머신러닝에서 Feature와 Label의 관계를 가장 잘 설명한 것은?
① Feature는 예측 목표값, Label은 입력 정보이다.
② Feature는 입력 정보, Label은 모델이 예측하려는 정답이다. ☑
③ Feature는 모델의 출력, Label은 모델의 입력이다.
④ Feature와 Label은 동일한 데이터 속성을 의미한다.

정답: ②

다음 중 지도학습(Supervised Learning) 의 특징으로 옳은 것은?
① 입력 데이터만 있고 정답이 없는 데이터를 학습한다.
② 입력과 정답(라벨)을 함께 이용해 예측 규칙을 학습한다. ☑
③ 테스트 데이터에만 정답이 포함되어 있다.
④ 새로운 데이터 예측은 불가능하다.

정답: ②

회귀(Regression) 문제에서 예측 오류를 평가하기 위한 대표적인 지표는?
① Accuracy
② MSE ☑
③ Precision
④ Recall

정답: ②

분류(Classification) 문제에서 "실제 양성인데 음성으로 예측한 경우"를 나타내는 것은?
① True Positive
② False Positive
③ True Negative
④ False Negative ☑

정답: ④

오버피팅(Overfitting)에 대한 설명으로 가장 올바른 것은?

① 테스트 데이터에서만 성능이 좋은 현상이다.
② 훈련 데이터의 잡음까지 학습해 일반화가 떨어지는 현상이다. ☑
③ 학습이 충분히 되지 않아 중요한 패턴을 놓친다.
④ 데이터의 분포가 변해서 생긴 오류이다.

정답: ②

다음 중 "언더피팅(Underfitting)"에 해당하는 상황은?
① 모델이 너무 단순해서 중요한 패턴을 잡지 못함 ☑
② 모델이 너무 복잡해서 잡음까지 학습
③ 테스트 데이터 성능만 좋음
④ 훈련 데이터가 과도하게 많음

정답: ①

결정계수($R^2$)의 값이 1에 가까울수록 의미하는 것은?
① 예측력이 낮다.
② 모델이 데이터를 잘 설명한다. ☑
③ 오류가 크다.
④ 예측값이 평균보다 못하다.

정답: ②

단답형 (7문항)

머신러닝에서 규칙을 직접 코딩하지 않고, 데이터로부터 규칙을 학습하는 과정을 무엇이라 하는가?
정답: 학습(Learning)

"입력(Feature)과 출력(Label) 관계를 표현할 수 있는 모든 후보 함수들의 집합"을 무엇이라 하는가?
정답: 가설공간(Hypothesis Space)

회귀 문제에서 예측값과 실제값의 차이를 제곱해 평균낸 값을 무엇이라 하는가?

정답: 평균제곱오차(Mean Squared Error, MSE)

Accuracy(정확도)만으로는 판단이 어려운 이유는 데이터의 ( )이 심할 때 오해를 일으키기 때문이다.
정답: 불균형(imbalanced)

"양성이라 예측한 것 중 실제 양성의 비율"을 나타내는 지표는?
정답: 정밀도(Precision)

"실제 양성 중 예측도 양성인 비율"을 의미하는 지표는?
정답: 재현율(Recall or Sensitivity)

Precision과 Recall의 조화평균으로 계산되는 지표는?
정답: F1-score

서술형 (6문항)

Feature와 Label의 차이점을 정의하고, 각각의 예시를 하나씩 들어 설명하시오.
정답 예시:
Feature는 모델이 예측에 사용하는 입력 정보이며, 예측의 근거가 되는 데이터이다.
Label은 모델이 예측하려는 정답으로, 학습의 목표값이다.
예: 이메일 스팸 분류에서 제목·내용·발신자는 Feature, 스팸/정상 여부는 Label이다.

지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)의 차이를 설명하시오.
정답 예시:
지도학습은 입력 데이터와 정답(Label)을 함께 사용하여 예측 규칙을 학습하는 반면,
비지도학습은 정답이 없는 데이터로 데이터 간의 구조나 군집을 찾는 학습이다.

오버피팅(Overfitting)과 언더피팅(Underfitting)의 차이 및 해결방안을

설명하시오.
정답 예시:
오버피팅은 모델이 너무 복잡해 훈련 데이터의 잡음까지 학습하는 현상이며,
언더피팅은 모델이 너무 단순해 중요한 패턴을 놓치는 현상이다.
오버피팅은 교차검증·정규화·데이터 증가로 완화하고,
언더피팅은 모델 복잡도 증가나 학습시간 연장으로 개선한다.

회귀(Regression)와 분류(Classification)의 주요 차이점을 서술하시오.
정답 예시:
회귀는 연속적인 수치(Label)를 예측하는 문제이며,
분류는 이산적인 범주(Label)를 예측하는 문제이다.
예: 집값 예측은 회귀, 스팸메일 판별은 분류 문제이다.

평균제곱오차(MSE)와 RMSE의 차이를 설명하시오.
정답 예시:
MSE는 예측값과 실제값의 제곱 오차 평균이며, 단위가 제곱이다.
RMSE는 MSE의 제곱근으로 계산되어 원래 데이터 단위와 동일하게 해석이 가능하다.

일반화(Generalization)의 의미를 설명하고, 왜 머신러닝 모델에서 중요한지 서술하시오.
정답 예시:
일반화란 학습에 사용되지 않은 새로운 데이터에서도 높은 예측 성능을 유지하는 능력이다.
훈련 데이터에만 최적화되면 실제 환경에서 성능이 급격히 저하될 수 있으므로,
모델의 실질적인 성능 평가는 일반화 성능으로 판단한다.

감추기

모델 복잡도가 증가함에 따라 훈련 오류는 계속 하강하는 반면, 테스트 오류 곡선은 일반적으로 U자형을 띱니다. 이 U자형 곡선의 양쪽 끝(높은 오류)에서 발생하는 현상에 대한 설명으로 옳은 것은?

A. 정답
모델이 지나치게 복잡하면 오버피팅(Overfitting)으로 인해 테스트 오류가 상승합니다.
B.
언더피팅(Underfitting)은 테스트 성능이 증가하는 시점부터 발생합니다.
C.
모델이 지나치게 단순하면 오버피팅(Overfitting)으로 인해 테스트 오류가 상승합니다.
D.
두 현상 모두 훈련 오류와 테스트 오류가 모두 하강하는 구간에서 발생합니다.

검증셋(홀드아웃) 접근법의 주요 단점 중 하나는 '훈련셋만으로 모델을 적합하므로, 전체 데이터로 학습했을 때보다 성능이 낮게 추정될 수 있다'는 것입니다. 이는 검증 기반 테스트 오류 추정치가 무엇을 하는 경향이 있음을 의미합니까?
A.
훈련 오류를 과소 추정한다.
B.
훈련 오류와 테스트 오류의 차이를 과소 추정한다.
C.
테스트 오류를 과소 추정한다.
D. 정답
테스트 오류를 과대 추정한다.

K-겹 교차 검증(K-Fold Cross-Validation)의 단계 중, K개의 MSE를 평균하여 최종 테스트 오류를 추정하기 직전에 수행되는 K번의 반복 과정의 핵심 원리는 무엇입니까?
A.
K개의 모델을 각각 다른 초기값으로만 학습시켜 지역 최솟값에 빠지는 것을 방지하는 것입니다.
B. 정답
각 그룹(폴드)이 번갈아 검증셋이 되고, 나머지 K-1개 그룹이 훈련셋이 되도록 반복하는 것입니다.
C.

각 폴드에 동일한 모델 복잡도를 적용하여 훈련하는 것입니다.
D.
훈련셋과 검증셋의 크기를 다르게 설정하여 오차의 가변성을 높이는 것입니다.

K-겹 교차 검증에서 K의 값이 데이터 전체 관측치 수 n과 같을 때(K=n) 발생하는 특별한 교차 검증 방법은 무엇이며, 이 경우 훈련셋과 검증셋은 각각 어떻게 구성됩니까?
A. 정답
Leave-One-Out 교차 검증(LOOCV): 훈련셋은 관측치 하나만 제외한 나머지 전부(n−1), 검증셋은 제외한 1개 관측치
B.
검증셋 접근(Hold out): 훈련셋과 검증셋의 비율은 K:n−K로 무작위 분할
C.
K-Fold 교차 검증: 훈련셋은 $₩frac{n}{K}$개, 검증셋은 $n - ₩frac{n}{K}$개
D.
Leave-One-Out 교차 검증(LOOCV): 훈련셋은 n개의 관측치 전부, 검증셋은 n개의 폴드 각각

훈련 오류는 계속 하강하고 테스트 오류는 U자형 곡선을 보이는 경우, 우리가 모델 학습의 목표로 삼아야 하는 것은 무엇입니까?
A.
훈련 오류와 테스트 오류의 차이를 최대화하는 것.
B.
언더피팅이 발생할 때 모델 학습을 중단하는 것.
C. 정답
테스트 오류를 최소화하는 지점(U자형 곡선의 최저점)에서 모델 복잡도를 선택하는 것.
D.
훈련 오류를 0에 가깝게 만드는 것.

비지도 학습(Unsupervised Learning)의 정의를 가장 잘 설명하는 것은 무엇입니까?
A. 정답

레이블(정답) 없이 데이터의 잠재 서브그룹, 구조, 패턴을 찾아내는 학습.
B.
훈련 데이터의 노이즈까지 학습하여 과적합(Overfitting)을 최소화하는 학습.
C.
레이블(정답)이 있는 데이터로 학습하여 새로운 입력에 대한 예측 모델을 구축하는 학습.
D.
환경과 상호작용하며 보상을 통해 최적의 행동을 학습하는 학습.

다음 중 비지도 학습의 대표적인 과제 목록에 속하지 않는 것은 무엇입니까?
A.
차원 축소 / 시각화
B.정답
가격 예측 / 악성 종양 예측
C.
밀도 추정 / 이상치 탐지
D.
군집화(Clustering)

클러스터링(Clustering)의 궁극적인 목표를 가장 정확하게 설명한 것은 무엇입니까?
A.
클러스터의 수를 미리 정해(K) 모든 관측치가 적어도 하나의 군집에 속하도록 하는 것입니다.
B.정답
클러스터 내부는 서로 유사하고, 클러스터 간은 서로 상이하도록 데이터를 하위 집단으로 분할하는 것입니다.
C.
전체 데이터의 분산을 최대화하여 데이터 간의 차이를 강조하는 것입니다.
D.
클러스터 내 모든 관측치에 대해 MSE의 합을 최소화하는 것입니다.

K-means 클러스터링 알고리즘은 전역(Global) 최솟값을 보장하지 못하고 초기값에 따라 지역 최솟값으로 수렴할 수 있는 특성이 있습니다. 이러한 특성 때문에 권장되는 조치는 무엇입니까?
A.
하나의 초기값으로만 시도하되, 군집 내 관측치 인덱스 집합 Ck를 매 단계 무작위로 변경합니다.
B.정답
서로 다른 초기 클러스터 레이블(초기값)에서 시작하여 K-means를 여러 번 시도하는 것이 권장됩니다.
C.
전역 최솟값을 찾을 때까지 클러스터 수를 K에서 K+1로 증가시키며 시도합니다.
D.
K-means 대신 항상 전역 최솟값을 보장하는 계층적 군집법을 사용해야 합니다.

클러스터링을 수행할 때 '스케일링(Scaling)' 또는 '표준화(Standardization, 평균 0 표준편차 1로 변환)'가 필요한 가장 주된 이유는 무엇입니까?
A.정답
입력 변수 간의 단위 차이가 클러스터링 거리 계산에 미치는 영향을 줄이고 모든 변수를 동등하게 반영하기 위해.
B.
비지도 학습 과제 중 하나인 차원 축소를 먼저 수행하기 위한 전처리 과정이기 때문에.
C.
K-means 알고리즘이 지역 최솟값으로 수렴하는 것을 방지하기 위해.
D.
클러스터링의 결과를 덴드로그램으로 시각화하기 쉽게 만들기 위해.

선형 회귀의 기본 목표는 무엇인가?
① 잔차를 최대화하는 것
② RSS를 최소화하는 것
③ 로그우도를 최소화하는 것
④ 정확도를 최대화하는 것

→ 정답: ②

단순 선형 회귀의 한계로 옳지 않은 것은?
① 변수 간 관계가 하나만 고려된다.
② 비선형 관계를 잘 표현할 수 있다.
③ 설명 변수가 1개뿐이다.
④ 복잡한 관계 설명이 어렵다.
→ 정답: ②

로지스틱 회귀는 어떤 문제를 해결하기 위해 사용되는가?
① 회귀
② 분류
③ 군집화
④ 차원 축소
→ 정답: ②

로지스틱 회귀에서 예측값이 0.5보다 크면 일반적으로 어떻게 분류하는가?
① 0으로 분류
② 1로 분류
③ 무시함
④ 다시 [계산함
→ 정답: ②

Deep Neural Network의 정의는?
① hidden layer가 1개
② hidden layer가 2개 이상
③ 입력층만 존재
④ 출력층만 존재
→ 정답: ②

활성화 함수(activation function)의 주된 목적은?
① 학습률을 조정하기 위함
② 비선형성을 부여하기 위함
③ 손실함수를 계산하기 위함

④ 입력데이터를 정규화하기 위함
→ 정답: ②

다음 중 비선형 활성화 함수가 아닌 것은?
① sigmoid
② tanh
③ ReLU
④ 선형함수 f(x)₩=xf(x)=xf(x)₩=x
→ 정답: ④

손실 함수(Loss Function)의 역할은?
① 모델 복잡도를 측정한다.
② 모델의 예측오차를 측정한다.
③ 데이터 분포를 추정한다.
④ 변수 간 상관관계를 제거한다.
→ 정답: ②

sigmoid, tanh, arctan 함수의 출력 범위를 각각 쓰시오.
→ 정답:
sigmoid: (0, 1)
tanh: (-1, 1)
arctan: (-π/2, π/2)

확률적 경사하강법(SGD)이 일반 경사하강법보다 가지는 장점을 서술하시오.
→ 정답:
계산 속도가 빠르고,
메모리 사용량이 적으며,
지역 최소값(local minima) 에 빠질 가능성을 줄여 더 나은 전역 최소점을 찾을 수 있다.

감추기

◇ 1. 손실함수와 학습

Q1. 손실 함수(Loss Function)의 역할은 무엇이며, 모델 학습 과정에서 이 함수를 최소화하는 이유를 설명하시오.

Q2. 손실 함수의 값이 작을수록 모델이 잘 학습되었다는 의미인 이유를 실제 예측 문제 관점에서 설명하시오.

◇ 2. 경사 하강법과 최적화

Q3. 경사 하강법(Gradient Descent)의 기본 원리를 간단히 설명하고, 학습률(learning rate)이 너무 크거나 작을 때 각각 어떤 문제가 발생할 수 있는지 말하시오.

Q4. 전체 경사 하강법(Batch Gradient Descent), 확률적 경사 하강법(SGD), 미니배치 경사 하강법의 차이를 비교하시오.

Q5. Convex 함수와 Non-convex 함수의 차이를 그래프 형태와 최적화 난이도 측면에서 설명하시오.

◇ 3. 신경망 구조 이해

Q6. Shallow Neural Network와 Deep Neural Network의 구조적 차이와 각각의 장단점을 설명하시오.

Q7. 활성화 함수(Activation Function)가 비선형 함수여야 하는 이유를 "표현력" 관점에서 설명하시오.

◇ 4. 학습 과정과 역전파

Q8. 역전파(Backpropagation)의 주요 단계들을 순서대로 나열하고, 각 단계의 목적을 간단히 요약하시오.

Q9. 신경망 학습 과정에서 역전파가 필요한 이유를 "손실 최소화" 관점에서 설명하시오.

◇ 5. 확률적 경사 하강법의 특징

Q10. 확률적 경사 하강법(SGD)이 지역 최소점(Local Minimum)을 탈출할 가능성이 높은 이유를 설명하시오.