{\sigma}\mathbb{1}_{i=j}, \tag{12}$$

where $\mathbb{1}_c$ is 1 when c is true and 0 otherwise.

iii. Show that

$$\frac{\partial \sigma}{\partial x_j} = \frac{1}{\sigma d}x_j \tag{13}$$

iv. Show that

$$\frac{\partial L}{\partial \sigma} = \sum_{i=1}^{d} \frac{\partial L}{\partial f_i}\left(-\frac{x_i-\mu}{\sigma^2}\right). \tag{14}$$

v. Show that

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{d} \frac{\partial L}{\partial f_i} \left(\frac{-1}{d} \right) + \frac{\partial L}{\partial \sigma} \left(-\frac{\mu}{\sigma} \right). \tag{15}$$

3. Suppose we have a data set organized as a matrix X where each row vector is a sample point. We know that the first principal component of X is a vector w_1 such that

$$w_1 = \operatorname*{argmax}_{w} \frac{w^\top X^\top X w}{w^\top w} \tag{16}$$

   (a) Show that if w_1 is the optimal solution for $\max_w \frac{w^\top X^\top X w}{w^\top w}$, then $a w_1$ is also an optimal solution for any $a \neq 0$.

   (b) Suppose we rotate the entire data set by a rotation matrix R, where $R^\top R = \mathrm{I}$. Show that if w_1 is the first principal component of X, then $R w_1$ is the first principal component of the rotated data set XR.