HW2 Handwritten Assignment

Yi-Lun Lee
b11901057@ntu.edu.tw

Che-Wei Tsai
f13945039@ntu.edu.tw

October 2025

Tools You Need to Know/Learn

- Backpropagation, Layer Normalization, Chain Rule
- Classification using GMM, Bayes' Rule
- AdaBoost, GradientBoost
- Bias–Variance Decomposition, Bregman Divergence, Kullback–Leibler Divergence
- Gradient Descent Convergence

You are expected to know (or learn through this homework) the definitions and usages of the above concepts, and are encouraged to discuss with TAs during TA hour if you encounter difficulties.

Homework Policy

- The homework is graded out of 100 points.
- The official submission deadline will follow the schedule announced on NTU COOL.
- Homework may be handwritten or typed (e.g. using LaTeX), but must be submitted in **PDF format**.
- If you discuss the homework with classmates, you should state their student IDs in your submission.
- Plagiarism is strictly prohibited. Serious violations will be dealt with according to NTU regulations.

Problem 1 : (Layer Normalization) (20%)

Layer normalization is a popular normalization technique used in deep neural networks, particularly in **transformer** architectures. Unlike batch normalization which normalizes across the batch dimension, layer normalization normalizes across the feature dimension for each individual sample. The alogrithm can be written as below:

Algorithm 1 Layer Normalization

Input Feature vector from a single data point $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ where d is the feature dimension
Output $\mathbf{y} = LN_{\gamma,\beta}(\mathbf{x})$

procedure LayerNormalize($\mathbf{x}$, γ, β)
  $\mu \leftarrow \frac{1}{d}\sum_{i=1}^{d} x_i$ ▷ layer mean
  $\sigma^2 \leftarrow \frac{1}{d}\sum_{i=1}^{d}(x_i - \mu)^2$ ▷ layer variance
  for $i \leftarrow 1$ to d **do**
    $\hat{x}_i \leftarrow \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$ ▷ normalize
    $y_i \leftarrow \gamma \hat{x}_i + \beta$ ▷ scale and shift
  end for
  return
end procedure

During training we need to backpropagate the gradient of loss ℓ through this transformation, as well as compute the gradients with respect to the parameters γ, β.

(a) (15pts) Towards this end, please write down the close form expressions for $\frac{\partial \ell}{\partial x_i}$, $\frac{\partial \ell}{\partial \gamma}$, $\frac{\partial \ell}{\partial \beta}$ in terms of x_i, μ, σ^2, $\hat{x}_i$, y_i (given by the forward pass) and $\frac{\partial \ell}{\partial y_i}$ (given by the backward pass).
- Hint: You may first write down the close form expressions of $\frac{\partial \ell}{\partial \hat{x}_i}$, $\frac{\partial \ell}{\partial \sigma^2}$, $\frac{\partial \ell}{\partial \mu}$, and then use them to compute $\frac{\partial \ell}{\partial x_i}$, $\frac{\partial \ell}{\partial \gamma}$, $\frac{\partial \ell}{\partial \beta}$.

(b) (5pts) Please explain why layer normalization needs learnable parameters γ and β to scale and shift after normalization.
- Hint: What if different features should have different importance levels?

Problem 2 : (Classification with Gaussian Mixture Model) (20%)

In this question, we tackle the binary classification problem through the generative approach, where we assume the data point X (viewed as a $\mathbb{R}^d$-valued r.v.) and its label Y (viewed as a $\{\mathcal{C}_1, \mathcal{C}_2\}$-valued r.v.) are generated according to the generative model (paramerized by θ) as follows:

$$\mathbb{P}_\theta[X = \mathbf{x}, Y = \mathcal{C}_k] = \pi_k f_{\boldsymbol{\mu}_k, \Sigma_k}(\mathbf{x}) \quad (k \in \{1, 2\}) \tag{1}$$

where $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ for which

$$f_{\boldsymbol{\mu}_k, \Sigma_k}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right).$$

Now suppose we observe data points $\mathbf{x}_1, ..., \mathbf{x}_N$ and their corresponding labels $y_1, ..., y_N$, and $\pi_1 + \pi_2 = 1$.

(a) (i) (2pts) Please write down the likelihood function $L(\theta)$ that describes how likely the generative model would generate the observed data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$.

(ii) (4pts) (Find the maximum likelihood estimate $\theta^* = (\pi_1^*, \pi_2^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma_1^*, \Sigma_2^*)$ that maximizes the likelihood function $L(\theta)$.

(iii) (2pts) Write down $\mathbb{P}_\theta[Y = \mathcal{C}_1 | X = \mathbf{x}]$ and $\mathbb{P}_\theta[X = \mathbf{x} | Y = \mathcal{C}_1]$ in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$. What are the physical meaning of the aforementioned quantities?

(iv) (2pts) Express $\mathbb{P}_\theta[Y = \mathcal{C}_1 | X = \mathbf{x}]$ in the form of $\sigma(z)$, where $\sigma(\cdot)$ denotes the sigmoid function, and express z in terms of $\theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ and x.

(b) (10pts) Suppose we pose an additional constraint that the covariance matrices of the two Gaussian distributions are identical, namely $\Sigma_1 = \Sigma_2 = \Sigma$, in which the generative model is parameterized by $\vartheta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. Redo questions (a) under such setting.

Problem 3 : (Multiclass AdaBoost) (20%)

Let $\mathcal{X}$ be the input space, $\mathcal{F}$ be a collection of multiclass classifiers that map from $\mathcal{X}$ to $[1, K]$, where K denotes the number of classes. Let $\{(x_i, \hat{y}_i)\}_{i=1}^m$ be the training data set, where $x_i \in \mathcal{X}$ and $\hat{y}_i \in [1, K]$. Given $T \in \mathbb{N}$, suppose we want to find functions

$$g_{T+1}^k(x) = \sum_{t=1}^{T} \alpha_t f_t^k(x), \quad k \in [1, K]$$

where $f_t \in \mathcal{F}$ and $\alpha_t \in \mathbb{R}$ for all $t \in [1, T]$. Here for $f \in \mathcal{F}$, we denote $f^k(x) = \mathbf{1}\{f(x) = k\}$, where $\mathbf{1}(\cdot)$ is an indicator function, as the k'th element in the one-hot representation of $f(x) \in [1, K]$. The aggregated classifier $h : \mathcal{X} \to [1, K]$ is defined as

$$x \mapsto \underset{1 \le k \le K}{\operatorname{argmax}}\ g_{T+1}^k(x)$$

Please apply gradient boosting to show how the functions f_t and coefficients α_t are computed with an aim to minimize the following loss function

$$L((g_{T+1}^1, \cdots, g_{T+1}^K) = \sum_{i=1}^{m} \exp\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{T+1}^k(x_i) - g_{T+1}^{\hat{y}_i}(x_i)\right)$$

(10 pts for f_t and 10 pts for α_t, and derivation process is needed.)

Problem 4 : (Bias–Variance Decomposition for Bregman Divergences) (20%)

Definition 1 (Strict Convexity)**.** *A differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ is said to be* strictly convex *if for all distinct $p, q \in \mathbb{R}^d$ and all $\lambda \in (0, 1)$,*

$$\phi(\lambda p + (1 - \lambda)q) < \lambda\phi(p) + (1 - \lambda)\phi(q).$$

Definition 2 (Bregman Divergence)**.** *The* Bregman divergence *associated with $\phi : \mathbb{R}^d \to \mathbb{R}$ is defined as*

$$D_\phi(p\|q) \;=\; \phi(p) - \phi(q) - \langle \nabla\phi(q),\, p - q\rangle,$$

for $p, q \in \mathbb{R}^d$.

(a) (4pts) Prove that strict convexity implies uniqueness in the sense that

$$D_\phi(p\|q) = 0 \quad \Longleftrightarrow \quad p = q.$$

(b) (6pts) Let $d = 1$. Suppose the true value is a fixed scalar $y \in \mathbb{R}$, and a learning algorithm produces a random prediction $\hat{Y}$ (due to training data or algorithm randomness). Denote the mean prediction by

$$\bar{Y} = \mathbb{E}[\hat{Y}].$$

Prove that

$$\mathbb{E}\big[D_\phi(y\|\hat{Y})\big] = D_\phi(y\|\bar{Y}) + \mathbb{E}\big[D_\phi(\bar{Y}\|\hat{Y})\big].$$

is true if and only if $\langle \nabla\phi(\bar{Y}) - \mathbb{E}(\nabla\phi(\hat{Y})),\, y - \bar{Y}\rangle = 0$

(c) (5pts) Let $d = 1$. Verify the decomposition for the special case $\phi(t) = f_1(t), t \in \mathbb{R}$, find $f_1(t)$ that can recover the classical MSE bias–variance identity

$$\mathbb{E}[(\hat{Y} - y)^2] = (\bar{Y} - y)^2 + \mathbb{E}[(\hat{Y} - \bar{Y})^2].$$

(d) (5pts) Let $\phi(t) = f_2(t),\ t \in (\mathbf{0}, \infty)^d$. Denote each component of the vectors $p = [p_1, p_2, \ldots, p_d]^T$ and $q = [q_1, q_2, \ldots, q_d]^T$. Show that we can properly select $\phi = f_2$ such that

$$D_\phi(p\|q) = \sum_{i=1}^{d} \left(p_i \log \frac{p_i}{q_i} - p_i + q_i \right).$$

Furthermore, if p and q represent probability mass functions (so $\sum_{i=1}^d p_i = \sum_{i=1}^d q_i = 1$), show that in this case the Bregman divergence reduces to the Kullback–Leibler divergence:

$$D_\phi(p\|q) = \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i}.$$

(hint: For (c) and (d), you may use the result of (b))

Problem 5 : (Gradient Descent Convergence) (20%)

Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Also, f is β-smoothness and α-strongly convex.

$$\beta\text{-smoothness} : \beta > 0, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

$$\alpha\text{-strongly convex} : \alpha > 0, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \geq \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Then we propose a gradient descent algorithm

(a) Find a initial $\boldsymbol{\theta}^0$.

(b) Let $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)$ $\forall n \geq 0$, where $\eta = \frac{1}{\beta}$.

The following problems lead you to prove the gradient descent convergence.

(a) (8pts) Prove the property of β-smoothness function

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \leq \frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$$

i. Define $g : \mathbb{R} \to \mathbb{R}, g(t) = f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$. Show that $f(\boldsymbol{y}) - f(\boldsymbol{x}) = \int_0^1 g'(t)\, dt$.

ii. Show that $g'(t) = \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))^T(\boldsymbol{y} - \boldsymbol{x})$.

iii. Show that $|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})| \leq \int_0^1 |(\nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}))^T(\boldsymbol{y} - \boldsymbol{x})|\, dt$.

iv. By Cauchy-Schwarz inequality and the definition of β-smoothness, show that $|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})| \leq \frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$, hence we get

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \leq \frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$$

(b) (4pts) Let $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})$ and use 1., prove that

$$f\left(\boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})\right) - f(\boldsymbol{x}) \leq -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2$$

and

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2,$$

where $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$.

(c) (2pts) Show that $\forall n \geq 0$,

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 = \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 + \eta^2\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)\|_2^2 - 2\eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)^T(\boldsymbol{\theta}^n - \boldsymbol{\theta}^*),$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$.

(d) (4pts) Use 2. and the definition of α-strongly convex to prove $\forall n \geq 0$

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 \leq (1 - \frac{\alpha}{\beta})\|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2,$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$.

(e) (2pts) Use the above inequality to show that $\boldsymbol{\theta}^n$ will converge to $\boldsymbol{\theta}^*$ when n goes to infinity.