

빅데이터의 이해, 현대의서양문화

2019학년도 2 학기

4 학년 2 교시

※ 정답 하나만을 골라 반드시 컴퓨터용 사인펜으로 OMR 답안지에 표기할 것.	학 과		감독관	인
	학 번	-	성 명	

1과목	빅 데이터 의 이 해 (1~25)
출제위원 : 방송대 이궁희	
출제범위 : 교재 1~10장 (R 프로그램 실습 제외)	

1. 빅데이터의 3V가 아닌 것은? (4점)

- ① 데이터의 속도
- ② 데이터의 다양성
- ③ 데이터의 규모
- ④ 데이터의 정확도

2. 빅데이터의 출현배경이 아닌 것은? (2점)

- ① 개인정보보호법의 시행
- ② 센서의 보급 확대
- ③ SNS의 보급 확대
- ④ 새로운 데이터 유형을 다루는 신기술 등장

3. 1 제타 바이트(ZB)는 몇 기가바이트(GB)인가? (2점)

- ① 10^6 GB
- ② 10^9 GB
- ③ 10^{12} GB
- ④ 10^{15} GB

4. “데이터 과학은 (), 해킹기술과 관련 주제 전문지식이 종합된 분야이다.”에서 ()에 가장 적합한 단어는? (4점)

- ① 프로그래밍
- ② 통계학
- ③ 인문학
- ④ 경영학

5. 다음 중 데이터과학자가 가져야할 소프트스킬을 모두 고르시오. (4점)

- ㉠ R프로그래밍 작성 능력
- ㉡ 데이터베이스와 하둡 기술
- ㉢ 데이터로부터 가치를 발견할 수 있는 통찰력
- ㉣ 다른 전문가와 소통할 수 있는 능력

- ① ㉠, ㉡
- ② ㉠, ㉢
- ③ ㉡, ㉣
- ④ ㉢, ㉣

6. 다음 중 빅데이터 경영이라고 할 수 없는 것은? (3점)

- ① 차량운행기록장치의 데이터에 기반한 보험료 산출
- ② 인터넷으로 주민등록등본을 발급
- ③ 점포별·상품별 데이터 분석을 통한 패션 트렌드 예측 및 생산
- ④ 추천시스템을 통해 영화를 개별적으로 추천

7. 필립 코틀러 등이 말하는 디지털 시대의 마케팅 트렌드를 모두 고르시오. (3점)

- ㉠ 고객 취향과 요구를 감지, 대응하는데 주력하는 마케팅
- ㉡ 새로운 고객을 확보하는데 주력하는 마케팅
- ㉢ 거래를 성사시키는 마케팅
- ㉣ 고객을 개별적으로 대하는 맞춤형 마케팅

- ① ㉠, ㉡
- ② ㉠, ㉣
- ③ ㉡, ㉣
- ④ ㉠, ㉢, ㉣

8. 분산 시스템에 대한 설명 중 가장 바른 것은? (4점)

- ① 분산 시스템은 저비용으로 고성능을 얻을 수 있으나 신뢰성은 낮다.
- ② 분산 시스템은 저비용으로 고성능을 얻을 수 없으며 신뢰성은 낮다.
- ③ 분산 시스템은 저비용으로 고성능을 얻을 수 있고 신뢰성도 높다.
- ④ 분산 시스템은 저비용으로 고성능을 얻을 수 없으나 신뢰성이 높다.

9. 다음 중 하둡에 대한 설명으로 가장 바른 것은? (3점)

- ① 오픈소스이며, 낮은 가격의 서버로 분산처리시스템을 구성할 수 있다.
- ② 오픈소스가 아니지만, 낮은 가격의 서버로 분산처리시스템을 구성할 수 있다.
- ③ 오픈소스가 아니며, 낮은 가격의 서버로 분산처리시스템을 구성할 수 없다.
- ④ 오픈소스이며, 낮은 가격의 서버로 분산처리시스템을 구성할 수 없다.

10. “맵리듀스 프로그래밍 모형에서는 (㉠)와 (㉡)의 쌍을 다루는 맵과 리듀스 함수를 작성하여 빅데이터를 처리한다.” ㉠, ㉡에 가장 알맞은 단어는? (3점)

- ① 키, 인메모리
- ② 시스템, 인메모리
- ③ 키, 값
- ④ 시스템, 값

11. HTML0이나 XML로 작성된 웹페이지에서 사용자가 정한 저장 규칙에 따라 데이터를 분해, 저장하는 것을 무엇이라 부르는가? (3점)

- ① 웹 스크래핑(web scraping)
- ② URL
- ③ 파싱(Parsing)
- ④ API

12. MIT Billion Price Project를 기반으로 작성된 온라인 물가지수에 대한 설명으로 가장 바른 것은? (3점)

- ① 온라인물가지수는 인터넷의 가격정보를 수집해서 국가통계인 소비자물가지수보다 빠르게 작성할 수 있다.
- ② 온라인물가지수는 시장의 가격정보를 조사해서 국가통계인 소비자물가지수보다 빠르게 작성할 수 있다.
- ③ 온라인물가지수는 인터넷의 가격정보를 수집해서 국가통계인 소비자물가지수보다 빠르게 작성할 수 없다.
- ④ 온라인물가지수는 시장의 가격정보를 조사해서 국가통계인 소비자물가지수보다 빠르게 작성할 수 없다.

13. 구글 Ngram에서 이용한 데이터는? (2점)

- ① 디지털화 된 책
- ② 인터넷 검색
- ③ 신문
- ④ SNS

14. 데이터 시각화의 전문가 Edgar Tufte는 데이터 시각화의 8가지 원칙을 세웠는데 다음 중에서 이에 해당한 것을 모두 고르시오. (4점)

- ㉠ 데이터 그 자체를 보여주는 것이 중요
- ㉡ 데이터는 몇 가지 단계로 깊이 들어가 자세히 살펴볼 수 있어야 함.
- ㉢ 너무 많은 숫자나 문자를 작은 화면에 보여주려고 하지 않음.

- ① 가, 나
② 가, 다
③ 나, 다
④ 가, 나, 다

15. 하나의 어절에서 의미를 갖는 최소 단위를 분석해 내는 기법은 무엇인가? (2점)

- ① 프로세싱
- ② 스테밍(stemming)
- ③ 토큰화(tokenizing)
- ④ 형태소분석

16. 벌점화 최소제곱추정량에 대한 설명 중 가장 바르게 기술된 것은?
(2점)

- ① 편의가 생기지만 평균제곱오차를 줄일 수 있다.
- ② 분산이 커지지만 평균제곱오차를 줄일 수 있다.
- ③ 편의가 없어지지만 평균제곱오차가 커진다.
- ④ 분산이 작아지지만 평균제곱오차가 커진다.

17. 라쏘 벌점함수를 이용한 추정에 대한 설명 중 가장 바르게 기술된 것은? (2점)

- ① 축소 추정을 통해 예측력은 높아지지만 불필요한 변수를 제거하지 못함으로써 모형의 해석력이 낮아진다.
- ② 축소 추정을 통해 예측력은 낮아지지만 불필요한 변수를 제거함으로써 모형의 해석력이 높아진다.
- ③ 축소 추정을 통해 예측력은 높아지고 불필요한 변수를 제거함으로써 모형의 해석력이 높아진다.
- ④ 축소 추정을 통해 예측력은 낮아지고 불필요한 변수를 제거하지 못함으로써 모형의 해석력이 낮아진다.

18. “배깅방법은 주어진 데이터에 대해서 여러 개의 () 데이터를 생성하고 이의 예측모형을 만든 후 결합하여 만드는 방법이다.”
() 내에 가장 알맞은 단어는? (3점)

- ① 붓스트랩
- ② 나무
- ③ 1차 근방
- ④ 벤치마크

19. 랜덤포레스트 방법에 대해 가장 바르게 기술한 것은? (3점)

- ① 배깅보다 무작위성이 작은 방법이며, 최종결과를 해석하기 어렵다.
- ② 배깅보다 무작위성이 작은 방법이며, 최종결과를 해석하기 쉽다.
- ③ 배깅보다 무작위성이 큰 방법이며, 최종결과를 해석하기 어렵다.
- ④ 배깅보다 무작위성이 큰 방법이며, 최종결과를 해석하기 쉽다.

20. 다음 중 유사성 척도에 대해 가장 바르게 기술된 것은? (2점)

- ① 범주형 변수에 대한 척도로는 유클리디언 거리, cosine 유사성과 연속형 변수에 대한 척도로는 Jaccard 계수가 있다.
- ② 범주형 변수에 대한 척도로는 유클리디언 거리와 연속형 변수에 대한 척도로는 Jaccard 계수, cosine 유사성이 있다.
- ③ 연속형 변수에 대한 척도로는 유클리디언 거리와 범주형 변수에 대한 척도로는 Jaccard 계수, cosine 유사성이 있다.
- ④ 연속형 변수에 대한 척도로는 유클리디언 거리, cosine 유사성 등이 있고, 범주형 변수에 대한 척도로는 Jaccard 계수가 있다.

※ (21~25) 편의점의 거래내역을 보고 물음에 답하시오.

고객번호	품목
1	우유, 라면, 바나나
2	커피, 우유, 김밥, 바나나
3	김밥, 생수, 바나나
4	커피, 라면
5	김밥, 우유, 생수, 커피

21. “김밥을 구매하면 바나나를 구매한다.”의 지지도는? (2점)

- $$\begin{array}{ll} \textcircled{1} \frac{1}{5} & \textcircled{2} \frac{2}{5} \\ \textcircled{3} \frac{3}{5} & \textcircled{4} \frac{4}{5} \end{array}$$

22. “김밥을 구매하면 바나나를 구매한다.”의 신뢰도는? (3점)

- $$\begin{array}{ll} \textcircled{1} \quad \frac{1}{2} & \textcircled{2} \quad \frac{2}{3} \\ \textcircled{3} \quad \frac{2}{4} & \textcircled{4} \quad \frac{2}{5} \end{array}$$

23. 최소지지도를 넘는 빈발품목집합에서 높은 지지도를 갖는 품목 집합을 찾는 방법은? (3점)

- ① 협력적 정화방법
- ② 선형적 알고리즘
- ③ 앙상블 추천방법
- ④ 계층적 추천방법

24. 다음 중 프라이버시에 속하는 것으로 가장 바른 것은? (2점)

- ① 정치적 신념
- ② 성명
- ③ 이메일
- ④ 소속 직장명

25. 다음 중 데이터 수집단계에서 이용자가 익명으로 이용할 수 있도록 하는 익명화 기술이 아닌 것은? (2점)

- ① 데이터 마스킹
- ② 데이터 범주화 방식
- ③ 가명처리
- ④ 방화벽 기술