머신러닝 CH 01~04
핵심 난이도 문제 리뷰

어려운 문제 100선 — 문제 · 정답 · 해설

대전대학교 컴퓨터공학과 | 박상돈

목차

CH 01-02 | Q5

KNN 예측 방식

01

k-최근접 이웃(KNN) 알고리즘의 예측 방식으로 올바른 것은?

① 수학적 함수를 학습

② 의사결정 트리 학습

③ 전체 데이터를 저장, 가장 가까운 이웃 다수결로 예측

④ 확률 분포 학습

⑤ 가중치 반복 업데이트

CH 01-02 Q5 정답 및 해설

정답

③

해설

KNN은 별도 학습 과정 없이 전체 데이터를 메모리에 저장 후, 새 데이터의 k개 최근접 이웃 다수결로 예측.

CH 01-02 | Q6

KNN k값과 적합도

02

KNN에서 n_neighbors를 너무 작게(1)와 너무 크게(전체) 설정했을 때 각각의 문제는?

① 과소/과대적합

② 과대/과소적합

③ 데이터누수/샘플링편향

④ 둘 다 과대적합

⑤ 둘 다 과소적합

정답

② 과대적합 / 과소적합

해설

k 작으면 노이즈에 민감(과대적합), k 크면 전체 다수결에 의존(과소적합). '작으면 과대, 크면 과소'.

CH 01-02 | Q7

n_neighbors=49 다수결

03

도미 35마리, 빙어 14마리에서 n_neighbors=49로 설정하면?

① 정확히 분류

② 오류 발생

③ 모두 도미로 예측

④ 모두 빙어로 예측

⑤ 랜덤 예측

CH 01-02 Q7 정답 및 해설

정답

③ 모든 입력을 도미(다수 클래스)로만 예측

해설

전체 49개 참고 시 도미 35 > 빙어 14로 항상 도미 예측. 정확도 35/49 ≈ 0.714.

CH 01-02 | Q8

KNN 실무적 단점

04

k-최근접 이웃 알고리즘의 실무적 단점으로 가장 적절한 것은?

① 범주형 데이터 불가

② 데이터 많으면 메모리·예측 시간 크게 증가

③ 이진 분류만 가능

④ GPU 필수

⑤ 전처리 불가

정답

② 데이터가 많으면 메모리 사용량과 예측 시간이 크게 증가한다

해설

전체 데이터를 메모리에 저장하고, 새 데이터마다 모든 훈련 데이터와의 거리를 계산해야 함.

05

훈련 정확도 0.99, 테스트 정확도 0.55인 모델의 문제는?

① 과소적합

② 과대적합

③ 샘플링 편향

④ 전처리 오류

⑤ 모델 선택 오류

정답

② 과대적합

해설

훈련↑ 테스트↓ = 과대적합의 전형. 훈련 데이터 노이즈까지 학습하여 일반화 실패.

타깃 리스트 만들기

06

도미 35마리=1, 빙어 14마리=0으로 fish_target을 올바르게 설정한 것은?

① [1,35]+[0,14]

② [1]*35+[0]*14

③ [35]+[14]

④ list(range(1,36))+list(range(0,14))

⑤ np.array([1,0])*49

CH 01-02 Q11 정답 및 해설

정답

② [1]*35 + [0]*14

해설

[1]*35는 1이 35개인 리스트, [0]*14는 0이 14개인 리스트. +로 합치면 정답 리스트 완성.

CH 01-02 | Q13

산점도의 실무적 의미

07

모델링 전에 산점도를 그려보는 실무적 이유는?

① 정확도 미리 계산

② 데이터 분포·이상치·분리 가능성 파악

③ fit() 필수 입력

④ 자동 전처리

⑤ 훈련 시간 단축

정답

② 데이터의 분포, 이상치, 클래스 간 분리 가능성을 시각적으로 파악

해설

적절한 전처리 방법과 알고리즘 선택을 위해 데이터를 먼저 눈으로 확인하는 것이 중요.

CH 01-02 | Q17

샘플링 편향

08

훈련/테스트 세트에 샘플이 골고루 섞이지 않아 치우치는 현상은?

① 과대적합

② 과소적합

③ 샘플링 편향

④ 데이터 누수

⑤ 편향-분산 트레이드오프

정답

③ 샘플링 편향(sampling bias)

해설

클래스 비율이 원본과 다르게 치우치는 현상. stratify로 방지 가능.

CH 01-02 | Q18

순서대로 분할 시 정확도 0

09

도미 35 + 빙어 14를 순서대로 앞 35/뒤 14로 나누면 테스트 정확도가 0인 이유?

① 데이터 부족

② 테스트에 빙어만, 훈련에 빙어 없음

③ KNN 버그

④ 스케일 차이

⑤ k값 오류

정답

② 테스트 세트에 빙어만 있고 훈련 세트에 빙어가 없어서

해설

훈련에 빙어가 없으므로 빙어를 인식 못함 → 정확도 0.0. 반드시 랜덤 분할 필요.

CH 01-02 | Q20

리스트 vs 넘파이 연산

10

파이썬 리스트 a=[1,2,3]의 a*2와 넘파이 배열 b의 b*2의 차이는?

① 둘 다 [2,4,6]

② a*2=[1,2,3,1,2,3], b*2=[2,4,6]

③ a*2 오류

④ 둘 다 반복

⑤ 결과 반대

정답

② a*2는 [1,2,3,1,2,3], b*2는 [2,4,6]

해설

리스트 *는 반복(repeat), 넘파이 *는 원소별 곱셈(element-wise).

CH 01-02 | Q22

슬라이싱 범위

11

fish_data[0:5]에서 실제로 선택되는 인덱스 범위는?

① 0~5

② 1~5

③ 0~4

④ 0~3

⑤ 1~4

정답

③ 0, 1, 2, 3, 4

해설

[start:end]는 start부터 end-1까지 선택. 5는 포함되지 않음.

concatenate vs column_stack

12

np.concatenate()와 np.column_stack()의 차이는?

① 동일

② concatenate는 행 방향, column_stack은 열 방향 연결

③ concatenate만 넘파이

④ column_stack은 3차원만

⑤ 숫자/문자열 전용

정답

② concatenate는 첫 번째 축(행)을 따라, column_stack은 두 번째 축(열)을 따라 연결

해설

concatenate: 배열을 이어붙임(기본 axis=0). column_stack: 1차원 배열을 열로 세워 나란히 연결 → 2차원.

CH 01-02 | Q26

데이터 누수

13

테스트 세트 정보가 모델 훈련에 포함되면 발생하는 문제는?

① 샘플링 편향

② 데이터 누수 — 실제보다 높은 성능

③ 과소적합

④ 브로드캐스팅 오류

⑤ 특성 공학 오류

정답

② 데이터 누수(data leakage)

해설

테스트 정보 유입 → 실제보다 높은 성능 → 배포 후 성능 하락. 전처리 시 전체 통계값 사용이 대표적 원인.

CH 01-02 | Q27

stratify 역할

14

train_test_split에서 stratify=fish_target의 역할은?

① 시간순 정렬

② 테스트 크기 지정

③ 클래스 비율 유지 분할

④ 시드 고정

⑤ 표준화

정답

③ 클래스 비율에 맞게 훈련/테스트 세트를 나눈다

해설

원본 데이터의 클래스 비율이 훈련/테스트 세트에도 동일하게 유지됨.

CH 01-02 | Q28

train_test_split 반환 수

15

train_test_split()에 2개의 배열을 전달하면 반환되는 배열의 개수는?

① 2개

② 3개

③ 4개

④ 5개

⑤ 6개

정답

③ 4개

해설

각 배열이 훈련/테스트로 나뉘어 총 4개: train_input, test_input, train_target, test_target.

CH 01-02 | Q30

스케일 차이와 KNN

16

길이 25cm, 무게 150g인 도미를 KNN이 빙어로 잘못 예측한 원인은?

① 훈련 부족

② k 너무 작음

③ 스케일 차이로 무게가 거리 지배

④ 데이터 오류

⑤ 과소적합

정답

③ 두 특성 스케일이 달라 무게가 거리 계산을 지배

해설

길이(10~40) vs 무게(0~1000). 유클리디안 거리에서 무게 차이만 반영됨. 표준화 필수.

CH 01-02 | Q31

표준점수 공식

17

표준점수(z-score)를 구하는 공식은?

① (값-최솟값)/(최댓값-최솟값)

② (값-평균)/표준편차

③ 값/최댓값

④ (값-중앙값)/사분위범위

⑤ log(값)

정답

② (특성값 - 평균) / 표준편차

해설

평균을 빼서 0 중심으로, 표준편차로 나눠서 스케일을 통일.

CH 01-02 | Q32

테스트 세트 전처리 원칙

18

테스트 세트를 전처리할 때 반드시 지켜야 하는 원칙은?

① 테스트 통계값

② 전체 통계값

③ 훈련 세트 통계값

④ 개별 정규화

⑤ 전처리 안 함

정답

③ 훈련 세트의 평균과 표준편차를 사용한다

해설

미래 데이터의 통계를 미리 알 수 없으므로 훈련 세트 기준으로 변환.

CH 01-02 | Q33

axis=0의 의미

19

np.mean()에서 각 열(특성)별 평균을 구하려면?

① axis=1

② axis=0

③ axis=-1

④ axis=None

⑤ dim=0

CH 01-02 Q33 정답 및 해설

정답

② axis=0

해설

axis=0은 행 방향으로 연산 → 각 열의 통계값. axis=1이면 각 행별 연산.

20

넘파이의 브로드캐스팅이란?

① 네트워크 전송

② 파일 저장

③ 크기 다른 배열 간 자동 사칙연산 확장

④ 시각화

⑤ 정렬

정답

③ 크기가 다른 배열 간에 자동으로 사칙연산을 확장하여 수행

해설

(42,2) 배열에서 (2,) 배열(평균)을 빼면 자동 확장. 표준화 계산의 핵심 원리.

CH 01-02 | Q35

kneighbors() 반환값

21

KNeighborsClassifier의 kneighbors() 메서드가 반환하는 것은?

① 예측 결과

② 정확도

③ 이웃까지의 거리와 인덱스

④ 훈련 데이터 전체

⑤ 클래스 확률

정답

③ 이웃까지의 거리와 이웃 샘플의 인덱스

해설

distances, indexes 두 배열 반환. 어떤 훈련 샘플이 이웃으로 선택되었는지 확인 가능.

CH 01-02 | Q36

fit() 재호출

22

fit()을 같은 모델 객체에 다시 호출하면?

① 추가 학습

② 오류

③ 이전 학습 잃고 새로 훈련

④ 변화 없음

⑤ 앙상블

정답

③ 이전에 학습한 모든 것을 잃고 새로 훈련된다

해설

fit()은 매번 처음부터 새로 훈련. 추가 학습은 SGD의 partial_fit().

CH 01-02 | Q37

거리 기반에서 전처리 중요성

23

거리 기반 알고리즘에서 데이터 전처리가 특히 중요한 이유는?

① 훈련 속도

② 메모리 절약

③ 스케일 차이가 거리 계산 왜곡

④ 해석력 향상

⑤ 과대적합 방지

정답

③ 특성 간 스케일 차이가 거리 계산 결과를 왜곡하기 때문

해설

스케일이 큰 특성이 거리를 지배. 표준화하면 모든 특성이 동등하게 기여.

CH 01-02 | Q38

재현성

24

머신러닝 실험에서 재현성이 중요한 이유는?

① 속도 향상

② 동료 공유·검증, 디버깅 시 동일 조건 재현

③ 메모리 절약

④ 정확도 향상

⑤ 전처리 자동화

정답

② 실험 결과를 동료와 공유·검증하고 디버깅 시 동일 조건 재현

해설

random_state 고정으로 매번 같은 결과. 논문, 리뷰, 디버깅에 필수.

CH 01-02 | Q40

shape 해석

25

arr = np.array([[1,2],[3,4],[5,6]])의 arr.shape 결과는?

① (2,3)

② (3,2)

③ (6,)

④ (1,6)

⑤ (3,3)

정답

② (3, 2)

해설

3개 행, 2개 열의 2차원 배열.

n_neighbors=50 정확도

26

n_neighbors=50 KNN(사과 30, 귤 20) 훈련 정확도가 0.6인 이유?

① 데이터 부족

② 전체 참고 → 항상 사과 예측

③ 데이터 오류

④ score() 버그

⑤ 짝수 k

정답

② 전체 50개를 참고하므로 항상 사과로만 예측 → 30/50 = 0.6

해설

사과(30표) > 귤(20표). 모든 샘플을 사과로 예측하여 30개만 맞춤.

CH 01-02 | Q46

정확도 1.0 미만 지점

27

n_neighbors를 5~49까지 증가시킬 때, 처음 1.0 미만이 되는 값은?

① 5

② 20

③ 36

④ 42

⑤ 49

정답

③ 36

해설

35까지 1.0 유지, 36에서 이웃 수 과다로 오분류 시작.

CH 01-02 | Q50

fruit_data 길이와 첫 원소

28

fruit_data를 만든 후 len(fruit_data)와 fruit_data[0]의 결과는?

① 50과 [151.0]

② 100과 151.0

③ 50과 [151.0, 14.2]

④ 2와 [151.0,…]

⑤ 50과 (151.0, 14.2)

정답

③ 50과 [151.0, 14.2]

해설

50개의 [무게, 당도] 쌍. 첫 번째 사과의 [151.0, 14.2].

CH 01-02 | Q54

predict 이중 대괄호

29

kn.predict([[200, 13]])에서 이중 대괄호를 쓰는 이유?

① 문법 필수

② predict()가 2차원 배열 요구

③ 속도

④ 버그

⑤ 확률 반환

정답

② predict()는 2차원 배열(샘플 × 특성)을 입력으로 요구

해설

[[200, 13]]은 1샘플 × 2특성의 2차원. [200, 13]은 1차원이라 오류.

CH 01-02 | Q56

최적 k 찾기 코드

30

정확도가 처음 1.0 미만이 되는 k를 찾는 올바른 코드 패턴은?

① 매번 새 객체

② kn.n_neighbors=n → score → break

③ fit만 반복

④ while문

⑤ print만

정답

② for문에서 kn.n_neighbors=n → score → 1.0 미만이면 break

해설

속성 직접 변경 후 score 확인. 1.0 미만이면 저장하고 break.

CH 01-02 | Q59

다수 클래스 비율 = 정확도

31

귤 30개, 사과 20개일 때 n_neighbors=50 모델의 정확도는?

① 0.4

② 0.5

③ 0.6

④ 0.8

⑤ 1.0

정답

③ 0.6

해설

다수 클래스 귤(30개)로 전부 예측 → 30/50 = 0.6. 비율이 같으면 결과 동일.

CH 01-02 | Q60

클래스 불균형

32

1000개 중 990정상/10불량에서 모두 '정상' 예측 시 정확도와 문제점?

① 0.01

② 0.99 — 불량 미탐지

③ 0.50

④ 0.99 — 우수

⑤ 계산불가

정답

② 정확도 0.99이지만 불량을 하나도 못 찾아 쓸모없음

해설

990/1000=0.99. 불균형 데이터에서는 정밀도·재현율도 확인 필요.

CH 01-02 | Q29

random_state 역할

33

train_test_split()의 random_state 매개변수의 역할은?

① 테스트 비율 지정

② 동일 값이면 매번 같은 분할

③ 정렬 기준

④ k값 지정

⑤ 표준화 기준

정답

② 동일한 값을 주면 매번 같은 방식으로 데이터를 나눈다

해설

난수 시드 고정 → 실험 재현성 확보.

CH 01-02 | Q23

2차원 배열 열 선택

34

2차원 넘파이 배열에서 모든 행의 첫 번째 열만 선택하는 방법?

① arr[0]

② arr[:,0]

③ arr[0,:]

④ arr[:][0]

⑤ arr.col(0)

CH 01-02 Q23 정답 및 해설

정답

② arr[:,0]

해설

':'는 모든 행, 0은 첫 번째 열. 넘파이의 인덱싱 기본 문법.

KNN 회귀 예측 방법

35

k-최근접 이웃 회귀에서 새 샘플의 타깃값을 예측하는 방법은?

① 가장 가까운 이웃 하나의 값

② 가장 가까운 k개 이웃 타깃의 평균

③ 전체 훈련 데이터 평균

④ 가장 먼 이웃의 값

⑤ 이웃 중앙값

정답

② 가장 가까운 k개 이웃 타깃값의 평균

해설

KNN 분류는 다수결, 회귀는 평균. 이웃 100, 80, 60이면 평균 80이 예측값.

CH 03 | Q7

R^2가 음수

36

R^2가 음수가 될 수 있는 경우는?

① 예측이 항상 큼

② 예측이 타깃 평균보다 못한 성능

③ 데이터 과다

④ 특성 하나

⑤ 절대 불가

정답

② 예측이 타깃의 평균보다도 못한 성능일 때

해설

R^2 분자 > 분모이면 음수. 평균 예측보다 못하다는 의미.

CH 03 | Q10

과소적합 정의

37

테스트 점수 > 훈련 점수이거나 두 점수가 모두 낮은 경우?

① 과대적합

② 과소적합

③ 정상

④ 데이터 누수

⑤ 샘플링 편향

정답

② 과소적합

해설

모델이 너무 단순하여 훈련 데이터조차 적절히 학습하지 못한 상태.

CH 03 | Q11

과소적합 해결 — k줄이기

38

KNN에서 과소적합을 해결하려면 k를 어떻게?

① 늘린다

② 줄인다

③ 0으로

④ 전체로

⑤ 변경 안 함

정답

② k를 줄인다

해설

k↓ → 국지적 패턴에 민감 → 모델 복잡도↑ → 과소적합 해결.

39

reshape(-1, 1)에서 -1의 의미는?

① 행 수를 1로

② 열 수 자동

③ 해당 차원 자동 계산

④ 1차원 변환

⑤ 음수 제거

정답

③ 나머지 원소 개수에 맞춰 해당 차원을 자동 계산

해설

열 1로 고정, 행은 원소 수에 맞춰 자동. 사람이 샘플 수를 세지 않아도 됨.

CH 03 | Q17

KNN 회귀 외삽 한계

40

KNN 회귀가 훈련 범위 밖에서 약한 근본적 이유?

① 버그

② 이웃 평균만 사용 → 새 추세 불가

③ 시간 부족

④ 특성 부족

⑤ 스케일링 미적용

정답

② 가장 가까운 이웃의 타깃 평균만 사용하므로 새 추세를 만들지 못함

해설

50cm든 100cm든 같은 이웃 그룹의 같은 평균 1033g이 나옴.

CH 03 | Q20

선형 회귀의 외삽 장점

41

선형 회귀가 KNN 회귀보다 외삽에 강한 이유?

① 더 많은 데이터 저장

② 방정식을 학습해 직선 연장 가능

③ 항상 정확도 높음

④ 내부적으로 KNN 사용

⑤ 스케일링 불필요

정답

② 관계식(방정식)을 학습했으므로 직선을 연장하여 예측 가능

해설

y = ax + b라는 식이 있으므로 훈련 범위 밖에서도 직선을 연장해 예측.

CH 03 | Q24

모델 파라미터 정의

42

coef_, intercept_ 같은 값을 무엇이라 부르는가?

① 하이퍼파라미터

② 모델 파라미터

③ 전처리 변수

④ 손실값

⑤ 규제 파라미터

정답

② 모델 파라미터

해설

모델이 특성에서 학습한 파라미터. 사람이 지정하는 alpha, k는 하이퍼파라미터.

다항 회귀 학습 클래스

43

사이킷런에서 다항 회귀를 훈련할 때 사용하는 모델 클래스는?

① PolynomialRegression

② LinearRegression

③ PolynomialFeatures

④ Ridge

⑤ KNeighborsRegressor

정답

② LinearRegression

해설

PolynomialRegression은 존재하지 않음. 다항 특성은 PolynomialFeatures가 만들고, 학습은 LinearRegression.

44

다항 회귀가 곡선인데 '선형 회귀'인 이유?

① 입력이 직선

② 계수에 대해 선형

③ 단순

④ 결과가 직선

⑤ 특성 하나

정답

② 계수(a, b, c)에 대해 선형이므로

해설

y = a×길이² + b×길이 + c는 길이에 비선형이지만 a,b,c에 선형 → LinearRegression으로 학습 가능

.

특성 공학 정의

45

특성 공학(feature engineering)의 정의?

① 데이터 수집

② 기존 특성으로 새 특성 생성

③ 하이퍼파라미터 조정

④ 테스트 세트 생성

⑤ 모델 평가

정답

② 기존의 특성을 사용해 새로운 특성을 뽑아내는 작업

해설

길이², 길이×높이 등 기존 특성 조합으로 새 특성을 만드는 것.

변환기 메서드 조합

46

사이킷런 변환기가 공통으로 제공하는 메서드 조합은?

① fit()+predict()

② fit()+transform()

③ fit()+score()

④ train()+test()

⑤ compile()+evaluate()

정답

② fit() + transform()

해설

변환기는 fit()으로 학습하고 transform()으로 변환. 모델(추정기)은 fit()+predict()+score().

CH 03 | Q34

규제의 목적

47

규제(regularization)의 목적은?

① 훈련 속도

② 과도한 학습 제어

③ 데이터 증가

④ 특성 증가

⑤ 테스트 축소

정답

② 모델이 훈련 세트를 과도하게 학습하지 못하도록 제어

해설

계수의 크기를 작게 만들어 보편적인 패턴을 학습하게 유도.

라쏘 vs 릿지 핵심 차이

48

라쏘가 릿지와 구별되는 가장 큰 특징?

① 항상 높은 성능

② 일부 계수를 완전히 0으로 만들 수 있음

③ 규제 없음

④ 분류 전용

⑤ 하이퍼파라미터 없음

정답

② 일부 계수를 완전히 0으로 만들 수 있다

해설

L1 규제(절댓값)의 수학적 특성. 55개 중 40개가 0이 된 교재 예제.

alpha 하이퍼파라미터

49

릿지와 라쏘에서 규제 강도를 조절하는 하이퍼파라미터는?

① n_neighbors

② degree

③ alpha

④ max_iter

⑤ random_state

정답

③ alpha

해설

alpha↑ → 규제↑ → 계수 더 줄임 → 과소적합 방향. alpha↓ → 선형회귀에 가까워짐.

CH 03 | Q42

규제 전 표준화 이유

50

규제 적용 전에 특성을 표준화해야 하는 이유?

① 속도

② 시각화

③ 스케일 다르면 규제가 불공정

④ 문법 필수

⑤ R^2 인위적 향상

정답

③ 특성마다 스케일이 다르면 규제가 계수를 공정하게 제어하지 못하므로

해설

스케일이 큰 특성의 계수는 원래 작고 반대도 마찬가지. 표준화 없으면 규제가 불공정.

모델 vs 하이퍼파라미터

51

모델 파라미터와 하이퍼파라미터의 차이?

① 사람 지정/모델 학습

② 모델 학습/사람 지정

③ 동일

④ 모델은 상수

⑤ 하이퍼는 자동 조정

정답

② 모델 파라미터는 모델이 학습하고, 하이퍼파라미터는 사람이 사전에 지정

해설

coef_, intercept_ = 모델 파라미터. alpha, k = 하이퍼파라미터.

CH 03 | Q47

특성 > 샘플 → 과대적합

52

샘플 42개, 특성 55개일 때 과대적합이 쉬운 이유?

① 짝수

② 샘플보다 특성 많아 개별 학습 가능

③ k 작음

④ 테스트 없음

⑤ 표준화 적용

정답

② 샘플 수보다 특성 수가 많아 각 샘플을 개별적으로 외울 수 있어서

해설

참새 비유: 55번 쏠 수 있으면 42마리를 하나씩 다 맞출 수 있다.

음수 무게 예측

53

다음 선형 회귀 학습 결과에서 5cm 농어의 무게가 음수로 예측되는 이유는?

lr.coef_ = [39.01], lr.intercept_ = -709.02

① 데이터 오류

② 절편이 큰 음수이므로 작은 입력에서 음수

③ KNN의 한계

④ 과대적합

⑤ MAE가 높아서

정답

② y = 39.01×5 - 709.02 ≈ -514g. 절편이 큰 음수이므로 작은 입력에서 음수가 됨

해설

직선 모델 y = 39.01x - 709.02에서 x=5를 대입하면 195.05 - 709.02 = -513.97g. 무게가 음수라는 현실 불가능한 결과가 나오며, 이것이 단순 직선 모델의 한계이고 다항 회귀가 필요한 이유.

CH 03 | Q55

column_stack 후 shape

54

column_stack((train_input**2, train_input)) 후 shape는? (input은 42×1)

① (42,1)

② (42,2)

③ (84,1)

④ (42,3)

⑤ (2,42)

정답

② (42, 2)

해설

(42,1)과 (42,1)을 열 방향으로 붙이면 (42, 2).

CH 03 | Q58

PolynomialFeatures 특성 수

55

3개 특성 + degree=2 + include_bias=False → 특성 수?

① 3

② 6

③ 9

④ 10

⑤ 15

정답

③ 9개

해설

원래 3 + 제곱 3(a^2,b^2,c^2) + 교호 3(ab,ac,bc) = 9.

CH 03 | Q60

극단적 과대적합

56

훈련 0.9999, 테스트 -144의 해석?

① 훌륭

② 과소적합

③ 심각한 과대적합

④ 문제없음

⑤ 스케일링 해결

정답

③ 심각한 과대적합

해설

55개 특성으로 42개 샘플 학습 → 훈련 완벽, 테스트 최악.

CH 03 | Q62

변환 결과

57

poly.transform([[2,3]]) 결과는? (bias=False, degree=2)

① [[2. 3.]]

② [[1. 2. 3. 4. 6. 9.]]

③ [[2. 3. 4. 6. 9.]]

④ [[4. 6. 9.]]

⑤ [[2. 3. 6.]]

정답

③ [[2. 3. 4. 6. 9.]]

해설

원래(2,3) + 제곱(4,9) + 교호(6). bias 제외.

CH 03 | Q65

라쏘 계수 0인 특성 수

58

np.sum(lasso.coef_ == 0) = 40의 의미?

① 0 아닌 계수 합

② 계수가 0인 특성 개수

③ 전체 합

④ MAE

⑤ alpha

정답

② 계수가 정확히 0인 특성의 개수

해설

55개 중 40개 계수가 0 → 실제 사용 특성 15개. 라쏘의 특성 선택 효과.

ConvergenceWarning 해결

59

라쏘의 ConvergenceWarning을 해결하기 위해 조정하는 매개변수?

① alpha

② degree

③ max_iter

④ n_neighbors

⑤ random_state

정답

③ max_iter

해설

반복 횟수가 부족할 때 발생. max_iter를 10000 등으로 늘려 해결.

CH 03 | Q74

차수 초과 항

60

a,b,c에 degree=3 적용 시 포함되지 않는 항?

① 1

② $a \times b$

③ $a^2 \times b$

④ $a \times b^3$

⑤ a^3

CH 03 Q74 정답 및 해설

정답

④ $a \times b^3$ (차수 1+3=4 > 3)

해설

각 항의 지수 합이 degree 이하여야 함. $a \times b^3$=4차로 초과.

CH 03 | Q80

라쏘 특성 선택

61

라쏘를 특성 선택 용도로 쓸 수 있는 이유?

① 항상 최고

② 계수를 0으로 만듦

③ PF 포함

④ PCA

⑤ CV 내장

정답

② 불필요한 특성의 계수를 0으로 만들어 자동 제거 효과

해설

L1 규제의 수학적 특성. 55개 → 15개만 남긴 교재 예제.

CH 03 | Q86

최적 alpha 선택

62

alpha별 릿지 결과에서 최적은?

0.001→훈련0.99/테스트0.85, 0.1→0.99/0.98, 100→0.80/0.79

① 0.001

② 0.1

③ 100

④ 차이없음

⑤ 판단불가

정답

② 0.1 — 두 점수 높고 차이 최소

해설

0.001은 과대적합, 100은 과소적합. 0.1이 균형.

interaction_only

63

PolynomialFeatures에서 interaction_only=True이면 제외되는 항?

① 교호항

② 거듭제곱 항(a^2,b^2)

③ 원래 특성

④ 절편

⑤ 모든 새 특성

정답

② 거듭제곱 항이 제외되고 교호항만 추가

해설

같은 특성의 거듭제곱은 빠지고 서로 다른 특성 간 곱셈만 남음.

과대적합 틀린 설명

64

과대적합/과소적합 설명 중 틀린 것?

① 과대적합→훈련 높음

② 과대적합→테스트도 높음

③ 과소적합→훈련 낮을 수 있음

④ 과소적합→모델 단순

⑤ 과대적합→데이터 외움

정답

② '과대적합이면 테스트도 높다'가 틀림

해설

과대적합은 훈련↑ 테스트↓. 테스트도 높으면 정상 모델.

CH 03 | Q100

예측 입력 형태 일치

65

다항 모델에 50cm 예측 시 [[50**2, 50]]으로 넣는 이유?

① 정확도 향상

② 훈련 시 (길이², 길이) 구조이므로 동일하게

③ predict 요구

④ 제곱근+원래값

⑤ 아무값 추가

정답

② 훈련 시 (길이², 길이) 순서로 구성했으므로 동일 구조 입력

해설

특성 구조가 일치해야 계수가 올바르게 적용됨.

특성 2개 회귀 형태

66

특성이 2개인 다중 회귀에서 학습되는 모델의 기하학적 형태는?

① 직선

② 곡선

③ 평면

④ 점

⑤ 구

정답

③ 평면

해설

특성 1개=직선, 2개=평면, 3개 이상=초평면.

CH 03 | Q69

evaluate() 존재 여부

67

사이킷런 모델 클래스에서 제공하지 않는 메서드?

① fit()

② predict()

③ score()

④ evaluate()

⑤ get_params()

정답

④ evaluate()

해설

evaluate()는 케라스(딥러닝)에서 사용. 사이킷런 기본 패턴은 fit/predict/score.

68

로지스틱 회귀에 대한 올바른 설명?

① 회귀 문제 전용

② 이름은 회귀이지만 분류 모델

③ 비선형 방정식 학습

④ KNN의 별명

⑤ 비지도 학습

정답

② 이름은 회귀이지만 실제로는 분류 모델이다

해설

선형 방정식 학습 후 시그모이드/소프트맥스로 확률 변환하여 분류.

69

로지스틱 회귀가 학습하는 것은?

① 가장 가까운 이웃

② 의사결정 트리

③ 선형 회귀와 동일한 선형 방정식

④ 확률 분포

⑤ 클러스터 중심

정답

③ 선형 회귀와 동일한 선형 방정식

해설

선형 방정식 학습 후 시그모이드/소프트맥스로 확률 변환.

CH 04 | Q5

시그모이드 0.5 초과

70

이진 분류에서 시그모이드 출력이 0.5보다 크면?

① 음성

② 양성

③ 판단 불가

④ 두 클래스 모두

⑤ 클래스 없음

정답

② 양성 클래스

해설

0.5 초과 → 양성, 0.5 이하 → 음성.

CH 04 | Q6

시그모이드 z=0

71

시그모이드에 z=0을 넣으면?

① 0

② 0.25

③ 0.5

④ 1

⑤ -1

CH 04 Q6 정답 및 해설

정답

③ 0.5

해설

$1/(1+e^{0}) = 1/2 = 0.5$. $z=0$이 양성/음성 경계.

소프트맥스 합=1

72

소프트맥스 함수의 핵심 특징?

① 합=0

② 합=1

③ 음수 가능

④ z 하나만

⑤ 이진만

정답

② 출력값의 합이 항상 1

해설

합이 1이므로 각 출력을 클래스별 확률로 해석 가능.

CH 04 | Q13

클래스 정렬 순서

73

사이킷런에서 타깃값을 전달하면 클래스 순서가 어떻게 정렬?

① 입력순

② 알파벳순 자동 정렬

③ 빈도순

④ 랜덤

⑤ 숫자 크기순

정답

② 알파벳 순으로 자동 정렬

해설

classes_ 속성에 정렬 결과 저장. predict_proba() 열 순서도 이와 동일.

CH 04 | Q15

coef_ 행 수

74

다중 분류에서 coef_ 행 수를 결정하는 것?

① 샘플 수

② 특성 수

③ 클래스 수

④ 에포크

⑤ alpha

정답

③ 클래스 수

해설

클래스마다 방정식 하나씩 학습. coef_ 행=클래스, 열=특성.

C 매개변수

75

LogisticRegression에서 규제를 제어하는 매개변수는?

① alpha

② C

③ penalty

④ max_iter

⑤ tol

CH 04 Q17 정답 및 해설

정답

② C

해설

C↓=규제↑ (alpha와 반대). 기본값 1.

CH 04 | Q18

C vs alpha 반대

76

C와 alpha의 관계?

① 둘 다 커지면 규제↑

② C↓=규제↑, alpha↑=규제↑ (반대)

③ 동일

④ 분류/회귀 전용

⑤ 무관

정답

② C는 작을수록, alpha는 클수록 규제가 강해진다(반대)

해설

수학적으로 C는 규제항의 역수.

CH 04 | Q20

기본 규제 방식

77

LogisticRegression의 기본 규제 방식?

① L1

② L2

③ 없음

④ L1+L2

⑤ 드롭아웃

정답

② L2 규제 (릿지)

해설

기본적으로 계수 제곱 기준 규제.

CH 04 | Q27

확률적의 의미

78

SGD에서 '확률적'이란?

① 확률 예측

② 무작위/랜덤의 기술적 표현

③ 분포 학습

④ 정확 계산

⑤ 결정적

정답

② 무작위하게 또는 랜덤하게라는 기술적 표현

해설

랜덤하게 하나의 샘플을 골라 경사를 따라 내려감.

79

에포크(epoch)의 정의?

① 하나 샘플 사용

② 처음부터 재훈련

③ 훈련 세트를 한 번 모두 사용하는 과정

④ 테스트 평가

⑤ 하이퍼파라미터 조정

정답

③ 확률적 경사 하강법에서 훈련 세트를 한 번 모두 사용하는 과정

해설

보통 수십~수백 에포크 수행.

CH 04 | Q32

손실 함수 정의

80

손실 함수에 대한 올바른 설명?

① 클수록 좋음

② 얼마나 엉터리인지 측정, 작을수록 좋음

③ 파라미터 수 측정

④ 데이터 크기

⑤ 정확도와 동일

정답

② 알고리즘이 얼마나 엉터리인지 측정하는 기준, 값이 작을수록 좋다

해설

경사 하강법으로 손실 함수를 최소화하는 방향으로 학습.

CH 04 | Q33

손실 함수 조건

81

손실 함수가 갖추어야 할 수학적 조건?

① 정수만

② 미분 가능

③ 항상 양수

④ 최대 1

⑤ 불연속

정답

② 미분 가능해야 한다

해설

경사 하강법은 기울기(미분)를 이용. 미분 불가면 방향 알 수 없음.

82

정확도를 손실 함수로 못 쓰는 이유?

① 복잡

② 불연속→미분 불가

③ 항상 0

④ 회귀만

⑤ 음수 가능

정답

② 정확도는 불연속적이라 미분이 불가능

해설

뚝뚝 끊기는 값 → 미분 불가 → 연속적인 로지스틱 손실 사용.

CH 04 | Q35

이진 분류 손실 함수

83

이진 분류에서 사용하는 손실 함수 이름?

① 평균 제곱 오차

② 로지스틱(=이진 크로스엔트로피) 손실

③ 힌지 손실

④ MAE 손실

⑤ R^2 손실

정답

② 로지스틱 손실 함수(= 이진 크로스엔트로피 손실 함수)

해설

두 이름은 같은 함수. 타깃이 1일 때 예측이 1에 가까우면 손실↓.

SGDClassifier loss 기본값

84

SGDClassifier의 loss 기본값?

① log

② squared_loss

③ hinge

④ cross_entropy

⑤ mse

정답

③ hinge

해설

hinge는 SVM 손실. 로지스틱 회귀는 loss='log'로 지정 필요.

CH 04 | Q40

partial_fit 메서드

85

SGDClassifier에서 점진적 학습(추가 학습)을 수행하는 메서드?

① fit()

② predict()

③ partial_fit()

④ score()

⑤ transform()

정답

③ partial_fit()

해설

기존 가중치 유지하면서 1 에포크씩 이어서 훈련.

partial_fit vs fit

86

partial_fit()과 fit()의 차이?

① partial이 리셋

② partial은 이어서, fit은 새로

③ fit이 느림

④ 동일

⑤ partial은 테스트용

정답

② partial_fit()은 1 에포크씩 이어서, fit()은 처음부터 새로

해설

점진적 학습 = partial_fit. 일반 학습 = fit.

87

조기 종료(early stopping)의 정의?

① 시작 안 함

② 과대적합 전에 멈춤

③ 과소적합에서 멈춤

④ 손실=0 멈춤

⑤ 에포크=1

정답

② 과대적합이 시작하기 전에 훈련을 멈추는 것

해설

테스트 점수가 꺾이기 직전이 최적. 딥러닝에서도 핵심 규제 기법.

88

에포크 수와 과대/과소적합의 관계?

① 적으면 과대

② 적으면 과소, 많으면 과대

③ 무관

④ 많을수록 항상 좋음

⑤ 적을수록 좋음

정답

② 에포크 적으면 과소적합, 많으면 과대적합 가능성 높음

해설

적으면 학습 부족, 많으면 과도한 학습.

89

힌지 손실은 어떤 알고리즘의 손실 함수?

① 로지스틱 회귀

② KNN

③ 서포트 벡터 머신

④ 랜덤 포레스트

⑤ 신경망

정답

③ 서포트 벡터 머신(SVM)

해설

SGDClassifier 기본 loss='hinge'가 SVM을 구현.

CH 04 | Q57

이진 양성 클래스

90

classes_=['Bream','Smelt']일 때 양성 클래스?

① Bream

② Smelt

③ 둘 다

④ 없음

⑤ 랜덤

정답

② SmeIt (두 번째)

해설

사이킷런은 classes_의 두 번째를 양성으로 취급.

decision_function+expit

91

decision_function의 z값에 expit을 적용하면?

① R^2

② 절댓값

③ predict_proba 양성 확률과 동일

④ 손실값

⑤ 표준화

정답

③ predict_proba()의 양성 클래스 확률과 동일

해설

z → 시그모이드(expit) → 양성 확률 = predict_proba 두 번째 열.

CH 04 | Q60

coef_ shape 해석

92

coef_.shape=(7,5), intercept_.shape=(7,)에서 7과 5?

① 샘플,에포크

② 클래스,특성

③ 특성,클래스

④ 이웃,샘플

⑤ 에포크,배치

정답

② 7은 클래스 수, 5는 특성 수

해설

7종 생선 각각에 5개 특성의 계수. 방정식 7개.

CH 04 | Q61

softmax axis=1

93

softmax에서 axis=1을 쓰는 이유?

① 열 방향

② 각 행(샘플)에 대해 계산

③ 전체

④ 합

⑤ 삭제

정답

② 각 행(샘플)에 대해 소프트맥스를 계산하기 위해

해설

axis 없으면 전체 배열에 적용되어 잘못된 결과.

CH 04 | Q85

C 증가 효과

94

C를 1→20으로 올리면?

① 규제↑ 단순

② 규제↓ 자유 학습

③ 에포크 변경

④ 클래스 제한

⑤ 학습률 증가

정답

② 규제가 약해져 더 자유롭게 학습

해설

C↑=규제↓. 교재에서 '규제 완화를 위해 C=20'으로 설정.

CH 04 | Q88

SGD 잘못된 설명

95

경사 하강법에 대해 잘못된 것?

① 손실은 샘플 하나 기준

② SGD는 하나씩

③ 미니배치는 여러 개

④ SGDClassifier는 배치 경사 하강법

⑤ 배치는 전체

정답

④ SGDClassifier는 확률적 경사 하강법이지, 배치가 아님

해설

S=Stochastic. 이름 자체가 '확률적'.

96

다중 분류에서 클래스마다 방정식을 학습하는 이유?

① 속도

② 각 z값이 있어야 소프트맥스로 확률 계산

③ 정확도 보완

④ 제한

⑤ 시그모이드 미사용

정답

② 각 클래스의 z값이 있어야 소프트맥스로 확률을 계산할 수 있기 때문

해설

소프트맥스는 여러 z값 입력이 필요. 7클래스=7방정식.

시그모이드 0.5 판정

97

시그모이드 출력이 정확히 0.5일 때?

① 양성

② 음성

③ 보류

④ 에러

⑤ 랜덤

정답

② 음성 클래스

해설

'0.5보다 크면 양성'이므로 정확히 0.5는 '초과'에 해당하지 않아 음성.

98

소프트맥스 함수의 계산 과정?

① 각 z를 합으로 나눔

② e^z를 계산 후 e^z 전체 합으로 나눔

③ 로그 후 합산

④ 절댓값

⑤ 시그모이드 개별 적용

정답

② 각 z값에 e^z를 계산한 후, 전체 합으로 나눈다

해설

e_sum = Σe^zi, si = e^zi / e_sum. 지수+정규화.

99

predict_proba()와 decision_function()의 차이?

① proba=z, decision=확률

② proba=확률, decision=z

③ 동일

④ 훈련/예측용

⑤ 회귀/분류용

정답

② predict_proba()는 확률, decision_function()은 z값

해설

z값에 시그모이드/소프트맥스 적용하면 확률(proba).

fit+partial_fit 제공 클래스

100

fit()과 partial_fit()을 모두 제공하는 클래스?

① LinearRegression

② KNeighborsClassifier

③ SGDClassifier

④ StandardScaler

⑤ PolynomialFeatures

정답

③ SGDClassifier

해설

SGD 기반 클래스만 점진적 학습(partial_fit) 가능.

수고하셨습니다!

100문제 복습 완료 — 틀린 문제는 반드시 다시 풀어보세요

34	33	33
CH 01-02	CH 03	CH 04