2022년 33회 기출복원 인쇄용 (50문제)

총점 47/50

1. 데이터 사이언티스트가 갖추어야 하는 역량 중 소프트 스킬에 해당하지 않는 것은? 1/1

- 통찰력 있는 분석
- 다분야 간 협력
- 설득력 있는 전달
- 빅데이터 이론적 지식 ✓

의견 보내기

하드 스킬
- Machine Learning, Modeling, Data Technical Skill
- 빅데이터에 대한 이론적 지식 : 관련 기법에 대한 이해와 방법론 습득
- 분석 기술에 대한 숙련 : 최적의 분석 설계 및 노하우 축적

소프트 스킬
- 통찰력 있는 분석 : 창의적 사고, 호기심, 논리적 비판
- 설득력 있는 전달 : Storytelling, Visualization
- 다분야 간 협력 : Communication

2. 데이터 크기를 작은 것부터 큰 것 순서로 올바르게 나열한 것은? 1/1

- PB < EB < ZB < YB ✓
- YB < ZB < EB < PB
- PB < YB < EB < ZB
- PB < ZB < EB < YB

의견 보내기

KB < MB < GB < TB < PB < EB < ZB < YB (Peta < Exa < Zetta < Yotta)

✓ **3. 데이터베이스의 특징이 아닌 것은?** 1/1

- ○ USB 등 컴퓨터가 접근할 수 있는 저장 매체에 저장할 수 있다
- ○ 데이터는 공동으로 이용 된다
- ○ 데이터가 중복되어 있지 않다
- ◉ 정형 데이터만 저장할 수 있다 ✓

의견 보내기

데이터베이스의 특징
- 통합된 데이터 : 데이터베이스에서 같은 내용의 데이터가 중복되어 있지 않다는 것을 의미
- 저장된 데이터 : 자기디스크나 자기테이프 등과 같이 컴퓨터가 접근할 수 있는 저장매체에 저장되는 것을 의미
- 공용 데이터 : 여러 사용자에게 서로 다른 목적으로 데이터베이스의 데이터를 공동으로 이용되는 것을 의미
- 변화되는 데이터 : 새로운 데이터의 추가, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 정확한 데이터를 유지해야 한다는 것을 의미

✓ **4. 미래의 빅데이터 관점에서 볼 때 사물인터넷(IoT)과 가장 관련이 큰 것은?** 1/1

- ◉ 모든 사물의 데이터화 ✓
- ○ 모든 사물의 독립화
- ○ 모든 사물의 그래픽화
- ○ 모든 사물의 정형화

의견 보내기

사물인터넷(IoT)
IoT란, 인터넷에 연결되어 IoT 애플리케이션이나 네트워크에 연결된 장치, 또는 산업 장비 등의 다른 사물들과 데이터를 공유할 수 있는 수많은 '사물'을 말합니다. 인터넷에 연결된 장치는 내장 센서를 사용하여 데이터를 수집하고, 경우에 따라 그에 맞게 반응합니다

✓ **5. 빅데이터가 가져온 변화로 맞지 않은 것은?** 1/1

- (●) 서비스 산업이 확대되고 제조업의 생산성이 감소되었다 ✓
- ( ) 빅데이터 시대에는 데이터 획득비용이 기하급수적으로 감소하고 모든 곳에서 데이터가 넘쳐나 사용자 전수조사가 가능해 졌다
- ( ) 가능한 한 많은 데이터를 모으고 그 데이터를 다양한 방식으로 조합해 숨은 정보를 찾아낸다
- ( ) 데이터의 질보다 양을 강조하게 되었다

의견 보내기

빅데이터는 각종 비즈니스, 공공기관 대국민 서비스, 경제 성장에 필요한 '정보'를 제공하여, 산업 전반의 생산성을 향상시킬 것으로 기대된다

✓ **6. 다음 중 빅데이터 위기 요인과 해결 방안을 잘못 연결된 것을 고르시오.** 1/1
가. 사생활 침해 -> 동의제를 책임제로 전환
나. 책임원칙의 훼손 -> 알고리즘 허용
다. 데이터의 오용 -> 결과 기반 책임 원칙

- ( ) 가, 나
- ( ) 가, 다
- (●) 나, 다 ✓
- ( ) 가, 나, 다

의견 보내기

- 책임원칙의 원칙 -> 기존의 책임원칙을 강화할 수 밖에 없다
- 데이터의 오용 -> 데이터 알고리즘에 대한 접근권 허용 및 객관적 인증방안을 도입 필요성 제기

✓ **7. 데이터 NoSQL 저장방식과 관련이 없는 도구는?** 1/1

- MongoDB
- HBase
- Redis
- MySQL ✓

의견 보내기

NoSQL : MongoDB, Apache Hbase, Redis
RDBMS : MySQL(오픈소스 RDBMS), Oracle Database(상용 RDBMS)

✓ **8. 빅데이터 특징 중 올바르지 않는 것은?** 1/1

- 비즈니스 상황에서는 인과관계를 모르고 상관관계 분석만으로 충분한 경우가 많다
- 사전처리에서 사후처리 시대로 변화하였고, 사전처리의 대표적인 예로는 표준화된 문서 포맷을 들 수 있다
- 표본조사의 중요성이 높아졌다 ✓
- 데이터 수가 증가함에 따라 몇 개의 오류데이터가 대세에 영향을 주지 못하는 경향이 증가하고 있다

의견 보내기

빅데이터가 가져온 본질적인 변화
- 사전처리 -> 사후처리
- 표본조사 -> 전수조사
- 질(Quality) -> 양(Quantity)
- 인과관계 -> 상관관계

✕ **9. 조직의 의사결정을 위한 데이터 집합체로 데이터 통합, 시계열성, 비소멸성 등의 특징을 가지고 있는 것은?** ···/1

데이터 웨어하우스 ✕

정답

데이터웨어하우스

Data WareHouse

의견 보내기

데이터웨어하우스 : 기업 내의 의사결정 지원 애플리케이션을 위한 정보를 제공하는 하나의 통합된 데이터 저장 공간으로 데이터의 통합, 데이터의 시계열성, 데이터 주제 지향적, 비소멸성의 특징을 가지고 있다.

✓ **10. 다양한 유형의 데이터를 다루는 통계학과 마이닝을 넘어서는 학문, 데이터 공학, 수학, 통계학, 컴퓨터 공학 등 해당 분야의 전문 지식을 종합한 학문은?** 1/1

데이터 사이언스 ✓

의견 보내기

데이터 사이언스 : 데이터로부터 의미 있는 정보를 추출해내는 학문으로 정형, 반정형, 비정형의 다양한 유형의 데이터를 대상으로 하며, 분석 뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함한 포괄적 개념이다

✓ **1. 빅데이터 분석 방법론에서 분석 기획 단계의 task로 적절하지 않은 것은?** 1/1

- ○ 비즈니스 이해 및 범위 설정
- ○ 프로젝트 정의 및 계획 수립
- ○ 프로젝트 위험 계획 수립
- ◉ 필요 데이터 정의 ✓

의견 보내기

- 분석 기획 단계 : 비즈니스 이해 및 범위 설정, 프로젝트 정의 및 계획 수립, 프로젝트 위험 계획 수립
- 데이터 준비 단계 : 필요 데이터 정의, 데이터 스토어 설계, 데이터 수집 및 적합성 점검

✓ **2. 빅데이터 분석방법론의 계층적 프로세스 모델에 대한 설명으로 적절하지 않는 것은?** 1/1

- ◉ Task는 단계를 구성하는 단위 활동으로 input, output로 구성된 단위 프로세스이다 ✓
- ○ Phase(단계)는 최상위 단계로 프로세스 그룹을 통하여 완성된 단계별 산출물을 생성한다.
- ○ Step(스탭)은 마지막 계층으로 WBS(Work Breakdown Structure)의 워크패키지에 해당한다
- ○ Phase, Task, Step 계층이 있다

의견 보내기

계층적 프로세스 모델(Stepwised Process Model)
- 단계(Phase) : 최상위 계층, 프로세스 그룹을 통해 완성된 단계별 산출물을 생성, 각 단계는 기준선(Baseline)으로 설정되어 관리, 버전관리(Configuration Management)등을 통해 통제
- 태스크(Task) : 단계를 구성하는 단위 활동, 물리적/논리적 단위의 품질 검토의 항목이 될 수 있으며, 각 단계(Phase)는 여러 개의 태스크(Task)로 구성됨
- 스텝(Step) : 마지막 계층으로 WBS의 워크패키지(Work Package)에 해당하며, 입력자료 처리 및 도구, 출력자료로 구성된 단위 프로세스이다

✓ **3. 분석 과제 우선순위 선정 매트릭스에 관한 설명 중 가장 적절하지 않은 것은?** 1/1

- (●) 시급성의 판단기준은 전략도 중요도와 비용범위에 따라 난이도는 분석수준과 복잡도 평가로 구분한다. ✓
- ( ) 데이터 분석 과제를 추진할 때 우선 고려해야 하는 요소는 전략도 중요도에 따른 시급성이 가장 중요한 요소이다.
- ( ) 난이도는 해당 기업의 현 상황에 따라 조율할 수 있다.
- ( ) 사분면 영역에서 가장 우선적인 분석 과제 적용이 필요한 영역은 3사분면 영역이다.

의견 보내기

- 시급성의 판단 기준: 전략적 중요도 및 목표가치
- 난이도의 판단 기준: 데이터 획득/저장/가공 비용, 분석 적용 비용, 분석 수준

✓ **4. 분석 마스터플랜 수립 시 우선 순위 고려사항에 해당하지 않는 것은 ?** 1/1

- ( ) 전략적 중요도
- ( ) 비즈니스 성과 및 ROI
- ( ) 실행 용이성
- (●) 데이터 필요 우선 순위 ✓

의견 보내기

분석 마스터 플랜 수립 시 고려 요소
- 우선순위 고려 요소 : 전략적 중요도, ROI(투자자본수익률), 실행 용이성
- 적용 범위/방식 고려 요소 : 업무 내재화 적용 수준, 분석 데이터 적용 수준, 기술 적용 수준

✓ **5. 하향식 접근 방식의 순서로 올바른 것은?** 1/1

(가) 문제탐색 단계 (나) 문제 정의 (다) 해결방안탐색 (라) 타당성 검토

- 나 – 가 – 다 - 라
- 가 – 나 – 다 - 라 ✓
- 가 – 나 – 라 - 다
- 나 – 가 – 라 – 다

의견 보내기

하향식 접근법의 데이터 분석 기획 단계
- Problem Discovery 문제 탐색
- Problem Definition 문제 정의
- Solution Search 해결 방안 탐색
- Feasibility Study 타당성 검토

✓ **6. 다음 중 상향식 접근방식의 특징으로 적절하지 않은 것은?** 1/1

- 상향식 접근 방식의 데이터 분석은 지도학습 방법에 의해 수행된다 ✓
- 문제의 정의 자체가 어려운 경우 사용하는 방식이다
- 디자인 싱킹(Design Thinking)의 발산 단계에 해당한다.
- 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식이다

의견 보내기

상향식 접근 방식
- 문제의 정의 자체가 어려운 경우 상향식 접근 방식 사용
- 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식
- 상향식 접근 방식의 데이터 분석은 비지도학습(Unsupervised Learning) 방법에 의해 수행됨
- 디자인 싱킹(Design Thinking)의 발산 단계에 해당함
- 인사이트 도출 후 반복적인 시행착오를 통해 수정하며 문제를 도출하는 일련의 과정

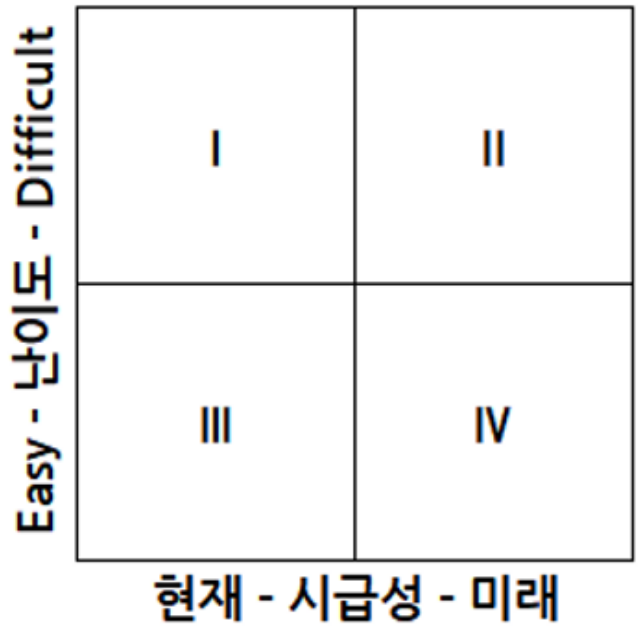
✓ **7. 다음 중 ROI 관점에서의 분석 과제에 대한 우선순위 평가 기준에 대한 설명 중 적절하지 않은 것은?** 1/1

- 분석 난이도는 분석 준비도와 성숙도에 따라 해당 기업의 분석 수준을 파악하고 이를 바탕으로 결정된다
- (●) 시급성이 높고 난이도가 높은 분석 과제는 우선 순위 기준이 높다 ✓
- 난이도에 우선 순위 기준을 놓으면 시급성 높고 난이도 쉬운 과제가 가장 먼저 수행되어야 한다
- 시급성이 높고 난이도가 높은 분석과제는 경영진에 의해 조정 가능하다

의견 보내기

분석 과제 우선순위 평가 기준 : 시급성이 높고 난이도가 낮은 분석 과제의 우선 순위 기준이 높다

✓ **8. 포트폴리오 사분면 분석을 통한 과제 우선순위를 선정하는 기법 중 분석 과제의 적용 우선순위를 '시급성'에 둔다면 결정해야 할 우선순위는?** 1/1



- (●) Ⅲ - Ⅳ - Ⅱ ✓
- ( ) Ⅰ - Ⅱ - Ⅲ
- ( ) Ⅱ - Ⅳ - Ⅰ
- ( ) Ⅲ - Ⅰ - Ⅱ

의견 보내기

우선순위를 '시급성'에 둔다면 Ⅲ - Ⅳ - Ⅱ순서 진행
우선순위를 '난이도'에 둔다면 Ⅲ - Ⅰ - Ⅱ순서 진행

✓ **9. 다음은 어떤 분석 모델 프로세스에 관한 것인가?** 1/1

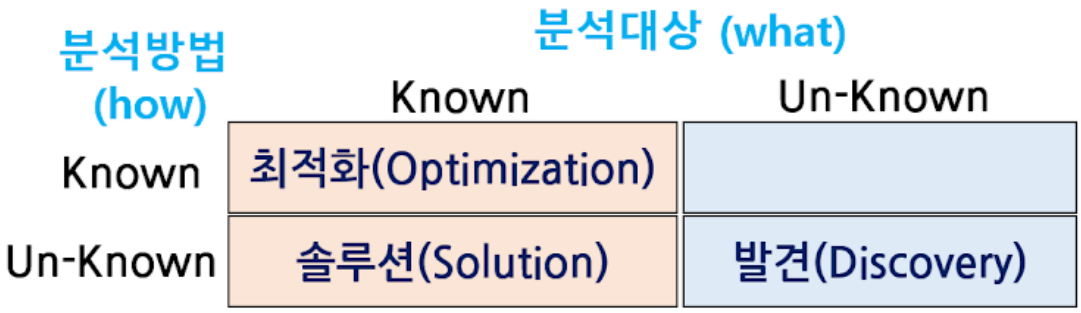
반복을 통하여 점증적으로 개발하는 방법으로써 처음 시도하는 프로젝트에 적용이 용이하지만, 반복에 대한 관리체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있다.

나선형 모델 ✓

의견 보내기

- 폭포수 모델: 단계를 순차적으로 진행하는 방법, 이전 단계가 완료되어야 다음 단계로 순차 진행하는 하향식 진행
- 나선형 모델: 반복을 통해 점증적으로 개발, 반복에 대한 관리 체계가 효과적으로 갖춰지지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있음
- 프로토타입 모델: 사용자 요구사항이나 데이터를 정확히 규정하기 어렵고 데이터 소스도 명확히 파악하기 어려운 상황에서 사용, 일단 분석을 시도해보고 그 결과를 확인해가면서 반복적으로 개선해 나가는 방법, 신속하게 해결책 모형제시, 상향식 접근방법에 활용

✓ **10. 다음 중 빈 칸에 들어갈 올바른 용어는 무엇인가?** 1/1



통찰 ✓

의견 보내기

분석 주제 유형 4가지
- Optimization : 분석 대상 및 분석 방법을 이해하고 현 문제를 최적화의 형태로 수행함
- Solution : 분석 과제는 수행되고, 분석 방법을 알지 못하는 경우 솔루션을 찾는 방식으로 분석 과제를 수행함
- Insight : 분석 대상이 불분명하고, 분석 방법을 알고 있는 경우 인사이트 도출
- Discovery : 분석 대상, 방법을 모른다면 발견을 통해 분석 대상 자체를 새롭게 도출함

✕ **1. Accuracy와 Precision에 대한 설명으로 틀린 것은?** 0/1

- ◯ Accuracy는 실제가 True 인 것 중 예측도 True인 것을 의미한다
- ◯ Precision은 예측이 True 인 것 중 실제도 True인 것을 의미한다
- ◯ Precision은 TP / (TP + FP)으로 구할 수 있다
- ◯ Accuracy는 (TP + TN) / (TP + FP + FN + TN)으로 구할 수 있다

의견 보내기

- Sensitivity/Recall은 실제가 True 인 것 중 예측도 True인 것을 의미한다
- Precision은 예측이 True 인 것 중 실제도 True인 것을 의미한다. 식) TP / (TP + FP)
- Accuracy는 전체 예측 중 옳은 예측의 비율이다. 식) (TP + TN) / (TP + FP + FN + TN)

✓ **2. 확률변수 X가 확률질량함수를 갖는 이산형 확률변수 인 경우 그 기댓값으로 옳은 것은?** 1/1

$$E(x) = \sum xf(x)$$

◉ 옵션 1 ✓

$$E(x) = \int xf(x)$$

○ 옵션 2

$$E(x) = E[(x-\mu)^2]$$

○ 옵션 3

$$E(x) = x^3 - x^2$$

○ 옵션 4

의견 보내기

기댓값
이산형 확률변수 x의 기댓값 : $E(X) = \sum x \cdot f(x)$
연속형 확률변수 x의 기댓값 : $E(X) = \int x \cdot f(x)$

✓ **3. 아래 오분류표를 이용하여 구한 F1 값은 얼마인가?** 1/1

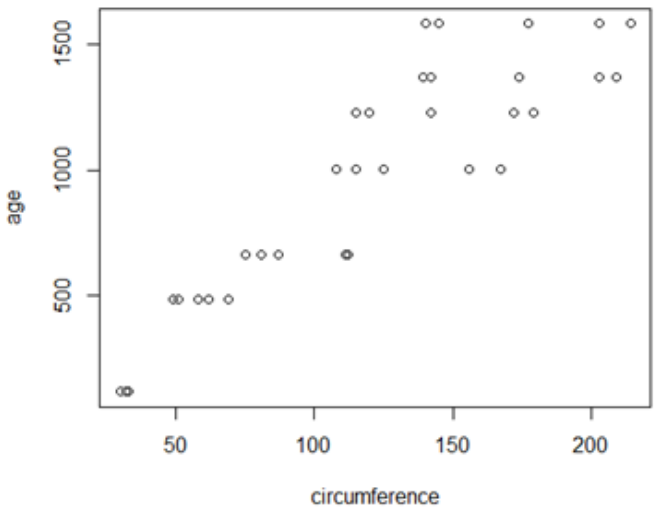
		예측치		합계
		TRUE	FALSE	합계
실제값	TRUE	200	300	500
	FALSE	300	200	500
합계		500	500	1000

- 0.5
- 0.4 ✓
- 0.2
- 0.3

의견 보내기

*F1 = 2 * (Precision * Recall) / (Precision + Recall)*
Precision : TP / (TP + FP) = 200/500 = 0.4
Recall : TP / (TP+ FN) = 200/500 = 0.4
*F1 = 2 * (0.4 * 0.4) / (0.4 + 0.4) = 0.4*

✓ **4. 오렌지 데이터에 대한 산점도 관련 설명으로 틀린 것은?** 1/1



- ○ 양의 상관 관계임을 알 수 있다
- ○ Circumference 데이터는 10 ~ 230 정도의 범위 인 것을 알 수 있다
- ◉ 종별로 관계를 나타낼 수 있다 ✓
- ○ Age 데이터는 10 ~ 1600 정도의 범위인 것을 알 수 있다

의견 보내기

위의 산점도는 종별로 나뉘어 그려지지 않아 산점도를 보고 종별로 나누어진 것은 알 수 없다

✓ **5. 다음 중 K-means 군집의 단점으로 가장 부적절한 것은?** 1/1

- ○ 볼록한 형태가 아닌 군집이 존재하면 성능이 떨어진다
- ○ 사전에 주어진 목적이 없으므로 결과 해석이 어렵다
- ○ 잡음이나 이상값에 영향을 많이 받는다
- ◉ 한 번 군집이 형성되면 군집내 객체들은 다른 군집으로 이동할 수 없다 ✓

의견 보내기

K-means는 비계층적 군집의 종류이며 4 번의 한 번 군집이 형성되면 군집내 객체들은 다른 군집으로 이동 할 수 없는 것은 '계층적 군집 ' 의 대표적 특징이며, K-means는 다른 군집으로 이동할 수 있다.
볼록한 형태(non-convex)가 아닌 군집의 대표적인 것이 초승달 모양이다. 구글을 검색해 보세요!

✓ **6. 다음 설명 중 연관 규칙의 단점이 아닌 것은?** 1/1

- ○ 분석 품목 수가 증가하면 분석 계산이 기하급수적으로 증가한다
- ○ 너무 세분화된 품목을 가지고 연관규칙을 찾으려면 의미 없는 분석 결과가 도출된다
- ◉ 품목 간에 구체적으로 어떠한 영향을 주는지 해석하기 어렵다 ✓
- ○ 상대적 거래량이 적으면 규칙 발견 시 제외되기 쉽다

의견 보내기

연관 규칙의 단점
분석 품목 수가 증가하면 분석 계산이 기하급수적으로 증가함
너무 세분화된 품목을 가지고 연관규칙을 찾으려면 의미 없는 분석 결과가 도출됨
상대적 거래량이 적으면 규칙 발견 시 제외되기 쉬움

연관 규칙의 장점
조건반응(if-then)으로 표현되는 연관 분석의 결과를 이해하기 쉬움
강력한 비목적성 분석 기법이며, 분석 계산이 간편함

✓ **7. 아래 데이터셋 A, B간의 유사성을 유클리드 거리로 계산하면 얼마인가?** 1/1

	키	몸무게
A	165	65
B	170	70

- ○ 50
- ○ 25
- ◉ $\sqrt{50}$ ✓
- ○ 10

의견 보내기

3-94. 계층적 군집의 거리
유클리드 거리는 차이의 제곱의 합에 대한 제곱근이다
$\sqrt{((165-170)2+(65-70)2)} = \sqrt{(25+25)} = \sqrt{50}$

2는 제곱표시로 보아주세요 ^^"

✓ **8. 스피어만 상관계수에서 사용하는 척도는?** 1/1

- ○ 명목척도
- ◉ 서열척도 ✓
- ○ 등간척도
- ○ 비율척도

의견 보내기

스피어만 상관계수
대상자료는 서열척도 사용, 두 변수 간의 비선형적인 관계를 나타낼 수 있음
연속형 외에 이산형도 가능함
스피어만 상관 계수는 원시 데이터가 아니라 각 변수에 대해 순위를 매긴 값을 기반으로 함

✓ **9. SOM은 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도 형태로 형상화 하는 방법이다. 다음 중 SOM 방법에 대한 설명으로 부적절한 것은?** 1/1

- SOM은 입력변수의 위치 관계를 그대로 보존한다는 특징이 있다. 이러한 SOM의 특징으로 인해 입력 변수의 정보와 그들의 관계가 지도상에 그대로 나타난다
- (●) SOM을 이용한 군집분석은 인공신경망의 역전파 알고리즘을 사용함으로써 수행 속도가 빠르고 군집의 성능이 매우 우수하다 ✓
- SOM 알고리즘은 고차원의 데이터를 저차원의 지도 형태로 형상화하기 때문에 시각적으로 이해하기 쉬울 뿐 아니라 변수의 위치 관계를 그대로 보존하기 때문에 실제 데이터가 유사하면 지도상 가깝게 표현된다
- 입력층과 2차원의 격자 형태의 경쟁층으로 이루어져 있다

의견 보내기

SOM vs 신경망
- 신경망은 역전파 알고리즘이지만, SOM은 전방패스를 사용해 속도가 매우 빠르다
- 신경망 모형은 연속적인 layer로 구성된 반면, SOM은 2차원의 그리드(격자)로 구성된다
- 신경망 모형은 에러 수정을 학습 하는 반면 SOM은 경쟁학습 실시한다

✓ **10. 과대적합에 대한 설명으로 가장 부적절한 것은?** 1/1

- (●) 생성된 모델이 훈련 데이터에 너무 최적화되어 학습하여 테스트데이터의 작은 변화에 민감하게 반응하는 경우는 발생하지 않는다 ✓
- 학습데이터가 모집단의 특성을 충분히 설명하지 못할 때 자주 발생한다
- 변수가 너무 많아 모형이 복잡할 때 생긴다
- 옵션과대적합이 발생할 것으로 예상되면 학습을 빠르게 종료하는 방법으로 과대적합을 방지할 수 있다

의견 보내기

과대적합(Overfitting) : 생성된 모델이 훈련 데이터에 너무 최적화되어 학습하여 테스트데이터의 작은 변화에 민감하게 반응하는 경우 발생한다

✓ **11. 아래는 피자와 햄버거의 거래 관계를 나타낸 표로, Pizza/Hamburgers는 피자/햄버거를 포함한 거래수를 의미하고, (Pizza)/(Hamburgers)는 피자/햄버거를 포함하지 않은 거래 수를 의미한다.**
아래 표에서 피자구매에 대해 설명한 것으로 가장 적절한 것은 무엇인가? 1/1

	Pizza	(Pizza)	합계
Hamburgers	2000	500	2500
(Hamburgers)	1000	1500	2500
합계	3000	2000	5000

- ○ 지지도가 0.6으로 전체 구매 중 햄버거와 피자가 같이 구매되는 경향이 높다
- ○ 정확도가 0.8로 햄버거와 피자의 구매 관련성은 높다
- ◉ 향상도가 1보다 크므로 햄버거와 피자는 연관성이 매우 높다. ✓
- ○ 연관규칙 중 “햄버거 -> 피자” 보다 “피자 -> 햄버거”의 신뢰도가 더 높다

의견 보내기

햄버거 -> 피자, 피자 -> 햄버거
지지도 : P(A∩B) = 햄버거 & 피자/ 전체 = 2000/5000 = 0.4
*향상도 : P(B|A)/P(B) = P(A∩B) / (P(A)*P(B))*
*P(햄버거 ∩ 피자) / P(햄버거)*P(피자)*
*0.4 / (0.5 * 0.6) = 0.4/0.3 = 1.333*

신뢰도 : P(B|A) = P(A∩B) / P(A)
햄버거 -> 피자 신뢰도 : 햄버거 & 피자/ 햄버거 = 2000/2500
피자 -> 햄버거 신뢰도 : 햄버거 & 피자/ 피자 = 2000/3000

✓ **12. 앙상블모형(Ensemble)이란 주어진 자료로부터 여러 개의 예측 모형을 만든 후 이러한 예측 모형들을 결합하여 하나의 최종 예측 모형을 만드는 방법을 말한다. 다음 중 앙상블 모형에 대한 설명으로 적절하지 않은 것은?** 1/1

- 배깅은 주어진 자료에서 여러 개의 붓스트랩(Bootstrap) 자료를 생성하고 각 붓스트랩 자료에 예측 모형을 만든 후 결합하여 최종 모형을 만드는 방법이다
- (●) 부스팅은 배깅의 과정과 유사하여 재표본 과정에서 각 자료에 동일한 확률을 부여하여 여러 모형을 만들어 결합하는 방법이다 ✓
- 랜덤 포레스트(Random Forest)는 의사결정나무모형의 특징인 분산이 크다는 점을 고려하여 배깅보다 더 많은 무작위성을 추가한 방법으로 약한 학습기들을 생성하고 이를 선형 결합해 최종 학습기를 만드는 방법이다
- 앙상블 모형은 훈련을 한 뒤 예측을 하는데 사용하므로 교사학습법(Supervised Learning) 이다

의견 보내기

앙상블(Ensemble) 모형 - 부스팅(Boosting)
- 이전 모델의 결과에 따라 다음 모델 표본 추출에서 분류가 잘못된 데이터에 가중치(weight)를 부여하여 표본을 추출함
- 맞추기 어려운 문제를 맞추는데 초점이 맞춰져 있고, 이상치(Outlier)에 약함

✓ **13. 정규분포 신뢰수준 95%일 때 에 대한 설명으로 가장 적절하지 않는 것은?** 1/1

- 표본크기가 커질수록 신뢰구간이 좁아진다. 이는 정보가 많을수록 추정량이 더 정밀하다는 것을 의미한다
- 99% 신뢰수준에 대한 신뢰구간이 95% 신뢰수준에 대한 신뢰구간보다 길다
- 신뢰수준은 모수값이 정해져 있을 때 다수 신뢰구간 중 모수값을 포함하는 신뢰구간이 존재할 확률을 말한다
- (●) 신뢰수준 95% 의미는 추정값이 신뢰구간에 존재할 확률이 95%라 할 수 있다 ✓

의견 보내기

신뢰수준 95% 의미 : n번 반복 추출하여 산정하는 신뢰구간들 중에서 평균적으로 95%는 모수 값을 포함하고 있을 것이라는 의미이다

✓ **14. 회귀방정식의 변수선택 방법에 대한 설명으로 적절하지 않는 것은?** 1/1

- (●) 전진선택법은 변수를 추가해도 영향을 받지 않는다 ✓
- ( ) 후진제거법은 독립변수 후보 모두를 포함한 모형에서 시작한다
- ( ) 단계별 선택법은 기준 통계치에 가장 도움이 되지 않는 변수를 삭제하거나, 모델에서 빠져 있는 변수 중에서 기준 통계치를 가장 개선 시키는 변수를 추가한다
- ( ) 회귀모델에서 변수 선택을 위한 판단 기준에는 Cp, AIC, BIC 등이 있으며 값이 작을수록 좋다

의견 보내기

- 전진선택법 : Forward Selection, 절편만 있는 모델에서 출발해 기준 통계치를 가장 많이 개선시키는 변수를 차례로 추가하는 방법이다.
- 후진제거법 : Backward Elimination, 독립변수 후보 모두를 포함한 모형에서 출발해 제곱합의 기준으로 가장적은 영향을 주는 변수로부터 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하고, 이때 모형을 선택하는 방법이다.

✓ **15. 주성분분석에 대한 설명 중 적절하지 않은 것은?** 1/1

```
> data_1 <- prcomp(data, scale=TRUE)
> data_1
Standard deviations (1, .., p=4):
[1] 1.4154072  1.3086525  0.4377899  0.3039594

Rotation (n * k) = (4 * 4)
                         PC1        PC2        PC3        PC4
               x1  0.2388128 -0.6895993  0.5325178  0.4287728
               x2  0.4604720 -0.5393126 -0.5603653 -0.4278997
               x3  0.6038420  0.3514805 -0.3277028  0.6359616
               x4  0.6052345  0.3317472  0.5431634 -0.4781303

> summary(data_1)
Importance of components:
                                 PC1        PC2        PC3        PC4
Standard deviation           1.4154     1.3087    0.43779    0.3040
Proportion of Variance       0.5008     0.4281    0.04791    0.0231
Cumulative Proportion        0.5008     0.9290    0.97690    1.0000
```

- ○ 제 3변수까지 사용하면 97.69%의 누적 비율을 갖게 된다.
- ○ 제 2변수는 42.81%의 분산 비율을 갖는다
- ○ 변수들의 scale이 많이 다른 경우 특정 변수가 전체적인 경향을 좌우하기 때문에 상관계수 행렬을 사용하여 분석하는 것이 좋다.
- ◉ PC2의 로딩벡터는 모두 양의 방향을 가지고 있다 ✓

의견 보내기

PC2의 로딩벡터 중에는 음수가 있으며 이것은 음의 방향을 의미한다.

✓ **16. 표본추출시 발생하는 오차에 관한 설명 중 잘못된 설명은?** 1/1

- 표본 오차(Sampling Error)는 모집단의 일부인 표본에서 얻은 자료를 통해 모집단 전체의 특성을 추론함으로써 생기는 오차를 의미한다
- 비표본 오차(non-sampling error)는 표본크기가 증가함에 따라 증가한다
- (●) 표본 편의(Sampling Bias)는 표본추출방법에서 기인하는 오차를 의미하고, 표본추출 방법에 의해 최소화하거나 없앨 수 있다 ✓
- 표본 오차는 표본의 크기를 증가시키고, 표본 선택 방법을 엄격히 하여 줄일 수 있다

의견 보내기

표본 편의(Sampling Bias) : 표본추출방법에서 기인하는 오차를 의미하며, 확률화를 통해 최소화하거나 없앨 수 있다

✓ **17. 양성 나온 사람 중에 실제 질병이 있는 사람의 확률은 무엇인가? 이때, 양성인 사람은 0.2, 실제 질병이 있는 사람은 0.1, 검사 결과 양성인 사람은 0.9 이다** 1/1

- 0.09
- (●) 0.45 ✓
- 0.18
- 0.5

의견 보내기

베이지안 확률
*P(질병|양성) = P(양성|질병) * P(질병) / P(양성) = (0.9 * 0.1) / 0.2 = 0.45*

✓ **18. 다음 표를 사용한 재현율에 관련된 설명으로 틀린 것은?** 1/1

Confusion Matrix		Predict	
		FALSE	TRUE
Actual	FALSE	30	70
	TRUE	10	40

- ○ 재현율(Recall)을 민감도(Sensitivity)라고도 한다
- ◉ 재현율은 3/10 이다 ✓
- ○ 재현율은 TP/(TP+FN) 이다
- ○ 재현율과 정밀도(Precision)을 사용해 F1Score를 구한다

의견 보내기

재현율(Recall)
- Sensitivity라고도 하며, 실제 True인 것 중 예측도 True로 된 것의 비율을 말한다
- Recall : TP/(TP+FN) 로 표에서는 40/50 = 4/5 이다
- Precision : TP / (TP + FP)
*- F1 Score : 2 * (Precision * Recall) / (Precision + Recall)*

✓ **19. 주성분 분석에 대한 설명 중 올바른 것은?** 1/1

- ○ 독립변수들과 주성분과의 거리인 '정보손실량'을 최대화하거나 분산을 최소화 한다
- ○ 상관관계가 있는 변수들을 선형 결합에 의해 상관관계가 있는 새로운 변수(주성분)를 만들고 분산을 최소화하는 변수로 축약한다
- ◉ 여러 개의 양적변수(Quantitative variable)들 사이의 분산-공분산 관계를 이용하여 변수들의 선형결합(linear combination)으로 표현하는 기법이다 ✓
- ○ 정규화 전후의 주성분 결과는 동일하다

의견 보내기

주성분 분석
-상관관계가 있는 변수들을 선형 결합에 의해 상관관계가 없는 새로운 변수(주성분)를 만들고 분산을 극대화하는 변수로 축약함
- 주성분은 변수들의 선형결합으로 이루어져 있음
- 독립변수들과 주성분과의 거리인 '정보손실량'을 최소화하거나 분산을 최대화 함
- 주성분 분석은 척도에 영향을 받으므로 정규화 전후의 주성분 결과는 다르다

✓ **20. 주성분 분석에서 주성분 수를 선택할 때 고려하지 않아도 되는 것은?** 1/1

- ○ Scree Plot
- ◉ 개별 고윳값의 분해 가능 여부 ✓
- ○ 성분들이 설명하는 분산의 비율
- ○ 고윳값(Eigenvalue)

의견 보내기

주성분 결정 기준
- 성분들이 설명하는 분산의 비율 : 누적 분산 비율이 70~90% 사이가 되는 주성분 개수 선택
- 고윳값(Eigenvalue) : 분산의 크기를 나타내며, 고윳값이 1보다 큰 주성분만 사용함
- Scree Plot : 고윳값을 가장 큰 값에서 가장 작은 값을 순서로 정렬해 보여줌 (1보다 큰 값 사용)

✓ **21. 분해 시계열 분석에 대한 설명 중 옳지 않은 것은?** 1/1

- ○ 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법이다
- ○ 추세요인은 자료의 그림을 그렸을 때 그 형태가 오르거나 내리는 등 자료가 어떤 특정한 형태를 취할 때이다
- ○ 계절요인은 고정된 주기에 따라 자료가 변화하는 경우이다
- ◉ 이동평균법은 최근 관측치에 더 높은 가중치를 부여하여 이동 평균을 계산하는 방법이다 ✓

의견 보내기

- 분해시계열 분해요인 : 추세, 계절(고정된 주기), 순환(알려지지 않은 주기), 불규칙 (오차에 해당하는 요인)
- 이동평균법 : 일정 기간의 관측치에 모두 동일 가중치를 부여하여 이동 평균을 계산하는 방법으로 계절 성분과 불규칙 성분을 제거함
- 지수평활법 :전체 시계열 자료를 이용하여 평균을 구하고, 최근 시계열에 더 큰 가중치를 적용하는 방법, 지수평활법을 사용하여 얻은 예측값은 과거 관측값의 가중평균(weighted average)

✓ **22. 의사결정나무모형에 관한 내용으로 적절하지 않은 것은?** 1/1

- ○ 의사결정나무의 목적은 새로운 데이터를 분류(classification)하거나 해당 범주의 값을 예측(Prediction)하는 것이다
- ○ 목표변수 유형에 따라 범주형 분류나무(Classification Tree)와 연속형 회귀나무(Regression Tree)로 분류된다
- ◉ 분리 변수의 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받지 않는다 ✓
- ○ 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나간다.

의견 보내기

의사결정나무모형에서 분리 변수 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받는다

✓ **23. 다음 회귀분석에 결과에 대한 설명으로 틀린 것은 무엇인가?** 1/1

```
Call:
lm(formula = wage ~ age + jobclass, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-107.234  -24.751   -6.311   16.308  197.278

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             76.6298     2.8320   27.06   <2e-16 ***
age                      0.6447     0.0638   10.11   <2e-16 ***
jobclass2. Information  15.9214     1.4732   10.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.16 on 2997 degrees of freedom
Multiple R-squared:  0.07435,   Adjusted R-squared:  0.07373
F-statistic: 120.4 on 2 and 2997 DF,  p-value: < 2.2e-16
```

- (●) age 변수는 wage에 대해 유의하지 않다 ✓
- ( ) 종속변수는 wage 이다
- ( ) jobclass는 범주형 변수이다
- ( ) 데이터 개수가 3000개 이다

의견 보내기

age 변수의 P-value가 0.05 보다 작은 값이므로 wage 변수에 대해 유의하다고 판단된다.

✓ **24. 다음 Orange 나무에 대한 나이 및 둘레에 대한 분석 결과로 옳지 않은 것은?** 1/1

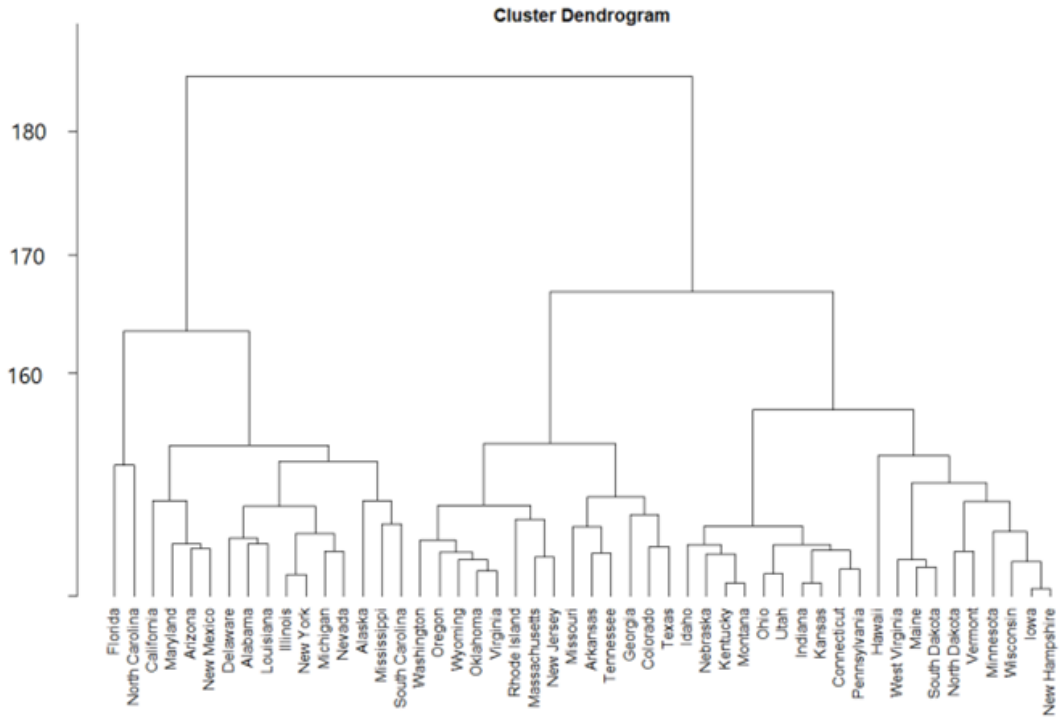
```
> summary(Orange)
 Tree       age         circumference
 3:7   Min.   : 118.0   Min.   : 30.0
 1:7   1st Qu.: 484.0   1st Qu.: 65.5
 5:7   Median :1004.0   Median :115.0
 2:7   Mean   : 922.1   Mean   :115.9
 4:7   3rd Qu.:1372.0   3rd Qu.:161.5
       Max.   :1582.0   Max.   :214.0
```

- ○ circumference의 Median은 115이다
- ◉ 나무 age와 circumference가 유의한 관계를 가진다. ✓
- ○ age의 IQR은 888이다
- ○ Tree의 종류는 5가지이며 각 종류당 7개의 sample이 존재한다

의견 보내기

통계적인 유의미를 검정하기 위해서는 F통계량과 같은 검정 통계량을 구하여 검정해야 한다

✓ **25. 다음 덴드로그램에서 height가 160일 때의 군집 개수는? (숫자 하나만 입력하세요)** 1/1



4 ✓

✓ **26. 귀무가설이 실제로 사실이어서 채택하여야 함에도 불구하고 이를 기각하는 오류를 무엇이라 하는가?** 1/1

제1종 오류 ✓

의견 보내기

제 1종 오류 : 귀무가설이 참인데 기각하게 되는 오류
제 2종 오류 : 귀무가설이 거짓인데 채택하는 오류

✗ **27. 신경망 모형에서 표준화 지수함수로 불리며, 출력 값 z가 여러 개로 주어지고, 목표치가 다 범주인 경우 각 범주에 속할 사후확률을 제공하여 출력 노드에 주로 사용되는 함수는 무엇인가?** ···/1

softmax 함수 ✗

정답

Softmax

softmax

소프트맥스

의견 보내기

- Sigmoid : 연속형 0~1, Logistic 함수라 불리기도 함, 선형적인 멀티-퍼셉트론에서 비선형 값을 얻기 위해 사용
- Softmax : 모든 logits의 합이 1이 되도록 output을 정규화, sigmoid 함수의 일반화된 형태로 결과가 다 범주인 경우 각 범주에 속할 사후 확률을(posterior probability) 제공하는 활성화 함수

✓ **28. 오분류표 용어 중 실제로 False일 때 예측이 적중하는 경우를 무엇이라고 하는가?** 1/1

특이도 ✓

의견 보내기

- 특이도(Specificity) = TN / (TN + FP)
- 실제로 False 인 것들 중 예측이 False 로 된 경우의 비율

✓ **29. 은닉층이 다층인 신경망을 학습하다 보면 역전파 과정에서 초기 부분의 입력층으로 갈수록 기울기 변화가 점차적으로 작아지는 현상은?** 1/1

기울기 소실 ✓

의견 보내기

기울기 소실은 다층신경망에서 은닉층이 많아 인공신경망 기울기 값을 베이스로 하는 역전파 알고리즘으로 학습시키려고 할 때 발생하는 문제이다.

✓ **30. 로지스틱 회귀모형에서 exp(x1)의 의미는 나머지 변수가 주어질 때 x1이 한 단위 증가할 때마다. 성공(Y=1)의 ( )가 몇 배 증가 하는지를 나타낸다. ( )에 들어가는 용어는?** 1/1

오즈 ✓