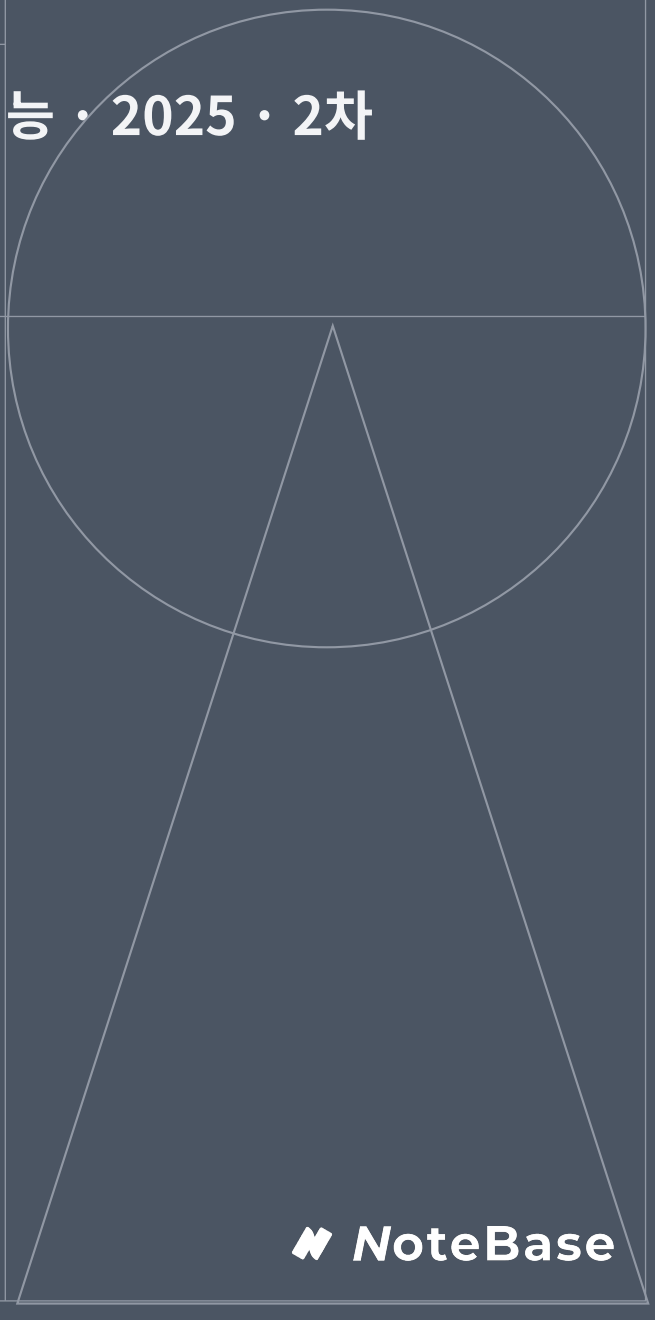
기출OX 해설집

공무원/국가직/7급 · 인공지능 · 2025 · 2차



NoteBase

기출OX 해설집

공무원/국가직/7급 · 인공지능 · 2025 · 2차

공무원/국가직/7급 · 인공지능 · 2025 · 2차

문 1. 다음 설명에 해당하는 기계 학습의 유형은?

어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법

① 강화 학습(reinforcement learning)
② 지도 학습(supervised learning)
③ 비지도 학습(unsupervised learning)
④ 앙상블 학습(ensemble learning)

정답 1번

출제 포인트

이 문제는 보상을 최대화하도록 행동을 선택하는 강화 학습의 정의와 다른 학습 유형의 차이를 구별하는 문제이다.

핵심 개념

강화 학습은 에이전트가 환경과 상호작용하며 보상을 기준으로 최적 행동이나 행동 순서를 학습하는 방법이다. 지도 학습은 정답이 붙은 데이터를 이용해 입력과 출력의 대응 관계를 학습하는 방법이다. 비지도 학습은 정답 라벨 없이 데이터의 구조나 군집을 찾는 방법이다. 앙상블 학습은 여러 모델의 예측을 결합해 성능을 높이는 방법이다.

NoteBase

선지별 해설

① O: 제시문은 에이전트, 환경, 상태 인식, 행동 선택, 보상 최대화라는 강화 학습의 핵심 요소를 그대로 설명한다.
② X: 틀린 부분은 보상을 최대화하는 행동 선택을 지도 학습으로 보는 표현이다. 지도 학습은 입력과 정답 쌍을 학습해 분류나 회귀를 수행하는 방법이지, 보상 기반으로 정책을 학습하는 방법이 아니다.
③ X: 틀린 부분은 보상 최대화 문제를 비지도 학습으로 보는 표현이다. 비지도 학습은 라벨 없는 데이터의 패턴, 군집, 차원 축소를 다루며 에이전트의 행동-보상 구조를 전제로 하지 않는다.
④ X: 틀린 부분은 행동 선택 문제를 앙상블 학습으로 보는 표현이다. 앙상블 학습은 여러 모델을 결합해 예측 성능을 높이는 기법이며, 보상을 최대화하는 정책 학습 자체를 뜻하지 않는다.

오답 포인트

보상 최대화라는 표현을 놓치면 단순히 학습 데이터의 유무로만 판단하여 지도 학습이나 비지도 학습으로 오인하기 쉽다. 이 문항의 결정 단서는 에이전트가 환경과 상호작용하며 보상을 기준으로 행동을 고른다는 점이다.

다시 나오면 이렇게 판단

문장에 에이전트, 환경, 상태, 행동, 보상, 정책 최적화가 함께 나오면 강화 학습으로 판단한다. 반대로 정답 라벨이 나오면 지도 학습, 라벨 없이 구조 탐색이 나오면 비지도 학습, 여러 모델 결합이 나오면 앙상블 학습으로 구분된다.

암기 포인트

보상으로 행동을 고르면 강화 학습, 정답으로 함수를 맞추면 지도 학습, 구조를 찾으면 비지도 학습, 모델을 묶으면 앙상블 학습이다.

NoteBase

문 2. 이진 분류기의 성능평가에 대한 설명으로 옳은 것은?

① 분류기가 양성으로 판정한 것들 중에서 실제 양성인 것의 비율을 재현율(recall)이라고 한다.

② 실제 음성인 것들 중에서 분류기가 음성으로 판정한 것의 비율을 특이도(specificity)라고 한다.

③ 실제 양성인 것들 중에서 분류기가 양성으로 판정한 것의 비율을 위양성율(false positive rate)이라고 한다.

④ 분류기는 민감도(sensitivity) 값이 작을수록 좋고 위양성율 값이 클수록 좋다.

정답 2번

출제 포인트

이 문제는 이진 분류 성능지표인 정밀도, 재현율, 특이도, 위양성률, 민감도의 정의를 정확히 구분하는지를 묻는다.

핵심 개념

정밀도는 분류기가 양성이라고 예측한 것 중 실제 양성의 비율이다. 재현율과 민감도는 실제 양성 중에서 양성으로 맞힌 비율이다. 특이도는 실제 음성 중에서 음성으로 맞힌 비율이다. 위양성률은 실제 음성 중에서 양성으로 잘못 판정한 비율이다.

선지별 해설

① X: '분류기가 양성으로 판정한 것들 중에서 실제 양성인 것의 비율'은 재현율이 아니라 정밀도이다. 재현율은 실제 양성 중에서 양성으로 맞힌 비율이다. 틀린 표현은 '재현율'이고, 올바르게는 '정밀도'이다.
② O: 실제 음성인 것들 중에서 분류기가 음성으로 판정한 비율은 참음성률이며 특이도와 같은 뜻이다. 따라서 설명이 정확하다.
③ X: 실제 양성인 것들 중에서 분류기가 양성으로 판정한 비율은 위양성률이 아니라 재현율 또는 민감도이다. 위양성률은 실제 음성인 것들 중에서 양성으로 잘못 판정한 비율이다. 틀린 표현은 '위양성률'이고, 올바르게는 '재현율 또는 민감도'이다.
④ X: 민감도는 실제 양성을 놓치지 않는 능력이므로 일반적으로 클수록 좋다. 위양성률은 실제 음성을 양성으로 잘못 잡는 비율이므로 작을수록 좋다. 틀린 표현은 '민감도 값이 작을수록 좋고 위양성률 값이 클수록 좋다'이고, 올바르게는 '민감도는 클수록 좋고 위양성률은 작을수록 좋다'이다.

오답 포인트

정밀도와 재현율은 둘 다 양성과 관련되어 보여 헷갈리기 쉽지만 분모가 다르다. ①은 예측 양성을 분모로 쓰는 정밀도를 재현율로 바꾸어 놓았고, ③은 실제 양성을 분모로 쓰는 재현율을 위양성률로 바꾸어 놓은 것이 함정이다.

NoteBase

다시 나오면 이렇게 판단

성능지표 문제는 먼저 분모가 예측 양성, 실제 양성, 실제 음성 중 무엇인지 확인하면 된다. 예측 양성이 분모이면 정밀도, 실제 양성이 분모이면 재현율이나 민감도, 실제 음성이 분모이면 특이도나 위양성률을 판단할 수 있다.

암기 포인트

정밀도는 예측 양성 기준, 재현율과 민감도는 실제 양성 기준, 특이도와 위양성률은 실제 음성 기준이다.

NoteBase

문 3. 운전면허 시험 성적에서 합격 6명, 불합격 4명으로 구성된 데이터 세트의 지니 계수(Gini index) 값은?

① 0.25

② 0.48

③ 0.52

④ 1

정답 2번

출제 포인트

이 문제는 두 클래스 비율이 주어진 데이터 세트의 지니 계수 $Gini = 1 - \sum p_i^2$를 계산하는 문제이다.

조건 정리

주어진 값: 합격 6명, 불합격 4명, 전체 10명이다.
구해야 할 값: 데이터 세트의 지니 계수이다.
주의할 조건: 지니 계수는 각 클래스의 비율로 계산하며, 제곱합 자체가 아니라 1에서 제곱합을 뺀 값이다.

사용할 공식

지니 계수는 $Gini = 1 - \sum p_i^2$이다.
이 문제는 합격과 불합격의 비율만 주어진 2개 클래스 분류이므로 각 클래스 비율의 제곱합을 구한 뒤 1에서 빼면 된다.

계산 과정

합격 비율은 $6/10 = 0.6$, 불합격 비율은 $4/10 = 0.4$이다.
각 비율의 제곱은 $0.6^2 = 0.36$, $0.4^2 = 0.16$이다.
따라서 지니 계수는 $1 - (0.36 + 0.16) = 1 - 0.52 = 0.48$이다.

선지별 해설

① X: 계산 결과와 다르다. 지니 계수는 0.48이므로 0.25는 맞지 않는다. 틀린 표현은 값 자체이며, 올바르게는 0.48이다.
② O: 합격 비율 0.6과 불합격 비율 0.4를 $1 - \sum p_i^2$에 대입하면 0.48이므로 맞다.
③ X: 0.52는 $0.6^2 + 0.4^2$인 제곱합 자체이다. 지니 계수는 그 값을 1에서 뺀 0.48이다. 틀린 표현은 최종값 처리이며, 올바르게는 0.52가 아니라 0.48이다.
④ X: 1은 이 문항의 계산 결과가 아니다. 지니 계수는 클래스 비율 제곱합을 뺀 값이므로 이 데이터에서는 0.48이다. 틀린 표현은 값 자체이며, 올바르게는 0.48이다.

NoteBase

오답 포인트

가장 흔한 실수는 ③처럼 클래스 비율의 제곱합 0.52를 구해 놓고 마지막에 1에서 빼지 않는 경우이다. ①처럼 전혀 다른 지표 값으로 착각하거나 ④처럼 지니 계수의 범위를 극단값 1과 혼동하는 실수도 나온다.

다시 나오면 이렇게 판단

분류 불순도 문항은 먼저 각 클래스 비율을 구한 뒤, 지니 계수인지 엔트로피인지 공식을 구분해서 대입하면 된다. 지니 계수이면 항상 $1 - \sum p_i^2$ 구조인지 마지막에 다시 확인하면 된다.

암기 포인트

지니 계수는 클래스 비율 제곱합을 1에서 뺀 값이다.

NoteBase

문 4. 선형 회귀에서 평균제곱오차(MSE)를 최소화하는 계수를 찾는 방법은?

① 주성분 분석법

② 선형 판별 분석법

③ 최소우도법

④ 최소제곱법

정답 4번

출제 포인트

선형 회귀에서 평균제곱오차를 최소화하는 계수 추정 방법과 차원 축소·분류·우도 개념을 구분하는 문제이다.

핵심 개념

최소제곱법은 예측값과 실제값의 제곱오차합을 가장 작게 만드는 회귀계수를 구하는 방법이다.
주성분 분석법은 데이터 분산을 가장 잘 설명하는 축을 찾는 차원 축소 기법이다.
선형 판별 분석법은 클래스 간 분리를 크게 만드는 분류 기법이다.
우도 기반 추정은 관측 데이터를 가장 그럴듯하게 만드는 모수를 찾는 접근이지만, '최소우도법'처럼 우도를 최소화하는 방식은 표준 개념이 아니다.

NoteBase

선지별 해설

① X
주성분 분석법은 변수들의 분산 구조를 요약하는 새 축을 찾는 방법이다.
틀린 부분은 "MSE를 최소화하는 계수"이다.
실제로는 주성분 분석은 차원 축소 기법이고, 선형 회귀 계수를 직접 구하는 표준 방법은 최소제곱법이다.

② X
선형 판별 분석법은 집단을 가장 잘 구분하는 선형 경계를 찾는 분류 방법이다.
틀린 부분은 "선형 회귀에서"이다.
실제로는 선형 판별 분석법은 분류 문제에 쓰이고, 회귀의 MSE 최소화 계수 추정법은 최소제곱법이다.

③ X
우도 기반 추정은 보통 우도를 크게 만드는 최대우도법으로 표현한다.
틀린 부분은 "최소우도법"이다.
실제로는 우도를 최소화하는 표준 추정법은 아니며, 이 문항에서 묻는 MSE 최소화 회귀계수 추정법은 최소제곱법이다.

④ O
최소제곱법은 잔차 제곱합을 최소화하는 회귀계수를 구하는 방법이다.
평균제곱오차는 잔차 제곱합을 표본 수로 나눈 값이므로, 잔차 제곱합을 최소화하는 계수는 평균제곱오차도 함께 최소화한다.

오답 포인트

주성분 분석과 선형 판별 분석은 모두 선형대수 기반 기법이라 회귀와 혼동하기 쉽다. 또한 ③처럼 '최대우도법'을 '최소우도법'으로 바꿔 놓은 단어 함정에 걸리기 쉽다.

다시 나오면 이렇게 판단

문장에서 묻는 핵심 동사가 무엇인지 먼저 본다. 'MSE를 최소화한다'처럼 오차 최소화가 나오면 회귀의 손실함수 최적화 문제이므로 최소제곱법을 우선 떠올리면 된다. 반대로 분산 설명이면 PCA, 클래스 분리면 LDA, 우도 최대화면 최대우도법으로 구분된다.

암기 포인트

회귀 오차 최소화는 최소제곱법, 분산 축 찾기는 PCA, 클래스 분리는 LDA이다.

NoteBase

문 5. 시간성을 갖는 데이터에 유용하고 문맥 의존성을 효율적으로 처리할 수 있는 신경망은?

① CNN

② KNN

③ MLP

④ RNN

정답 4번

출제 포인트

이 문제는 시계열 데이터와 문맥 의존성을 처리하는 신경망 구조를 다른 모델과 구분하는 문제이다.

핵심 개념

RNN은 이전 시점의 은닉상태를 다음 시점으로 넘기면서 순차 데이터를 처리하는 순환 신경망이다. 키워드는 시계열, 순서, 문맥 유지이다.
CNN은 합성곱 필터로 국소 영역의 공간적 패턴을 추출하는 신경망이다. 키워드는 이미지, 공간 구조, 지역 특징이다.
MLP는 입력에서 출력으로만 흐르는 전방향 완전연결 신경망이다. 키워드는 고정 길이 입력, 독립 처리, 순환 없음이다.
KNN은 가까운 이웃의 거리로 분류·회귀를 수행하는 비모수 학습 알고리즘이다. 키워드는 거리 기반, 이웃 투표, 비신경망이다.

선지별 해설

① X: CNN은 공간적 국소 패턴 추출에 강한 구조이므로 이미지나 격자형 데이터에 더 적합하다. 틀린 부분은 시간성과 문맥 의존성 처리에 특화된 구조로 본 점이다. 올바르게는 CNN은 시계열의 장기 문맥 처리보다 공간 특징 추출에 적합하다.
② X: KNN은 신경망이 아니라 거리 기반의 최근접 이웃 알고리즘이다. 틀린 부분은 신경망 후보처럼 제시된 점이다. 올바르게는 KNN은 순차 문맥을 내부 상태로 유지하지 못하는 비신경망 분류·회귀 기법이다.
③ X: MLP는 층을 순방향으로만 통과하므로 이전 시점 정보를 상태로 누적하지 못한다. 틀린 부분은 문맥 의존성을 효율적으로 처리할 수 있다고 볼 수 있는 여지가 없다는 점이다. 올바르게는 MLP는 고정된 특징 벡터 입력에는 쓸 수 있으나 시계열 문맥 처리는 별도 구조가 필요하다.
④ O: RNN은 이전 시점의 정보를 은닉상태에 반영해 다음 시점 계산에 사용하므로 시간성을 갖는 데이터와 문맥 의존성 처리에 적합하다. 문장, 음성, 시계열처럼 순서가 중요한 데이터에서 강점을 가진다.

NoteBase

오답 포인트

CNN과 RNN을 모두 신경망이라는 이유만으로 묶어 보면 헷갈리기 쉽다. 이 문항의 함정은 모델 이름이 아니라 처리 대상의 구조이다. 시간 순서와 앞뒤 문맥이 핵심이면 RNN 쪽이고, 공간적 배열과 국소 특징이 핵심이면 CNN 쪽이다. KNN은 이름이 비슷하지만 애초에 신경망이 아니라는 점도 자주 놓친다.

다시 나오면 이렇게 판단

문항에 시간성, 순서, 이전 상태, 문맥, 시퀀스라는 표현이 보이면 먼저 순환 구조가 있는지 확인하면 된다. 상태를 다음 시점으로 넘기는 구조이면 RNN 계열이고, 각 입력을 독립된 고정 벡터처럼 다루면 MLP나 CNN 쪽이다.

암기 포인트

RNN은 순서를 기억하고, CNN은 공간을 훑고, MLP는 고정 입력을 통과시키고, KNN은 이웃 거리를 본다.

NoteBase

문 6. 탐색 방법에 대한 설명으로 옳지 않은 것은?

① 맹목적 탐색은 문제의 상태 공간 정보를 이용한다.

② 깊이 우선 탐색은 너비 우선 탐색에 비해 메모리에 대한 부담이 적지만 최단 경로를 찾는다는 보장이 없다.

③ 너비 우선 탐색은 목표 상태를 찾을 때까지 생성된 모든 노드를 메모리에서 관리하기 때문에 메모리 비용이 크다.

④ 반복적 깊이 심화 탐색은 탐색 깊이 한계를 0부터 점차 증가시키면서 깊이 우선으로 탐색하는 방법이고 최단 경로 찾는 것을 보장한다.

정답 1번

출제 포인트

맹목적 탐색과 정보이용 탐색의 차이, 깊이 우선 탐색과 너비 우선 탐색의 특성, 반복적 깊이 심화 탐색의 성질을 구분하는 문제이다.

핵심 개념

맹목적 탐색은 문제에 대한 휴리스틱이나 상태 공간에 대한 추가 지식 없이 초기 상태, 연산자, 목표 검사 같은 기본 정보만으로 탐색하는 방법이다. 깊이 우선 탐색은 한 경로를 끝까지 내려가며 탐색하므로 메모리 사용이 적지만 최단 경로를 항상 보장하지 않는다. 너비 우선 탐색은 같은 깊이의 노드를 순서대로 확장하므로 메모리 사용량이 크지만 단위 비용 조건에서는 최단 경로를 찾는다. 반복적 깊이 심화 탐색은 깊이 제한을 0부터 점차 늘리며 깊이 우선 탐색을 반복하는 방식이다.

NoteBase

선지별 해설

① X: 틀린 부분은 "상태 공간 정보를 이용한다"이다. 맹목적 탐색은 상태 공간에 대한 추가 정보나 휴리스틱 정보를 이용하지 않고 기본 문제 정의만으로 탐색한다. 올바르게는 "맹목적 탐색은 문제 특화 정보 없이 탐색한다"이다.
② O: 깊이 우선 탐색은 현재 경로의 노드만 주로 저장하므로 너비 우선 탐색보다 메모리 부담이 작다. 대신 먼저 찾은 해가 가장 얕은 해라는 보장이 없으므로 최단 경로를 보장하지 않는다.
③ O: 너비 우선 탐색은 목표를 찾을 때까지 생성된 프런티어 노드를 폭넓게 관리해야 하므로 메모리 비용이 크다. 같은 깊이의 노드를 모두 보관하는 구조가 이 선지의 핵심 근거이다.
④ O: 반복적 깊이 심화 탐색은 깊이 한계를 0, 1, 2처럼 점차 증가시키면서 깊이 우선 탐색을 반복한다. 깊이 한계를 얕은 쪽부터 늘려 가므로 너비 우선 탐색과 마찬가지로 항상 가장 얕은 해를 먼저 발견한다. 여기서 "최단 경로 보장"은 모든 간선의 비용이 같은 균일 비용(step cost = 1) 전제에서 성립하는 성질이며, 이는 너비 우선 탐색의 최단성이 성립하는 조건과 정확히 동일하다. 균일 비용에서는 가장 얕은 해가 곧 최단 경로이므로, 깊이 한계를 0부터 늘려 가장 얕은 해를 먼저 찾는 반복적 깊이 심화 탐색도 최단 경로를 보장한다. 선지에 비용 조건이 명시되어 있지 않으나, 이 선지는 별도의 비용 가중이 없는 표준적인 탐색 맥락(균일 비용)을 전제한 교과서적 서술이며, ②·③에서 너비 우선 탐색의 최단성을 같은 전제로 다루는 것과 일관되므로 옳은 설명으로 판정한다.

오답 포인트

①은 맹목적 탐색과 휴리스틱 탐색을 바꿔 놓은 함정이다. 수험생이 "탐색이니까 상태 공간 정보를 쓴다"라고 막연히 생각하면 틀리기 쉽지만, 맹목적 탐색의 핵심은 추가 정보 없이 기계적으로 확장한다는 점이다.

다시 나오면 이렇게 판단

탐색 알고리즘 문제는 먼저 "추가 정보 사용 여부", 그다음 "메모리 사용량", 마지막으로 "최단 경로 보장 여부"를 순서대로 대조하면 빠르게 판단할 수 있다. 맹목적 탐색은 추가 정보 없음, DFS는 저메모리 비최단, BFS는 고메모리 최단, IDS는 DFS 방식 반복으로 정리된다.

암기 포인트

맹목적은 정보 없음, DFS는 깊게 가서 메모리 작음, BFS는 넓게 가서 메모리 큼, IDS는 깊이 제한을 늘리며 최단 성질을 가져간다.

NoteBase

문 7. 추천 시스템(recommender system)의 유형 중 협업 기반 추천과 내용 기반 추천에 대한 설명으로 옳지 않은 것은?

① 내용 기반 추천은 메모리 기반 방법과 모델 기반 방법이 있고 콜드 스타트(cold start) 문제가 발생한다.

② 아이템 기반 추천은 이전에 구매한 아이템을 기반으로 그 상품과 유사한 다른 상품을 추천하는 방법이다.

③ 사용자 기반 추천은 나와 비슷한 성향을 지닌 사용자의 데이터를 기반으로 상품을 추천하는 방법이다.

④ 추천 시스템은 콘텐츠의 내용에 기반하거나 사람들의 성향 정보를 취득한 후 개인화된 맞춤 항목을 추천한다.

정답 1번

출제 포인트

이 문제는 내용 기반 추천과 협업 필터링의 구분, 그리고 협업 필터링의 사용자 기반 추천과 아이템 기반 추천의 의미를 비교하는 문제이다.

핵심 개념

내용 기반 추천은 아이템 자체의 속성이나 특징을 분석해 사용자가 선호할 만한 비슷한 아이템을 추천하는 방식이다. 협업 필터링은 여러 사용자의 구매, 평점, 클릭 같은 상호작용 데이터를 이용해 추천하는 방식이다. 사용자 기반 추천은 나와 비슷한 사용자를 찾는 방식이고, 아이템 기반 추천은 내가 선호한 아이템과 비슷한 아이템을 찾는 방식이다.

NoteBase

선지별 해설

① X: 이 선지가 틀린 핵심 이유는 '메모리 기반 방법과 모델 기반 방법'을 내용 기반 추천의 특징으로 귀속시킨 점이다. 메모리 기반과 모델 기반은 협업 필터링의 대표적 분류이지 내용 기반 추천의 분류가 아니므로, 이 오귀속만으로 선지는 옳지 않은 설명이 된다. 콜드 스타트는 새로운 사용자나 아이템에 대한 데이터가 없어 추천이 어려운 문제로, 상호작용 데이터에 전적으로 의존하는 협업 필터링에서 가장 전형적으로 거론된다. 다만 내용 기반 추천에서도 새 아이템의 속성 정보가 충분하지 않거나 새 사용자의 선호 이력이 없을 때 콜드 스타트가 발생할 수 있으므로, 콜드 스타트가 내용 기반 추천에 '아예 없다'고 단정할 수는 없고 다만 '덜 전형적'이라고 보는 편이 정확하다. 따라서 이 선지에서 결정적으로 틀린 부분은 콜드 스타트 언급이 아니라 메모리 기반과 모델 기반을 내용 기반 추천에 잘못 붙인 점이다. 올바르게는 '협업 기반 추천은 메모리 기반 방법과 모델 기반 방법이 있고, 내용 기반 추천은 아이템의 속성을 이용해 유사 항목을 추천한다'라고 정리하는 편이 맞다.
② O: 아이템 기반 추천은 사용자가 이전에 구매하거나 선호한 아이템을 기준으로 그 아이템과 유사한 다른 상품을 추천하는 방식이다. 유사도 계산의 대상이 사용자끼리가 아니라 아이템끼리라는 점이 핵심이다.
③ O: 사용자 기반 추천은 나와 비슷한 취향이나 행동 패턴을 보인 다른 사용자를 찾고, 그 사용자가 선호한 상품을 추천하는 방식이다. 비슷한 사용자를 먼저 찾는다는 점에서 아이템 기반 추천과 구별된다.
④ O: 추천 시스템은 아이템의 콘텐츠 특징을 활용하는 내용 기반 추천을 쓰거나, 사용자들의 선호와 행동 데이터를 활용하는 협업 기반 추천을 써서 개인화된 항목을 추천한다. 추천의 정보 원천을 포괄적으로 설명한 문장으로 타당하다.

오답 포인트

출제자는 '내용 기반 추천'과 '협업 필터링의 세부 분류'를 일부러 섞어 놓았다. 메모리 기반, 모델 기반, 사용자 기반, 아이템 기반은 협업 필터링 쪽 용어인데 이를 내용 기반 추천의 특징처럼 붙이면 오답이 된다.

다시 나오면 이렇게 판단

추천의 근거가 아이템의 장르, 키워드, 속성처럼 콘텐츠 자체이면 내용 기반 추천이다. 추천의 근거가 다른 사용자들의 구매, 평점, 클릭 같은 집단 행동 데이터이면 협업 필터링이다. 그다음 협업 필터링 안에서 비슷한 사용자를 찾으면 사용자 기반, 비슷한 아이템을 찾으면 아이템 기반으로 판단한다.

암기 포인트

내용 기반은 아이템 특징, 협업 필터링은 사용자 행태이며 협업 필터링 안에서 사용자 기반과 아이템 기반이 갈린다.

NoteBase

문 8. 규칙 기반 시스템에 대한 설명으로 옳지 않은 것은?

① 경합 해소(conflict resolution)는 여러 규칙을 동시에 실행할 경우 모순이 발생할 수 있기 때문에 실행 가능한 규칙들 중에서 한 개를 선택하는 것을 의미한다.

② 전향 추론(forward chaining)은 규칙의 조건부를 만족시키는 사실이 있을 때 규칙의 결론부를 실행하거나 처리한다.

③ 후향 추론(backward chaining)은 특정 사실을 확인하기 위해 해당 사실을 결론부에 포함하는 규칙의 조건부가 만족하는지 확인해 가는 방식으로 추론한다.

④ 추론 엔진(inference engine)은 경합 해소 → 패턴 매칭(pattern matching) → 규칙 실행(rule execution)의 과정을 반복하여 추론한다.

정답 4번

출제 포인트

이 문제는 규칙 기반 시스템의 추론 방식과 추론 엔진의 처리 순서를 정확히 구분하는지를 묻는 문제이다.

핵심 개념

규칙 기반 시스템은 지식베이스의 규칙을 추론 엔진이 적용하여 결론을 도출하는 방식이다. 전향 추론은 이미 주어진 사실에서 출발해 적용 가능한 규칙의 결론을 확장하는 데이터 주도 방식이다. 후향 추론은 확인하려는 목표 사실에서 출발해 그 목표를 만들어 내는 규칙의 조건이 성립하는지 거꾸로 확인하는 목표 주도 방식이다. 패턴 매칭은 조건을 만족하는 규칙을 찾는 단계이고, 경합 해소는 동시에 적용 가능한 여러 규칙 중 실제로 실행할 규칙을 고르는 단계이다.

선지별 해설

① O. 경합 해소는 동시에 실행 가능한 규칙이 여러 개일 때 우선순위나 전략에 따라 하나를 선택하는 단계이므로 설명이 맞다.
② O. 전향 추론은 규칙의 조건부를 만족하는 현재 사실이 있으면 그 규칙의 결론부를 실행하여 새로운 사실을 추가하는 방식이므로 설명이 맞다.
③ O. 후향 추론은 확인하려는 사실을 목표로 두고, 그 사실을 결론으로 갖는 규칙의 조건이 참인지 역으로 추적하는 방식이므로 설명이 맞다.
④ X. 틀린 부분은 \"경합 해소→패턴 매칭→규칙 실행\"이라는 순서이다. 실제로는 패턴 매칭으로 조건을 만족하는 규칙을 찾고, 그다음 경합 해소로 실행할 규칙을 하나 선택한 뒤, 마지막에 규칙 실행이 이루어진다.

오답 포인트

전향 추론과 후향 추론의 방향만 외우고 추론 엔진의 내부 처리 순서를 흐릿하게 기억하면 ④를 맞는 설명처럼 보기 쉽다. 이 문항의 함정은 구성 요소 이름은 모두 맞지만 순서를 뒤바꿔 놓았다는 점이다.

NoteBase

다시 나오면 이렇게 판단

규칙 기반 시스템 문제는 먼저 전향 추론이 사실→결론인지, 후향 추론이 목표→조건인지 확인하면 된다. 이어서 추론 엔진은 패턴 매칭→경합 해소→규칙 실행의 순서로 작동하는지만 점검하면 선지를 빠르게 판별할 수 있다.

암기 포인트

전향은 사실에서 결론으로 가고, 후향은 목표에서 조건으로 거슬러 가며, 추론 엔진 순서는
매칭→해소→실행이다.

NoteBase

문 9. A^* 알고리즘에 대한 설명으로 옳지 않은 것은?

① A^* 알고리즘의 평가함수는 $f(n) = g(n) + h(n)$으로 표현한다.

② 휴리스틱 함수 $\hat{h}(n)$이 실제 남은 비용을 과대평가하지 않으면 허용성(admissibility)을 만족한다.

③ 시작 노드에서 노드 n까지의 실제 비용 $g(n)$은 탐색이 진행됨에 따라 감소할 수 있다.

④ 휴리스틱 함수 $\hat{h}(n)$이 허용성을 만족하면 A^* 알고리즘은 항상 최적해를 찾는다.

정답 3번

출제 포인트

이 문제는 A* 알고리즘의 평가함수 $f(n) = g(n) + h(n)$와 $g(n)$, $h(n)$, 허용적 휴리스틱의 의미를 정확히 구분하는지를 묻는 문제이다.

핵심 개념

$g(n)$은 시작 노드에서 현재 노드 n까지 실제로 누적된 경로 비용이다. 에지 비용이 음이 아니라는 일반적인 가정에서 하나의 경로를 따라 노드를 확장해 나갈수록 $g(n)$은 비음수 비용이 계속 더해지므로 증가할 뿐 감소하지 않는다. $h(n)$은 현재 노드에서 목표 노드까지 남은 비용의 추정값이다. $f(n)$은 $g(n) + h(n)$로 계산하는 총 평가값이다. 허용성은 $h(n)$이 실제 남은 최소 비용을 넘지 않는 성질이다.

선지별 해설

① O. A*는 현재까지의 실제 비용 $g(n)$과 앞으로의 추정 비용 $h(n)$을 합한 $f(n) = g(n) + h(n)$을 기준으로 노드를 선택한다.

② O. 허용적 휴리스틱은 목표까지의 실제 남은 최소 비용을 과대평가하지 않는 휴리스틱이다. 선지의 설명은 허용성의 정의와 일치한다.

③ X. 틀린 부분은 "감소할 수 있다"이다. 핵심 논거는 비음수 비용 가정이다. 에지 비용이 음이 아니라는 일반적인 가정에서는 시작점부터 출발해 하나의 경로를 따라 노드를 확장해 나갈수록 0 이상의 비용만 계속 더해지므로, 그 경로 위에서 누적 비용 $g(n)$은 단조 증가할 뿐 결코 줄어들지 않는다. 또한 A*는 OPEN 리스트에서 f 값이 가장 작은 노드를 꺼내 CLOSED로 옮기는데, 비음수 비용 가정 아래에서 일단 CLOSED로 확정된 노드의 g 값은 이미 최단 누적 비용이므로 그 뒤로 감소하지 않는다. 한편 A*가 OPEN 리스트에 남아 있는 동일 노드에 대해 더 짧은 경로를 새로 발견하면 그 노드에 기록된 g 값을 더 작은 값으로 갱신하기는 한다. 그러나 이는 같은 노드에 도달하는 서로 다른 경로 중 더 나은 경로의 비용으로 기록을 교체하는 것일 뿐, 어느 하나의 경로 위에서 이미 지불한 누적 비용 자체가 줄어드는 것은 아니다. 선지가 말하는 "탐색이 진행됨에 따라 $g(n)$이 감소"한다는 의미는 비음수 비용 가정 아래에서 성립하지 않는다.

④ O. 허용성을 만족하는 휴리스틱을 사용하면 A*는 최적해를 보장한다. 따라서 허용적 휴리스틱과 최적성의 연결을 설명한 선지이다.

NoteBase

오답 포인트

출제자는 $g(n)$과 $h(n)$의 역할을 뒤섞는 함정을 넣었다. 특히 ③처럼 실제 누적 비용 $g(n)$을 탐색 중 계속 바뀌는 추정치처럼 읽으면 틀리기 쉽다. A*에서 변동 가능성이 핵심인 것은 남은 비용 추정 $h(n)$이지, $g(n)$의 개념 자체는 이미 누적된 실제 비용이다.

다시 나오면 이렇게 판단

A* 선지는 먼저 g, h, f의 뜻을 분리해서 읽으면 판단이 빨라진다. g는 이미 든 실제 비용, h는 앞으로 들 비용의 추정, f는 둘의 합이다. 그다음 허용성은 "실제 남은 비용을 넘지 않는가"만 확인하면 된다. 누적 실제 비용이 감소한다는 진술이 나오면 우선 오답 후보로 본다.

암기 포인트

A*는 $f = g + h$, g는 실제 누적비용, h는 남은 비용 추정, 허용성은 과대평가 금지이다.

NoteBase

문 10. 이진 분류 문제에 대한 혼동행렬(confusion matrix)에서 각각 계산한 민감도, 정확도(accuracy), 위양성율 값을 모두 곱한 값은?

예측값: 양성, 음성
실제값 양성: 양성 40, 음성 10
실제값 음성: 양성 15, 음성 35

① $\frac{3}{80}$
② $\frac{32}{275}$
③ $\frac{3}{55}$
④ $\frac{9}{50}$

정답 4번

출제 포인트

이 문제는 혼동행렬에서 민감도, 정확도, 위양성률의 정의를 구분하여 각각 계산한 뒤 곱하는 문제이다.

조건 정리

주어진 값은 실제 양성·예측 양성 40, 실제 양성·예측 음성 10, 실제 음성·예측 양성 15, 실제 음성·예측 음성 35이다.
구해야 할 값은 민감도, 정확도, 위양성률을 각각 계산한 뒤 그 곱이다.
주의할 조건은 민감도의 분모는 실제 양성 집단 $TP + FN$, 정확도의 분모는 전체 표본 수, 위양성률의 분모는 실제 음성 집단 $FP + TN$이라는 점이다.

사용할 공식

민감도는 $\frac{TP}{TP+FN}$, 정확도는 $\frac{TP+TN}{TP+FP+FN+TN}$, 위양성률은 $\frac{FP}{FP+TN}$이다.
이 공식을 쓰는 이유는 표가 실제값과 예측값으로 나뉜 전형적인 혼동행렬이므로 각 성능지표를 정의식 그대로 계산하면 되기 때문이다.

계산 과정

1. 민감도는 실제 양성 중 양성으로 맞힌 비율이므로 $\frac{40}{40+10} = \frac{40}{50} = \frac{4}{5}$이다.
2. 정확도는 전체 100개 중 맞힌 75개의 비율이므로 $\frac{40+35}{40+10+15+35} = \frac{75}{100} = \frac{3}{4}$이다.
3. 위양성률은 실제 음성 중 양성으로 잘못 예측한 비율이므로 $\frac{15}{15+35} = \frac{15}{50} = \frac{3}{10}$이다.
4. 세 값을 곱하면 $\frac{4}{5} \times \frac{3}{4} \times \frac{3}{10} = \frac{9}{50}$이다.

NoteBase

선지별 해설

① X: 보기의 $\frac{3}{80}$은 계산 결과와 다르다. 이 문장에서 틀린 표현은 "세 값을 모두 곱한 값이 $\frac{3}{80}$이다"이고, 올바르게는 "세 값을 모두 곱한 값이 $\frac{9}{50}$이다"이다.
② X: 보기의 $\frac{32}{275}$는 민감도 $\frac{4}{5}$, 정확도 $\frac{3}{4}$, 위양성률 $\frac{3}{10}$을 곱한 값과 일치하지 않는다. 틀린 표현은 "곱한 값이 $\frac{32}{275}$이다"이고, 올바르게는 "곱한 값이 $\frac{9}{50}$이다"이다.
③ X: 보기의 $\frac{3}{55}$도 계산 결과와 다르다. 틀린 표현은 "곱한 값이 $\frac{3}{55}$이다"이고, 올바르게는 "곱한 값이 $\frac{9}{50}$이다"이다.
④ O: 계산 결과가 $\frac{9}{50}$이므로 옳다. 민감도의 분모를 실제 양성 50, 위양성률의 분모를 실제 음성 50으로 잡으면 정확히 일치한다.

오답 포인트

민감도와 위양성률은 모두 분모가 전체 표본 수가 아니라 특정 실제 집단이라는 점이 핵심 함정이다. 특히 위양성률을 $\frac{FP}{TP+FP}$처럼 정밀도의 분모와 혼동하거나, 정확도를 구한 뒤 다른 지표도 전체 100으로 나누면 오답이 된다.

다시 나오면 이렇게 판단

혼동행렬 문제는 먼저 TP, FN, FP, TN를 칸에 대응시킨 뒤 각 지표의 분모가 무엇인지 확인하고 계산한다. 마지막에는 약분 전후 값이 보기와 일치하는지 한 번 더 검산하면 된다.

암기 포인트

민감도는 실제 양성 기준, 위양성률은 실제 음성 기준, 정확도는 전체 기준이라는 구분이 핵심이다.

NoteBase

문 11. 기계 학습에 대한 설명으로 옳지 않은 것은?

① 군집화와 밀도추정 문제는 일반적으로 비지도 학습 방식을 사용한다.

② 지도 학습의 학습 데이터는 입력과 기대 출력 정보의 쌍으로 구성된다.

③ 데이터 증강은 입력 데이터를 저차원 데이터로 압축하는 인코더 형태의 모델을 통해 학습한다.

④ 프로그램을 명시적으로 작성하지 않고 컴퓨터가 학습할 수 있는 능력을 갖추게 하는 기술이다.

정답 3번

출제 포인트

이 문제는 지도 학습, 비지도 학습, 데이터 증강의 의미를 구분할 수 있는지를 묻는 문제이다.

핵심 개념

지도 학습은 입력과 정답 출력의 쌍으로 학습하는 방식이다.
비지도 학습은 정답 레이블 없이 데이터의 구조나 분포를 찾는 방식이다.
데이터 증강은 기존 학습 데이터를 변형하거나 새롭게 만들어 학습 표본을 늘리는 기법이다.
오토인코더는 입력을 저차원 표현으로 압축했다가 복원하도록 학습하는 표현 학습 모델이다.

선지별 해설

① O - 군집화는 레이블 없이 비슷한 데이터끼리 묶는 문제이고, 밀도추정도 데이터의 분포 자체를 추정하는 문제이므로 일반적으로 비지도 학습에 해당한다.
② O - 지도 학습은 입력과 기대 출력이 함께 주어진 학습 데이터로 함수 관계를 학습한다.
③ X - 틀린 부분은 "데이터 증강은 입력 데이터를 저차원 데이터로 압축하는 인코더 형태의 모델을 통해 학습한다"는 표현이다. 이는 데이터 증강이 아니라 오토인코더를 이용한 차원축소 또는 표현 학습에 가깝다. 올바르게는 데이터 증강은 회전, 이동, 자르기, 반전, 노이즈 추가 등으로 학습 데이터를 다양화하는 기법이다.
④ O - 기계 학습은 규칙을 일일이 명시적으로 프로그램하지 않아도 데이터로부터 패턴을 학습하게 하는 기술이라는 설명으로 타당하다.

오답 포인트

③은 "데이터를 변형해 표본을 늘리는 기법"인 데이터 증강과 "입력을 압축해 특징을 학습하는 모델"인 오토인코더를 서로 바꿔 놓은 선지이다. "저차원으로 압축"이라는 표현이 보이면 증강보다 차원축소나 표현 학습을 먼저 떠올려야 한다.

NoteBase

다시 나오면 이렇게 판단

선지에 레이블의 유무, 데이터 분포 탐색, 데이터 개수 확대, 저차원 압축 중 무엇이 핵심인지 먼저 본다. 레이블 쌍이면 지도 학습, 구조 탐색이면 비지도 학습, 표본을 늘리면 데이터 증강, 압축 표현을 학습하면 오토인코더로 판단한다.

암기 포인트

증강은 데이터를 늘리는 것이고, 오토인코더는 데이터를 압축하는 것이다.

NoteBase

문 12. 생성적 적대 신경망(generative adversarial networks)에서 생성자(generator)와 판별자(discriminator)에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. 생성자와 판별자는 적대적인 관계에서 경쟁하면서 훈련한다.
ㄴ. 생성자는 진짜 데이터를 만들면서 판별자가 가짜라고 속을 때까지 학습한다.
ㄷ. 판별자는 훈련 데이터를 진짜 데이터로 판별하고 생성자가 만든 데이터는 가짜 데이터로 판별하도록 학습한다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

출제 포인트

GAN에서 생성자와 판별자의 역할, 그리고 생성자가 무엇을 만들어 어떤 판정을 유도하는지를 구별하는 문제이다.

핵심 개념

생성자는 잠재공간의 입력으로부터 훈련 데이터와 유사한 가짜 데이터를 만들어 판별자가 이를 진짜로 오인하게 학습하는 모델이다.
판별자는 입력 데이터가 실제 훈련 데이터인지 생성자가 만든 가짜 데이터인지 구분하도록 학습하는 모델이다.
GAN은 두 모델이 서로 적대적으로 경쟁하면서 함께 성능을 높이는 구조이다.

NoteBase

선지별 해설

① X
ㄱ은 맞다. 생성자와 판별자는 서로 목적이 반대인 적대적 관계에서 경쟁하며 훈련한다.
ㄴ은 틀리다. 틀린 부분은 "진짜 데이터를 만들면서"와 "판별자가 가짜라고 속을 때까지"이다.
생성자는 진짜 데이터가 아니라 진짜처럼 보이는 가짜 데이터를 만들고, 목표는 판별자가 그 가짜를 진짜라고 오인하게 만드는 것이다.
따라서 ㄱ, ㄴ을 묶은 ①은 틀리다.

② O
ㄱ은 생성자와 판별자의 경쟁적 학습 구조를 설명하므로 맞다.
ㄷ은 판별자가 실제 훈련 데이터는 진짜로, 생성 데이터는 가짜로 분류하도록 학습한다는 뜻이므로 맞다.
따라서 ㄱ, ㄷ만 고른 ②가 옳다.

③ X
ㄴ이 틀리므로 ㄴ, ㄷ을 묶은 ③은 틀리다.
ㄷ 자체는 맞지만, ㄴ의 잘못된 표현 때문에 전체 선택지는 성립하지 않는다.
ㄴ을 고치면 "생성자는 가짜 데이터를 생성하면서 판별자가 이를 진짜로 오인하도록 학습한다"가 된다.

④ X
ㄱ과 ㄷ은 맞지만 ㄴ이 틀리다.
모든 진술이 옳아야 하는데 ㄴ이 포함되어 있으므로 ④는 틀리다.
ㄴ의 핵심 오류는 생성 데이터의 성격과 판별자가 속는 방향을 반대로 쓴 점이다.

오답 포인트

생성자의 목표를 "진짜 데이터를 만든다"로 오해하거나, 판별자가 속는 방향을 "가짜라고 판단한다"로 거꾸로 읽으면 ㄴ을 맞다고 보기 쉽다. GAN의 핵심은 생성자가 가짜 데이터를 더 진짜처럼 만들고, 판별자는 그 가짜를 가려내려 한다는 대립 구조이다.

다시 나오면 이렇게 판단

GAN 문제는 먼저 생성자가 만드는 것이 진짜 데이터인지 가짜 데이터인지, 다음으로 판별자가 최종적으로 무엇을 맞혀야 하는지 확인하면 된다. 생성자는 가짜를 진짜처럼 만들고, 판별자는 진짜와 가짜를 구분하는 쪽이면 맞는 진술이다.

암기 포인트

생성자는 가짜를 진짜처럼 만들고, 판별자는 진짜와 가짜를 가른다.

NoteBase

문 13. 역학조사에서 흡연 여부와 암 유병 여부를 조사하였더니 다음 표와 같이 계산되었다. 흡연하는 사람의 사건을 A로 나타내고 암환자의 사건을 B로 나타내었을 때, 결합 확률 $P(A \cap B)$의 값은?

	비흡연	흡연	합계(명)
정상인	35	15	50
암환자	7	3	10
합계(명)	42	18	60

① 0.05

② 0.15

③ 0.25

④ 0.33

정답 1번

출제 포인트

이 문제는 표에서 흡연이면서 암환자인 사람 수를 찾아 전체 조사 인원으로 나누는 결합확률 계산을 묻는다.

자료 읽는 순서

1. 사건 A는 흡연, 사건 B는 암환자로 각각 표의 열과 행에 대응시킨다.
2. 결합확률 $A \cap B$는 흡연 열과 암환자 행이 만나는 칸의 값 3명이다.
3. 전체 조사 인원 60명으로 나누어 $P(A \cap B) = \frac{3}{60} = 0.05$로 계산한다.

선지별 해설

① O: 근거 위치는 암환자 행, 흡연 열의 3명과 전체 합계 60명이다. 결합확률은 두 사건이 동시에 일어나는 경우를 전체 표본수로 나누므로 $\frac{3}{60} = 0.05$가 맞다.

② X: 0.15는 표의 흡연·암환자 동시 발생을 전체 60명으로 나눈 값이 아니다. 틀린 표현은 '결합확률이 0.15이다'이고, 올바르게는 '결합확률은 0.05이다'이다.

③ X: 0.25는 암환자이면서 흡연자인 3명을 전체 60명으로 나눈 결과와 맞지 않는다. 틀린 표현은 '결합확률이 0.25이다'이고, 올바르게는 '결합확률은 3/60으로 0.05이다'이다.

④ X: 0.33은 전체 60명을 기준으로 한 결합확률 값이 아니다. 틀린 표현은 '결합확률이 0.33이다'이고, 올바르게는 '결합확률은 두 사건의 교집합 3명을 전체 60명으로 나눈 0.05이다'이다.

NoteBase

오답 포인트

행과 열이 만나는 칸의 값 3명은 잘 찾고도 이를 흡연자 18명이나 암환자 10명으로 나누어 조건부확률처럼 계산하면 틀린다. 결합확률은 항상 전체 조사 인원 60명을 분모로 둔다.

다시 나오면 이렇게 판단

표에서 사건 A와 사건 B를 각각 행·열에 대응시킨 뒤, 교집합 칸의 값을 찾고 전체 합계로 나누면 결합확률을 바로 판단할 수 있다.

암기 포인트

결합확률은 교집합 칸 수를 전체 표본수로 나눈 값이다.

NoteBase

문 14. 빅데이터 분석을 위한 데이터 마이닝에 대한 설명으로 옳지 않은 것은?

① 정제되지 않은 대용량의 데이터로부터 의미 있는 상관관계, 패턴, 추세 등을 발견한다.

② 트랜잭션에서 항목 간의 불순도를 표현하기 위해 최소지지도와 최소신뢰도를 사용한다.

③ 저장된 데이터를 정제(cleaning), 통합(integration)하여 잡음과 불일치를 제거하고 데이터 웨어하우스에 저장한다.

④ 대규모 데이터에서 암묵적인, 이전에 알려지지 않은, 잠재적으로 유용할 것 같은 정보나 지식을 추출하는 체계적인 과정이다.

정답 2번

출제 포인트

데이터 마이닝의 정의와 연관규칙 분석에서 쓰는 평가 지표를 구분하는 문제이다.

핵심 개념

데이터 마이닝은 대규모 데이터에서 의미 있는 패턴과 지식을 추출하는 과정이다. 연관규칙 분석은 트랜잭션에서 항목들이 함께 나타나는 관계를 찾는 기법이며 대표 지표는 지지도, 신뢰도, 향상도이다. 불순도는 주로 의사결정나무에서 분할의 순도를 평가할 때 쓰는 개념이다.

선지별 해설

① O 데이터 마이닝은 정형 또는 비정형을 포함한 대용량 데이터에서 상관관계, 패턴, 추세 같은 유의미한 정보를 발견하는 과정이라는 설명에 해당한다.
② X 틀린 부분은 "항목 간의 불순도를 표현하기 위해"라는 표현이다. 최소지지도와 최소신뢰도는 불순도를 재는 지표가 아니라 항목 집합의 출현 빈도와 규칙의 조건부 신뢰 수준을 판단하는 기준이다. 올바르게는 "트랜잭션에서 항목 간의 연관성을 찾기 위해 최소지지도와 최소신뢰도를 사용한다"이다.
③ O 데이터 마이닝을 위해 저장 데이터를 정제하고 통합하여 잡음과 불일치를 줄인 뒤 데이터 웨어하우스 등에 적재하는 전처리 과정은 타당하다.
④ O 데이터 마이닝은 대규모 데이터에서 암묵적이고 이전에 알려지지 않았으며 잠재적으로 유용한 정보나 지식을 체계적으로 추출하는 과정이라는 대표적 정의에 부합한다.

오답 포인트

연관규칙의 지지도·신뢰도와 의사결정나무의 불순도 개념을 섞으면 2번을 맞는 설명처럼 보기 쉽다. 이 문항은 지표 이름이 아니라 지표가 무엇을 측정하는지까지 구분하는지를 묻는다.

NoteBase

다시 나오면 이렇게 판단

선지에 지표 이름이 나오면 먼저 그 지표가 어느 기법의 것인지 확인한다. 지지도·신뢰도는 연관규칙, 지니지수·엔트로피 같은 불순도는 의사결정나무로 연결하면 빠르게 판단할 수 있다.

암기 포인트

연관규칙은 지지도·신뢰도, 의사결정나무는 불순도라는 축으로 구분한다.

NoteBase

문 15. 다음 설명의 (가)에 들어갈 학습 방법은?

(가)은 학습하는 방법을 학습하는 것을 의미한다. 특정 작업을 인간과 같이 빠르게 학습하는 능력을 만들기 위해 동일한 분포를 이루는 여러 작업에 대한 학습 방법을 학습한 뒤, 유사한 작업을 적은 데이터로 빠르고 효과적으로 학습할 수 있게 한다.

① 연합 학습

② 메타 학습

③ 강화 학습

④ 표현 학습

정답 2번

출제 포인트

메타 학습의 핵심인 '학습하는 방법을 학습'과 소량 데이터 기반의 빠른 적응 능력을 다른 학습 방식과 구별하는 문제이다.

핵심 개념

메타 학습은 여러 유사 작업에서 공통되는 학습 전략 자체를 익혀 새로운 작업에 적은 데이터로 빠르게 적응하는 학습이다. 연합 학습은 데이터를 한곳에 모으지 않고 여러 참여자가 분산된 상태로 함께 학습하는 방식이다. 강화 학습은 보상을 최대화하도록 시행착오로 정책을 학습하는 방식이다. 표현 학습은 데이터에서 유용한 특징 표현을 자동으로 학습하는 방식이다.

선지별 해설

① X: 연합 학습은 여러 단말이나 기관이 데이터를 직접 공유하지 않고 각자 학습한 결과를 결합하는 분산 학습 방식이다. 제시문처럼 학습법 자체를 익혀 새로운 유사 작업에 적은 데이터로 빠르게 적응하는 개념이 아니다. 틀린 표현은 이를 연합 학습으로 본 점이다. 올바르게는 메타 학습이다.
② O: 메타 학습은 여러 작업에 대한 학습 경험을 바탕으로 새로운 유사 작업을 few-shot 상황에서도 빠르게 학습하도록 만드는 방식이다. 제시문의 '학습하는 방법을 학습'과 '적은 데이터로 빠르고 효과적으로 학습'이 메타 학습의 정의와 정확히 맞는다.
③ X: 강화 학습은 환경과 상호작용하며 보상 신호를 최대화하는 행동 정책을 학습하는 방식이다. 문제 설명은 보상 기반 의사결정이 아니라 여러 작업 간 공통 학습 전략과 빠른 적응을 말한다. 틀린 표현은 이를 강화 학습으로 본 점이다. 올바르게는 메타 학습이다.
④ X: 표현 학습은 원시 데이터로부터 유용한 특징 벡터나 잠재표현을 학습하는 방식이다. 이는 좋은 입력 표현을 만드는 데 초점이 있으며, 여러 작업을 통해 학습 절차 자체를 익혀 새 작업에 빠르게 적응하는 개념과는 다르다. 틀린 표현은 이를 표현 학습으로 본 점이다. 올바르게는 메타 학습이다.

NoteBase

오답 포인트

메타 학습은 '무엇을 표현하느냐'보다 '어떻게 빨리 배우느냐'에 초점이 있다. 연합 학습은 데이터 분산, 강화 학습은 보상, 표현 학습은 특징 추출이 핵심이므로, 제시문의 '적은 데이터로 빠른 적응'이라는 표현이 나오면 메타 학습을 우선 떠올려야 한다.

다시 나오면 이렇게 판단

여러 유사 작업에서 학습 경험을 축적한 뒤 새로운 작업을 적은 데이터로 빠르게 익힌다는 문장이 나오면 메타 학습으로 판단한다. 반대로 데이터 분산이면 연합 학습, 보상 기반이면 강화 학습, 특징 추출이면 표현 학습으로 구분한다.

암기 포인트

메타 학습은 여러 작업으로 학습법을 익혀 새 작업을 적은 데이터로 빠르게 적응시키는 학습이다.

NoteBase

문 16. 결정 트리 학습 알고리즘에 대한 설명으로 옳지 않은 것은?

① C4.5, C5.0, CHAID, CART 등의 알고리즘이 있다.

② 엔트로피는 어떤 집단에 각 부류의 데이터들이 균등한 비율로 섞여 있을수록 작아지는 특성이 있다.

③ ID3 알고리즘은 범주형 속성값을 갖는 학습 데이터로부터 엔트로피 개념을 사용하여 결정 트리를 만든다.

④ 정보 이득(information gain)은 특정한 속성이 원하는 분류 방식에 부합하게 데이터를 나누는지를 측정할 수 있는 척도이다.

정답 2번

출제 포인트

이 문제는 결정 트리의 대표 알고리즘과 엔트로피, 정보 이득의 의미를 정확히 구별할 수 있는지를 묻는다.

핵심 개념

결정 트리: 데이터를 속성값으로 반복 분할하여 분류 규칙을 만드는 지도학습 모형이다.
엔트로피: 한 집단 안의 클래스가 얼마나 불순하게 섞여 있는지를 나타내는 척도이며, 비율이 균등할수록 커지고 한쪽으로 몰릴수록 작아진다.
정보 이득: 어떤 속성으로 분할했을 때 엔트로피가 얼마나 감소하는지를 나타내는 값이다.
ID3: 엔트로피와 정보 이득을 사용해 분할 속성을 고르는 결정 트리 알고리즘이다.

선지별 해설

1번 O: C4.5, C5.0, CHAID, CART는 모두 대표적인 결정 트리 학습 알고리즘이다. 알고리즘 이름을 묻는 기본 사실로서 옳은 설명이다.
2번 X: 틀린 부분은 "각 부류의 데이터들이 균등한 비율로 섞여 있을수록 작아지는"이다. 엔트로피는 클래스 비율이 균등할수록 커지고, 한 클래스에 몰릴수록 작아진다. 올바르게는 "각 부류의 데이터들이 균등한 비율로 섞여 있을수록 커지는 특성이 있다"이다.
3번 O: ID3는 범주형 속성을 중심으로 데이터를 분할하며, 엔트로피와 정보 이득을 사용해 결정 트리를 만든다. 따라서 설명이 맞다.
4번 O: 정보 이득은 특정 속성으로 나누었을 때 목표 분류에 필요한 불확실성이 얼마나 줄어드는지를 보는 척도이다. 결국 원하는 분류에 더 잘 맞게 데이터를 나누는 속성을 평가하는 기준이라는 설명이 타당하다.

오답 포인트

엔트로피를 "불순도"로만 외우고 증가 방향을 거꾸로 기억하면 2번에 걸린다. 불순도가 높다는 말은 클래스가 더 균등하게 섞였다는 뜻이므로 엔트로피 값도 커진다.

NoteBase

다시 나오면 이렇게 판단

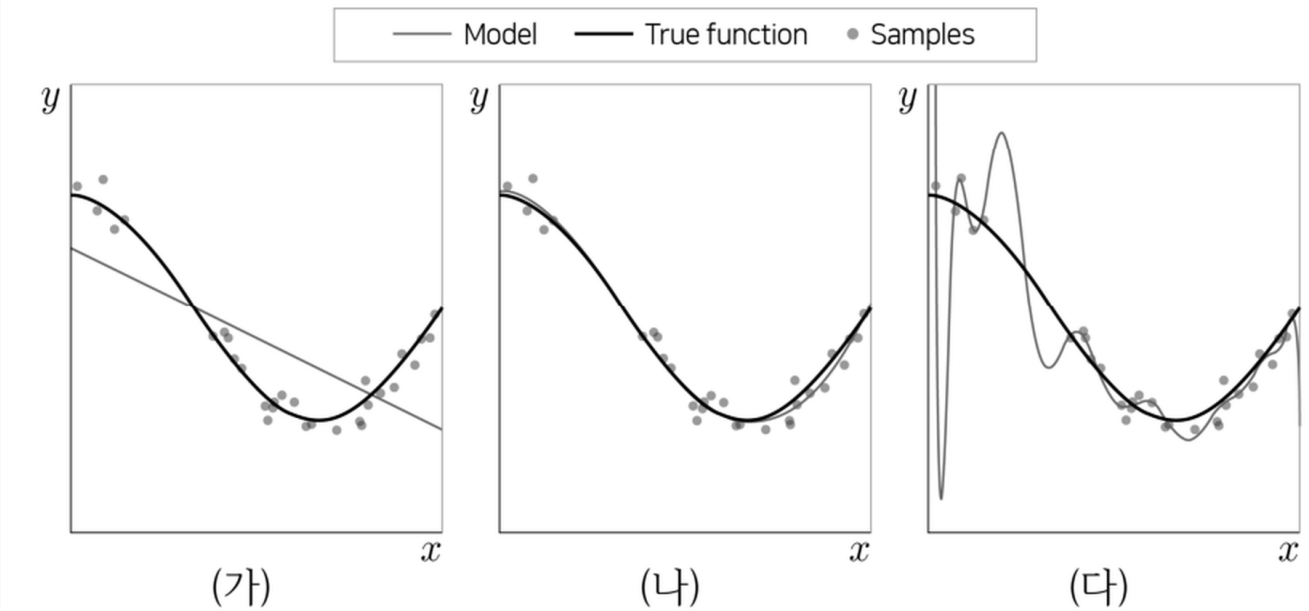
결정 트리 문항은 먼저 각 지표가 무엇을 재는지 확인하면 된다. 엔트로피는 현재 집단의 혼잡도, 정보 이득은 분할 후 혼잡도 감소량을 뜻하므로 "균등하게 섞임"이 나오면 엔트로피는 커진다고 판단한다.

암기 포인트

엔트로피는 균등할수록 크고, 정보 이득은 잘 나눌수록 크다.

NoteBase

문 17. 다음 그래프에 대한 설명으로 옳지 않은 것은? (단, x는 특징, y는 레이블이다)



① (가)는 모델의 표현력이 부족해 훈련 오차와 검증 오차가 모두 높게 유지되는 상황을 나타낸다.

② (나)는 훈련 오차와 검증 오차가 모두 낮고 서로 차이가 작아, 모델이 데이터에 일반화되고 있음을 보여 준다.

③ (다)는 훈련 오차가 낮고 검증 오차는 높게 유지되므로, 성능을 향상하기 위해서 모델 복잡도를 더 높여야 한다.

④ 기계 학습에서 과적합은 학습 데이터에서 성능이 뛰어나지만 일반화에 대한 성능을 보장할 수 없다.

정답 3번

출제 포인트

과소적합, 적절한 적합, 과적합 그래프를 보고 훈련 오차와 검증 오차의 관계 및 일반화 성능을 구분하는 문제이다.

NoteBase

선지별 해설

① O: (가)는 단순한 직선 모델이라 실제 곡선 관계를 담지 못한다. 훈련 데이터에도 잘 맞지 않으므로 훈련 오차와 검증 오차가 모두 큰 과소적합 설명으로 맞다.
② O: (나)는 모델이 실제 함수와 거의 일치하고 샘플에도 무리 없이 맞는다. 훈련 오차와 검증 오차가 모두 낮고 차이도 작아 일반화가 잘 된 상태라는 설명이 맞다.
③ X: (다)는 샘플을 과도하게 따라가 훈련 오차는 낮지만 검증 오차는 커지는 과적합 상태이다. 틀린 표현은 "성능을 향상하기 위해서 모델 복잡도를 더 높여야 한다"는 부분이다. 올바르게는 모델 복잡도를 낮추거나 정규화를 강화하거나 데이터를 늘려 일반화 성능을 개선해야 한다.
④ O: 과적합 모델은 학습 데이터에는 매우 잘 맞을 수 있지만 보지 못한 데이터에서는 성능이 떨어질 수 있다. 따라서 일반화 성능을 보장할 수 없다는 설명이 맞다.

오답 포인트

③은 앞부분의 과적합 진단 자체는 맞지만, 뒤에서 처방을 반대로 제시한 점이 함정이다. 과적합은 복잡도를 더 높인다고 해결되지 않고 오히려 악화되기 쉽다.

다시 나오면 이렇게 판단

그래프 문제에서는 먼저 모델 선이 실제 함수와 학습 샘플을 어느 정도까지 따라가는지 본다. 둘 다 못 맞추면 과소적합, 둘 다 적절하면 일반화, 샘플만 과도하게 따라가면 과적합으로 판단한다.

암기 포인트

과소적합은 복잡도 부족, 적정적합은 일반화 양호, 과적합은 훈련 오차↓·검증 오차↑이다.

NoteBase

문 18. 딥러닝 모델의 출력 노드가 3개(A, B, C)로 구성되어 있고, 각 출력 노드의 값이 A = 3.0, B = 2.0, C = 1.0이라고 하면 활성화 함수로 소프트맥스(softmax) 함수를 사용하였을 때 출력 노드 B를 변환한 근삿값은? (단, 자연 상수 e = 3.0으로 가정한다)

① 0.23

② 0.33

③ 0.67

④ 0.83

정답 1번

출제 포인트

이 문제는 소프트맥스 함수를 이용해 출력 노드 B의 정규화된 확률값을 계산할 수 있는지를 묻는 문제이다.

조건 정리

주어진 값은 A=3.0, B=2.0, C=1.0이다.
구해야 할 값은 소프트맥스 적용 후 B의 값이다.
주의할 조건은 각 노드값 자체가 아니라 지수화한 값 e^A, e^B, e^C를 전체 합으로 나누어야 한다는 점이다.

사용할 공식

소프트맥스에서 B의 출력은 $\frac{e^B}{e^A + e^B + e^C}$이다.
이 공식을 쓰는 이유는 소프트맥스가 각 출력값을 지수화한 뒤 전체 합으로 나누어 확률처럼 정규화하는 함수이기 때문이다.

계산 과정

1. $e = 3.0$이므로 $e^A = 3^3 = 27$, $e^B = 3^2 = 9$, $e^C = 3^1 = 3$이다.
2. 분모는 $27 + 9 + 3 = 39$이다.
3. 따라서 B의 소프트맥스 값은 $\frac{9}{39} = \frac{3}{13} \approx 0.23$이다.

NoteBase

선지 판단

① O: 계산 결과가 $\frac{9}{39} \approx 0.23$이므로 맞는 선택지이다.
② X: 0.33은 세 노드를 단순히 비슷하게 보거나 $\frac{1}{3}$처럼 평균적으로 처리한 값에 가깝다. 소프트맥스는 지수화 후 정규화하므로 올바르게는 0.23이다.
③ X: 0.67이면 B가 전체 확률의 3분의 2를 차지한다는 뜻인데, 실제로는 A의 지수값 27이 가장 커서 B가 그렇게 클 수 없다. 올바르게는 0.23이다.
④ X: 0.83은 B가 거의 대부분을 차지한다는 값인데 소프트맥스 계산에서 B의 분자는 9, 분모는 39이므로 나올 수 없다. 올바르게는 0.23이다.

오답 포인트

원래 노드값 3, 2, 1을 그대로 비율로 처리하거나, 지수화는 했더라도 전체 합으로 나누는 정규화를 빠뜨리면 오답이 된다. 특히 가장 큰 원점수 A=3.0과 B=2.0를 혼동하면 B의 확률을 실제보다 크게 잡기 쉽다.

다시 나오면 이렇게 판단

소프트맥스 문제는 각 값에 먼저 지수를 씌운 뒤, 구하려는 항의 지수값을 전체 지수합으로 나누는 순서로 판단한다. 계산 후에는 가장 큰 입력값이 가장 큰 확률이 되는지도 함께 검산한다.

암기 포인트

소프트맥스는 값 자체의 비율이 아니라 지수값의 비율이며, $\mathrm{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$이다.

NoteBase

문 19. 워드 임베딩(word embedding)에 대한 설명으로 옳지 않은 것은?

① N-gram은 문자열에서 N개의 연속된 요소이다.

② 원-핫 인코딩(one-hot encoding)은 단어 개수가 많아질수록 벡터를 저장하기 위한 공간이 계속해서 늘어난다는 단점이 있다.

③ TF-IDF는 단어마다 가중치를 부여하여 단어를 벡터로 변환하는 방법으로 TF는 특정 단어가 등장한 문서의 수를 의미한다.

④ 워드투벡터(Word2Vec)는 비슷한 컨텍스트에 등장하는 단어들이 유사한 의미를 지닌다는 이론에 기반을 두고 단어를 벡터로 표현해 주는 기법이다.

정답 3번

출제 포인트

이 문제는 N-gram, 원-핫 인코딩, TF-IDF, Word2Vec의 핵심 정의와 서로 다른 역할을 구별할 수 있는지를 묻는다.

핵심 개념

N-gram은 문자열이나 토큰열에서 연속해서 나타나는 N개의 요소이다.
원-핫 인코딩은 어휘 수만큼 차원을 두고 해당 단어 위치만 1로 표시하는 희소 벡터 방식이다.
TF-IDF는 문서 안에서 자주 나오지만 전체 문서에서는 드문 단어에 더 큰 가중치를 주는 표현 방식이다.
Word2Vec은 비슷한 문맥에서 함께 나타나는 단어가 비슷한 의미를 가진다는 분포 가설에 따라 밀집 벡터를 학습하는 기법이다.

선지별 해설

1번 O: N-gram은 문자 또는 단어 시퀀스에서 연속된 N개의 요소를 뜻하므로 설명이 맞다.
2번 O: 원-핫 인코딩은 단어 집합 크기만큼 차원이 필요하므로 단어 개수가 많아질수록 저장 공간이 커지는 단점이 있다.
3번 X: 틀린 부분은 "TF는 특정 단어가 등장한 문서의 수"라는 표현이다. TF는 한 문서 안에서 특정 단어가 몇 번 나타났는지를 뜻한다. 특정 단어가 몇 개의 문서에 등장했는지는 DF이며, IDF는 이런 문서 빈도를 이용해 흔한 단어의 가중치를 낮춘다. 올바르게는 "TF-IDF는 단어마다 가중치를 부여하는 방법이며 TF는 특정 문서 내에서의 단어 출현 빈도를 의미한다"이다.
4번 O: Word2Vec은 비슷한 컨텍스트에 등장한 단어들이 유사한 의미를 가진다는 생각에 기반해 단어를 벡터로 학습하므로 설명이 맞다.

NoteBase

오답 포인트

TF와 DF를 바꿔 놓는 함정이 핵심이다. TF는 문서 내부 빈도이고, 여러 문서에 걸친 등장 문서 수는 DF인데 3번은 이 둘을 뒤섞어 틀리게 만들었다.

다시 나오면 이렇게 판단

벡터화 기법 문제는 먼저 각 방법이 무엇을 기준으로 수치를 만드는지 확인하면 된다. 원-핫 인코딩은 차원 수, TF는 문서 내 출현 횟수, IDF는 전체 문서에서의 희소성, Word2Vec은 문맥 기반 의미 유사성으로 판단한다.

암기 포인트

TF는 한 문서 안의 빈도이고, DF는 문서 수이며, Word2Vec은 문맥으로 의미를 학습한다.

NoteBase

문 20. 컨볼루션 신경망(CNN)에 대한 설명으로 옳지 않은 것은?

① 학습 알고리즘은 기본적으로 경사 하강법을 사용하는 역전파 알고리즘이다.

② 컨볼루션 신경망을 분류 문제에 사용하는 경우, 마지막 층은 소프트맥스층으로 구성된다.

③ 컨볼루션 신경망에서 활성화 함수로 시그모이드를 사용하면 ReLU 함수에 비해 계산시간이 줄어든다.

④ 풀링층(pooling layer)은 합성곱층의 출력 데이터를 입력으로 받아서 특징맵의 크기를 줄이거나 특징을 추출하는 용도로 사용된다.

정답 3번

출제 포인트

이 문제는 CNN의 학습 방식, 분류 출력층, 활성화 함수, 풀링층 역할을 구분할 수 있는지를 묻는다.

핵심 개념

역전파는 오차를 뒤에서 앞으로 전달하며 가중치의 기울기를 계산해 경사 하강법 계열 최적화에 쓰이는 학습 절차이다.
소프트맥스는 다중 분류에서 각 클래스의 출력값을 확률처럼 정규화하는 출력층 함수이다.
ReLU는 음수 구간을 0으로 두는 단순한 활성화 함수이고, 시그모이드는 출력을 0과 1 사이로 압축하는 활성화 함수이다.
풀링층은 합성곱층의 출력 특징맵을 요약해 공간 크기를 줄이고 중요한 반응을 남기는 층이다.

선지별 해설

1번: O. CNN은 일반적으로 순전파로 출력을 계산한 뒤 역전파로 손실의 기울기를 구하고, 경사 하강법 계열 알고리즘으로 가중치를 갱신한다. 학습 알고리즘의 기본 틀을 올바르게 설명한 선지이다.
2번: O. CNN을 분류 문제에 쓰면 마지막 층에서 클래스별 점수를 만들고, 보통 소프트맥스층으로 이를 확률 형태로 변환한다. 다중 클래스 분류의 전형적 구조를 설명한 선지이다.
3번: X. 틀린 부분은 "시그모이드를 사용하면 ReLU 함수에 비해 계산시간이 줄어든다"이다. 실제로는 ReLU가 max 연산 기반이라 계산이 더 단순하고 빠르며, 시그모이드는 지수 연산이 들어가 상대적으로 계산 부담이 크다. 올바르게는 "CNN에서 활성화 함수로 ReLU를 사용하면 시그모이드보다 계산이 단순하고 기울기 소실 완화에 유리하다"이다.
4번: O. 풀링층은 합성곱층의 출력 특징맵을 입력으로 받아 공간 크기를 줄이고 중요한 특징을 요약한다. 차원 축소와 핵심 특징 보존이라는 역할 설명이 맞다.

오답 포인트

활성화 함수 문제에서 시그모이드와 ReLU의 출력 범위만 보고 계산 효율까지 혼동하면 3번에 걸리기 쉽다. 이 문항의 함정은 "계산시간이 줄어든다"라는 성능 비교 문장이다.

NoteBase

다시 나오면 이렇게 판단

CNN 선지는 층의 역할과 함수의 연산 특성을 분리해서 본다. 학습은 역전파, 분류 출력은 소프트맥스, 활성화 함수 계산 효율은 ReLU 우세, 풀링은 특징맵 축소로 대응하면 판단이 정리된다.

암기 포인트

역전파는 학습, 소프트맥스는 분류 출력, ReLU는 빠른 활성화, 풀링은 특징맵 축소이다.

NoteBase

문 21. 로지스틱 회귀에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. 이진 분류에 사용하지 않는 알고리즘이다.
ㄴ. 예측값을 변환할 때 시그모이드 함수를 사용한다.
ㄷ. 종속변수에 로그를 적용하는 로짓(logit) 변환을 사용한다.
ㄹ. 최적해를 구하기 위해서 최소제곱법을 사용한다.

① ㄱ, ㄴ
② ㄴ, ㄷ
③ ㄴ, ㄹ
④ ㄷ, ㄹ

정답 2번

출제 포인트

이 문제는 로지스틱 회귀의 용도, 시그모이드 함수, 로짓 변환, 추정 방법을 구분하는 문제이다.

핵심 개념

로지스틱 회귀는 사건 발생확률을 예측하는 분류 모델이다. 선형결합 값을 시그모이드 함수에 넣어 0과 1 사이의 확률로 바꾸고, 그 확률의 오즈에 로그를 취한 로짓을 선형식과 연결한다. 모수 추정은 보통 최소제곱법이 아니라 최대우도추정으로 수행한다.

선지별 해설

① X: ㄴ은 옳지만 ㄱ이 틀리다. 로지스틱 회귀는 대표적인 이진 분류 알고리즘이다. 틀린 표현은 '이진 분류에 사용하지 않는'이고, 올바르게는 '이진 분류에 사용되는'이다.
② O: ㄴ과 ㄷ이 모두 옳다. 로지스틱 회귀는 선형결합 값을 시그모이드 함수로 변환해 확률을 만들고, 그 확률의 오즈에 로그를 취한 로짓을 선형식과 연결한다.
③ X: ㄴ은 옳지만 ㄹ이 틀리다. 로지스틱 회귀의 모수 추정은 최소제곱법이 아니라 최대우도추정을 사용한다. 틀린 표현은 '최소제곱법'이고, 올바르게는 '최대우도추정으로 최적해를 구한다'이다.
④ X: ㄷ은 옳지만 ㄹ이 틀리다. 로짓 변환 사용은 맞지만 최적화 기준을 최소제곱으로 보는 것은 잘못이다. 틀린 표현은 '최소제곱법'이고, 올바르게는 '로그우도를 최대화하는 방식'이다.

오답 포인트

ㄷ의 로짓 설명과 ㄹ의 회귀라는 단어 때문에 선형회귀의 최소제곱법을 그대로 떠올리기 쉽다. 로지스틱 회귀는 회귀라는 이름을 쓰지만 실제 목적은 확률 기반 분류이며, 추정 기준도 최소제곱이 아니라 최대우도이다.

NoteBase

다시 나오면 이렇게 판단

로지스틱 회귀가 나오면 먼저 분류 모델인지, 출력값을 시그모이드로 확률화하는지, 추정은 최대우도인지 확인하면 된다. 최소제곱법이 보이면 선형회귀와 혼동시킨 함정으로 판단할 수 있다.

암기 포인트

로지스틱 회귀는 이진 분류, 시그모이드, logit, 최대우도추정의 조합이다.

NoteBase

문 22. 퍼지 집합론에서 영역 X의 특정 원소 x가 있을 때, 퍼지 소속 함수(membership function) $\mu_A(x)$는 x가 집합 A에 속하는 정도이다. 퍼지 집합 연산 중 합집합으로 옳은 것은?

① $\mu_A(x) + \mu_B(x)$

② $\mu_A(x) \times \mu_B(x)$

③ $\min(\mu_A(x), \mu_B(x))$

④ $\max(\mu_A(x), \mu_B(x))$

정답 4번

출제 포인트

이 문제는 퍼지 집합의 기본 연산에서 합집합과 교집합, 곱 연산을 구분하는지를 묻는 문제이다.

핵심 개념

퍼지 집합의 합집합은 원소가 두 집합 중 더 크게 속하는 정도를 취하는 연산이므로 $\max(\mu_A(x), \mu_B(x))$로 정의된다. 퍼지 집합의 교집합은 두 집합에 동시에 속하는 정도를 취하므로 $\min(\mu_A(x), \mu_B(x))$로 정의된다. $\mu_A(x) \times \mu_B(x)$는 대표적인 곱 기반 교집합이며, $\mu_A(x) + \mu_B(x)$는 기본 합집합 정의가 아니다.

선지별 해설

① X: $\mu_A(x) + \mu_B(x)$는 기본 퍼지 합집합 연산이 아니다. 틀린 표현은 단순 덧셈이다. 올바르게는 합집합은 $\max(\mu_A(x), \mu_B(x))$로 본다.
② X: $\mu_A(x) \times \mu_B(x)$는 합집합이 아니라 곱 기반의 교집합 해석에 가깝다. 틀린 표현은 합집합으로 본 점이다. 올바르게는 곱은 두 소속 정도를 함께 만족하는 정도를 줄이는 방향의 연산이다.
③ X: $\min(\mu_A(x), \mu_B(x))$는 합집합이 아니라 교집합이다. 틀린 표현은 합집합과 교집합을 바꾼 점이다. 올바르게는 교집합이 min이고 합집합이 max이다.
④ O: $\max(\mu_A(x), \mu_B(x))$는 두 집합 중 더 큰 소속 정도를 취하므로 퍼지 집합의 기본 합집합 정의에 해당한다.

오답 포인트

③과 ④의 min, max를 서로 바꾸어 기억하면 가장 쉽게 틀린다. ①의 단순 덧셈도 합쳐 보인다는 이유로 고르기 쉽지만 퍼지 기본 연산의 합집합 정의는 덧셈이 아니라 큰 값을 취하는 방식이다.

다시 나오면 이렇게 판단

퍼지 집합 연산 문제는 먼저 큰 값을 취하면 합집합, 작은 값을 취하면 교집합이라는 기준으로 본다. 곱이나 단순 합은 기본 max/min 정의와 구분해서 판단한다.

NoteBase

암기 포인트

퍼지 합집합은 max, 퍼지 교집합은 min이다.

NoteBase

문 23. 이진 분류 모델의 실제값과 예측값이 다음과 같이 계산되었을 때, F1 점수(score)의 근삿값은? (단, 1을 positive로 가정한다)

실젯값	1	0	1	0	1	0	1	0
예측값	1	0	0	0	1	1	0	0

① 0.33

② 0.5

③ 0.57

④ 0.75

정답 3번

출제 포인트

이 문제는 이진 분류의 혼동행렬에서 TP, FP, FN을 세고 $F1$ 점수를 계산하는 문제이다.

조건 정리

주어진 값: 실제값은 1, 0, 1, 0, 1, 0, 1, 0이고 예측값은 1, 0, 0, 0, 1, 1, 0, 0이다.
구해야 할 값: 1을 positive로 둘 때의 $F1$ 점수이다.
주의할 조건: $F1$은 정확도와 다르고, TP, FP, FN만으로 계산한다.

사용할 공식

공식은 $F1 = \frac{2PR}{P+R} = \frac{2TP}{2TP+FP+FN}$이다.
이 공식을 쓰는 이유: 문제에서 positive 클래스를 1로 지정했고, 정밀도와 재현율을 함께 반영한 분류 성능을 묻고 있기 때문이다.

계산 과정

1. 실제 1이고 예측도 1인 경우는 1번째, 5번째이므로 $TP = 2$이다.
2. 실제 0인데 예측이 1인 경우는 6번째이므로 $FP = 1$이다.
3. 실제 1인데 예측이 0인 경우는 3번째, 7번째이므로 $FN = 2$이다.
4. 따라서 $F1 = \frac{2TP}{2TP+FP+FN} = \frac{2\times 2}{2\times 2+1+2} = \frac{4}{7} \approx 0.5714$이다.

NoteBase

선지별 해설

① X: 0.33은 이 문항의 $F1$ 계산값이 아니다. 올바르게는 $\frac{4}{7} \approx 0.57$이다.
② X: 0.5는 $\frac{TP}{TP+FN} = \frac{2}{4}$로 계산한 재현율 값이다. $F1$은 재현율 하나만으로 정해지지 않고, 올바르게는 $\frac{4}{7} \approx 0.57$이다.
③ O: $F1 = \frac{4}{7} \approx 0.5714$이므로 0.57이 가장 가깝다.
④ X: 0.75는 이 문항의 $F1$ 공식으로 계산되지 않는다. 올바르게는 $\frac{4}{7} \approx 0.57$이다.

오답 포인트

재현율 $\frac{TP}{TP+FN}$만 계산해 0.5를 고르거나, 정확도와 $F1$을 같은 지표로 착각하는 실수가 자주 나온다. 이 문제는 먼저 TP, FP, FN을 정확히 세는지가 핵심이다.

다시 나오면 이렇게 판단

이진 분류 지표 문제는 positive 클래스를 먼저 확인한 뒤, 표를 보며 $TP \rightarrow FP \rightarrow FN$ 순서로 세고 마지막에 $F1 = \frac{2TP}{2TP+FP+FN}$을 바로 적용하면 된다.

암기 포인트

$F1$은 정밀도와 재현율의 조화평균이며, 빠른 계산식은 $\frac{2TP}{2TP+FP+FN}$이다.

NoteBase

문 24. 품사 태깅(POS tagging)에 대한 설명으로 옳지 않은 것은?

① 구문에 따라 주어진 문장에서 단어들의 역할을 파악하여 계층적인 트리 구조로 변환한다.

② 문장의 각 단어에 품사의 범주 및 문법적 기능에 대한 태그를 붙인다.

③ 컴퓨터가 읽을 수 있는 형태로 텍스트를 모아놓은 말뭉치(corpus)를 사용한다.

④ 규칙 기반 품사 태깅은 언어 정보를 이용하여 수작업으로 만든 규칙들을 사용한다.

정답 1번

출제 포인트

이 문제는 품사 태깅과 구문 분석의 기능 차이를 구별할 수 있는지를 묻는다.

핵심 개념

품사 태깅은 문장 속 각 단어에 품사와 문법적 기능 태그를 붙이는 작업이다. 구문 분석은 문장 성분 사이의 관계를 파악해 구문 트리나 의존 구조로 표현하는 작업이다. 말뭉치는 컴퓨터가 처리할 수 있도록 수집한 텍스트 자료이며, 규칙 기반 품사 태깅은 수작업 규칙과 언어 지식을 이용한다.

선지별 해설

① X. '주어진 문장에서 단어들의 역할을 파악하여 계층적인 트리 구조로 변환한다'는 설명은 품사 태깅이 아니라 구문 분석에 해당한다. 틀린 표현은 '계층적인 트리 구조로 변환한다'이다. 올바르게는 '문장의 각 단어에 품사와 문법 기능 태그를 부착한다'이다.
② O. 품사 태깅은 각 단어에 명사, 동사 같은 품사 범주와 문법적 기능 정보를 붙이는 작업이므로 맞는 설명이다.
③ O. 품사 태깅은 말뭉치 기반으로 학습하거나 평가하는 경우가 많으므로, 컴퓨터가 읽을 수 있는 형태의 텍스트 집합인 말뭉치를 사용한다는 설명은 타당하다.
④ O. 규칙 기반 품사 태깅은 형태소 패턴, 주변 문맥, 사전 정보 같은 언어 지식을 바탕으로 사람이 만든 규칙을 적용하는 방식이므로 맞다.

오답 포인트

품사 태깅과 구문 분석이 모두 문장 구조를 다룬다는 이유로 같은 단계로 섞어 보면 틀린다. 특히 ①은 '역할'과 '트리 구조'라는 표현 때문에 그럴듯해 보이지만, 트리 구조는 구문 분석의 대표 결과물이다.

다시 나오면 이렇게 판단

선지에 '품사 부착, 태그, 명사·동사 분류'가 나오면 품사 태깅으로 보고, '트리, 구문 구조, 의존 관계'가 나오면 구문 분석으로 구분한다.

NoteBase

암기 포인트

품사 태깅은 단어별 태그 부착, 구문 분석은 문장 구조 트리 구성이다.

NoteBase

문 25. 연관 규칙 마이닝(association rule mining)에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. 향상도가 1.0보다 크면, 항목집합 사이의 연관성 정도와 유용성이 높다는 것을 의미한다.
ㄴ. 신뢰도는 전체 거래 건수 대비 항목집합 X와 Y를 모두 포함하는 거래 건수의 비율이다.
ㄷ. 지지도는 항목집합 X를 포함하는 거래 중에서 Y도 포함하는 거래 건수의 비율이다.
ㄹ. 트랜잭션에서 어프라이어리(Apriori) 알고리즘으로 빈발 항목집합들을 찾은 다음, 이들로부터 연관 규칙을 만든다.

① ㄱ
② ㄱ, ㄹ
③ ㄴ, ㄷ
④ ㄴ, ㄷ, ㄹ

정답 2번

출제 포인트

이 문제는 연관 규칙의 평가 지표인 지지도·신뢰도·향상도의 정의와 Apriori 기반 규칙 생성 절차를 정확히 구분하는지를 묻는 문제이다.

핵심 개념

지지도는 전체 거래 중 항목집합 $X \cup Y$가 함께 나타나는 비율이다.
신뢰도는 X가 나타난 거래 중 Y도 함께 나타나는 조건부 비율이다.
향상도는 X와 Y의 동시 발생이 독립일 때보다 얼마나 더 자주 일어나는지를 보는 값이며, 1보다 크면 양의 연관이 있다고 본다.
Apriori는 최소 지지도를 만족하는 빈발항목집합을 먼저 찾고, 그 빈발항목집합으로부터 규칙을 생성한다.

NoteBase

선지별 해설

① X: ㄱ은 맞지만 ㄹ도 옳으므로 ㄱ만 고른 조합은 불완전하다. 향상도 1 초과는 양의 연관을 뜻하고, Apriori는 빈발항목집합을 찾은 뒤 규칙을 만든다.
② O: ㄱ과 ㄹ만 옳은 조합이다. ㄱ은 향상도 1 초과의 의미를 올바르게 설명했고, ㄹ은 Apriori의 절차를 정확히 설명했다.
③ X: ㄴ과 ㄷ이 모두 틀렸으므로 성립할 수 없는 조합이다. ㄴ의 틀린 표현은 '전체 거래 건수 대비'이고, 올바르게는 '항목집합 X를 포함하는 거래 중에서 Y도 포함하는 비율'이다. ㄷ은 '신뢰도'의 정의(항목집합 X를 포함하는 거래 중에서 Y도 포함하는 거래 건수의 비율)를 '지지도'라는 이름으로 잘못 붙여 놓았으므로 틀리다. 즉 설명 내용 자체는 신뢰도의 정의인데 지표 이름만 지지도로 바꿔 단 것이다.
④ X: ㄹ은 맞지만 ㄴ, ㄷ이 틀려서 조합 전체는 틀리다. ㄴ은 지지도를 신뢰도로 바꿔 말한 것이고, ㄷ은 신뢰도를 지지도로 바꿔 말한 것이다. 올바른 표현은 '지지도는 전체 거래 대비 X와 Y를 모두 포함하는 비율, 신뢰도는 X를 포함한 거래 중 Y도 포함하는 비율'이다.

오답 포인트

지지도와 신뢰도의 분모를 뒤바꾸면 쉽게 틀린다. '전체 거래 대비'이면 지지도이고, 'X가 나온 거래 중에서'이면 신뢰도인데, ㄴ과 ㄷ은 바로 이 분모를 서로 바꿔 놓은 함정이다.

다시 나오면 이렇게 판단

연관 규칙 문제에서는 먼저 분모를 본다. 전체 거래 수가 분모이면 지지도, 선행항 X를 포함한 거래 수가 분모이면 신뢰도이다. 그다음 향상도는 1을 기준으로 양의 연관 여부를 판단하고, 알고리즘은 '빈발항목집합 탐색 후 규칙 생성' 순서인지 확인하면 된다.

암기 포인트

지지도는 전체 기준, 신뢰도는 X 기준, 향상도는 1 기준, Apriori는 빈발항목집합 후 규칙 생성이다.

NoteBase