

모델 및 이론 QA

Q1. 선형회귀(Linear Regression)는 언제 사용하면 좋을까요?

A1.

선형회귀는 ****종속변수(연속형)****가 있고, ****설명변수(독립변수)****와 선형적인 관계가 있을 때 사용합니다.

- 예시: 광고비 지출에 따라 매출액이 증가하는지 분석하고 싶을 때.
- 전제 조건: 선형성, 정규성, 등분산성, 독립성. 이상치 민감하므로 사전 확인 필요

Q2. 로지스틱 회귀(Logistic Regression)는 어떤 경우에 쓰나요?

A2.

****종속변수가 범주형(특히 이진형)****일 때 사용합니다.

예시: 고객이 이탈할지(Yes/No), 질병 유무(Yes/No)를 예측

출력: 사건이 발생할 확률 (0~1)

활용: 마케팅 이탈 예측, 의료 진단 모델 등

Q3. 다항 로지스틱 회귀(Multinomial Logistic Regression)는 언제 쓰나요?

A3. 종속변수가 세 개 이상의 범주형 변수일 때 사용합니다.

예시: 고객 만족도 (낮음/보통/높음) 예측

Q4. 의사결정나무(Decision Tree)는 어떤 경우에 유용할까요?

A4. 해석이 쉬운 모델이 필요할 때, 변수 간 관계가 비선형이거나 상호작용이 있을 때

예시: 고객의 연령, 지역, 구매 이력으로 상품 추천

장점: 시각화, 해석 용이, 단점: 과적합 가능성 → 가지치기 필요

Q5. 랜덤포레스트(Random Forest)는 어떤 분석에 적합한가요?

A5. 고차원 데이터, 변수 간 복잡한 상호작용, 높은 정확도 필요할 때

랜덤하게 여러 개의 결정트리를 만들고 평균을 내 예측 → 과적합 줄이면서 성능 확보

예시: 고객 이탈 예측, 사기 탐지

Q6. XGBoost/LightGBM은 언제 쓰나요?

A6.

- 정형 데이터에서 최고의 성능이 필요할 때
 - 변수 수 많고, 결측치 처리 자동화가 필요한 경우
 - Kaggle 등에서 자주 사용되는 boosting 기반 앙상블 모델
- 차이점: LightGBM은 속도와 효율성, XGBoost는 안정성에 강점

Q7. 주성분 분석(PCA)은 언제 활용하나요?

A7.

- 차원 축소가 필요할 때 (변수가 너무 많을 때)
- 변수 간 다중공선성이 있을 때
- 시각화, 속도 향상, 노이즈 제거 목적

예시: 이미지, 유전자 데이터 등 고차원 데이터 전처리

Q8. 군집분석(K-means, Hierarchical Clustering)은 언제 유용한가요?

A8.

- 레이블 없는 데이터에서 자연스러운 그룹을 찾고 싶을 때
- 고객 세분화, 행동 그룹화에 적합

K-means: 속도 빠름, 단순

계층적 군집: 시각화(덴드로그램) 가능, 정교한 구조 분석

Q9. 주성분 회귀(Principal Component Regression, PCR)는 어떤 경우 쓰나요?

A9.

- 설명변수가 많고, **다중공선성(multicollinearity)**이 심한 경우
- PCA로 차원 축소 → 회귀 분석

예시: 유사한 지표가 많은 경제 데이터 분석

Q10. 베이지안 이론(Bayesian Inference)은 언제 사용하나요?

A10.

- 사전 지식(prior knowledge)을 모델에 반영하고 싶을 때
- 작은 표본 또는 불확실성이 클 때 유리

예시: 신약 효능 분석, A/B 테스트 개선

Q11. 생존분석(Survival Analysis)은 어떤 경우에 활용되나요?

A11.

- 시간에 따라 사건이 발생할 확률을 분석
- 이탈 시점, 사망 시점, 구매 시점 예측

대표적 기법: Kaplan-Meier, Cox 비례위험모형

예시: 고객이 서비스를 언제까지 유지할지 예측

Q12. 시계열 분석(ARIMA, Prophet 등)은 어떤 때 쓰나요?

A12.

- 시간 순서가 중요한 데이터 (월별 매출, 주가, 기온 등)
- 계절성, 추세성 고려하여 미래를 예측

ARIMA: 통계적, 안정된 데이터

Prophet: Facebook 개발, 실무 친화적, 휴일 효과도 반영 가능

Q13. 가설 검정(Hypothesis Testing)은 어떤 목적일 때 쓰나요?

A13. 특정 주장(가설)이 통계적으로 의미 있는지 판단

예시: 신약이 기존 약보다 효과 있는지, A와 B 광고 중 어느 것이 전환을 높는지

검정 예: t-test, ANOVA, 카이제곱 등

Q14. 다중회귀분석(Multiple Linear Regression)은 언제 쓰나요?

A14.

- 여러 독립변수가 종속변수에 영향을 미치는 상황
- 변수 간 상호작용 고려 가능

예시: 가격, 광고, 지역 요인이 판매량에 미치는 영향 분석

Q15. 혼합모형(Mixed Effect Model)은 어떤 상황에 적합한가요?

A15. 데이터에 계층 구조 또는 반복 측정이 있는 경우

예시: 학생이 학교에 속하고, 여러 번 시험을 보는 경우

고정효과 + 랜덤효과로 전체적인 경향과 개별 차이 모두 반영