기계학습원론 기말고사
2025 가을

강의자: 조재민

강의조교: 이재웅

이름: ________________________

학번: ________________________

2025년 12월 8일

문제 1. 명제의 참(O) 또는 거짓(X)을 판단하시오. (정답: +2점, 오답: -2점, 미기재: 0점)

1. 연속확률변수 X에서 임의의 실수 $x \in \mathbb{R}$에 대해 $P(X = x) = 0$이다.
2. 배깅(Bagging)에서 크기가 n인 데이터셋에 대해 복원 추출로 n개의 샘플을 뽑을 때, $n \to \infty$이면 각 데이터 포인트가 선택되지 않을 확률은 $1/e$로 수렴한다.
3. Disentangled representation의 각 잠재 차원(latent dimension)은 서로 간의 상관관계가 낮다.
4. Soft-Margin SVM에서 마진 바깥에 올바르게 분류된 데이터 포인트들은 슬랙 변수 (slack variable) $\xi_i = 0$을 가지며, 힌지 손실(hinge loss)이 0이므로 서포트 벡터에 영향을 미치지 않는다.
5. 임의의 함수 $f(x, y)$에 대해 항상 $\min_x \max_y f(x, y) \leq \max_y \min_x f(x, y)$가 성립한다.
6. 입력 벡터의 모든 성분에 같은 수를 곱하더라도 소프트맥스(Softmax) 함수를 적용한 결괏값은 변하지 않는다.
7. 퍼셉트론 학습 알고리즘(Perceptron Learning Algorithm, PLA)에서, 각 샘플에 대해 가중치 업데이트를 수행할 때마다 전체 훈련 데이터에 대한 오분류 개수는 항상 감소하거나 그대로 유지된다.
8. t-SNE의 하이퍼파라미터인 *Perplexity*는 높게 설정할수록 고차원 공간에서의 전역적인 구조를 더 반영하며, 이렇게 얻은 저차원 임베딩의 각 축은 원본 데이터의 구조를 잘 설명하는 축이다.
9. 다음 제약 최적화(Constrained optimization) 문제를 고려하자:

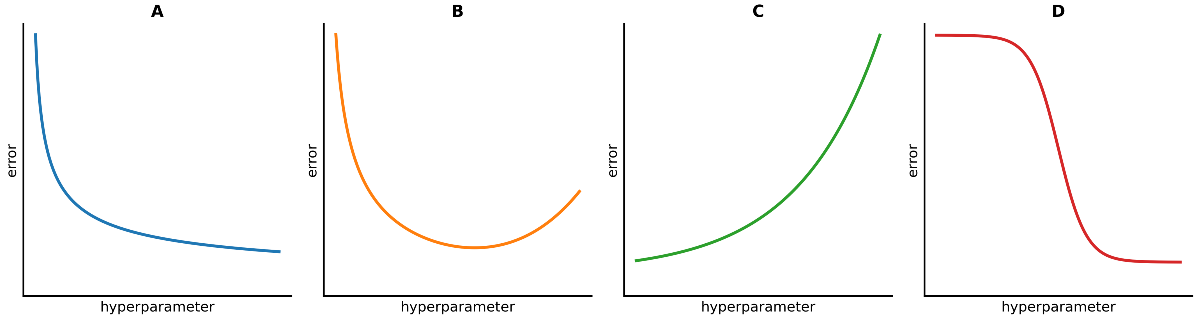
$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to } g_i(\mathbf{x}) = 0,\ h_j(\mathbf{x}) \leq 0.$$

   이에 대한 라그랑지안(Lagrangian)은 다음과 같다:

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) + \sum_j \mu_j h_j(\mathbf{x})$$

   KKT 조건(Karush-Kuhn-Tucker Conditions)을 만족하는 점 $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$에 대해, 다음이 항상 성립하는지 판단하시오. (단, 모든 수식 도출은 올바르다고 보아도 좋다.)
   (1) $\nabla_{\mathbf{x}} f(\mathbf{x}^\star) = \mathbf{0}$이다.
   (2) 제약 $h_j(\mathbf{x}) \leq 0$에 대해, $h_j(\mathbf{x}^\star) < 0$이면 $\mu_j^\star = 0$ 이고 $h_j(\mathbf{x}^\star) = 0$이면 $\mu_j^\star \geq 0$이다.

문제 2. **[8점]** 여러 머신러닝 알고리즘에 대해, 하이퍼파라미터 값에 따라 훈련 오차(Training Error) 또는 테스트 오차(Testing Error)가 변할 수 있는 몇 가지 가능한 방식은 다음과 같다:



아래에서 제시하는 각각의 하이퍼파라미터와 관련하여, 위 그림 중 어떤 것이 훈련 오차와 테스트 오차에 대한 가장 일반적인 경향을 나타내는지 고르시오.

1. 경사 하강법(Gradient Descent)의 학습률(Learning Rate, η) (반복 횟수는 고정):
   i. 훈련 오차 (Training Error): ○ A ○ B ○ C ○ D
   ii. 테스트 오차 (Testing Error): ○ A ○ B ○ C ○ D
2. 로지스틱 회귀(Logistic Regressions)에서의 L2 가중치 감쇠(Weight Decay) 계수 λ:
   i. 훈련 오차 (Training Error): ○ A ○ B ○ C ○ D
   ii. 테스트 오차 (Testing Error): ○ A ○ B ○ C ○ D
3. 배깅(Bagging) 앙상블에서의 모델 수(Number of Base Learners, B):
   i. 훈련 오차 (Training Error): ○ A ○ B ○ C ○ D
   ii. 테스트 오차 (Testing Error): ○ A ○ B ○ C ○ D
4. Soft-Margin SVM의 정규화 상수 C:
   i. 훈련 오차 (Training Error): ○ A ○ B ○ C ○ D
   ii. 테스트 오차 (Testing Error): ○ A ○ B ○ C ○ D

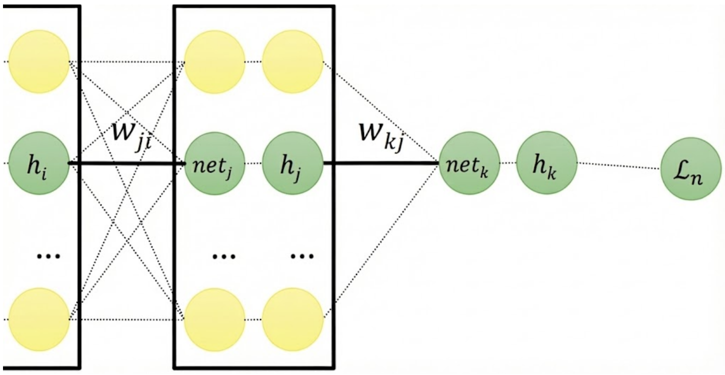
문제 3. [8점] 다음 머신러닝 모델 중 은닉층(Hidden Layer)이 없는 단일 층 신경망(1-layer Neural Network)을 적절한 목적함수로 충분히 학습하였을 때, 정확하게 등가 표현할 수 있는 모델을 모두 고르시오.

a. 선형 회귀 (Linear Regression)
b. 로지스틱 회귀 (Logistic Regression)
c. k-최근접 이웃 (k-Nearest Neighbors, k-NN)
d. 주성분분석(Principal Component Analysis, PCA)
e. t-SNE (t-Stochastic Neighbor Embedding)

문제 4. [8점] 다음은 클러스터링 방법론에 대한 설명이다. 옳은 문장을 모두 고르시오.

a. 좋은 클러스터링 결과를 얻기 위해서는 군집 내 유사도(Intra-class similarity)는 높이고, 군집 간 유사도(Inter-class similarity)은 낮춰야 한다.
b. Bisecting k-Means는 선택된 하나의 클러스터에 대해서만 k-Means ($k = 2$)를 수행하므로, 일반적인 k-Means보다 보통 더 효율적이다.
c. 병합적 계층 군집화(Agglomerative Hierarchical Clustering)는 초기 시드(Seed) 선택에 따른 무작위성이 없으므로, 동일한 데이터에 같은 연결 기준(Linkage criteria)를 적용하면 항상 같은 결과를 보장한다.
d. DBSCAN에서 탐색 반경 ϵ을 고정한 상태로 $MinPts$값을 증가시키면, 노이즈로 분류되는 데이터 포인트의 수는 증가하거나 같다.
e. GMM (Gaussian Mixture Models)을 학습시키기 위한 EM 알고리즘의 Expectation 단계에서는 현재의 파라미터(μ, Σ, π)를 고정한 상태에서 각 데이터가 각 가우시안 분포에 속할 확률을 계산한다.

문제 5. [**8점**] 다음과 같은 다층 신경망 구조를 고려하자. 출력은 단일 뉴런 h_k이며, $\sigma(\cdot)$는 시그모이드 함수이다:



$$\mathcal{L}_n = \frac{1}{2}(t_k - h_k)^2 \quad (t_k : \text{target})$$

$$\text{net}_k = \sum_j h_j w_{kj}, \quad h_k = \sigma(\text{net}_k)$$

$$\text{net}_j = \sum_i h_i w_{ji}, \quad h_j = \sigma(\text{net}_j)$$

한 개의 학습 데이터 $(\mathbf{x}, t_k)$에 대해 현재 각 뉴런의 출력과 가중치는 다음과 같다:

$$t_k = 1, \quad h_k = \frac{1}{2}, \quad h_j = \frac{1}{2}, \quad h_i = \frac{1}{2}, \quad w_{kj} = 2$$

다음의 값을 순서대로 구하시오.

1. $\delta_k = \dfrac{\partial \mathcal{L}_n}{\partial \text{net}_k}$
2. $\dfrac{\partial \mathcal{L}_n}{\partial w_{kj}}$
3. $\delta_j = \dfrac{\partial \mathcal{L}_n}{\partial \text{net}_j}$
4. $\dfrac{\partial \mathcal{L}_n}{\partial w_{ji}}$

문제 6. [10점] 어느 장학재단에서 다음 학생들의 학점, 논문실적, 추천서 데이터를 바탕으로 장학생 선발여부를 예측하는 의사결정 나무(Decision Tree)를 구축하고자 한다:

학점	**논문실적**	**추천서**	**장학생선발**
높음	유	유	선발
높음	유	무	선발
보통	무	유	선발
보통	유	유	탈락
낮음	무	유	탈락
낮음	무	무	탈락

각 속성(학점, 논문실적, 추천서)의 Information Gain, SplitInfo, GainInfo를 차례로 구하여 아래 표를 완성하고, GainInfo를 기준으로 의사결정 나무를 구성할 때 루트노드의 속성을 제시하시오. (단, $H(\frac{1}{3}, \frac{2}{3}) = \frac{14}{15}$, $\log_2 3 = \frac{8}{5}$ 로 계산한다.)

	Information Gain	SplitInfo	GainInfo
학점			
논문실적			
추천서			

루트노드 속성: ______________

문제 7. [**8점**] AdaBoost 모델을 학습하려한다. 약한 분류기(weak learner)는 데이터의 실제 레이블 y에 대해, 첫 번째 라운드와 두 번째 라운드에서 샘플을 각각 아래 표의 y_1과 y_2와 같이 예측한다:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y	$+$	$+$	$+$	$+$	$-$	$-$	$-$	$-$
y_1	$+$	$+$	$+$	$-$	$-$	$-$	$-$	$-$
y_2	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$-$

주어진 식 및 각 샘플의 초기 가중치를 참고하여, 이후 각 라운드에서 갱신되는 가중치를 계산하시오. (단, 갱신된 가중치는 정규화되어야 한다.)

$$\mathrm{err}_t = \frac{\sum_{i=1}^{n} w^{(i)} \mathbb{I}[h_t(\mathbf{x}^{(i)}) \neq y^{(i)}]}{\sum_{i=1}^{n} w^{(i)}}, \qquad \alpha_t = \frac{1}{2} \ln \frac{1 - \mathrm{err}_t}{\mathrm{err}_t}$$

$$w^{(i)} \leftarrow w^{(i)} \exp\big(-\alpha_t \, y^{(i)} h_t(\mathbf{x}^{(i)})\big)$$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
초기 가중치	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
라운드 1 이후 가중치								
라운드 2 이후 가중치								

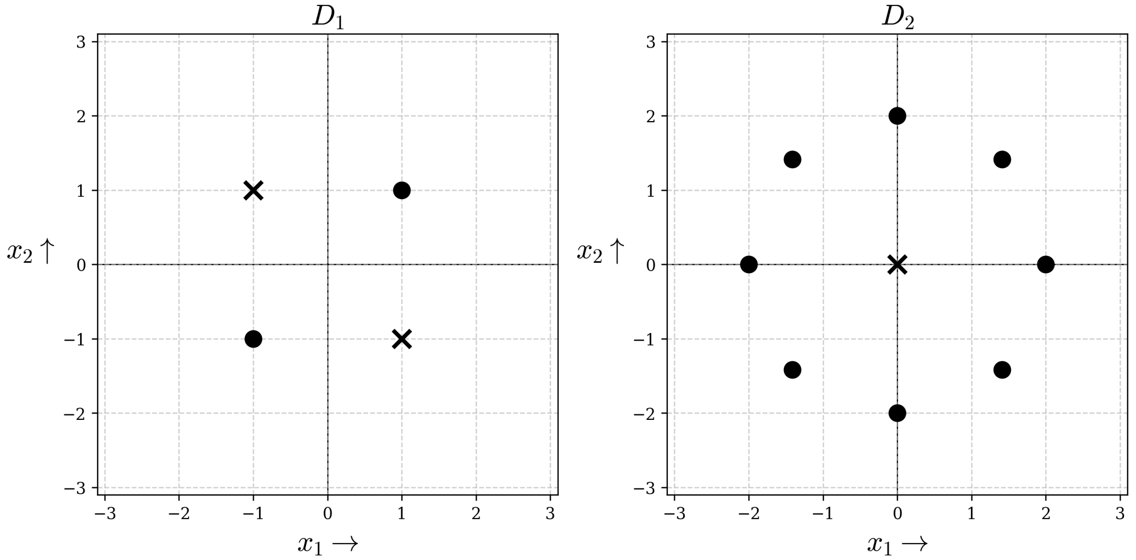
문제 8. [9점] 다음은 4개의 3차원 데이터셋 X이다. 주성분분석(Principal Component Analysis, PCA)을 사용하여 이 데이터를 1차원으로 축소하려 한다:

$$X = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & -1 \\ 0 & -2 & -1 \end{bmatrix}$$

> **Hint:** 3×3 행렬 A의 행렬식(determinant)은 다음과 같이 계산한다:
>
> $$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = a(ei - fh) - b(di - fg) + c(dh - eg)$$

1. 중앙화(mean-centered)한 데이터 X' 및 X'의 공분산 행렬 Σ를 구하시오. (단, 공분산은 $N = 4$로 나눈다.)
2. 공분산 행렬 Σ의 고유값(eigenvalue)들을 모두 구하고, 이 중 가장 큰 고유값에 해당하는 고유벡터(eigenvector)를 계산하시오. (단, 고유벡터는 크기가 1인 단위벡터로 정규화한다.)
3. 위에서 구한 주성분을 사용하여, 중앙화된 데이터 X'를 1차원으로 변환한 새로운 데이터셋 X_{new}를 구하시오.

문제 9. **[8점]** 다음과 같이 선형 분리되지 않는 두 데이터셋 D_1, D_2를 고려하자. 그림에서 '$\times$' 표시는 클래스 $y = -1$, '$\bullet$' 표시는 클래스 $y = +1$을 나타낸다:



이 데이터를 이진 분류하기 위해 다음과 같은 최적화 문제를 고려하자:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to } y_i\big(\mathbf{w}^\top \mathbf{z}_i + b\big) \geq 1 \quad (i = 1, \ldots, n)$$

여기서 $\mathbf{z}_i$는 입력 $\mathbf{x}_i$에 어떤 사상 함수 ϕ를 적용한 특징 벡터 $\mathbf{z}_i = \phi(\mathbf{x}_i)$이거나, 커널 함수 $k(\mathbf{x}, \mathbf{x}')$를 통해 간접적으로 정의되는 특징 표현이라고 하자.

다음과 같이 두 개의 사상 함수와 두 개의 커널 함수가 주어져 있다:

$$\phi_1(\mathbf{x}) = x_1^2 + x_2^2, \qquad \phi_2(\mathbf{x}) = \begin{bmatrix} x_1 + x_2 \\ x_1 x_2 \end{bmatrix},$$

$$k_1(\mathbf{x}, \mathbf{x}') = \exp\Big(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2\Big), \qquad k_2(\mathbf{x}, \mathbf{x}') = \big(1 + \mathbf{x}^\top \mathbf{x}'\big)^2.$$

위 네 가지 함수에 대해, 데이터셋 D_1과 D_2가 해당 특징 공간(또는 커널 공간)에서 완벽히 분류될 수 있는지 여부를 아래 표에 O(가능) 또는 X(불가능)로 표시하시오. (정답: +1점, 오답: -1점, 미기재: 0점)

	ϕ_1	ϕ_2	k_1	k_2
D_1				
D_2				

문제 10. [13점] 은닉층 1개(ReLU) + 출력층(Softmax, 클래스 수 $C = 2$)으로 구성된 다음 네트워크를 고려하자:

$$\begin{aligned}
\mathbf{z}^{(1)} &= W_1\mathbf{x} + \mathbf{b}_1 && (W_1 \in \mathbb{R}^{3\times 2}) \\
\mathbf{a}^{(1)} &= \mathrm{ReLU}\big(\mathbf{z}^{(1)}\big) \\
\mathbf{z}^{(2)} &= W_2\mathbf{a}^{(1)} + \mathbf{b}_2 && (W_2 \in \mathbb{R}^{2\times 3}) \\
\hat{\mathbf{y}} &= \mathrm{softmax}\big(\mathbf{z}^{(2)}\big), \quad \mathcal{L} && = -\sum_{k=1}^{C} y_k \log \hat{y}_k
\end{aligned}$$

데이터 및 레이블, 그리고 현재 네트워크의 가중치 및 편향은 다음과 같다:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, W_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 0 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, W_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

이때, 출력층 로짓(logit)에 대한 손실의 기울기는 $\dfrac{\partial \mathcal{L}}{\partial \mathbf{z}^{(2)}} = \hat{\mathbf{y}} - \mathbf{y}$임이 알려져 있다.

1. $\mathbf{a}^{(1)}$, $\hat{\mathbf{y}}$, $\mathcal{L}$을 구하시오. (단, $\ln 2 = 0.7$로 계산한다.)
2. 손실 $\mathcal{L}$에 대한 그래디언트 $\dfrac{\partial \mathcal{L}}{\partial W_2}$, $\dfrac{\partial \mathcal{L}}{\partial \mathbf{b}_2}$를 구하시오.
3. 손실 $\mathcal{L}$에 대한 그래디언트 $\dfrac{\partial \mathcal{L}}{\partial W_1}$, $\dfrac{\partial \mathcal{L}}{\partial \mathbf{b}_1}$를 구하시오.
4. 학습률 $\eta = 2.0$로 경사하강법을 한 번 수행한 뒤 W_1^{new}, $\mathbf{b}_1^{\mathrm{new}}$, W_2^{new}, $\mathbf{b}_2^{\mathrm{new}}$ 값을 구하시오.