CS280 Spring 2025 Assignment 1
Part A

Basics

February 25, 2025

Name:

Student ID:

1. ***Maximum Likelihood Estimation*** **(10 points).**

Consider a dataset $\mathcal{D}$ consisting of n independent and identically distributed samples:

$$\mathcal{D} = \left\{((x_1^1, x_2^1), y^1), ((x_1^2, x_2^2), y^2), \ldots, ((x_1^n, x_2^n), y^n)\right\}, \tag{1}$$

where $(x_1^i, x_2^i) \in \mathbb{R}^2$ are input features and $y^i \in \mathbb{R}$ is an output.
Assume that every output y^i in $\mathcal{D}$ is generated by inputting (x_1^i, x_2^i) into a model:

$$y = f_{\theta_1,\theta_2}(x_1, x_2) + \epsilon, \tag{2}$$

where the function f_{θ_1,θ_2} is a mapping from features $(x_1, x_2) \in \mathbb{R}^2$ to a value in $\mathbb{R}$, which has two parameters θ_1 and θ_2. Here we assume that the random noise $\epsilon \sim N(0, \sigma^2)$ is independent and distributed according to a Gaussian distribution with zero mean and variance σ^2.

(a) Show that the log likelihood of the data given the parameters is:

$$l(\mathcal{D}; \theta_1, \theta_2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i))^2 - n\log(\sqrt{2\pi}\sigma). \tag{3}$$

Recall the probability density function of the Gaussian distribution $N(\mu, \sigma^2)$ is:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{4}$$

(b) To find the maximum likelihood estimates of θ_1 and θ_2 using gradient descent, compute the gradient of the log likelihood with respect to θ_1 and θ_2. Express you answer in terms of:

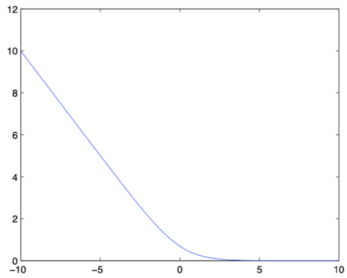
$$y^i, \;\; f_{\theta_1,\theta_2}(x_1^i, x_2^i), \;\; \frac{\partial}{\partial\theta_1}f_{\theta_1,\theta_2}(x_1^i, x_2^i), \;\; \frac{\partial}{\partial\theta_2}f_{\theta_1,\theta_2}(x_1^i, x_2^i)$$

(c) Given the learning rate η, what update rule would you use in gradient descent to *maximize* the likelihood.

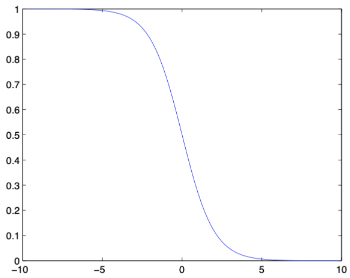
2. *Loss Function* (10 points).

Assume that a classifier is written as $H(x) = sign(F(x))$, where $H(x) : \mathbb{R}^d \rightarrow \{-1, 1\}$, $sign()$ is a sign function, and $F(x) : \mathbb{R}^d \rightarrow \mathbb{R}$. To obtain the parameters in $F(x)$, we need to minimize the loss function averaged over the training set: $\sum_i L(y^i F(x^i))$. Here L is a function of $yF(x)$. For example, for linear classifiers, $F(x) = w_0 + \sum_{j=1}^{d} w_j x_j$, and $yF(x) = y(w_0 + \sum_{j=1}^{d} w_j x_j)$.

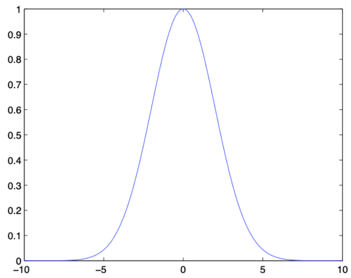
(a) Which loss functions below are appropriate to use in classification? For the ones that are not appropriate, explain why not. In general, what conditions does L have to satisfy in order to be an appropriate loss function? The x axis is $yF(x)$, and the y axis is $L(yF(x))$.



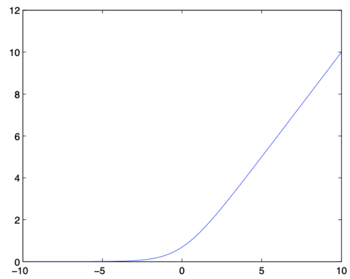
(a)



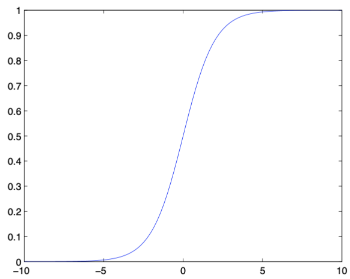
(b)



(c)



(d)



(e)

(b) Among the above loss functions appropriate to use in classification, which one is the most robust to outliers? Justify your answer.

(c) Let $F(x) = w_0 + \sum_{j=1}^{d} w_j x_j$ and $L(yF(x)) = \frac{1}{1+exp(yF(x))}$. Suppose you use gradient descent to obtain the optimal values for w_0 and w_j. Give the update rules for these parameters.