CS280 Spring 2025 Assignment 1
Part A

Basics

March 4, 2025

Name: Chen Xuanxin

Student ID: 2024233125

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:

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

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

1 Answer for Maximum Likelihood Estimation

(a) Derivation of Log-Likelihood

Given the noise $\epsilon \sim N(0, \sigma^2)$, the conditional distribution of each sample y^i is:

$$y^i \mid x_1^i, x_2^i \sim N\left(f_{\theta_1,\theta_2}(x_1^i, x_2^i), \sigma^2\right).$$

Using the probability density function of the Gaussian distribution (Equation (4)), the likelihood for a single sample is:

$$p(y^i \mid x_1^i, x_2^i, \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i))^2}{2\sigma^2}\right).$$

Since the samples are independent and identically distributed (i.i.d.), the total likelihood is the product of individual likelihoods:

$$L(\mathcal{D}; \theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i))^2}{2\sigma^2}\right).$$

Taking the logarithm and simplifying:

$$\begin{aligned} l(\mathcal{D}; \theta_1, \theta_2) &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i))^2}{2\sigma^2}\right)\right) \\ &= \sum_{i=1}^{n} \left(-\ln(\sqrt{2\pi}\sigma) - \frac{(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i))^2}{2\sigma^2}\right) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i))^2 - n\ln(\sqrt{2\pi}\sigma). \end{aligned}$$

Thus, the log-likelihood is:

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

(b) Gradient Computation

The partial derivative with respect to θ_1 is:

$$\begin{aligned} \frac{\partial l}{\partial \theta_1} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_1} \left(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i)\right)^2 \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2\left(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i)\right) \cdot \left(-\frac{\partial f_{\theta_1,\theta_2}}{\partial \theta_1}\right) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i)\right) \frac{\partial f_{\theta_1,\theta_2}}{\partial \theta_1}. \end{aligned}$$

Similarly, the partial derivative with respect to θ_2 is:

$$\frac{\partial l}{\partial \theta_2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i)\right) \frac{\partial f_{\theta_1,\theta_2}}{\partial \theta_2}.$$

The gradient vector is:

$$\nabla_{\theta_1,\theta_2} l = \left[\frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i)\right) \frac{\partial f_{\theta_1,\theta_2}}{\partial \theta_1}, \ \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i)\right) \frac{\partial f_{\theta_1,\theta_2}}{\partial \theta_2} \right]$$

(c) Gradient Descent Update Rule To maximize the likelihood, we use gradient ascent. The update rule for parameters is:

$$\theta_j^{(k+1)} = \theta_j^{(k)} + \eta \cdot \frac{\partial l}{\partial \theta_j}, \quad j = 1, 2.$$

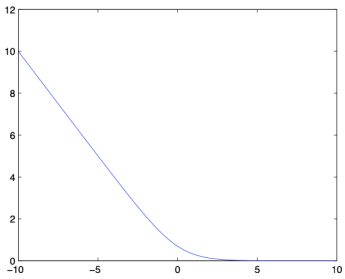
Substituting the gradient expressions, we get:

$$\theta_j^{(k+1)} = \theta_j^{(k)} + \frac{\eta}{\sigma^2} \sum_{i=1}^{n} \left(y^i - f_{\theta_1,\theta_2}(x_1^i, x_2^i)\right) \frac{\partial f_{\theta_1,\theta_2}}{\partial \theta_j}, \quad j = 1, 2$$

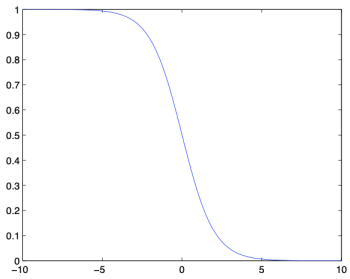
2. *Loss Function* (10 points).

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 $yF(x) = y(w_0 + \sum_{j=1}^{d} w_j x_j)$.

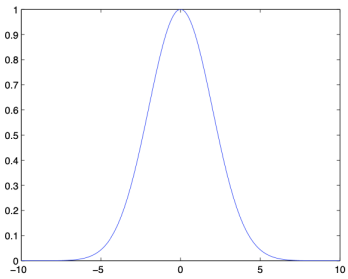
(a) Which loss functions below are appropriate to use in classification? For the ones that are not appropriate, explain why not. In general, what conditions does L have to satisfy in order to be an appropriate loss function? The x axis is $yF(x)$, and the y axis is $L(yF(x))$.



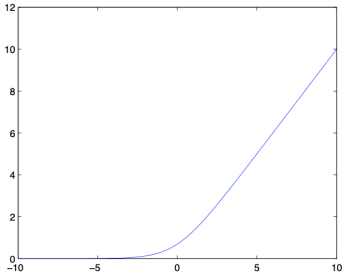
(a)



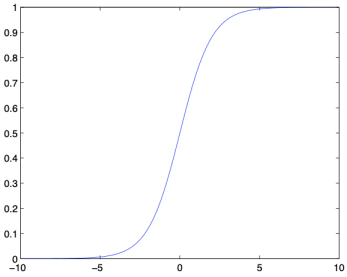
(b)



(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.

2 Answer for Loss Functions

(a) Appropriate loss functions

From the provided plots (a)–(e) (with the x-axis being $yF(x)$ and the y-axis being $L(yF(x))$), we see that only (a) and (b) are monotonically decreasing and approach 0 as $yF(x) \to +\infty$. These two satisfy the typical requirements for classification loss:

- $L(t) \geq 0$ for all t,
- $L(t)$ is decreasing in t,
- $L(t) \to 0$ as $t \to +\infty$ (correct classification with large margin),
- $L(t)$ becomes large when $t \ll 0$ (incorrect classification with large negative margin).

Analysis of subfigures (a)-(e):

Subfigure	Characteristics	Appropriate?
(a)	Exponentially decreasing curve: $L \approx 10 \to 0$ as $yF(x)$ increases	✓
(b)	Saturated loss: Flat regions for $\|yF(x)\| > 2$, sharp transition near $yF(x) = 0$	✓
(c)	Bell-shaped curve: Peaks at $yF(x) = 0$ (*opposite* to required monotonicity)	×
(d)	Increasing linear function: $L \propto yF(x)$ (*violates* the decreasing requirement)	×
(e)	Logistic-like curve but reversed sign: does not decrease as $yF(x)$ grows	×

Only (a) and (b) are *monotonically decreasing* in $yF(x)$ and approach 0 when $yF(x) \to +\infty$. Hence, (a) and (b) are appropriate for classification; (c), (d), and (e) are not.

(b) Robustness to outliers

Among the appropriate loss functions, (a) resembles the exponential loss, while (b) resembles a logistic-type loss. The exponential loss $\exp(-yF(x))$ grows very large for highly negative margins and is therefore more sensitive to outliers. The logistic loss saturates more smoothly, making it more robust to outliers. Hence, (b) is the most robust among them.

(c) Gradient descent updates for $L\big(yF(x)\big) = \frac{1}{1+\exp\left(yF(x)\right)}$

Let the training set be $\{(x^i,\, y^i)\}_{i=1}^n$, where $y^i \in \{-1, +1\}$, and define

$$F(x^i) \;=\; w_0 \;+\; \sum_{j=1}^{d} w_j\, x_j^i.$$

Then the loss for sample i is

$$L_i \;=\; L\big(y^i F(x^i)\big) \;=\; \frac{1}{1+\exp\big(y^i F(x^i)\big)}.$$

We want to minimize the sum $\displaystyle\sum_{i=1}^{n} L_i$.

Partial derivatives. Let $z_i = y^i F(x^i)$. Then

$$L_i \;=\; \frac{1}{1+e^{z_i}}, \quad \frac{\partial L_i}{\partial z_i} = \;-\frac{e^{z_i}}{\big(1+e^{z_i}\big)^2} = \;-L_i\,(1-L_i).$$

Note also that $\dfrac{\partial z_i}{\partial w_0} = y^i$, and $\dfrac{\partial z_i}{\partial w_j} = y^i\,x_j^i$.

Hence,

$$\frac{\partial L_i}{\partial w_0} = \;\frac{\partial L_i}{\partial z_i}\,\frac{\partial z_i}{\partial w_0} = \;-y^i\,L_i\,(1-L_i),$$

$$\frac{\partial L_i}{\partial w_j} = \;-y^i\,x_j^i\,L_i\,(1-L_i).$$

Gradient descent updates. Let α be the learning rate. Then the update rules for each iteration are:

$$w_0 \;\leftarrow\; w_0 \;-\; \alpha\sum_{i=1}^{n}\frac{\partial L_i}{\partial w_0} \;=\; w_0 \;+\; \alpha\sum_{i=1}^{n}\Big[y^i\,L_i\,\big(1-L_i\big)\Big],$$

$$w_j \;\leftarrow\; w_j \;-\; \alpha\sum_{i=1}^{n}\frac{\partial L_i}{\partial w_j} \;=\; w_j \;+\; \alpha\sum_{i=1}^{n}\Big[y^i\,x_j^i\,L_i\,\big(1-L_i\big)\Big],$$

where $L_i = \dfrac{1}{1+\exp\big(y^i F(x^i)\big)}$.