빅데이터 분석기사 필기 출제 요약 포인트

1과목 빅데이터 분석 기획

출제 유형

① 데이터 수집 기술 — 2가지 축으로 구분
수집 주기(실시간 vs 배치)와 데이터 형태(정형/반정형/비정형)를 교차해서 문제가 자주 출제
② HDFS
③ NoSQL 4가지 유형 — 대표 제품을 세트로 암기
④ 정형데이터 품질 기준
⑤ 하둡 에코시스템

1. 정성적 데이터와 정량적 데이터 구분(수치화 여부)★

정성적 데이터: 문자·범주·분류 중심
정량적 데이터: 수치·도형·문자

2. 정형 데이터와 반정형 데이터 특징 구분(구조화 정도)★★

정형 데이터: 고정된 스키마·표 형태·RDB 저장 용이
반정형 데이터: 스키마가 완전히 고정되지 않음(메타데이터), 태그·키값 구조 포함
예: XML·JSON·HTML·웹 로그 데이터

3. 데이터 저장방식 중 RDBMS와 NoSQL 도구 분류★★

RDBMS: Oracle·MySQL·PostgreSQL·SQL Server
NoSQL: MongoDB·Cassandra·HBase·Redis
RDBMS는 관계형·스키마 중심·NoSQL은 유연성·확장성 중심

4. 암묵지와 형식지의 상호작용 정의★

암묵지: 개인 경험·직관 속 지식·문서화 어려움
형식지: 문서·매뉴얼 등으로 표현 가능한 지식
상호작용: 공동화·표출화·연결화·내면화(SECI 모형)

5. DIKW 피라미드 정의★

Data → Information(의미) → Knowledge(예측) → Wisdom

6. 데이터베이스와 데이터 웨어하우스의 특징 구분★★★

데이터베이스(DB): 실시간 처리·통합성·저장된 데이터·운영 데이터·공용 데이터의 특징
데이터 웨어하우스(DW): 분석 처리(OLAP), 주제 중심·통합·비휘발·시계열 데이터 저장
무결성은 데이터의 정확성과 일관성에 관한 성질이며·DBMS는 제약조건 등을 통해 이를 유지

7. ODS(Operational Data Store)★

운영계 데이터의 통합 저장소·주로 정형 데이터 중심

8. 데이터 레이크(Data Lake)★★

정형·반정형·비정형 데이터를 원형 그대로 대량 저장하는 저장소
데이터를 저장할 때 미리 엄격한 스키마를 적용하지 않고·읽는 시점(Read)에 필요한 구조와 형식을 해석 → Schema-on-Read(스키마 온 리드)

9. 데이터베이스 설계 순서★

요구사항 분석 → 개념적 설계 → 논리적 설계 → 물리적 설계 → 구현

10. ETL 기능★

Extract: 원천 데이터 추출
Transform: 정제·변환·통합

Load: 저장소(DW 등)에 적재

핵심: 여러 소스의 데이터를 분석 가능한 형태로 통합

11. 빅데이터가 만들어내는 본질적인 변화★

사전처리 → 사후처리: 문제 발생 전에 예측하기보다·발생 후 빠르게 대응·최적화

표본조사 → 전수조사: 일부만 보는 것이 아니라 가능한 많은 데이터를 활용

질 → 양: 데이터의 정밀성만보다 데이터의 양과 패턴 발견이 중요

인과관계 → 상관관계: 왜 그런지 완벽히 몰라도·함께 나타나는 패턴을 활용 가능

12. 빅데이터의 위기요인과 통제방안★★

사생활 침해 → 비식별화·익명화

책임 원칙 훼손 → 책임성 확보

데이터 오용·남용 → 접근통제·보안강화·제도정비

13. 가트너의 비즈니스 분석 유형★

기술 분석(Descriptive Analytics): 무슨 일이 일어났는가

진단 분석(Diagnostic Analytics): 왜 일어났는가

예측 분석(Predictive Analytics): 앞으로 무엇이 일어날 것인가

처방 분석(Prescriptive Analytics): 무엇을 해야 하는가

14. 분석업무 수행 주체에 따른 3가지 조직구조 유형 구분★★

집중구조: 분석 전담 조직이 전사의 분석 업무를 통합 수행하는 구조

단점: 현업과 분리되어 이원화 문제가 발생할 수 있음

분산구조: 전사 공통 기준과 거버넌스 아래 각 부서가 분석 업무를 나누어 수행하는 구조

단점: 역할분담을 명확히 해야함

15. 데이터 사이언티스트 역량 중 하드스킬과 소프트스킬의 정의★

하드스킬: 데이터 처리·통계·프로그래밍·머신러닝 등 분석 수행에 필요한 기술적 역량

소프트스킬: 문제 정의·의사소통·협업·설득·비즈니스 이해 등 분석 결과를 활용하게 하는 비기술적 역량

16. 분석 준비도(Readiness)의 6개 영역(역량과 환경이 갖추어져 있는지를 평가)★★★

① 분석 업무: 분석을 실제 업무에 활용할 수 있는 수준인지 평가

② 인력 및 조직: 분석 전문 인력·조직체계·역할 분담이 갖춰져 있는지 평가

③ 분석 기법: 통계·데이터마이닝·머신러닝 등 분석 방법론과 기법 활용 수준 평가

④ 분석 데이터: 분석에 필요한 데이터의 확보·품질·통합 수준 평가

⑤ 분석 문화: 데이터 기반 의사결정을 수용하는 조직 문화가 형성되어 있는지 평가

⑥ IT 인프라: 분석 시스템·저장소·컴퓨팅 환경 등 기술적 기반이 마련되어 있는지 평가

17. 분석 성숙도(조직의 분석 활용 수준과 분석 역량 수준에 대한 진단)★★★

① 단계별 핵심

도입 단계: 분석 환경과 시스템 구축

활용 단계: 분석 결과를 실제 업무에 적용

확산 단계: 전사 차원으로 분석 공유·관리

최적화 단계: 분석 고도화로 혁신과 성과 창출

② 부문별 흐름

비즈니스 부문: 실적 분석 → 예측·시뮬레이션 → 전사 분석 → 최적화·실시간 분석

조직·역량 부문: 일부 부서·개인 의존 → 전문부서 운영 → 전사 확산·CoE → 경영진 활용·전략 연계

IT 부문: DW/DM/ETL → OLAP/대시보드 → 빅데이터·비주얼 분석 → 협업 환경·Sandbox → 프로세스 내재화

18. 분석 수준 진단 결과 4가지 유형★

준비형: 데이터·인력·조직·분석업무 등이 거의 적용되지 않은 상태

도입형: 분석업무·분석기법 등이 부족한 상태

정착형: 분석 요소가 기업 내부에서 제한적으로 활용되는 상태

확산형: 기업에 필요한 6가지 분석 구성요소를 고르게 갖춘 상태

19. 데이터 거버넌스: 전사 데이터의 표준화·통합·체계적 관리 체계★

20. 데이터 거버넌스 구성 요소 요약★

원칙: 지침·보안·품질

조직: 역할·책임

프로세스: 절차·모니터링·측정

21. 데이터 거버넌스 체계 요소★★★

표준화: 용어·명명 규칙·메타데이터·데이터사전
관리 체계: 원칙·프로세스·역할·책임
저장소 관리: 전사 저장소·통제·영향평가
표준화 활동: 주기적 점검·모니터링

22. 빅데이터를 수집·처리·분석해 가치 있는 정보를 제공하는 통합 IT 환경★

23. 하둡 에코시스템: 카프카·주키퍼·스파크 ★

카프카: 실시간으로 데이터를 전달·중개하는 역할
주키퍼(ZooKeeper): 분산 환경에서 서버 간 조정과 상태 관리를 담당
스파크(Spark): 고속 처리·인메모리 연산·범용 분석 엔진

24. 빅데이터 플랫폼은 역할에 따라 소프트웨어 계층·플랫폼 계층·인프라스트럭처 계층으로 구분★★★

소프트웨어 계층: 분석 실행
플랫폼 계층: 자원·작업 관리
인프라 계층: 저장·연산 기반 제공

25. 머신러닝 학습 유형★★

지도학습: 정답을 보고 학습
비지도학습: 정답 없이 패턴 탐색
강화학습: 보상을 받으며 행동 학습
준지도학습: 소량의 레이블 있는 데이터와 다량의 레이블 없는 데이터를 함께 활용
전이학습: 이미 학습된 모델이나 지식을 다른 과제에 이전하여 활용하는 학습 전략·접근 방식

26. 가명정보와 익명정보 차이★

가명정보: 추가정보와 결합하면 다시 개인 식별 가능성이 남아 있는 정보 → 개인정보에 해당
익명정보: 다시 식별할 수 없도록 처리된 정보 → 개인정보 아님
모든 개인정보를 무조건 익명처리하라는 뜻은 아니고·익명 또는 가명으로 처리해도 수집 목적을 달성할 수 있는 경우에 한해 적용됩니다.
이 경우에는 익명처리가 가능하면 익명으로·익명으로는 목적 달성이 어렵다면 가명으로 처리해야 한다고 규정
모든 개인정보를 무조건 익명처리하라는 뜻은 아니다.

27. 분석 주제 유형★★

Optimization: 대상 O·방법 O
Solution: 대상 O·방법 X
Insight: 대상 X·방법 O
Discovery: 대상 X·방법 X

28. 분석 마스터플랜 수립★★

우선순위 설정 시 전략적 중요도·비즈니스 성과·ROI·실행 용이성 등을 고려한다.
분석의 적용 범위와 방식도 함께 검토해야 한다.
즉·업무 내재화 여부·내부·외부 데이터 활용 범위·적용 기술 수준 등을 종합적으로 고려

29. ROI 관점에서 보는 빅데이터 4V 요약★

Volume·Variety·Velocity = 투자비용
Value = 비즈니스 효과

30. ROI 관점의 분석 과제 우선순위 평가★

시급성: 전략적으로 얼마나 급한가
난이도: 비용·범위상 얼마나 실행하기 쉬운가

31. 분석 과제 도출 방식★

하향식: 문제 주어짐 → 단계적으로 해결
상향식: 데이터 기반 → 문제 재정의·해결안 탐색

32. 하향식 접근방식(Top-down Approach): 전체 문제에서 출발해 세부적으로 내려가는 방식★

문제탐색 → 문제정의 → 해결방안 → 타당성 → 과제도출

33. 상향식 접근방식(Bottom Up Approach): 개별 사실에서 출발해 전체를 파악하는 방식★

프로세스 분류 → 프로세스 흐름 분석 → 분석 요건 식별 → 분석 요건 정의

34. 빅데이터 분석 방법론★★★

① 분석기획
비즈니스 이해 및 범위설정: SOW 작성·이해관계자 합의
프로젝트 정의 및 계획수립: 목표·KPI·목표수준·일정·조직·WBS 구체화
프로젝트 위험계획 수립: 위험 식별 및 대응계획 수립 → 회피·전이·완화·수용
② 데이터 준비
필요 데이터 준비: 내·외부 정형·반정형·비정형 데이터 정의
데이터 스토어 설계: RDBMS·Hadoop·NoSQL 등 저장 구조 설계
데이터 수집 및 정합성 점검: ETL·API·스크립트·크롤링 등으로 수집·검증
분석용 데이터 준비: 분석 목적에 맞는 데이터 추출·가공
③ 데이터 분석
텍스트 분석: 감성분석·토픽분석·오피니언 분석·SNA
탐색적 분석(EDA): 기초 통계량·분포·변수 관계 파악
모델링: 분류·예측·군집 등 모델 생성
모델평가 및 검증: 평가기준에 따라 성능·품질 검증
④ 시스템 구현
설계 및 구현: 알고리즘 설명서·시각화 보고서 기반으로 시스템·UI 설계 및 개발
시스템 테스트 및 운영: 단위 테스트·통합 테스트·시스템 테스트 수행
모델 발전 계획수립: 지속 운영과 성능 향상 계획 수립
⑤ 평가 및 전개
프로젝트 평가 보고: 성과를 정량·정성 평가
산출물·지식·프로세스를 지식 자산화
최종보고서 작성

35. CRISP-DM 단계와 태스크 요약★★★

①. 업무 이해(Business Understanding)
비즈니스 목적 파악 / 상황 파악 / 데이터 마이닝 목표 설정 / 프로젝트 계획 수립
② 데이터 이해(Data Understanding)
초기 데이터 수집 / 데이터 기술 분석 / 데이터 탐색 / 데이터 품질 확인
③ 데이터 준비(Data Preparation)
분석용 데이터셋 선택 / 데이터 정제 / 데이터 통합 / 데이터 포맷팅
④ 모델링(Modeling)
모델링 기법 선택 / 모델 테스트 계획 설계 / 모델 작성 / 모델 평가
⑤ 평가(Evaluation)
분석 결과 평가 / 모델링 과정 평가 / 모델 적용성 평가
⑥ 전개(Deployment)
전개 계획 수립 / 모니터링 및 유지보수 계획 수립 / 프로젝트 종료 보고서 작성 / 프로젝트 리뷰

36. KDD 분석 방법론★★

데이터 셋 선택: 목표 데이터 선정
데이터 전처리: 잡음·이상값·결측치 처리
데이터 변환: 변수 선택·차원 축소
데이터 마이닝: 패턴 발견·분류·예측
결과 평가: 해석·평가·활용

37. 데이터 전처리 기법 요약★★

데이터 정제: 결측값을 채우거나 이상치를 제거하여 데이터 신뢰도를 높이는 작업
데이터 통합: 여러 데이터를 합쳐 하나의 일관된 데이터로 만드는 작업
데이터 축소: 데이터 크기를 줄이되 분석 결과는 최대한 유지하는 작업
데이터 변환: 데이터 마이닝 효율 향상을 위해 형태를 변환·변형하는 작업

38. 분석 절차 프로세스★

문제 인식 → 연구 조사 → 모형화 → 데이터 수집 → 데이터 분석 → 분석 결과 제시

39. 데이터 유형별 수집 기술★

① DBMS(정형 데이터)
적합한 수집 방식: SQL·ETL·DB to DB·Sqoop·FTP
주의: 크롤링은 웹 문서용이므로 DBMS 수집 방식으로는 부적절
② 웹(Web 문서)
적합한 수집 방식: Crawling·Scraping
이유: 웹페이지에 있는 문서·텍스트·HTML 정보를 가져오는 방식이기 때문

③ 센서 데이터 / 실시간 대량 데이터
적합한 수집 방식: Open API·API
④ FTP
의미: 파일 단위 전송 방식
적합한 대상: 파일 형태 데이터
부적합한 경우: 지속적으로 실시간 생성되는 웹 로그 수집

40. 데이터 변환 기술★

① 평활화(Smoothing): 데이터나 이미지의 노이즈를 제거하고 부드럽게 만드는 기법
② 집계(Aggregation): 여러 데이터를 요약·통합하여 규모나 해상도를 바꾸는 기법
여러 속성을 하나로 축소하거나 유사 객체를 묶음
③ 일반화(Generalization): 민감한 정보를 보호하기 위해 데이터를 더 큰 범주로 단순화하는 기법
예: 나이 → 연령대·주소 → 시·도/국가 수준
④ Rescaling: 데이터를 일정한 범위나 기준에 맞게 변환하는 기법
정규화: 0~1 범위로 변환
표준화: 평균 0·표준편차 1로 변환
⑤ 속성 생성(Attribute / Feature): 기존 데이터를 바탕으로 새로운 속성(특징)을 만드는 기법

41. 개인정보 비식별화 기술(범주화), 데이터 마스킹도 출제 예상★★

① 범주화: 개별 값을 대표 범주값으로 바꾸는 방법
예: 35세 → 30대
② 랜덤 올림(랜덤 라운딩): 수치 데이터를 임의 기준으로 올림 또는 절사
예: 42·45·49세 → 40 또는 40대
③ 범위 방법: 수치 데이터를 특정 범위(구간)로 표현하는 방법
예: 3,300만원 → 3,000만원~4,000만원
④ 제어 올림: 올림 후에도 행·열 합계를 맞추도록 조정

42. 정형 데이터 품질기준★★

완전성 = 누락 없음
유일성 = 중복 없음
유효성 = 형식·범위 맞음
일관성 = 서로 모순 없음
정확성 = 사실과 규칙에 맞음

43. 데이터 프로파일링★

데이터 프로파일링은 데이터의 구조·내용·품질을 통계적으로 분석하여 품질 문제를 발견하고 개선점을 찾는 절차
주로 정형 텍스트 데이터와 비정형 콘텐츠의 메타데이터 품질 진단에 활용

44. 분산 파일시스템 ★★

여러 대의 머신에 데이터를 나누어 저장하고 하나의 파일 시스템처럼 관리하는 방식

45. 구글 파일 시스템(GFS)★

대용량 데이터를 청크 단위로 분산 저장하고 복제하여 관리하는 구글의 분산 파일 시스템

46. 하둡의 HDFS(Hadoop Distributed File System)★★

대용량 데이터를 블록 단위로 여러 서버에 분산·복제 저장하는 하둡의 분산 파일 시스템
HDFS는 데이터 안정성을 위해 복제(Replication)를 사용합니다. 파일 생성 시 복제 수를 지정할 수 있고·나중에 변경도 가능합니다. 따라서 "복제를 허용하지 않는다", "복제 횟수는 사용자가 변경할 수 없다"는 설명은 틀립니다.
HDFS는 GFS(Google File System)를 기반으로 설계된 하둡의 분산 파일시스템

47. NameNode / DataNode 구조★

NameNode: 파일시스템 메타데이터와 블록 복제 관련 의사결정을 담당
DataNode: 실제 데이터 블록 저장

48. NoSQL★

NoSQL은 분산 병렬처리에 적합한 확장성(Scale-Out)을 제공하는 비관계형 데이터베이스
전통적인 RDBMS처럼 복잡한 조인 연산 중심 구조에는 적합하지 않지만·대용량 데이터를 빠르게 처리하기 위해 설계되었다.

49. NoSQL 4가지 유형★★

① Key-Value NoSQL
키(Key)-값(Value) 쌍으로 저장

구조가 단순하여 확장성 우수·응답속도 빠름
예: Redis·DynamoDB
② Document NoSQL
데이터를 문서 형태(JSON 등)로 저장
계층적 트리 구조를 가짐
예: MongoDB·SimpleDB·CouchDB
③ Column NoSQL
테이블 개념은 있지만 행(Row)보다 컬럼(Column) 중심으로 저장
대용량 분산 저장에 적합
예: Cassandra·HBase
④ Graph NoSQL
데이터를 그래프 구조(노드-관계)로 저장
관계 표현과 연결 분석에 적합
예: Neo4j

50. 아파치 스쿱**(Apache Sqoop)★**

정형 데이터 전송 도구: 로그나 텍스트 같은 반정형 데이터 수집 도구인 플룸(Flume)과 대비되어·관계형 데이터베이스(RDBMS)의 정형 데이터 수집을 담당하는 도구

2과목 빅데이터 탐색

출제 유형

① 변수 선택 (래퍼/임베디드)
② PCA VS SVD VS 요인분석 VS 선형판별분석
③ 기댓값·분산 성질
④ 데이터 전처리
⑤ 확률분포
⑥ 가설검정(유의수준·유의확률·제1종 오류·제2종 오류)
⑦ 확률 표본 추출
⑧ 척도

51. 데이터 정제 (Data Cleansing)★★

데이터 정제는 데이터의 정확성을 높이기 위해 오류나 불필요한 정보를 바로잡는 과정
결측치(Missing value): 식별 후 제거하거나 다른 값으로 대체(대치/보정)합니다.
이상치(Outlier): 탐지 후 적절한 처리 방법으로 제거하거나 보정합니다.
노이즈 제거란: 데이터에 포함된 불규칙한 오차·이상한 흔들림·불필요한 변동을 줄여서 데이터의 본래 패턴이나 의미를 더 잘 드러나게 하는 것을 의미
☞ 정규화는 데이터의 범위와 스케일을 조정하는 기법이지·노이즈 제거 기법은 아니다.

52. 결측치 처리와 결측치 대치의 구분★★

단순 대치법: 결측값을 채움·범주형 변수는 최빈값 대치
다중 대치법: 결측값을 여러 방식으로 채움
회귀 대치법: 예측해서 채움·회귀 대치법은 변수 간 관계가 있어야 효과적
완전 삭제법: 채우는 것이 아니라 결측이 있는 행 자체를 제거

53. 상자그림(Box Plot) 해석(단변량 이상치 검색)★★

① 상자그림으로 확인 가능:
중위값(Median), 사분위수(Q1·Q3), 사분위범위(IQR), 이상값(Outlier)
자료의 퍼짐 정도를 대략적으로 비교(Q2가 Q1 가까울 때 오른쪽 꼬리 분포)
② 상자그림으로 직접 확인 어려움
관측치 수
통계적 유의성
결측치 수

54. 변수 선택(래퍼 방법(Wrapper Method)★★★

① 전진 선택법
→ 중요한 변수를 하나씩 추가
→ 추가만 가능·제거 불가
② 후진 제거법
→ 덜 중요한 변수를 하나씩 제거
→ 제거만 가능·다시 추가 불가
③ 단계 선택법
→ 변수를 추가도 하고 제거도 함
→ 새 변수가 들어오면 기존 변수의 유의성이 약해질 수 있으므로 재검토

55. 변수 선택(임베디드 메소드(Embedded Method))★★★

① Ridge → L2 규제
회귀계수의 제곱합에 패널티
② Lasso → L1 규제
회귀계수의 절댓값 합에 패널티
③ Elastic Net → L1 + L2 결합
Lasso와 Ridge를 함께 사용

56. 차원의 저주(Curse of Dimensionality)★

변수 수가 증가할수록 데이터가 희소해지고·학습이 어려워짐
차원이 커질수록 계산량 증가
저장 공간 증가
모델이 복잡해져 과대적합 위험 증가
차원 수에 비해 학습데이터가 부족해지면 성능 저하

즉·자주 틀리게 내는 부분은
→ 학습데이터 수가 차원의 수보다 적어져 문제 발생

57. 차원축소(Dimensionality Reduction)의 목적★

차원축소는 정보를 최대한 유지하면서 변수를 줄이는 것
노이즈 제거
특징 추출
시각화 용이
계산 효율 향상
다중공선성 완화 가능
→ 차원축소는 데이터 단순화와 효율성 향상 목적이지·설명력 증가가 핵심은 아님

58. PCA(주성분분석)★★★

기존 변수들의 선형결합으로 새로운 변수(주성분) 생성
분산이 가장 큰 방향을 찾음
주성분끼리는 서로 직교
차원축소·시각화·잡음 제거 등에 활용
고유값 분해(Eigendecomposition)를 이용한 PCA (정방행렬 필수)
자주 틀리게 내는 포인트
→ PCA는 설명력 증가 목적이 아님
→ PCA 결과는 해석이 직관적이지 않음

59. 요인분석(Factor Analysis)★

변수들 사이의 상관관계를 바탕으로 공통요인을 찾음
변수들 내부에 숨겨진 구조를 파악
독립변수/종속변수 구분 없음

60. SVD(특이값분해)★

임의의 행렬을 세 행렬의 곱으로 분해
차원축소 기법
정방행렬만 가능한 것이 아님

61. 파생변수★★

파생변수는 기존 변수로부터 새로운 변수를 만들어내는 것
기존 변수의 조합
변환
구간화
날짜·시간 분해 등을 통해 새로운 의미를 가진 변수를 만드는 것
종속변수는 파생변수 생성에 직접 쓰면 안 됨
자주 틀리게 내는 포인트
→ 결측치 대체는 파생변수 생성이 아님
→ 변수 이름만 변경하는 것도 파생변수 생성이 아님

62. Box-Cox 변환★

분포를 정규분포에 가깝게 만듦
분산을 안정화
양수 데이터에만 적용 가능
로그변환을 포함한 일반화된 형태로 이해 가능
자주 틀리게 내는 포인트
→ 파생변수 생성 목적이 아님
→ 범주형 데이터 처리용이 아님
→ 음수 데이터에는 직접 적용 불가

63. 로그변환★

값의 범위 차이가 클 때 사용
큰 값을 상대적으로 줄여 왜도 완화
양수 데이터에 주로 적용
정규성 개선에 도움 가능

64. 인코딩 핵심 포인트★

① Label Encoding

범주형 값을 숫자로 일대일 대응
순서 없는 범주에도 숫자가 부여될 수 있어 주의
② Ordinal Encoding
순서 정보가 있는 범주형 변수에 사용
순서를 유지하며 숫자 부여
③ One-Hot Encoding
각 범주를 0/1 더미변수로 변환
④ Target Encoding
범주를 타깃 변수의 평균값 등 통계값으로 대체
표준편차값으로 대체하는 방식이라고 하면 틀린 설명으로 자주 출제

65. 스케일링과 범주화 구분★

이 문제의 핵심은 스케일링은 수치 범위를 조정하는 것이고,
범주화는 값을 구간이나 범주로 나누는 것이라는 점입니다.
스케일링: 값의 크기 조정
Min-Max Normalization: X' = (X - Xmin) / (Xmax - Xmin)
Z-Score Standardization: Z = (X - 평균) / 표준편차
RobustScaler: X' = (X - 중앙값) / IQR
MaxAbsScaler: X' = X / max(|X|)
범주화: 연속형 값을 구간으로 나눔
예: 나이 → 청년/중년/노년
범주화는 스케일링 방법이 아님

66. 불균형 데이터의 핵심 문제★

불균형 데이터를 그대로 학습하면 모델이 다수 클래스로 치우쳐 예측하기 쉽다.
Accuracy(정확도)는 높게 보일 수 있지만
소수 클래스의 Precision·Recall은 낮아질 수 있다.

67. 불균형 데이터의 해결 방안★

① 언더샘플링: 다수 클래스 데이터를 줄이는 방법
장점: 불균형 완화
단점: 정보 손실 가능
② 오버샘플링: 소수 클래스 데이터를 늘리는 방법
장점: 정보 손실 줄임
단점: 과적합 가능
③ 임계값 이동(Threshold-moving): 학습 단계가 아니라 예측/검정 단계에서
분류 기준을 조정하는 방법
④ 비용 민감 학습 / 가중치 조정
소수 클래스 오분류에 더 큰 비용 또는 가중치를 부여
Weight Balancing 가능

68. EDA에서 확인하는 주요 항목★

① 결측치와 이상치 확인
② 데이터 분포 확인
③ 기초통계량 확인
④ 변수 간 관계 확인
자주 틀리게 내는 포인트
→ 단순 시각화나 추론통계와 동일하지 않으며 복잡한 전처리가 항상 필요한 것도 아니다.

69. 공분산(Covariance) 관련 핵심★★

독립이면 공분산은 0
하지만 공분산이 0이라고 해서 반드시 독립은 아님
공분산은 단위의 영향을 받으므로 관계 강도 비교에는 한계가 있음

70. 상관분석★★

① 수치형 변수
→ 피어슨 상관계수(Pearson Correlation)
→ 두 변수 간의 선형 관계의 강도와 방향을 측정
② 서열척도 변수
→ 스피어만 상관계수(Spearman Correlation)
→ 두 변수 간의 순위 기반 단조 관계(비선형 가능)의 강도와 방향을 측정

71. 편상관분석(Partial Correlation Analysis)★

다른 변수의 영향을 통제한 뒤 두 변수 간 순수한 상관관계를 보는 것

72. 변동계수(CV)

변동계수는 평균 대비 표준편차의 크기를 나타내며·단위가 다른 자료의 상대적 산포 비교에 사용한다.

73. 왜도(Skewness)★★

분포가 어느 한쪽으로 치우친 정도
즉·꼬리 방향의 비대칭성
양의 왜도 분포에서는 보통
평균 > 중앙값 > 최빈값·오른쪽 꼬리 분포임

74. 표본평균과 표본분산 계산★

예: 2·4, 6·8, 10
표본평균: $\bar{x} = (2 + 4 + 6 + 8 + 10) / 5 = 6$
표본분산: $s^2 = \Sigma(xi - \bar{x})^2 / (n - 1)$
값을 대입하면
$s^2 = \{(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2\} / 4$
$= (16 + 4 + 0 + 4 + 16) / 4 = 40 / 4 = 10$
분모가 n이 아니라 (n-1)인 이유
표본평균을 계산할 때 이미 데이터 하나에 대한 정보가 평균에 의해 제약을 받으므로 자유도 1이 감소한다.
표본분산의 분모를 n-1로 나누는 것은 자유도 1 감소를 반영하여 모분산의 불편추정량이 되도록 하기 위한 것이다.

75. 히스토그램★

계급값(또는 계급구간 폭)의 크기에 따라 히스토그램 해석이 달라짐
① 계급값이 너무 클 때
구간이 너무 넓어져서 데이터가 지나치게 뭉뚱그려짐
분포의 세부 구조가 잘 보이지 않음
봉우리 개수·치우침·이상치 등을 놓칠 수 있음
② 계급값이 너무 작을 때
구간이 너무 잘게 나뉘어 막대 수가 과도하게 많아짐
전체적인 분포 형태를 파악하기 어려움
잡음(noise)이 많아 보여 해석이 어려워질 수 있다는 점이 핵심

76. 단계 구분도★

지역별 통계값을 일정한 구간으로 나누고·각 구간에 서로 다른 색이나 명암을 적용하여 값의 크기 차이를 표현하는 주제도

77. 카토그램★

지역의 실제 면적이 아니라 특정 통계값에 비례하도록 크기를 왜곡해 표현하는 주제도이다.

78. 주성분분석 biplot 해석★★

주성분분석 biplot은 보통 두 가지를 한 그림에 함께 보여줌
점: 관측치(개별 데이터)
화살표: 원래 변수가 주성분 공간에서 가지는 방향과 기여 정도를 나타낸다.
화살표의 길이는 해당 원래 변수가 주성분 공간에서 얼마나 잘 표현되는지·그리고 주성분에 얼마나 크게 기여하는지를 나타냄

79. 텍스트 전처리의 용어 정의★★

① 토크나이징(Tokenizing)
문장을 의미 있는 최소 단위(토큰)로 나누는 작업
토큰 = 분리된 단어 또는 단위
② 품사 태깅(POS Tagging)
각 단어에 품사 정보를 부여하는 작업
명사·동사·형용사·부사 등으로 구분
③ 어간 추출(Stemming)
단어의 형태를 단순화하여 공통 어간으로 통합·형태가 달라도 유사 의미 단어를 묶음
④ 표제어 추출(Lemmatization)
단어를 사전적 기본형으로 변환
문법적·의미적 정보를 더 고려

80. 사회연결망 분석의 대표 지표★

먼저 큰 틀에서 자주 묻는 포인트는 다음 3가지입니다.

중심성(Centrality): 개별 노드가 얼마나 중심적인가
밀도(Density): 네트워크 전체가 얼마나 촘촘한가
중심화(Centralization): 특정 소수 노드에 중심이 얼마나 집중되어 있는가
① 연결정도 중심성(Degree Centrality)
한 노드에 직접 연결된 노드 수
연결이 많을수록 중심성이 높음
② 매개 중심성(Betweenness Centrality)
다른 노드들 사이를 이어주는 중개·브로커 역할 정도
③ 근접 중심성(Closeness Centrality)
다른 모든 노드까지의 거리를 기준으로 중심성 측정
전체 네트워크에 얼마나 빠르게 도달할 수 있는가
④ 위세 중심성(Eigenvector Centrality)
중요한 노드와 연결되어 있을수록 중심성이 높음

81. 조건부 확률★★

조건부 확률은 어떤 사건 B가 발생했다는 조건 아래에서 사건 A가 일어날 확률
$P(A|B) = P(A \cap B)/P(B)$ $(P(B)>0)$
조건부 확률은 교집합 확률과 연결
$P(A \cap B) = P(B) \cdot P(A|B)$ 또는
$P(A \cap B) = P(A) \cdot P(B|A)$
독립사건과의 구분
$P(A|B) = P(A)$ 또는 $P(A \cap B) = P(A)P(B)$

82. 베이즈 정리★★

베이즈 정리는 어떤 결과(증거)가 주어졌을 때·원인이 될 사건의 확률을 다시 계산하는 방법
$P(A|B) = P(B|A)P(A)/P(B)$
$P(A|B)$: 사후확률
$P(B|A)$: 가능도
$P(A)$: 사전확률
$P(B)$: 증거확률
사전확률은 기존 믿음·가능도는 조건부 관측 확률·사후확률은 새로운 정보 반영 후의 확률

83. 초기하분포 vs 이항분포★

① 이항분포
복원추출
독립
성공확률 일정
② 초기하분포
비복원추출
종속
성공확률 변함

84. 포아송 분포★

일정한 단위 시간 또는 단위 공간 안에서 어떤 사건이 발생하는 횟수를 나타내는 이산확률분포
포아송 분포는 평균과 분산이 모두 λ로 서로 같다.

85. 확률변수의 기댓값과 분산★

$E(aX+b) = aE(X)+b$
$Var(aX) = a^2Var(X)$ (상수배는 제곱해서 반영)
$Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y)$, 두 확률변수가 독립이 아닐 때

86. F분포의 특징★

F분포는 두 개의 카이제곱분포를 각각 자유도로 나눈 값의 비로 만들어지는 분포
두 집단 분산 비교·분산분석(ANOVA), 회귀모형 전체 유의성 검정에 사용

87. 카이제곱분포의 특징★

카이제곱분포는 표준정규분포를 따르는 독립 확률변수들을 제곱해서 더한 분포
분산 검정·적합도 검정·독립성 검정·동질성 검정 등에 사용됨

88. 중심극한정리★★

모집단의 분포 형태와 관계없이·표본크기가 충분히 크면 표본평균의 분포는 정규분포에 가까워진다는 것

중심극한정리는 연속형 변수에만 적용된다 → 틀림
모집단이 정규분포가 아니면 표본평균은 정규분포를 따를 수 없다 → 틀림

89. 신뢰수준★

신뢰수준은 같은 방법으로 구간추정을 반복했을 때·그 구간들 중 모수를 포함하는 비율
예를 들어 신뢰수준 95%라면,
같은 방식으로 표본을 반복 추출하여
매번 신뢰구간을 만들면
그중 약 95%의 구간이 모수를 포함한다는 뜻
이미 계산된 특정 한 구간이 95% 확률로 모수를 포함한다는 뜻으로 해석하면 부정확한 해석
신뢰수준이 높을수록 신뢰구간은 넓어진다.
표본크기 증가 → 표준오차 감소 → 구간 좁아짐
표본크기 감소 → 표준오차 증가 → 구간 넓어짐

90. 모평균 신뢰구간★

모표준편차를 알거나·문제에서 Z값을 주는 경우:
$\bar{X} \pm Z(\alpha/2) \times (\sigma / \sqrt{n})$
모분산을 모를 때 t분포:
$\bar{X} \pm t(\alpha/2 \cdot n-1) \times (s / \sqrt{n})$

91. 유의수준(significance level·α)★

귀무가설이 참인데도 이를 잘못 기각하는 제1종 오류를 범할 최대 허용 확률
유의수준이 0.05라는 것은·귀무가설이 실제로 참일 때 이를 잘못 기각할 최대 허용 오류를 5%로 설정했다는 의미

92. 유의확률(p-value)★

귀무가설이 참이라는 가정하에서·현재 관측된 결과와 같거나 그보다 더 극단적인 결과가 나올 확률
유의확률이 0.05라는 것은·귀무가설이 참이라고 할 때 현재 관측된 결과와 같거나 그보다 더 극단적인 결과가 나올 확률이 5%라는 의미
p-value ≤ 0.05이면 귀무가설 기각 = 통계적으로 유의하다고 판단한다.
= 관측된 결과가 우연만으로 나타났다고 보기 어렵다는 뜻
= 단지 귀무가설을 기각할 만큼 충분한 증거가 없었다는 의미
p-value > 0.05이면 귀무가설 기각하지 않음 = 통계적으로 유의하지 않다고 판단한다.

93. 제1종 오류(Type I Error)★

제1종 오류는 참인 귀무가설을 잘못 기각하는 오류이다.
유의수준은 제1종 오류를 허용하는 최대 확률

94. 제2종 오류(Type II Error)★

귀무가설이 실제로 거짓인데도 이를 기각하지 못하는 오류
β는 이러한 제2종 오류가 발생할 확률

95. 검정력(Power)★

귀무가설이 실제로 거짓일 때 이를 올바르게 기각할 확률을 의미
검정력 = 1 - β

96. 귀무가설(H0)★

차이가 없거나 효과가 없거나 변화가 없다는 기본 가설

97. 대립가설(H1 또는 Ha)★

차이가 있거나 효과가 있거나 변화가 있다는 가설
분석자가 입증하거나 주장하고 싶은 가설

98. 단측검정과 양측검정의 해석★

양측검정: 차이의 존재 여부 확인
단측검정: 차이의 방향까지 지정하여 확인
가설검정에서 대립가설은 검정 목적에 따라 양측검정·좌측 단측검정·우측 단측검정 등 여러 형태로 설정될 수 있으므로·하나의 귀무가설과 대립가설만이 존재한다고 단정할 수 없음

99. 유의수준 α를 미리 고정하고 가설검정하는 이유★

유의수준은 사후 해석 기준이 아니라 사전 의사결정 기준
유의수준 α를 고정하는 이유는 제1종 오류를 허용 가능한 수준으로 통제하고·객관적이고 일관된 기준에 따라 귀무가설의 기각 여부를 판단하기 위해서이다.

100. 모수검정(Parametric Test)와 비모수검정(Nonparametric Test)★★

① 모수검정

모집단이 특정 분포(주로 정규분포)를 따른다고 가정하고,

평균·분산 같은 모수(parameter)를 이용하여 검정하는 방법

가정이 충족되면 검정력이 높음

② 비모수검정

모집단의 특정 분포를 가정하지 않고·자료의 순위(rank)나 부호·범주 등을 이용하여 검정하는 방법

101. 확률적 표본추출★★

층화추출: 층 내부는 동질적·층 간에는 이질적

군집추출: 군집 내부는 이질적·군집 간에는 동질적

단순임의추출: 모두 같은 확률

계통추출: 모집단에서 처음 시작점은 무작위로 선택하고·그 이후에는 일정한 간격 K로 추출

102. 척도★★

명목척도: 분류·행정구역

서열척도: 분류 + 순서·선호도 조사

등간척도: 분류 + 순서 + 간격·온도

비율척도: 분류 + 순서 + 간격 + 절대영점·소득·척도 중 정보량이 가장 큼

3과목 빅데이터 모델링

출제유형

① 분석모형 선정 — 데이터 유형(정형/반정형/비정형), 지도학습·비지도학습·강화학습·준지도학습 구분
② 데이터 분할 방법 — 홀드아웃 vs K-Fold vs 부트스트랩의 차이점 구분이 핵심
③ 하이퍼파라미터 vs 파라미터 구분 — 분석자가 직접 설정하는 값(하이퍼파라미터)과 모델이 학습으로 도출하는 값(파라미터)의 개념 구분 문제
④ 지도학습 분류 또는 회귀모형 차이
⑤ 군집분석 — 계층적 vs 비계층적 군집 차이·K-means 프로세스·군집 수 k 결정 방법·SOM과 인공신경망 차이점이 핵심
⑥ 앙상블 기법 — 배깅 vs 부스팅 개념 차이·랜덤포레스트
⑦ 시계열 분석
⑧ 딥러닝 — CNN 은닉계층 구조·피처맵 계산·RNN 구조·오토인코더가 출제

103. 공분산분석(ANCOVA)★

집단 간 차이를 비교하되·다른 연속형 변수의 영향은 보정하여 살펴보는 방법
예) 학습방법(A반·B반)에 따른 시험점수를 비교하되·사전학습 수준의 영향은 통제

104. 다변량분산분석(MANOVA)★

하나의 독립변수에 대해 2개 이상의 종속변수를 동시에 분석하여 집단 간 차이를 검정하는 방법
예) 교수법(A·B, C)에 따라 수학·영어·과학 점수의 평균 차이가 전체적으로 유의한지 분석

105. 정준상관분석★

둘 이상의 변수로 이루어진 두 집합 간의 상관관계를 동시에 분석하는 방법
학업 관련 변수집합(수학·영어·과학)과 생활습관 변수집합(수면시간·공부시간·결석일수) 사이의 전체적 관련성 분석

106. 모형 복잡도와 편향·분산의 방향·과대적합/과소적합과의 연결★★★

편향(Bias)은 모델의 예측값이 실제값에서 벗어나는 정도를 의미한다.
분산(Variance)은 학습데이터가 달라질 때 모델 예측이 변하는 정도를 의미한다.
① 모형이 복잡할수록 편향은 작아지고 분산은 커진다.
② 모형이 단순할수록 편향은 커지고 분산은 작아진다.
③ 편향-분산 상충관계란 편향을 줄이면 분산이 커지고·분산을 줄이면 편향이 커지는 관계를 의미
④ 과대적합은 모형이 너무 복잡하여 편향은 낮고 분산은 높은 상태이다.
⑤ 과소적합은 모형이 너무 단순하여 편향은 높고 분산은 낮은 상태이다.

107. 데이터 분할★★★

① 데이터 분할의 목적
모델의 일반화 성능을 평가한다.
과적합 여부를 확인한다.
새로운 데이터에 대한 예측 성능을 검증한다.
② 데이터 분할의 기본 구성
가. 학습세트(Training set): 모델 학습에 사용
나. 검증세트(Validation set): 모델 선택 및 하이퍼파라미터 조정에 사용
다. 평가세트(Test set): 최종 성능 평가에 사용
③ Holdout 방식
데이터를 한 번 분할하여 사용하는 방법
간단하고 빠르지만·데이터가 적으면 평가 결과의 변동성이 커질 수 있다.
④ k-fold 교차검증
데이터를 k개로 나누고·각 부분집합을 번갈아 검증용으로 사용하여 평균 성능을 평가하는 방법
⑤ 부트스트랩(Bootstrap)
복원추출을 통해 여러 표본을 생성하고·포함되지 않은 OOB 데이터를 검증에 활용하는 방법
⑥ 부트스트랩의 포함 비율
특정 관측치가 한 번도 뽑히지 않을 확률은 (1-1/n)^n이며·n이 클 때 약 0.368이 된다.
따라서 약 36.8%는 표본에 포함되지 않고·약 63.2%는 적어도 한 번 포함된다.

108. 회귀분석★★★

① 개별 회귀계수의 유의성 검정은 t통계량을 사용한다.
② 전체 회귀모형의 유의성 검정은 F통계량을 사용한다.
③ 다항회귀는 독립변수의 차수를 높여 비선형 관계를 설명하는 회귀모형이다.
④ 다중공선성은 독립변수 간의 선형관계가 높을 때 발생한다.
⑤ 주성분 분석(PCA)은 변수 선택이 아니라 차원축소를 위한 기법이다.

⑥ 수정된 결정계수는 불필요한 변수가 추가되면 감소할 수 있다.
⑦ 종속변수의 유형에 따라 선형회귀와 로지스틱회귀를 구분한다.
⑧ 회귀분석 모형 구축 절차는 독립변수·종속변수 설정 → 회귀계수 추정 → 개별 회귀계수 유의성 검정 → 모형 유의성 검정의 순서로 이해해야 한다

109. 로지스틱 회귀분석★★

① 로지스틱 회귀분석은 범주형 종속변수를 설명하기 위한 회귀모형이다.
② 결과는 확률로 표현할 수 있으나·회귀계수 자체는 로그오즈의 변화량을 의미한다.
③ 로지스틱 회귀분석은 선형회귀분석과 달리 오차의 정규성 및 등분산성 가정을 직접 요구하지 않는다.
④ 선형회귀분석에서 전체 모형의 유의성은 F검정·개별 회귀계수의 유의성은 t검정으로 판단한다.
⑤ 로지스틱 회귀분석에서 전체 모형의 유의성은 우도비 검정·개별 회귀계수의 유의성은 Wald 검정으로 판단한다.
⑥ 로지스틱 회귀계수를 지수변환한 e^β는 오즈비로 해석된다

109. 의사결정나무★★

① 목표변수의 형태에 따라 분리기준이 달라진다.
② 이산형 목표변수에서는 지니지수·엔트로피 지수·카이제곱 통계량 등을 사용할 수 있다.
③ 연속형 목표변수에서는 F-통계량을 분리기준으로 사용할 수 있다.
④ 의사결정나무는 부모노드보다 자식노드의 순수도(동질성)가 높아지는 방향으로 분리된다.
⑤ 뿌리노드만 남는 경우는 유의미한 분할이 불가능하다는 의미이다.
⑥ 정지규칙은 트리의 과도한 성장을 막기 위해 분할을 멈추는 조건이다.
⑦ 의사결정나무는 설명력과 해석력이 높고 시각적으로 이해하기 쉽다.
⑧ 의사결정나무는 정규성·등분산성 가정이 필요 없는 비모수적 방법이다.
⑨ 변수 간 교호작용을 비교적 잘 파악할 수 있다.
⑩ 트리가 지나치게 복잡해지면 과적합 위험이 커질 수 있다.
⑪ 스케일링·정규화 등의 전처리가 필수는 아니다.

110. 인공신경망★★★

① 인공신경망의 학습은 가중치를 갱신하는 과정이다.
② 가중치는 입력 정보의 중요도를 조절하고·편향은 활성화 기준을 이동시키는 역할을 한다.
③ 출력값은 입력값과 가중치의 곱의 합에 편향을 더해 계산한다.
④ 활성함수는 비선형성을 부여하여 복잡한 패턴을 학습할 수 있게 한다.
⑤ 선형 활성함수를 여러 층 쌓아도 전체적으로 하나의 선형변환에 불과하다.
⑥ 소프트맥스 함수는 출력값을 확률값으로 변환하며·전체 합이 1이 되도록 한다.
⑦ 소프트맥스 함수는 주로 다중분류 문제의 출력층에서 사용된다.
⑧ 역전파는 오차를 출력층부터 역방향으로 전달하며 기울기를 계산하는 과정이다.
⑨ 경사하강법은 계산된 기울기를 이용하여 가중치를 갱신하는 방법이다.
⑩ 학습률은 가중치 갱신의 크기를 결정하는 값이다.
⑪ ReLU 함수는 양수 구간에서 일정한 기울기를 가져 기울기 소실 문제를 완화할 수 있다.

111. SVM의 기본 개념★

① 서포트 벡터 머신은 초평면을 이용하여 클래스를 분리하고·클래스 간 마진을 최대화하는 분류 모형이다.
② SVM은 분류뿐 아니라 회귀 문제(SVR)에도 적용할 수 있다.
③ 커널 트릭을 사용하면 비선형 데이터도 처리할 수 있다.
④ SVM은 마진 최대화를 통해 일반화 성능이 우수한 편이다.
⑤ 다만·학습 속도가 느릴 수 있고 계산 비용이 커질 수 있다.
⑥ 특히 대규모 데이터에서는 학습 시간이 길어질 수 있다.
⑦ SVM은 C 값과 커널 선택의 영향을 크게 받으므로 하이퍼파라미터 최적화가 중요하다.
⑧ C 값은 오분류 허용 정도와 마진의 크기 사이를 조절하는 파라미터이다.

112. 연관성 규칙★★

① 지지도(Support)는 전체 거래 중 특정 항목집합이 함께 등장한 비율이다.
② 신뢰도(Confidence)는 X가 발생한 거래 중 Y도 함께 발생한 비율로·조건부확률의 개념이다.
③ 향상도(Lift)는 X와 Y의 동시 발생이 우연인지 실제 연관성에 의한 것인지를 판단하는 지표이다.
④ 향상도 해석에서 Lift가 1보다 크면 양의 연관·1이면 독립·1보다 작으면 음의 연관으로 본다.
⑤ Apriori 알고리즘은 빈발항목집합을 탐색한 뒤 이를 기반으로 연관규칙을 생성하는 대표적 알고리즘이다.
⑥ 연관성 분석은 숨겨진 패턴을 발견하는 데 유용하고 해석이 직관적이라는 장점이 있다.
⑦ 다만 품목 수가 많아지면 계산량이 증가하고·불필요하거나 의미 없는 규칙이 많이 생성될 수 있다.

113. knn 알고리즘★

① KNN은 가장 가까운 k개의 이웃 데이터를 기준으로 새로운 데이터의 범주나 값을 예측하는 방법
② k는 참고할 이웃의 수를 의미하며·예측 결과와 성능에 큰 영향을 미친다.

③ k가 너무 작으면 과적합되기 쉽고·k가 너무 크면 과소적합될 수 있다.

114. 군집분석★★★

① 군집분석은 집단 내 유사성은 높이고 집단 간 이질성은 크게 하도록 데이터를 분류하는 비지도학습
② 군집분석은 크게 계층적 군집분석과 비계층적 군집분석으로 구분된다.
③ 계층적 군집분석은 군집 수를 사전에 정하지 않아도 되며·덴드로그램으로 표현할 수 있다.
④ 비계층적 군집분석은 일반적으로 군집 수를 사전에 지정해야 하며·대표적으로 K-means가 있다.
⑤ 병합 군집분석은 각 관측치를 하나의 군집으로 시작하여 가까운 군집끼리 순차적으로 합치는 방식
⑥ 계층적 군집분석에서는 유클리드 거리·맨해튼 거리·민코우스키 거리·마할라노비스 거리 등을 사용할 수 있다.
⑦ 민코우스키 거리는 p값에 따라 여러 거리척도를 일반화한 개념으로·p=1이면 맨해튼 거리·p=2이면 유클리드 거리와 관련된다.
⑧ K-means는 초기 중심점 설정·군집 할당·중심 재계산 과정을 반복하여 군집 내 제곱합을 최소화
⑨ 최적 군집 수는 엘보우 기법이나 실루엣 계수 등을 활용해 결정할 수 있다.
⑩ 가우시안 혼합모형은 군집을 확률적으로 할당하며·모수 추정에 EM 알고리즘을 사용한다.
⑪ DBSCAN은 밀도 기반 군집 알고리즘으로·군집 수를 미리 정하지 않아도 되고 이상치 탐지가 가능
⑫ SOM은 고차원 데이터를 저차원 격자에 시각화하는 비지도학습 기반 기법이다.

115. 범주형 자료분석★

① 카이제곱 검정은 범주형 변수들 사이의 독립성·동질성·적합도를 검정하는 데 사용된다.
② 카이제곱 독립성 검정은 두 범주형 변수가 서로 독립인지 여부를 판단하는 검정이다.
③ 카이제곱 동질성 검정은 여러 집단의 범주 분포가 동일한지 여부를 검정하는 방법이다.
④ 카이제곱 적합도 검정은 관측도수가 특정 이론적 분포에 적합한지 여부를 검정하는 방법이다.
⑤ 기대도수는 일반적으로 (행의합×열의합)/전체합으로 계산한다.
⑥ 기대도수는 귀무가설이 참일 때 각 셀에서 기대되는 이론적 빈도수를 의미한다. → 기준값의 역할

116. 시계열 분석★★

① 정상 시계열은 평균과 분산이 시간에 따라 일정하고·공분산은 시차에만 의존하는 시계열이다.
② 비정상 시계열은 평균이나 분산이 일정하지 않거나·공분산이 시차뿐 아니라 시점의 영향을 받는 시계열이다.
③ 백색잡음은 평균이 0이고 분산이 일정하며 자기상관이 없는 정상 시계열의 대표적 예이다.
④ AR 모형은 현재값이 자신의 과거값에 의존하는 모형이다.
⑤ MA 모형은 현재와 과거의 백색잡음의 선형결합으로 표현되는 모형이다.
⑥ ARIMA 모형은 AR·I, MA의 결합으로 이루어진다.
⑦ 차분은 비정상 시계열을 정상 시계열로 변환하는 데 사용된다.
⑧ ACF는 시차에 따른 자기상관 정도를 나타내는 함수이다.
⑨ 지수평활법은 최근 관측값에 더 큰 가중치를 부여하는 방법이다.
⑩ 시계열의 구성 요소에는 추세·계절성·순환성·불규칙성이 있다.

117. 나이브 베이즈 분류모형의 개념★★

① 나이브 베이즈 분류모형은 베이즈 정리에 기반한 확률적 분류모형이다.
② 각 속성은 주어진 클래스 하에서 서로 조건부 독립이라고 가정한다.
③ 사후확률이 가장 큰 클래스로 분류한다.
④ 장점은 구조가 단순하고 계산이 빠르며·적은 데이터에도 적용 가능하다는 점이다.
⑤ 단점은 독립성 가정이 깨질 경우 분류 성능이 저하될 수 있다는 점이다

118. 딥러닝★★★

① CNN은 필터를 이용한 합성곱 연산으로 입력 데이터의 국소적 특징을 추출하는 신경망이다.
② 합성곱 연산은 필터를 입력 위로 이동시키며 부분영역과의 곱을 합산하여 특징맵을 만드는 과정
③ 패딩은 입력 가장자리에 값을 추가하여 출력 크기 감소와 정보 손실을 줄이는 역할을 한다.
④ 스트라이드는 필터의 이동 간격을 의미하며·값이 커질수록 출력 크기는 작아진다.
⑤ CNN의 구성요소에는 합성곱층·풀링층·완전연결층이 있다.
⑥ RNN은 순차 데이터를 처리하는 신경망이지만·장기 의존성 문제와 기울기 소실 문제가 발생할 수 있다.
⑦ LSTM은 게이트 구조를 도입하여 중요한 정보를 오래 기억할 수 있도록 한 RNN의 확장 모형
⑧ Seq2Seq는 입력 시퀀스를 출력 시퀀스로 변환하는 Encoder-Decoder 구조의 모형이다.
⑨ 초기 Seq2Seq는 입력 정보를 하나의 컨텍스트 벡터에 압축하므로 긴 문장에서 정보 손실 한계가 있다.
⑩ 오토인코더는 입력을 자기 자신으로 복원하도록 학습하는 비지도학습 기반 신경망이다.
⑪ 오토인코더는 입력층과 출력층의 노드 수가 보통 같으며·은닉층을 통해 차원 축소에 활용될 수 있다.

119. 앙상블 기법★★★

① 배깅은 복원추출한 여러 표본으로 모델을 병렬 학습시키고·결과를 결합하여 분산을 줄이는 방법
② 랜덤 포레스트는 배깅 기반의 의사결정나무 앙상블 기법으로·데이터뿐 아니라 변수도 무작위로 선택하여 나무의 다양성을 높인다.
③ 부스팅은 이전 모형의 오류를 보완하도록 순차적으로 학습하여 편향을 줄이는 방법이다.
④ AdaBoost는 오분류된 데이터에 가중치를 부여하는 방식이다.

⑤ GBM은 잔차를 줄이는 방향으로 학습을 반복하는 방식이다.
⑥ XGBoost는 Level-wise 방식으로 트리를 성장시켜 비교적 안정적인 성능을 보인다.
⑦ LightGBM은 Leaf-wise 방식으로 트리를 성장시켜 빠르지만 과적합 위험이 있을 수 있다.
⑧ Stochastic Gradient Boosting은 각 단계에서 일부 데이터만 무작위로 사용하여 학습하는 방식
⑨ 스태킹은 여러 모델의 예측값을 활용해 메타 모델을 학습시키는 앙상블 기법이다.

120. 비모수 검정★

비모수 검정은 모집단의 분포를 가정하기 어렵거나 정규성 가정이 충족되지 않을 때 사용하는 검정 방법이다.
① 단일집단의 중위수 검정
부호검정
윌콕슨 부호순위 검정
② 대응표본인 2집단 비교
부호검정
윌콕슨 부호순위 검정
③ 독립인 2집단 비교
윌콕슨 순위합 검정
맨-휘트니 검정
K-S 검정

4과목 빅데이터 결과 해석

출제유형

① 평가지표: 회귀모형·분류모형 평가지표
② 과대적합 해결방안
③ 파라미터 최적화 기법: 경사하강법(Gradient Descent) 개념
④ 교차검증 vs 과적합 관계: 홀드아웃·K-Fold·LOOCV
⑤ 앙상블 융합: 하드 보팅 vs 소프트 보팅 차이
⑥ 정보 시각화 유형 분류
⑦ 인포그래픽 특징

121. 회귀모형의 평가지표★★★

① MAE는 오차의 절댓값을 평균한 지표로·해석이 직관적이고 이상치의 영향이 비교적 작다.
② MSE는 오차의 제곱을 평균한 지표로·이상치에 민감하다.
③ RMSE는 MSE에 제곱근을 취한 값으로·실제 데이터와 같은 단위를 가져 해석이 쉽다.
④ MAPE는 오차를 비율로 계산한 지표로·상대적 오차 비교에 유용하다.
⑤ 결정계수는 회귀모형의 설명력을 나타내는 지표로·1에 가까울수록 설명력이 높다.

122. 분류모형 평가지표 및 계산

① 분류모형의 평가지표는 혼동행렬을 기반으로 산출한다.
② 정확도는 전체 예측 중 올바르게 분류한 비율이다.
③ 정밀도는 양성으로 예측한 것 중 실제 양성의 비율이다.
④ 재현율은 실제 양성 중 양성으로 올바르게 분류한 비율이다.
⑤ F1-Score는 정밀도와 재현율의 조화평균이다.
⑥ F-베타 지표는 정밀도와 재현율에 서로 다른 가중치를 부여한 지표이다.
⑦ 카파 통계량은 우연에 의한 일치를 제외한 분류의 일치도를 나타낸다.

123. ROC AUC 해석 ★★★

① ROC 커브는 X축에 1-특이도·Y축에 민감도를 두어 나타낸 곡선이다.
② AUC는 ROC 커브 아래의 면적으로·값이 1에 가까울수록 분류 성능이 우수하다.
③ AUC가 0.5이면 무작위 추측과 같은 수준으로 해석한다.

124. Leave-One-Out Cross Validation (LOOCV) ★

① LOOCV는 전체 데이터 N개 중 1개를 테스트셋으로·나머지 N-1개를 훈련셋으로 사용하는 방법
② 이 과정을 모든 샘플에 대해 반복하므로 총 N번의 학습과 평가를 수행한다.
③ 장점은 모든 샘플을 한 번씩 테스트에 사용하므로 데이터 활용도가 높다는 점이다.
④ 단점은 모델을 N번 학습해야 하므로 계산 시간이 많이 소요된다는 점이다.

125. Leave-P-Out Cross Validation (LPOCV)★

① 전체 데이터 N개 중 p개를 테스트셋으로·나머지 N-p개를 훈련셋으로 사용하는 방법이다.
② 가능한 모든 p개 조합에 대해 반복하므로·LOOCV보다 경우의 수가 더 많아질 수 있다.

126. Monte Carlo Cross Validation (무작위 분할) ★

① 훈련셋과 테스트셋을 무작위 비율로 여러 번 반복 분할하여 성능을 평가하는 방법이다.
② 반복마다 분할이 달라질 수 있어 평가 결과가 달라질 수 있다.

127. Stratified K-Fold Cross Validation (층화 K-폴드)★

① 타깃 클래스의 비율을 유지하면서 K개의 폴드로 나누는 교차검증 방법이다.
② 불균형 데이터에서 주로 사용된다.
☞ 다양한 교차검증 기법이 존재하는 이유는 데이터 크기·클래스 분포·계산 효율성에 따라 각 방법의 장단점과 한계가 서로 다르기 때문이다.

128. 과대적합 해결방안★★★

① 데이터 증강 또는 데이터 추가 확보는 학습 데이터의 양을 늘려 과대적합을 완화하는 방법
② 가중치 규제는 가중치가 지나치게 커지지 않도록 패널티를 부여하여 모델의 복잡도를 억제하는 방법
③ 모델 복잡도 감소는 모델 구조를 단순하게 만들어 과대적합 가능성을 줄이는 방법
④ 드롭아웃은 학습 과정에서 일부 뉴런을 임의로 비활성화하여 일반화 성능을 높이는 기법

129. 경사하강법(Gradient Descent)의 종류 비교★★

① 경사하강법은 기울기 계산에 사용하는 데이터 양에 따라 구분된다.

② 배치 경사하강법은 전체 데이터를 사용하므로 안정적이지만 속도가 느리고 메모리 사용량이 크다.
③ 확률적 경사하강법은 샘플 1개를 사용하므로 빠르지만 업데이트 경로의 변동이 크다.
④ 미니배치 경사하강법은 일부 데이터만 사용하여 계산 효율성과 안정성을 함께 고려한 방법으로·가장 널리 사용된다.

130. 확률적 경사하강법(SGD) 기반 최적화 알고리즘 계층 구조★

SGD 기반 최적화 알고리즘은 기울기 방향 개선과 학습률 조절 개선의 관점에서 구분할 수 있다.
① 방향 개선 계열
Momentum: 이전 기울기의 방향을 일부 반영하여 관성을 부여하는 방법이다.
② 학습률 개선 계열
AdaGrad: 변수별로 학습률을 다르게 조정하는 방법이다.
많이 갱신된 변수는 학습률이 작아지고·적게 갱신된 변수는 학습률이 커진다.
단·학습이 진행될수록 학습률이 매우 작아질 수 있다.

131. 하이퍼파라미터 최적화★★

① 비정보적 탐색은 이전 탐색 결과를 다음 탐색에 반영하지 않고·정해진 규칙이나 무작위성에 따라 탐색하는 방법이다.
② 그리드 탐색은 사전에 정한 모든 조합을 탐색하는 방법이다.
③ 랜덤 탐색은 정해진 범위 내에서 무작위로 값을 선택하여 탐색하는 방법이다.
④ 정보적 탐색은 이전 탐색 결과를 바탕으로 유망한 탐색 영역을 예측하며 진행하는 방법이다.
⑤ 베이지안 최적화는 이전 탐색 결과를 반영하여 적은 횟수로 효율적으로 최적해를 찾는 방법

132. 정보 시각화 방법(Information Visualization) 분류 및 특징★★★

① 시간 시각화: 막대그래프·누적막대그래프·산점도·선그래프·계단식그래프·영역차트
② 분포 시각화: 파이차트·도우넛차트·누적막대그래프·트리맵·누적영역차트·히스토그램
③ 관계 시각화: 산점도·산점도 행렬·버블차트
④ 비교 시각화: 막대그래프·히트맵·평행 좌표계·스타 차트·체르노프 페이스
⑤ 공간 시각화: 지도매핑