ECE421 - Winter 2022
Homework Problems - Tutorial #3

Theme: Gradients and Logistic Regression

Due: February 6, 2022 11:59 PM

Question 1 (Gradient Computation)

For a scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$, the gradient evaluated at $w \in \mathbb{R}^d$ is

$$\nabla f(w) = \begin{bmatrix} \frac{\partial f(w)}{\partial w_1} & \cdots & \frac{\partial f(w)}{\partial w_d} \end{bmatrix}^\top \in \mathbb{R}^d.$$

Using this definition, compute the gradients of following functions, where $A \in \mathbb{R}^{d \times d}$ is *not* necessarily a symmetric matrix.

(i) $f(w) = w^\top A v + w^\top A^\top v + v^\top A w + v^\top A^\top w$, $v \in \mathbb{R}^d$

(ii) $f(w) = w^\top A w$

Compute the gradients of following functions using above definition and the chain rule.

(iii) $f(w) = \sum_{i=1}^d \log(1 + \exp(w_i))$

(iv) $f(w) = \sqrt{1 + \|w\|_2^2}$

Question 2 (Logistic Regression)

You are given a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, where $x_n \in \mathbb{R}^d, d \geq 1$, and $y_n \in \{+1, -1\}$. For $w \in \mathbb{R}^{d+1}$ and $x \in \mathbb{R}^{d+1}$, we wish to train a logistic regression model

$$h(x) = \theta(b + \textstyle\sum_{i=1}^d w_i x_i) = \theta(w^\top x), \tag{1}$$

where $\theta(z) = \dfrac{e^z}{1+e^z}, z \in \mathbb{R}$ is the logistic function. Following the arguments on page 91 of LFD, the in-sample error can be written as

$$E_{\text{in}}(w) = \frac{1}{N} \textstyle\sum_{n=1}^N \log\left[\dfrac{1}{P_w(y_n|x_n)}\right], \tag{2}$$

where

$$P_w(y|x) = \begin{cases} h(x) & y = +1 \\ 1 - h(x) & y = -1 \end{cases}. \tag{3}$$

(a) Show that $E_{\text{in}}(w)$ can be expressed as

$$E_{\text{in}}(w) = \frac{1}{N}\left(\textstyle\sum_{n=1}^N [\![y_n = +1]\!] \log\left[\dfrac{1}{h(x_n)}\right] + [\![y_n = -1]\!] \log\left[\dfrac{1}{1-h(x_n)}\right]\right), \tag{4}$$

where $[\![\text{argument}]\!]$ evaluates to 1 if the argument is true and 0 if it is false.

(b) Show that $E_{\text{in}}(w)$ can also be expressed as

$$E_{\text{in}}(w) = \frac{1}{N} \textstyle\sum_{n=1}^N \log(1 + \exp(-y_n w^\top x_n)). \tag{5}$$

(c) Use (5) to show that $\nabla E_{\text{in}}(w) = \dfrac{1}{N} \sum_{n=1}^N -y_n x_n \theta(-y_n w^\top x_n)$, and argue that a "misclassified" example contributes more to the gradient than a correctly classified one.

(d) Show that $\nabla E_{\text{in}}(w)$ can be expressed as

$$\nabla E_{\text{in}}(w) = \frac{1}{N} X^\top p, \tag{6}$$

for some expression p, where X is the data matrix you are familiar with from linear regression. What is p and how does it compare with the gradient of the in-sample error of linear regression?

Question 3 (Problem 4, Midterm 2017)

Consider the logistic regression setup as in the previous question. Suppose we are given a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$ with

$$x_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top, y_1 = 1 \quad \text{and} \quad x_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top, y_2 = -1.$$

We consider the l_2-regularized error as

$$E_{\text{in}}(w) = -\sum_{n=1}^{N} \log \left[P_w(y_n | x_n) \right] + \lambda \|w\|_2^2, \lambda > 0, \tag{7}$$

where

$$P_w(y|x) = \begin{cases} h(x) & y = +1 \\ 1 - h(x) & y = -1 \end{cases}, \tag{8}$$

and $h(x) = \dfrac{e^{w^\top x}}{1 + e^{w^\top x}} = \dfrac{1}{1 + e^{-w^\top x}}$.

(a) For $\lambda = 0$, find the optimal w that minimizes $E_{\text{in}}(w)$ and the minimum value of $E_{\text{in}}(w)$. (Hint: you are given x_n, y_n, so plug those values into the expression of the in-sample error).

(b) Suppose λ is a very large constant such that it suffices to consider weights that satisfy $\|w\|_2 \ll 1$. Since w has a small magnitude, we may use the Taylor series approximation

$$\log(1 + \exp(-y_n w^\top x_n)) \approx \log(2) - \frac{1}{2} y_n w^\top x_n. \tag{9}$$

Assuming the above approximation is exact, find w that minimizes $E_{\text{in}}(w)$ (it should be expressed in terms of λ).