기계학습원론 기말고사
2025 봄

강의자: 조재민
강의조교: 이재웅

June 12, 2025

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

1. 손실 함수 $E(\mathbf{w}) = \frac{1}{2n}\sum(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2$ 에 대한 경사 하강법에서 $\mathbf{w}$의 갱신식은 $\mathbf{w} \leftarrow \mathbf{w} + \eta\frac{1}{n}\sum(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})\mathbf{x}^{(i)}$ 이다. (단, η는 학습률이다.)
2. $P(C) > 0,\ P(A\cap C) > 0,\ P(B\cap C) > 0$인 사건 $A, B, C \subseteq \Omega$에 대해 $P(B \mid A\cap C) = \frac{P(A|B\cap C)\,P(B|C)}{P(A|C)}$가 항상 성립한다.
3. 선형 분리 가능한 데이터에 대해 퍼셉트론 학습 알고리즘(PLA)을 적용할 때, 학습률이 양의 상수일 경우 항상 유한 단계 내에 수렴한다.
4. 질병의 조기 발견을 위해서는 정밀도(precision)보다 재현율(recall)이 더 중요하다.
5. 다층 퍼셉트론(MLP)의 모든 층의 가중치(weight)를 동일한 상수로 초기화한다면, 역전파 및 경사 하강법을 적용해도 동일한 층 내에서는 서로 같은 가중치를 가진다.
6. 배깅(Bagging) 앙상블에서 모델 수를 무한히 늘리면 예측의 분산은 0으로 수렴한다.
7. 사전 분포가 정규 분포를 따르면 항상 MAP 추정치는 MLE와 일치한다.
8. 임의의 두 점 $x, y \in X$에 대하여 $f(y) \le f(x) + \nabla_x f(x)^\top (y - x)$를 만족하는 함수 $f : X \to \mathbb{R}$에서, 어떤 지점의 그래디언트가 0이면 그 지점은 항상 전역 최솟값이다.
9. 다음은 Soft-Margin SVM의 목적함수이다:

$$\min_{w,b,\{\xi_i\}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{subject to} \quad \begin{cases} y^{(i)}(w^\top x^{(i)} + b) \ge 1 - \xi_i \quad (\forall i), \\ \xi_i \ge 0 \quad (\forall i) \end{cases}$$

   라그랑지안을 정의한 뒤 최적화하여 얻은 쌍대 문제(dual problem)는 다음과 같다:

$$\mathcal{L}(w, b, \{\xi_i\}, \{\alpha_i\}, \{\mu_i\}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\big(y^{(i)}(w^\top x^{(i)} + b) - 1 + \xi_i\big) - \sum_{i=1}^{n}\mu_i\xi_i$$

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}(x^{(i)})^\top x^{(j)} \quad \text{subject to} \quad \begin{cases} 0 \le \alpha_i \le C \quad (\forall i), \\ \sum_{i=1}^{n}\alpha_i y^{(i)} = 0 \end{cases}$$

   이때, $\alpha_i = C$이고 $0 < \xi_i < 1$ 모든 데이터 포인트는 오분류된 데이터 포인트이다. (단, 모든 수식 도출은 올바르다고 보아도 좋다.)
10. 소프트맥스 회귀에서 입력 벡터 $x \in \mathbb{R}^d$, 원-핫 정답 벡터 $y \in \mathbb{R}^K$, 가중치 행렬 $W \in \mathbb{R}^{K\times d}$가 주어질 때, 샘플 (x, y)의 Cross-Entropy 손실 및 손실의 그래디언트는 다음과 같다:

$$\hat{y} = \mathrm{softmax}(Wx), \qquad L = -\sum_{k=1}^{K} y_k \log \hat{y}_k, \qquad \frac{\partial L}{\partial W} = (\hat{y} - y)\,x^\top.$$

    이때, $\frac{\partial L}{\partial W}$의 모든 행 벡터의 합은 원소의 합이 1인 d차원 벡터이다. (단, 모든 수식 도출은 올바르다고 보아도 좋다.)

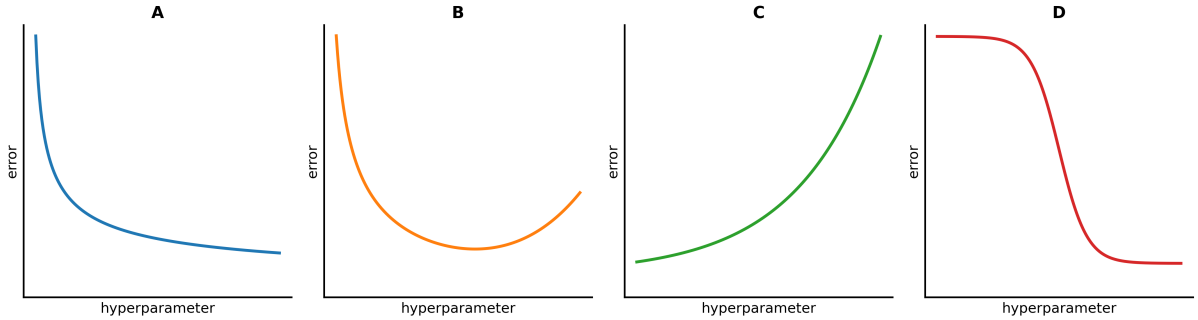
문제 2. [**8점**] 다음은 클러스터링 방법론에 대한 설명이다. 옳은 문장을 모두 고르시오.

a. k-Means는 이상치(outlier)에 민감하다.
b. DBSCAN, 계층적 군집화(Hierarchical Clustering)는 사전에 군집 수를 지정할 필요가 없다.
c. DBSCAN에서 $MinPts$를 증가시키면 중심점(Core point)의 수는 증가하거나 같다.
d. 계층적 군집화는 동일한 데이터라도 선택한 연결 기준(Linkage criteria)에 따라 서로 다른 덴드로그램(Dendrogram)을 형성한다.
e. 가우시안 혼합 모델(Mixture of Gaussians)의 Expectation 단계에서는 군집 할당 C를 고정한 상태에서 평균 벡터 $\boldsymbol{\mu}$, 공분산 행렬 $\boldsymbol{\Sigma}$ 등의 모수를 업데이트한다.

문제 3. [**8점**] 다음은 고차원 공간 및 차원축소 방법론에 대한 설명이다. 옳은 문장을 모두 고르시오.

a. 매니폴드 가설(manifold hypothesis)은 "실제 관측한 데이터가 고차원 공간 전체에 거의 균일하게 퍼져 있다"고 주장한다.
b. N차원 데이터에 대해 주성분분석(Principal Component Analysis)을 수행하면 최대 2개의 주성분을 얻는다.
c. PCA를 통해 N차원 데이터를 1차원으로 차원축소할 때, 분산을 최대화하는 길이가 1인 주성분은 유일하다.
d. t-SNE(t-Distributed Stochastic Neighbor Embedding)의 하이퍼파라미터인 Perplexity는 값이 클수록 지역적인 구조를, 값이 작을수록 전역적인 구조를 보존한 결과를 산출한다.
e. t-SNE에서는 고차원 및 저차원 상의 유클리디안 거리를 확률 분포로 모델링하여 둘 간의 발산을 최소화한다.

문제 4. [**10점**] 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. 결정 트리(Decision Tree)의 최대 깊이 D:
   i. 훈련 오차 (Training Error): ◯ A ◯ B ◯ C ◯ D
   ii. 테스트 오차 (Testing Error): ◯ A ◯ B ◯ C ◯ D
3. Soft-Margin SVM의 정규화 상수 C:
   i. 훈련 오차 (Training Error): ◯ A ◯ B ◯ C ◯ D
   ii. 테스트 오차 (Testing Error): ◯ A ◯ B ◯ C ◯ D
4. 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
5. 인공신경망(Neural Network)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

문제 5. **[10점]** 학습시간, 출석률, 과제제출여부를 가지고 합격여부를 예측하는 결정 나무(Decision Tree)를 생성하고자 한다.

학습시간	출석률	과제제출	합격여부
낮음	낮음	제출	합격
낮음	높음	미제출	불합격
높음	낮음	미제출	불합격
높음	높음	제출	합격
낮음	낮음	미제출	불합격
높음	높음	미제출	합격

1. 각 속성(학습시간, 출석률, 과제제출)의 정보 이득(Information Gain)을 구하시오. (단, $\log_2 3 = \frac{8}{5}$로 계산한다.)
2. 정보이득이 가장 큰 속성으로부터 분류를 시작하여 생성된 결정 나무를 그리시오. (단, 정보 이득 동일 시 출석률, 과제제출, 학습시간 순으로 우선순위를 가진다.)

문제 6. [**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 7. [**8점**] 다음 2차원 데이터 포인트가 주어져 있다.

$$x_1 = (8, 9),\ x_2 = (7, 8),\ x_3 = (1, 4),\ x_4 = (2, 2),$$
$$x_5 = (3, 1),\ x_6 = (5, 5),\ x_7 = (9, 7),\ x_8 = (4, 6).$$

k-Means 알고리즘으로 3개의 군집 C_1, C_2, C_3로 분할하려 한다. 초기 군집 중심(centroid)은 다음과 같이 주어진다:

$$c_1^{(0)} = (2, 3), \qquad c_2^{(0)} = (6, 7), \qquad c_3^{(0)} = (9, 6)$$

1. 첫 번째 할당-업데이트 단계를 마친 뒤 계산되는 새로운 군집 중심 $c_1^{(1)}$, $c_2^{(1)}$, $c_3^{(1)}$을 구하시오.
2. 위와 같은 2차원 데이터에 대해 할당-업데이트 단계를 총 t번 반복한다고 하자. 데이터 개수를 n, 군집 수를 k로 바꾸었을 때, 전체 알고리즘의 시간 복잡도를 t, n, k에 대한 Big-O 표기법으로 나타내시오.
3. k-Means 알고리즘이 최적화하려는 목적함수는 각 할당-업데이트 단계에서 절대로 증가하지 않는다. (O / X)

문제 8. [**16점**] 은닉층 1개(ReLU) + 출력층(Softmax, 클래스 수 $C = 2$)으로 구성된 다음 네트워크를 고려하자:

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

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

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

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 계산하시오.
3. 손실 $\mathcal{L}$에 대한 그래디언트 $\dfrac{\partial \mathcal{L}}{\partial W_1}$, $\dfrac{\partial \mathcal{L}}{\partial \mathbf{b}_1}$를 계산하시오.
4. 학습률 $\eta = 0.1$로 경사하강법을 한 번 수행한 뒤 W_1^{new}, $\mathbf{b}_1^{\text{new}}$, W_2^{new}, $\mathbf{b}_2^{\text{new}}$ 값을 구하시오.

문제 9. [**12점**] 다음 데이터를 이진 분류하려고 한다:

$$\text{-1 클래스: } (-1,-1) \qquad +1 \text{ 클래스: } (1,1)$$

이 데이터를 분류하기 위해 다음 최적화 문제를 고려하자:

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

이 때, $\mathbf{x} = [x_1, x_2]^\top$는 입력 벡터, $y_i \in \{+1, -1\}$는 각 데이터 포인트 $\mathbf{x}_i$의 클래스 레이블, $\mathbf{w}$는 분류기 파라미터(가중치 벡터), b는 편향 파라미터, n은 데이터 포인트 개수이다.

1. 위의 최적화 문제를 만족하는 결정 경계를 x_1, x_2에 관한 식으로 표현하시오.
2. 학습 데이터에 -1 클래스 데이터 포인트 $(-3,-3)$가 추가되었다($n = 3$). 이 때, 1번에서 얻은 결정경계에 대한 전체 힌지 손실(Hinge Loss)의 합은 (증가한다 / 감소한다 / 변하지 않는다).
3. 학습 데이터에 +1 클래스 데이터 포인트 $(-5,-5)$가 추가되었다($n = 4$). 사상 함수(Mapping Function)가 $\phi(\mathbf{x}) = \left[\, x_1 + x_2,\ x_1 x_2 \,\right]^\top$일 때, 위 최적화 문제를 통해 모든 데이터 포인트를 완벽히 분류할 수 있는지, 그리고 그 이유에 대해 설명하시오.