1 과목 | 빅데이터 분석 기획

01 데이터 수집과 관련된 표준 용어로 추출, 변환, 적재의 과정으로 구성된 기술로 올바른 것은?

① ETL
② Sensor Network
③ Crawling
④ Open API

02 딥러닝에 대한 설명으로 적절하지 않은 것은?

① Dropout은 과적합을 방지하기 위한 방법으로 데이터 학습 과정에서 유닛의 일부를 랜덤하게 누락시킨다.
② 딥러닝은 인공신경망을 사용하므로 각 hidden layer의 가중치를 통해 모형의 결과를 해석하기 쉽다.
③ 딥러닝 분석 수행 시 주로 sigmoid function을 Activation으로 사용한다.
④ 최적의 학습 결과를 찾기 위해 역방향으로 오차를 전파하면서 각 layer의 가중치를 갱신하는 오류역전파 알고리즘을 사용한다.

03 다음 중 빅데이터 분석 방법론의 개발 절차로 올바른 것은?

① 데이터 준비 – 분석 기획 – 데이터 분석 – 평가 및 전개 – 시스템 구현
② 분석 기획 – 데이터 준비 – 데이터 분석 – 평가 및 전개 – 시스템 구현
③ 분석 기획 – 데이터 준비 – 데이터 분석 – 시스템 구현 – 평가 및 전개
④ 데이터 준비 – 분석 기획 – 데이터 분석 – 시스템 구현 – 평가 및 전개

04 기존의 데이터를 학습시켜 새로운 데이터 입력 시 이를 예측하는 방법으로 분류나 회귀 문제에 적합한 것은?

① 강화학습
② 지도학습
③ 준지도학습
④ 비지도학습

05 개인정보 비식별 조치에 대한 익명성 검증 방법으로 적절하지 않은 것은?

① l-다양성은 민감한 정보의 분포를 낮추어 추론 가능성을 더욱 낮추는 기법이다.
② k-익명성은 특정인임을 추론할 수 있는지 여부를 검토, 일정 확률수준 이상 비식별되도록 하는 기법이다.
③ m-유일성은 원본 데이터와 동일한 속성 값의 조합이 비식별 결과 데이터에 최소 m개 존재해야 재식별 위험성이 낮다는 것이다.
④ t-근접성은 전체 데이터 집합의 정보 분포와 특정 정보의 분포 차이를 t 이하로 하여 추론을 방지한다.

06 개인정보 비식별화 방법으로 적절하지 않은 것은?

① 가명 처리
② 총계 처리
③ 데이터 범주화
④ 데이터 암호화

07 데이터의 기초 통계량과 분포를 확인하여 데이터를 이해하고 의미 있는 관계를 찾아내는 방법으로 올바른 것은?

① 기술통계
② 가설검정
③ 탐색적 데이터 분석
④ 데이터 시각화

08 분석 대상이 명확하지 않으나 분석 방법은 알고 있을 때 적용할 수 있는 문제 해결 방법으로 올바른 것은?

① 발견(Discovery)
② 솔루션(Solution)
③ 최적화(Optimization)
④ 통찰(Insight)

09 개인정보에 대하여 정보주체의 동의 없이 수집 및 이용이 가능한 경우로 적절하지 않은 것은?

① 학교에서 신임 교원 임용 시 후보자에 대한 범죄 이력 등을 조회할 수 있다.
② 병원에서 진료기록부 작성을 위해 개인정보를 기입하는 경우 가능하다.
③ 정보주체의 생명이나 신체 또는 재산상의 이익을 위하여 필요하다고 인정되는 상황으로 사전 동의를 구할 수 없을 만큼 급박한 경우 가능하다.
④ 통신사에서 고객에게 요금을 부과하기 위하여 조회하는 경우 가능하다.

10 정형 데이터의 품질 진단 방법으로 적절하지 않은 것은?

① 부가요소 정확성 분석
② 메타데이터 수집 및 분석
③ 칼럼 속성 분석
④ 값의 허용 범위 분석

11 탐색적 데이터 분석(EDA)에 대한 설명으로 적절하지 않은 것은?

① 데이터에 대한 이해 및 의미 있는 관계를 찾아낸다.
② 시각화 도구를 이용하여 데이터를 직관적으로 파악할 수 있다.
③ 분석을 위한 후보 모형들을 선정하는 과정이다.
④ 데이터를 다양한 관점으로 파악하는 과정이다.

12 데이터 분석 절차에서 복잡한 문제의 단순화를 통해 문제를 변수들 간의 관계로 정의하는 것으로 올바른 것은?

① EDA
② 문제 인식
③ 연구조사
④ 모형화

13 진단 분석에 대한 설명으로 올바른 것은?

① 원인은 무엇인지 파악하는 것이다.
② 앞으로 어떻게 될 것인지 파악하는 것이다.
③ 어떻게 대처해야 하는지 파악하는 것이다.
④ 무엇이 발생했는지 파악하는 것이다.

14 통계적 데이터 분석 시 추정치가 편파성을 일으키는 문제나 추정치의 타당도 문제가 발생할 수 있는 값으로 올바른 것은?

① 편차(deviation)
② 분산(variance)
③ 이상치(outlier)
④ 편향(bias)

15 데이터 유형별 데이터 수집 방법으로 적절하지 않은 것은?

① 센서데이터 : 센싱(sencing)
② 동영상 : 스트리밍(streaming)
③ DBMS : FTP
④ 웹 : 크롤링(Crawling)

16 데이터 분석 성숙도 모델의 성숙도 수준으로 적절하지 않은 것은?

① 도입단계
② 최적화단계
③ 확산단계
④ 파악단계

17 개인정보의 수집 시 정보주체에게 사전 고지하지 않아도 되는 항목으로 올바른 것은?

① 파기하는 내용
② 보유 및 이용 기간
③ 동의를 거부할 권리가 있다는 사실
④ 수집 및 이용 목적

18 상향식 접근 방식에 대한 설명으로 올바른 것은?

① 데이터를 활용하여 생각지도 못했던 인사이트 도출 및 시행착오를 통한 개선이 가능하다.
② 전통적 분석 과제 발굴 방식으로 근래의 문제들은 변화가 심하여 문제를 사전에 정확하게 정의하기 어렵다.
③ 비즈니스 모델 기반 문제 탐색, 외부 참조 모델 기반 문제 탐색, 분석 유즈케이스 정의를 통한 문제 탐색이 가능하다.
④ 동적인 환경에서 발산과 수렴 단계를 반복적으로 수행하며 상호 보완을 통해 분석의 가치를 극대화할 수 있다.

19 실세계에 존재하는 객체의 표현 값이 정확히 반영되어야 한다는 것을 뜻하는 품질 기준으로 올바른 것은?

① 유효성
② 일관성
③ 정확성
④ 무결성

20 전사 차원의 모든 데이터에 대하여 표준화된 관리 체계를 수립하는 것을 나타내는 용어로 올바른 것은?

① 데이터 아키텍처
② 데이터 컴플라이언스
③ 데이터 표준화
④ 데이터 거버넌스

21 박스 플롯을 통해서 알 수 없는 것은?

① 1사분위수
② 분산
③ 이상값
④ 최댓값

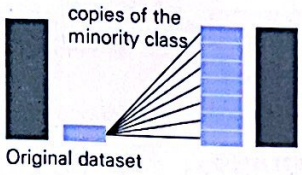
22 다음은 변수선택의 방법 중 단계적 선택법에 대한 설명이다. 잘못된 것은?

① 전진 선택법과 후진 선택법의 보완방법이다.
② 전진 선택법을 통해 가장 유의한 변수를 모형에 포함한다.
③ 나머지 변수들에 대해 후진 선택법을 적용하나 새롭게 유의하지 않은 변수들을 제거하지는 않는다.
④ 제거된 변수는 다시 모형에 포함하지 않으며 유의한 설명변수가 존재하지 않을 때까지 과정을 반복한다.

23 파생변수에 대한 설명으로 틀린 것은?

① 기존의 변수를 조합하여 새로운 변수를 만들어내는 것을 의미한다.
② 사용자가 특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여하는 변수로 매우 주관적일 수 있으므로 논리적 타당성을 갖출 필요가 있다.
③ 데이터의 특성을 파악하는 데 중점을 두어 특정상황에 유의미하도록 변수를 생성해야 한다.
④ 세분화, 고객행동 예측 등에 유용하게 사용된다.

24 다음은 어떤 학습 데이터 불균형에 대한 처리 방법이다. 옳은 것을 고르시오.



소수클래스의 복사본을 만들어, 대표클래스의 수만큼 데이터를 만들어 주는 것이다. 똑같은 데이터를 그대로 복사하는 것이기 때문에 새로운 데이터는 기존 데이터와 같은 성질을 갖게 된다.

① 언더샘플링(Undersampling)
② 오버샘플링(Oversampling)
③ 음수 미포함 행렬분해(NMF: Non-negative Matrix Factorization)
④ 특이값분해 (Singular Value Decomposition)

25 하나의 제품을 A, B, C 공장에서 각각 50%, 30%, 20%씩 물량을 나누어 생산하며 불량률은 1%, 2%, 3% 이라고 한다. 생산된 제품 중 하나를 선택했을 때 불량품이면 그 제품이 A공장에서 나왔을 확률은?

① $\frac{6}{17}$ ② $\frac{5}{17}$
③ $\frac{12}{17}$ ④ $\frac{11}{17}$

26 통계학과 학생들 100명을 대상으로 기말고사 시험의 결과가 평균이 80, 분산이 100인 정규분포를 보인다고 한다. 수강생중에서 어떤 학생이 80점에서 85점사이 점수를 받을 확률은 얼마인가? (단, P(Z≤0.5)=0.6915, P(Z≤0.0)=0.5000)

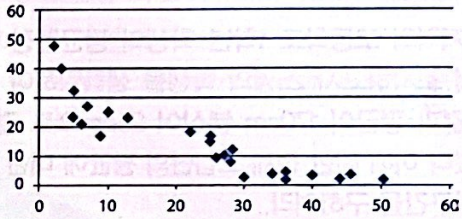
① 0.3457 ② 0.6915
③ 0.1915 ④ 0.7230

27 다음 확률함수에 대해서 최대가 되는 모수 θ 값은 얼마인가?

주어진 데이터 3,1,2,3,3에 대해서
$f(t;\theta)=\theta e^{-\theta t}$
여기서 $t \geq 0$ 이다.

① $\frac{5}{12}$ ② $\frac{1}{12}$
③ 1 ④ $\frac{1}{13}$

28 다음 아래의 산점도 그래프의 개형과 맞는 피어슨 상관계수 유형은 어느 것인가?



① $\rho = 0$
② $\rho = 1$
③ $-1 < \rho < 0$
④ $0 < \rho < 1$

29 다음 아래 설명은 어떤 분석에 대한 것인가?

- 자료의 값 대신 순위를 이용하는 경우의 상관계수로서, 데이터를 작은 것부터 차례로 순위를 매겨 서열순서로 바꾼 뒤 순위를 이용해 상관계수를 구한다.
- 두 변수 간의 연관 관계가 있는지 없는지를 밝혀 주며 자료에 이상점이 있거나 표본크기가 작을 때 유용하다.

① 피어슨 상관계수
② 스피어만 상관계수
③ 크론바흐 알파 계수 신뢰도
④ 단조상관계수

30 다음은 데이터의 시각화에 대한 설명이다. 아래 설명에 해당되는 차트는 무엇인가?

- 하나의 공간에 각각의 변수를 표현하는 몇 개의 축을 그려서 축에 해당되는 변수값을 연결하는 그래프이다.
- 각 변수마다 축시작점은 최소, 가장 먼 점은 최대값을 나타낸다.
- 연결되는 선의 모양이나 색을 다르게 하는 경우 여러 속성을 한번에 표현이 가능하다.

① 버블차트
② 스타차트
③ 히트맵
④ 산점도

31 한 지역의 고등학교 1학년 학생의 평균신장에 대해서 조사하고자 25명의 학생을 샘플링하여 조사한 결과 평균이 170cm 분산이 25cm으로 결과가 나왔다 이에 대한 전체 모집단의 평균에 대한 95% 신뢰구간을 구하여라.

① $169 \leq \mu \leq 171$
② $166.818 \leq \mu \leq 173.182$
③ $167.936 \leq \mu \leq 172.064$
④ $164.959 \leq \mu \leq 175.041$

32 기댓값을 나타내는 다음의 두 추정량이 있다.

$\hat{\theta}_1 = \frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3 + \frac{1}{4}X_4$, $\hat{\theta}_2 = \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3$
(단, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$)

다음 중 옳은 것은?

① 둘 다 불편추정량으로 $E(\hat{\theta}_1) = \mu$, $E(\hat{\theta}_2) = \mu/4$, $Var(\hat{\theta}_1) = \sigma^2$, $Var(\hat{\theta}_2) = \sigma^2/16$이고 분산의 효율성도 동일하다.
② 둘 다 불편추정량으로 $E(\hat{\theta}_1) = \mu$, $E(\hat{\theta}_2) = \mu$, $Var(\hat{\theta}_1) = \sigma^2$, $Var(\hat{\theta}_2) = \sigma^2/16$이고 $\hat{\theta}_1$이 $\hat{\theta}_2$보다 더 효율적이라고 말할 수 있다.
③ 둘 다 불편추정량으로 $E(\hat{\theta}_1) = \mu$, $E(\hat{\theta}_2) = \mu$, $Var(\hat{\theta}_1) = \sigma^2/4$, $Var(\hat{\theta}_2) = 3\sigma^2/8$이고 $\hat{\theta}_1$이 $\hat{\theta}_2$보다 더 효율적이라고 말할 수 있다.
④ 둘 다 불편추정량으로 $E(\hat{\theta}_1) = \mu$, $E(\hat{\theta}_2) = \mu/4$, $Var(\hat{\theta}_1) = \sigma^2/4$, $Var(\hat{\theta}_2) = 3\sigma^2/8$이고 $\hat{\theta}_1$이 $\hat{\theta}_2$보다 더 효율적이라고 말할 수 있다.

33 다음은 가설 검정의 결과로 채택 여부를 결정시에 관한 표이다 빈칸에 들어갈 내용으로 옳은 것은?

검정결과 \ 실제상황	H_0 귀무가설	H_1 대립가설
H_0 귀무가설 채택	success	(a)
H_0 귀무가설 기각	(b)	success

① (a) 제1종 오류, (b) 제2종 오류
② (a) 제2종 오류, (b) 제1종 오류
③ (a) 제1종 오류, (b) 제1종 오류
④ (a) 제2종 오류, (b) 제2종 오류

34 다음은 차원축소에 관한 설명이다. 틀린 것은?

① 복잡도의 축소(Reduce Complexity)에서 동일한 품질을 나타낼 수 있다면 효율성 측면에서 데이터 종류의 수를 줄여야 한다.
② 차원의 증가는 분석모델 파자미터의 증가 및 파자미터 간의 복잡한 관계의 증가로 분석결과의 과적합 발생의 가능성이 커진다.
③ 해석력(Interpretability)의 확보 측면에서 차원이 작은 간단한 분석모델일수록 내부구조 이해가 용이하고 해석이 쉬워진다.
④ 차원의 저주란 데이터분석 및 알고리즘을 통한 학습을 위해 차원이 증가하면서 학습 데이터의 증가를 수반하여 계산성능이 저하되는 현상을 말한다.

35 프로스포츠의 선수들의 연봉에 대한 분석 시 팀 전체의 연봉의 50% 이상을 소수의 선수들이 차지하는 경우가 많다. 이 경우 중심성 경향의 분석 시 용이한 통계량은 무엇인가?

① 평균
② 최빈값
③ 중앙값
④ 분산

36 다음은 층화 추출에서 각 층별로 표본을 배정하는데 있어서 한가지 방법을 설명한 것이다. 해당하는 표본배정법은?

> 추정량의 분산을 최소화 시키거나 주어진 분산의 범위 하에서 비용을 최소화 시키는 표본 배정 방법

① 비례 배분법 ② 고정 배분법
③ 네이만 배분법 ④ 최적 배분법

37 군집 불균형을 해결하는 방법에 대한 설명으로 틀린 것은?

① 가중치균형법을 이용하여 데이터 클래스의 균형이 필요한 경우로 각 클래스별 특정 비율로 가중치(Weight)를 주어서 불균형을 해결한다.
② 대표클래스(Majority Class)의 일부만을 선택하고, 소수클래스(Minority Class)는 최대한 많은 데이터를 사용하는 방법인 언더샘플링(Under Sampling)으로 해결한다.
③ 소수클래스의 복사본을 만들어, 대표클래스의 수만큼 데이터를 만들어 데이터를 추가하여 불균형을 해결하는 오버샘플링(Over Sampling)이 있다.
④ 데이터에 대한 임계값을 설정하여 임계값을 조절하면서 데이터를 선택하여 불균형을 해소한다.

38 모집단과 표본의 통계량에 대한 설명 중 틀린 것은?

① 표본분포의 평균은 모집단의 평균 μ와 동일하다.
② 모집단의 표준편차가 σ이면 표본분포의 표준편차는 $\sigma/\sqrt{n}$이라고 정의한다. 특히 표본평균의 표본분포는 $N(\mu,\sigma^2/n)$인 정규분포를 따른다.
③ 모집단의 크기가 무한대에 한해서 표본평균의 표준오차는 $\sigma/\sqrt{n}$로 정의한다.
④ 동일한 모집단의 표준편차에서 표본의 크기가 커지면 커질수록 표준오차는 늘어나는 경향이 있다.

39 다음 중 성격이 다른 분포는?

① 지수분포
② 정규분포
③ 이항분포
④ F-분포

40 모집단이 정규분포를 따를 때 표본크기에 따른 표본분포에 관한 내용으로 틀린 것은?

① 표본의 크기가 30이상이면 표본은 정규분포를 따른다.
② 표본의 크기와 상관없이 정규분포를 따른다.
③ 표본의 크기가 30미만이면 표본은 T 분포를 따른다.
④ 표본의 크기가 커질수록 표준오차는 줄어든다.

3과목 | 빅데이터 모델링

41 전체 독립변수 중에서 종속변수와의 상관관계가 적은 변수를 점진적으로 분석모형에서 제외하는 방법은?

① 후진소거
② 전진선택
③ 차원축소
④ 주성분 분석

42 딥러닝과 관련된 설명으로 틀린 것은?

① 드롭아웃 : 신경망에서 은닉층의 뉴런을 임의로 삭제하면서 학습한다.
② 오차역전파 : 오차를 입력층에서 출력층으로 전달 연쇄법칙을 통해 가중치와 편향을 업데이트한다.
③ 활성화 함수 : 입력신호의 총합을 출력신호로 변환한다.
④ 손실 함수 : 신경망이 출력한 값과 실제 값과의 오차에 대한 함수이다.

43 입력층이 (5,5), 필터가 (3,3)이며 스트라이드(stride)는 1, 패딩(padding)이 0인 값의 특징맵(Feature Map)의 크기는?

① (3, 3) ② (4, 4)
③ (5, 5) ④ (6, 6)

44 회귀분석의 진단과 관련하여 틀린 설명은?

① 선형성 : 독립변수와 종속변수가 선형적이어야 한다.
② 잔차 정규성 : 잔차의 기댓값은 0이며 정규분포를 이루어야 한다.
③ 잔차 독립성 : 잔차들은 서로 독립적이어야 한다.
④ 다중 공선성 : 다중 회귀 분석을 수행할 경우 2개 이상의 독립변수 간에 상관관계로 인한 문제가 없어야 한다.

45 SVM의 특징으로 잘못된 설명은?

① 분류, 회귀, 특이점 판별에 활용되는 지도학습 기법이다.
② 데이터가 많은 경우에도 학습 처리속도가 빠르다.
③ 선형 또는 비선형 분류가 가능하다.
④ 예측 정확도가 높은 편이다.

46 다차원 척도법과 거리가 먼 키워드는?

① 근접성 ② 유사성
③ 시각화 ④ 연속성

47 비용함수(손실함수)에 L1-norm(규제항)을 더한 규제 이름은?

① Lasso
② Ridge
③ ShrinkageNet
④ ElasticNet

48 SVM의 주요 요소로 맞지 않는 것은?

① 초평면 ② 특징맵
③ 마진 ④ 서포트벡터

49 독립변수가 연속형이면서 종속변수가 범주형인 조건을 가진 분석기법은?

① 로지스틱 회귀 ② 선형 회귀
③ 시계열 분석 ④ 나이브 베이지안

50 다음 분류 모델 해석에서 맞는 설명은?

		실제값	
		일반인	암환자
예측값	일반인	60	0
	암환자	10	30

① 재현율(Recall)은 0.75이다.
② 정확도(Accuracy)는 0.9이다.
③ 정확도(Accuracy)가 높을수록 좋은 모델이라고 할 수 있다.
④ 정밀도(Precision)는 0.5이다.

51 앙상블 분석에서 기법과 알고리즘이 잘못 기술된 것은?

① 배깅: 부트스트랩
② 배깅: 랜덤포레스트
③ 부스팅: GBM
④ 배깅: Adaboost

52 $P(A)$, $P(B)$, $P(C)$, $P(x|A)$, $P(x|B)$, $P(x|C)$를 이용해서 $P(B|x)$를 나타낸 것은?

① $P(B|x) = \frac{P(x|B)}{P(B)}$

② $P(B|x) = \frac{P(x|B)P(B)}{P(x|A)P(A)+P(x|B)P(B)+P(x|C)P(C)}$

③ $P(B|x) = \frac{P(x|B)P(B)}{P(x|A)P(x)+P(x|B)P(x)+P(x|C)P(x)}$

④ $P(B|x) = \frac{P(x|B)P(B)}{P(x|A)+P(x|B)+P(x|C)}$

53 교차검증에서 전체 데이터를 학습 데이터와 테스트 데이터, 검증 데이터로 나누는 기법은?

① k-폴드 교차검증
② Holdout 교차검증
③ 계층별 k-폴드 교차검증
④ 셔플링 교차검증

54 다음 중 비지도학습 적용에 적합한 경우는?

① 상품 구매 패턴분석
② SNS 기반 선호 브랜드 그룹 분석
③ 실시간 스팸 메일 분류
④ CCTV 통한 얼굴 자동 인식

55 다음 중 지도학습 분류 분석 예시에 해당되는 것은?

① 유동인구에 따른 절도 범죄율 관계 분석
② 마케팅 캠페인 집행 후 매출액 추이 변화 분석
③ 전염병 확진자 수에 따른 마스크 판매량 추이 분석
④ 색상비율에 따라 사람들이 느끼는 감정변화 분석

56 한 놀이공원에서 고객들로부터 다양한 놀이기구에 대한 선호도를 조사하여 놀이기구별로 주제 테마파크를 재구성하려고 한다 이럴 때 사용되는 분석기법으로 타당한 것은?

① 군집분석
② 다층판별분석
③ 요인분석
④ 분산분석

57 시계열 모형에 대해서 설명한 것 중 옳은 것은?

① 백색잡음은 아무런 패턴이 남아있지 않은 무작위한 움직임(진동)을 보이는 데이터를 말한다.
② 자기회귀모형은 관찰기간의 제한이 없이 모든 시계열 데이터를 사용하며 최근 시계열에 더 많은 가중치를 주며 추세를 찾는 방법을 말한다.
③ 정상성은 시계열 데이터가 평균과 분산이 일정하지 않은 경우를 지칭한다.
④ 이동평균은 과거로부터 현재까지 시계열 자료를 대상으로 일정기간(관측기간)을 시계열을 이동하면서 분산을 계산하는 방법이다.

58 정의된 구조가 없으며 고정된 필드에 저장되지 않는 데이터를 뜻하는 것은?

① 반정형 데이터
② 비정형 데이터
③ 분산형 데이터
④ 질적 데이터

59 랜덤 포레스트의 장점과 거리가 먼 것은?

① 분류와 회귀 모두 이용할 수 있다.
② 의사결정나무의 쉽고 직관적인 특징을 가진다.
③ 데이터 수가 많아져도 빠른 수행속도를 나타낸다.
④ 예측의 변동성이 적으며 과적합을 방지한다.

60 k-폴드 교차검증의 장점으로 틀린 설명은?

① 모든 데이터셋을 훈련으로 사용할 수 있다.
② 모든 데이터셋을 평가로 사용할 수 있다.
③ 모델 훈련/평가 소요시간이 상대적으로 짧다.
④ 테스트 데이터에 과적합되는 현상을 방지할 수 있다.

61 다음 중 표본추출 방법에 대한 설명으로 맞지 않는 것은?

① 단순무작위 추출은 표본을 난수를 사용하여 무작위로 추출하는 것으로 편향성을 제거한다.
② 계통추출은 모집단에서 추출간격을 설정하고 설정 간격에서 무작위로 추출한다.
③ 모집단의 다양한 특성을 표현하기 위해서 각 집단내에 특징 집단을 나누고, 해당 집단에서 표본을 추출하는 방법을 층화추출이라고한다.
④ 군집추출을 시행하는 경우 단순무작위 추출보다 편향성이 감소한다.

62 매개변수가 하이퍼파라미터와의 다른 차이점은?

① 모델 내부에서 결정되는 변수이다.
② 모델 최적화를 위해 사용자가 직접 세팅하는 변수이다.
③ 은닉층의 뉴런 개수도 포함된다.
④ 절대적인 최적값이 존재하지 않는다.

63 다음 시각화 도구 중 2개 이상의 변수 사이의 관계를 표현하기 적합한 것은?

① 막대그래프 ② 도넛차트
③ 파이차트 ④ 스캐터 플롯

64 시간에 따른 값의 변화를 표현하기에 적합하지 않은 도구는?

① 막대그래프 ② 스타차트
③ 플로팅 차트 ④ 꺾은선 그래프

65 불균형 데이터 처리기법 중 맞지 않는 것은?

① 언더샘플링 ② 오버샘플링
③ 데이터 증강 ④ 임계값 조정

66 ROC 곡선에 관한 설명으로 틀린 것은?

① X, Y가 모두 [0, 1] 범위이다.
② 군집분석 모델의 성능을 평가하는 지표로 사용된다.
③ Y축은 민감도이다.
④ ROC 곡선의 하단 면적을 AUC라고 한다.

67 분류모델 평가지표에 해당되는 지표는?

① Adjusted R^2 ② MAPE
③ RMSE ④ AUC

68 다음 중 딥러닝의 하이퍼파라미터 종류와 관계 없는 것은?

① 학습률
② 배치크기
③ 은닉층의 뉴런개수
④ 가중치

69 주어진 데이터를 k개의 클러스터로 분할 군집하는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 군집분석 기법은?

① 계층적 군집분석 ② DBSCAN
③ K-평균 군집분석 ④ GMM

70 Precision(정밀도)가 95%이고 재현율(Recall)이 90%일때의 F1점수를 구하시오.

① 91.2% ② 92.4%
③ 93.5% ④ 95.5%

71 다층 퍼셉트론에 대한 설명 중 틀린 것은?

① 다층 퍼셉트론을 통해 비선형 영역 표현도 가능하다.
② 활성화 함수인 계단 함수를 이용한다.
③ 가중치와 편향을 매개변수로 설정한다.
④ 입력층과 출력층 사이에 은닉층은 별도로 존재하지 않는다.

72 관측값들이 어떤 이론적 분포를 따르고 있는지를 검정하는 방법으로 한 개의 요인을 대상으로 하는 것은?

① 적합도 검정
② 분포도 검정
③ 독립성 검정
④ 동질성 검정

73 인포그래픽의 특징 중 잘못된 것은?

① 전달하려는 메시지를 통계나 그래픽을 사용하여 간결하게 구성한다.
② 시각적으로 이해하기 쉽게 구성한다.
③ 복잡한 데이터는 시각화가 복잡하다.
④ 보는 사람에게 흥미와 관심을 유발한다.

74 분류, 회귀에 사용되는 학습기법은 무엇인가?

① 준지도 학습 ② 지도 학습
③ 강화 학습 ④ 비지도 학습

75 다음 중 분석 모형 진단 평가에 대한 설명으로 옳지 않은 것은?

- 참긍정(TP, True Positive)
- 참부정(TN, True Negative)
- 거짓긍정(FP, False Positive)
- 거짓부정(FN, False Negative)

① 실제 Positive인 대상 중에 실제와 예측 값이 일치하는 비율은 재현율(Recall)을 사용한다.
② 특이도(Specificity)는 전체 실제거짓 중 거짓예측한 비율이며 TN / (TN + FP)식으로 나타낸다.
③ 전체 예측된 긍정 중 거짓긍정한 비율을 정밀도(Precision)라고 하며 TP / (TP + FP)이다.
④ 정확도(Accuracy)는 (TP + TN) / (TP + FP + TN + FN)이다.

76 신경망모델에서 은닉층의 뉴런을 임의로 삭제하면서 학습하는 방법으로 적은 뉴런만으로 훈련한 뒤 테스트 시에 전체 뉴런을 사용하면 정답을 보다 잘 찾을 수 있어 과적합을 방지할 수 있는 기법명은?

① 가중치 규제
② 가중치 초기화
③ 드롭아웃
④ 하이퍼파라미터 튜닝

77 다음 분석결과 활용 방법에 대한 설명으로 맞지 않는 것은?

① 분석모형은 시간이 지나면서 성능이 떨어질 수 있다.
② 분석 데이터의 크기가 커지면 검증과정을 생략해도 신뢰성 높은 결과를 얻을 수 있다.
③ 분석 모형의 성능을 높이기 위해 리모델링을 수행한다.
④ 데이터셋의 특성이 달라지는 경우 새롭게 분석 모형을 구축해야 한다.

78 k-평균 군집분석에서 최적의 k값을 선택하기 위해 군집 간 분산과 전체 분산 간의 비율이 완곡하게 줄어드는 기법은?

① 엘보우(Elbow) ② 실루엣(Silhouette)
③ 분산 최적화 ④ 오차율 최소화

79 선형회귀분석 기반 예측모델과 관계가 있는 항목은?

① 잔차 분석 ② 로짓 변환
③ 크로스 엔트로피 ④ 확률 분포

80 회귀분석 모형 진단에서 표본의 실제값에 대한 회귀식의 설명력에 대한 것은?

① 적합도 검정 ② 유의성 검정
③ 회귀 테스트 ④ 잔차분석

1 과목 | 빅데이터 분석 기획

01 정성적 데이터에 대한 설명으로 적절하지 않은 것은?

① 객체 하나가 함축된 의미를 내포하고 있다.
② 반정형 데이터와 비정형 데이터로 구성되어 있다.
③ 주로 주관적 내용을 담고 있다.
④ 문자나 언어로 표현되어 통계 분석 시 어려움이 있다.

02 암묵지와 형식지에 대한 설명으로 적절하지 않은 것은?

① 암묵지는 어떠한 시행착오나 다양하고 오랜 경험을 통해 개인에게 체계화되어 있다.
② 형식지는 공통화 및 연결화 과정을 통해 암묵지가 구체화되어 외부로 표현된 것이다.
③ 암묵지는 외부에 표출되지 않은 무형의 지식으로 그 전달과 공유가 어렵다.
④ 형식지는 형상화된 유형의 지식으로 그 전달과 공유가 쉽다.

03 데이터 활용 기술에 대한 설명으로 적절하지 않은 것은?

① OLTP는 호스트 컴퓨터와 온라인으로 접속된 여러 단말 간 처리 형태의 하나로 데이터베이스의 데이터를 수시로 갱신하는 프로세싱을 의미한다.
② OLAP는 정보 위주의 분석 처리를 하는 것으로 트랜잭션 데이터를 분석해 제품의 판매 추이, 구매 성향 파악, 재무 회계 분석 등을 프로세싱하는 것을 의미한다.
③ 데이터베이스는 다양한 비즈니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근하여 의사결정에 활용할 수 있는 정보를 얻을 수 있게 하는 기술이다.
④ 데이터 마이닝은 대용량의 데이터로부터 인사이트를 도출할 수 있는 방법론이다

04 빅데이터의 특징에 대한 설명으로 적절하지 않은 것은?

① 단일 데이터만으로는 가치가 크지 않지만 다른 데이터들과 연계할 때 크게 증가한다.
② 최근에는 3Vs(규모, 유형, 속도) 외에 빅데이터 분석을 통해 얻을 수 있는 가치와 데이터에 대한 품질의 중요성이 강조되고 있다.
③ 품질은 데이터의 신뢰성, 정확성, 타당성 보장이 필수적이며, 고품질의 데이터에서 고수준 인사이트 도출이 가능하다.
④ 빅데이터 용어가 사용된 초기에 가트너 그룹은 3Vs로 빅데이터의 특징을 설명하였다.

05 빅데이터 활용을 위한 테크닉에 대한 설명으로 적절하지 않은 것은?

① 연관규칙분석은 독립변수가 종속변수에 미치는 영향을 분석할 때 사용한다.
② 유형분석은 문서를 분류하거나 조직을 그룹화할 때 사용한다.
③ 유전 알고리즘은 최적화가 필요한 문제를 생물 진화의 과정을 모방하여 점진적으로 해결책을 찾는 방법이다.
④ 소셜네트워크분석은 특정인과 다른 사람의 관계를 파악하고 영향력 있는 사람을 분석할 때 사용한다.

06 데이터 권리 시대에 대한 설명으로 적절하지 않은 것은?

① 데이터의 원래 소유자인 개인이 자신의 데이터에 대한 권리를 보유하고 있으며 스스로 행사할 수 있어야 한다는 마이데이터(My Data)가 등장하였다.
② 데이터 소비자의 역할과 활용 역량을 높이기 위한 데이터 리터러시 프로그램의 중요성이 커지고 있다.
③ 데이터 연결과 데이터 권리는 개인 데이터가 완전하게 보호되며, 개인은 자신의 데이터를 완전하게 통제할 수 있다는 믿음이 보편화되어야 한다.
④ 개인은 데이터를 만들고 자신이 만든 데이터를 기반으로 비즈니스 모델을 구상할 수 있으며, 기업들은 개인 데이터 사용에 제약을 받게 됨으로써 고객 접점을 상실하게 될 수 있다.

07 빅데이터 수집 기술에 대한 설명으로 적절하지 않은 것은?

① 크롤링은 무수히 많은 컴퓨터에 분산 저장되어 있는 문서를 수집하여 검색 대상의 색인으로 포함시키는 기술이다.
② ETL은 다양한 원천 데이터를 취합해 추출하고 공통된 형식으로 변환하여 데이터 웨어하우스에 적재하는 과정이다.
③ 센서 네트워크는 조직 내부에 있는 웹 서버나 시스템의 로그를 수집하는 기술이다.
④ ODS는 다양한 DBMS 시스템에서 추출한 데이터를 통합적으로 관리한다.

08 빅데이터 플랫폼에 대한 설명으로 적절하지 않은 것은?

① 분산시스템은 네트워크상에 분산 되어 있는 컴퓨터를 단일 시스템인 것처럼 구동하는 기술이다.
② 하둡은 분산 처리 환경에서 대용량 데이터 처리 및 분석을 지원하는 오픈 소스 소프트웨어 프레임워크이다.
③ 맵리듀스는 구글에서 개발한 방대한 양의 데이터를 신속하게 처리하는 프로그래밍 모델로 효과적인 병렬 및 분산 처리를 지원한다.
④ NoSQL은 기존의 RDBMS 트랜잭션 속성인 원자성, 일관성, 독립성, 지속성을 보장하는 비관계형 데이터베이스이다.

09 NoSQL의 데이터 모델에 대한 설명으로 적절하지 않은 것은?

① 관계형데이터베이스의 ACID 특성을 모두 지원하며, 성능과 확장성을 높이는 데이터 모델을 지원한다.
② 키-값(key-value) 데이터베이스는 단순한 데이터 모델에 기반을 두고 있어 관계형 데이터베이스보다 확장성이 뛰어나고 질의 응답시간이 빠르다.
③ 열 기반(column-oriented) 데이터베이스는 칼럼과 로우는 확장성을 보장하기 위하여 여러 개의 노드로 분할되어 저장 및 관리된다.
④ 문서 기반(document-oriented) 데이터베이스는 문서의 내부 구조에 기반을 둔 복잡한 형태의 데이터 저장을 지원하고 이에 따른 최적화가 가능하다.

10 빅데이터 분석절차에 대한 설명으로 적절하지 않은 것은?

① 일반적인 분석 절차는 문제 인식, 연구조사, 모형화, 데이터 수집, 데이터 분석, 분석 결과 제시 단계로 구성되어 있다.
② 분석 방법론을 구성하는 최소 요건이다.
③ 상황에 따라 단계를 추가할 수도 있으며 생략 가능 하다.
④ 문제에 대한 구체적 정의가 없다면 통계 기반의 전통적 데이터 분석을 수행할 수 없으므로 문제에 대한 구체적 정의가 필요하다.

11 데이터 분석 방법에 대한 설명으로 적절하지 않은 것은?

① 회귀는 독립변수가 종속변수에 미치는 영향을 분석할 때 사용하는 방법이다.
② 분류는 학습 데이터 셋을 학습시켜 새로 추가되는 데이터가 속할 만한 데이터 셋을 찾는 지도학습 방법이다.
③ 군집화는 특성이 비슷한 데이터를 하나의 그룹으로 분류하는 방법으로 지도학습의 한 방법이다.
④ 텍스트 마이닝은 분류나 군집화 등 빅데이터에 숨겨진 의미 있는 정보를 발견하는데 사용하기도 한다.

12 인공지능 기술에 대한 설명으로 적절하지 않은 것은?

① 인공지능은 사람이 생각하고 판단하는 사고 구조를 구축하려는 전반적인 노력이다.
② 기계학습은 인공지능의 연구 분야 중 하나로 인간의 학습 능력과 같은 기능을 축적된 데이터를 활용하여 실현하고자 하는 기술 및 방법이다.
③ 딥러닝은 기계학습 방법 중 하나로 컴퓨터가 많은 데이터를 이용해 사람처럼 스스로 학습할 수 있도록 인공신경망 등의 기술을 이용한 기법이다.
④ 강화학습의 초점은 학습 과정에서의 성능이며 이는 탐색과 이용의 균형을 맞춤으로써 제고되며, 시뮬레이션 데이터 생성, 누락 데이터 생성, 패션 데이터 생성 등에 응용할 수 있다.

13 개인정보와 관련된 설명으로 적절하지 않은 것은?

① 개인정보보호법은 당사자의 동의 없는 개인정보 수집 및 활용하거나 제 3자에게 제공하는 것을 금지하는 등 개인정보보호를 강화한 내용을 담아 제정한 법률이다.
② 개인정보의 처리 위탁은 개인정보처리자의 업무를 처리할 목적으로 제 3자에게 이전되는 것이다.
③ 개인정보의 제3자 제공은 개인정보가 제 3자에게 이전되거나 공동으로 처리하게 하는 것이다.
④ 상대방의 동의 없이 개인정보를 제 3자에게 제공하면 5년 이하의 징역이나 5,000만원 이하의 벌금에 처할 수 있다.

14 개인정보비식별화 방법에 대한 설명으로 적절하지 않은 것은?

① 가명처리는 값을 대체 시 규칙이 노출되어 역으로 쉽게 식별할 수 없도록 주의해야 한다.
② 범주화 과정에서 특정 속성을 지닌 개인으로 구성된 단체의 속성 정보를 공개하는 것은 그 집단에 속한 개인의 정보를 공개하는 것과 마찬가지이므로 주의해야 한다.
③ 삭제는 데이터 공유나 개방 목적에 따라 데이터 셋에 구성된 값 중 필요 없는 값 또는 개인식별에 중요한 값을 삭제하는 방법이다.
④ 마스킹은 개인을 식별하는데 기여할 확률이 높은 주요 식별자를 보이지 않도록 처리하는 방법이다.

15 탐색적 데이터 분석에 대한 설명으로 적절하지 않은 것은?

① 분석용 데이터셋에 대한 정합성 검토, 데이터 요약, 데이터 특성을 파악하고 모델링에 필요한 데이터를 편성한다.
② 다양한 관점으로 평균, 분산 등 기초 통계량을 산출하여 데이터의 분포와 변수간의 관계 등 데이터 자체의 특성과 통계적 특성을 파악한다.
③ 정형, 비정형, 반정형 등 모든 내외부 데이터를 대상으로 데이터의 속성, 오너, 관련 시스템 담당자 등을 포함한 데이터 정의서를 작성한다.
④ 시각화를 탐색적 데이터 분석을 위한 도구로 활용하여 데이터의 가독성을 명확히 하고 데이터의 형상 및 분포 등 데이터 특성을 파악한다.

16 데이터 거버넌스 체계에 대한 설명으로 적절하지 않은 것은?

① 데이터 표준 용어 설정은 표준 단어 사전, 표준 도메인 사전, 표준 코드 등으로 구성되며, 각 사전 간 상호 검증이 가능한 점검 프로세스를 포함한다.
② 데이터 관리 체계는 표준 데이터를 포함한 메타 데이터와 데이터 사전의 관리 원칙 수립 및 이에 근거한 항목별 상세 프로세스를 수립한다.
③ 저장소는 데이터 관리 체계 지원을 위한 Workflow 및 관리용 Application을 지원하여야 한다.
④ 메타 데이터 및 데이터 사전 구축과 같은 표준화 활동을 주기적으로 진행한다.

17 분석 성숙도 모델에 대한 설명으로 적절하지 않은 것은?

① 데이터 분석 능력 및 데이터 분석 결과 활용에 대한 조직의 성숙도 수준을 평가하여 현재 상태를 점검하는 방법이다.
② 총 6가지 영역을 대상으로 현재 수준을 파악한다.
③ 비즈니스 부문, 조직 및 역량 부문, IT 부문 총 3개 부문을 대상으로 실시한다.
④ 성숙도 수준에 따라 도입단계, 활용단계, 확산단계, 최적화단계로 구분한다.

18 분석 문제 정의 방법에 대한 설명으로 적절하지 않은 것은?

① 하향식 접근 방식은 문제가 주어지고 이에 대한 해법을 찾기 위하여 각 과정이 체계적으로 단계화되어 수행하는 방식이다.
② 프로토타이핑 접근법의 경우 진화적 프로토타입보다 실험적 프로토타입에 가깝다고 볼 수 있다.
③ 상향식 접근 방식은 문제의 정의 자체가 어려운 경우 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식이다.
④ 동적인 환경에서 발산과 수렴 단계를 반복적으로 수행하며 상호 보완을 통해 분석의 가치를 극대화하는 혼합방식을 통해 최적의 의사결정을 할 수 있다.

19 빅데이터 분석 방법론에 대한 설명으로 적절하지 않은 것은?

① 응용 서비스 개발을 위한 단계, 태스크, 스텝 3계층으로 구성되었다.
② 분석 기획, 데이터 준비, 데이터 분석, 시스템 구현, 평가 및 전개 5단계로 구성되었다.
③ 비즈니스 이해 및 범위 설정은 데이터 준비 단계의 한 태스크로 프로젝트의 범위를 명확하게 파악하기 위해 구조화된 명세서를 작성한다.
④ 모델링은 데이터 분석 단계의 한 태스크로 개발된 모형을 활용하기 위해 상세한 알고리즘 설명서 작성과 모니터링 방안이 필요하다.

20 분석 프로젝트 속성에 대한 설명으로 적절하지 않은 것은?

① 분석 프로젝트는 도출된 결과의 재해석을 통한 지속적인 반복과 정교화가 수행되는 경우가 대부분이다.
② 분석 프로젝트는 데이터 크기, 데이터 복잡도, 속도, 분석 모형의 복잡도, 정확도와 정밀도를 추가적으로 고려하여야 한다.
③ 분석 결과를 활용하는 측면에서는 정확도가 중요하며, 분석 모형의 안정성 측면에서는 정밀도가 중요하다.
④ 정확도와 정밀도는 항시 Trade off 관계에 있다.

21 다음은 결측값에 대한 처리방법을 설명한 것이다. 어떠한 방법에 대한 설명인지 바르게 짝지어진 것을 고르시오.

ㄱ. 관측 또는 실험으로 얻어진 데이터의 평균으로 결측치를 대치해서 사용한다. 이러한 대치법은 효율성의 향상 측면에는 장점이 있으나 통계량의 표준오차가 과소 추정되는 단점이 있다.
ㄴ. 전체표본을 몇 개의 대체군으로 분류하여 각 층에서의 응답자료를 순서대로 정리한 후 결측값 바로 이전의 응답을 결측치로 대치한다. 응답값이 여러 번 사용될 가능성이 단점이다.

① ㄱ. 평균 대치법 ㄴ. 회귀 대치법
② ㄱ. 단순확률 대치법 ㄴ. 최근방 대치법
③ ㄱ. 평균 대치법 ㄴ. 최근방 대치법
④ ㄱ. 단순확률 대치법 ㄴ. 평균 대치법

22 다음은 어떤 변수 선택법에 대한 설명인가?

- 영 모형에서 시작, 모든 독립변수 중 종속변수와 단순상관계수의 절댓값이 가장 큰 변수를 분석모형에 포함시키는 것을 말한다.
- 부분 F 검정(F test)을 통해 유의성 검증을 시행. 유의한 경우는 가장 큰 F 통계량을 가지는 모형을 선택하고 유의하지 않은 경우는 변수선택 없이 과정을 중단한다.
- 한번 추가된 변수는 제거하지 않는 것이 원칙이다.

① 전진 선택법
② 후진 선택법
③ 단계적 선택법
④ 통계적 선택법

23 요인분석(PCA)의 특징에 대한 설명으로 틀린 것은?

① 가장 작은 분산의 방향들이 주요 중심 관심으로 가정한다.
② 본래의 변수들의 선형결합으로만 고려한다.
③ 차원의 축소는 본래의 변수들이 서로 상관이 있을 때만 가능하다.
④ 스케일에 대한 영향이 크다. 즉 PCA 수행을 위해선 변수들 간의 스케일링이 필수이다.

24 어떤 주어진 데이터의 기술적 통계량에 대한 분석 결과 Mean 〈 Median 〈 Mode의 위치를 가지는 형태의 분포를 정규분포형태로 변환하는 방법으로 옳은 것은?

① 순위를 데이터로 범주를 나누어 상대비교로 나누어 정렬한다.
② 모든 데이터를 최소값 0 최대값 1로 그리고 다른 값은 0과 1 사이 값으로 변환한다.
③ Negative Skew 경우로 ln(X)를 통한 변환을 이용한다.
④ Positive Skew 경우로 X^n을 통한 변환을 이용한다.

25 다음 중 오버샘플링에 대한 설명으로 옳은 것은?

(가) 다수 클래스 데이터에서 일부만 사용하는 방법이다.
(나) 소수 클래스 데이터를 증가시키는 방법이다.
(다) 소수클래스(Minority Class)의 복사본을 만들어, 대표클래스(Majority Class)의 수만큼 데이터를 만들어 주는 것이다.
(라) 데이터에서 loss를 계산할 때 특정 클래스의 데이터에 더 큰 loss 값을 갖도록 하는 방법이다.

① 가, 나 ② 나, 다
③ 다, 라 ④ 나, 라

26 다음 데이터 12, 20, 23, 25, 30에 대해서

$$A = \frac{1}{n}\sum_{i=1}^{n}|x_i - B|$$

최소화 값 A와 최소값을 만들어주는 데이터 또는 통계량 B는 얼마인가?

① A = 4.3 B = 12 (최소값)
② A = 4.4 B = 30 (최대값)
③ A = 4.8 B = 22 (산술평균)
④ A = 4.6 B = 23 (중앙값)

27 다음은 비확률표본 추출법 중 하나를 설명한 내용이다. 어떠한 방법에 대한 설명인가?

- 조사자가 나름의 지식과 경험에 의해 모집단을 가장 잘 대표한다고 여겨지는 표본을 주관적으로 선정하는 방법이다.
- 추출된 표본은 조사자의 주관적 판단에 의해서 표본이 추출되기 때문에 그 표본을 통해 얻은 추정치의 정확성에 대해 객관적으로 평가할 수 없다.
- 표본의 크기가 작은 경우에 조사의 오차를 좌우하는 요인은 추정량의 분산이 될 수 있다.

① 판단추출법(Judgement Sampling)
② 할당추출법(Quota Sampling)
③ 편의추출법(Convenience Sampling)
④ 눈덩이추출법(Snowball Sampling)

28 대한민국 30, 40대 직장인들의 30%는 음주 및 스트레스로 인해 간에 이상이 있는 것으로 알려져 있다. 간기능 검사 시 10% 비율로 잘못 진단할 수 있다고 할 때, 임의의 직장인이 간기능 검사 시 실제 간기능에 문제가 없음에도 불구하고 이상이 있음을 나타낼 확률은 얼마인가?

① 10.0%　② 20.6%
③ 34.0%　④ 53.1%

29 다음 아래와 같은 함수가 정의되어 있다고 할 때, 아래 함수가 연속확률밀도함수가 되기 위한 상수 A(단 A>0)의 값을 정하고 P(x<1/2)인 값을 각각 순서대로 구하시오.

$$f(x)=\begin{cases} Ax^2, & 0<x<1 \\ 0, & otherwise \end{cases}$$

① 1, $\frac{1}{3}$
② 2, $\frac{2}{3}$
③ 3, $\frac{1}{8}$
④ 4, $\frac{1}{16}$

30 어떤 부품의 수명은 평균 300시간을 가지고 (β=300)인 지수분포를 따른다.

$$f(x)=\frac{1}{300}e^{-x/300}, \quad x>0$$

이 부품이 100시간동안 고장나지 않았을 때, 앞으로 400시간동안 고장나지 않고 작동할 확률은?

① $e^{-\frac{5}{3}}$　② $e^{-\frac{4}{3}}$
③ e^{-1}　④ $e^{-\frac{1}{3}}$

31 스튜던트 t 분포에서 자유도에 대한 설명으로 틀린 것은?

① 자유도는 자료집단의 변수 중에서 자유롭게 선택될 수 있는 변수의 수를 말한다.
② 스튜던트 t 분포는 분포의 모양은 Z-분포와 유사하다. 종 모양으로서 t=0에 대하여 대칭을 이루는데 t-곡선의 모양을 결정하는 것은 자유도이다.
③ 자유도가 클수록 정규분포의 종 모양을 가지게 된다.
④ 자유도가 1보다 클 때 스튜던트 t 분포에서 기대값은 1이다.

32 다음 설명 중 틀린 것은?

① 표본의 크기가 클수록(표본 수 30 이상) 정규분포를 따른다.
② 표본의 크기가 작고 모 표준편차를 모르는 경우는 t 분포를 따른다.
③ 표본의 크기가 큰 경우 근사적으로 정규분포를 따르게 된다는 것이 대수의 법칙(law of Large Number)이다
④ 표본의 크기가 작고 모 표준편차를 아는 경우는 정규분포를 따른다.

33 편향에 대한 설명으로 옳은 것은?

① 기대하는 추정량과 모수의 비율을 편향(bias)이라고 한다.
② 분산은 평균에 대한 편차로 이상값에 대한 영향이 적은 대표적 편의추정량이다.
③ 불편추정량(Unbiased Estimator)은 B0=0, 즉 편향이 0이 되는 상황의 추정량 θ를 불편추정량이라고 한다.
④ 표본평균은 이상치의 영향으로 값의 변화가 커지므로 대표적인 불편추정량이 아니다.

34 다음 각 분포에 대한 설명으로 틀린 것은?

① 카이제곱분포의 확률밀도함수는 $f(x;k)=\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}x^{\frac{k}{2}-1}e^{-x/2}$ $\{x\geq 0\}$이고 기댓값은 k, 분산 2k이다.
② t 분포에서 자유도가 커지면 커질수록 분포의 형태는 정규분포를 따르게 되므로 평균<중앙값<최빈값의 순으로 나타나는 분포의 모습을 따르게 된다.
③ 포아송 분포의 기댓값과 분산은 동일하다.
④ 정규분포는 평균을 중심으로 좌우로 표준편차의 3배 이상 떨어진 값은 거의 취하지 않는다.

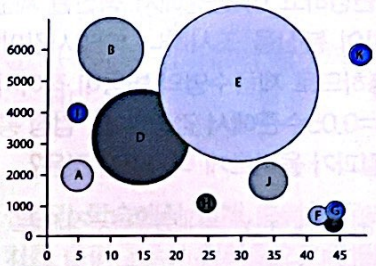
35 어떤 기업이 신입사원선발 대한 직무능력시험을 본 결과에 대해서 전체 응시자 중 100명을 뽑아 조사한 결과 평균이 90, 분산이 16이었다고 한다면 이 시험에 대한 전체 모평균의 신뢰구간을 95%수준에서 구하시오. (소수점 둘째자리에서 반올림)

① $89.41 \leq \mu \leq 90.59$
② $86.41 \leq \mu \leq 95.59$
③ $89.22 \leq \mu \leq 90.78$
④ $89.34 \leq \mu \leq 90.66$

36 두 대선 후보의 지지율을 조사하기 위하여 층화표본추출에 의해 각 나이대별로 지지율 조사를 하고자 한다. 유권자기준 20, 30, 40, 50, 60, 70대 이상으로 나누어 조사를 실시하고자 하는데 95% 신뢰수준으로 추정오차가 1% 이내가 되기 위한 각 나이대별 필요 표본크기는 얼마인가?

① 9604명 이상 되어야 한다.
② 6724명 이상 되어야 한다.
③ 2704명 이상 되어야 한다.
④ 1807명 이상 되어야 한다.

37 다음 아래와 같은 차트의 특징으로 볼 수 없는 것은?



① x, y값의 위치를 표시하는 산점도에 점의 위치에 해당하는 제3의 변수값을 원의 크기로 표현한 그래프로 한 번에 3개의 변수를 비교해볼 수 있다.
② 원(버블)은 면적으로 표현되어야 하며, 반지름이나 지름으로 표현되면 실제 값보다 너무 크게 원이 그려질 수 있어서 주의해야 한다.
③ 국가나 지역에 따른 값의 분포를 표현하는데 매우 유리하다.
④ 데이터 분포와 관계에 대한 정보를 색으로 표현한 그래프이다. 데이터를 식별하기 위해 각각의 칸마다 색으로 수치의 정도를 표현한다.

38 다음 중 상관계수에 대한 설명으로 부적절한 것을 고르시오.

① 피어슨 상관계수는 서열자료인 두 변수들의 상관관계를 측정하는데 사용한다.
② 상관계수 0은 두 변수 간 상관관계가 없음을 의미한다.
③ 스피어만 상관계수는 두 변수 간 상관관계가 선형관계가 아닌 경우도 고려할 수 있다.
④ 상관계수가 1에 가까울수록 두 변수 간 상관관계가 높음을 의미한다.

39 어느 기계회사의 생산제품 수명은 분산이 1200시간인 정규분포를 따른다. 새로운 공정설계에 의하여 일부를 변경하고 이 공정에서 생산된 제품 30개를 추출하여 분산을 조사하니 1050시간이었다. 공정을 변경하므로 제품수명의 변동이 적어지는지 유의수준 α=0.05수준에서 검정할 때, 검정 통계량과 검정의 결과가 올바르게 짝지어진 것은?

① 사용검정 통계량 χ^2, 새로운 공정으로 변경하더라도 제품수명의 변동은 적어지지 않는다.
② 사용검정 통계량 χ^2, 새로운 공정으로 변경하면 제품수명의 변동에 차이가 있다.
③ 사용검정 통계량 t, 새로운 공정으로 변경하더라도 제품수명의 변동은 적어지지 않는다.
④ 사용검정 통계량 t, 새로운 공정으로 변경하면 제품수명의 변동에 차이가 있다.

40 다음 중 잘못된 설명은?

① 가설검정은 모집단에 대해 어떤 가설을 설정하고 그 모집단으로부터 추출된 표본을 분석함으로써 그 가설이 틀리는지 맞는지 타당성 여부를 결정(검정)하는 통계적 기법이다.
② 제1종 오류(Type I Error)는 귀무가설이 참일 때 귀무가설을 기각하도록 결정하는 오류이며 우리가 말하는 유의 수준이 곧 1종 오류의 확률이다.
③ 임계치(Critical Value)는 주어진 p-value에서 귀무가설의 채택과 기각에 관련된 의사결정을 할 때, 그 기준이 되는 점이다.
④ 귀무가설의 기각여부는 p-value와 유의수준 α의 크기에 달려 있다. 즉 p-value가 작을수록 그리고 유의수준 α의 값이 클수록 귀무가설을 기각할 수 있다.

3 과목 | 빅데이터 모델링

41 지도학습의 종류 기법으로 세부설명과 맞지 않는 것은?

① 분류 : 랜덤 포레스트
② 회귀 : 다중 회귀분석
③ 분류 : SVM
④ 회귀 : 로지스틱 회귀분석

42 준지도학습의 종류인 GAN은 적대적 생성모델로 2가지 모델이 존재한다. 맞게 나열한 것은?

① 생성모델, 학습모델
② 생성모델, 환경모델
③ 생성모델, 판별모델
④ 생성모델, 특징모델

43 강화학습이란 주어진 환경에서 ( )를/을 최대화하도록 에이전트를 학습하는 기법이다. 괄호에 알맞은 것은?

① 자극 ② 보상
③ 목표치 ④ 예측률

44 다음은 의사결정나무의 구성요소를 설명한 것이다. 틀린 것은?

a. 가지 : 하나의 마디로부터 끝마디까지 연결된 마디들
b. 깊이 : 가지를 이루는 마디의 개수
c. 뿌리마디 : 나무줄기 끝에 있는 마디
d. 자식마디 : 하나의 마디로부터 분리된 2개이상의 마디

① a ② b
③ c ④ d

45 카이제곱 검정은 관찰된 빈도가 기대되는 빈도와 비교하여 유의미하게 다른지를 검증하는 기법으로 의사결정나무에 적용될 때 관측도수와 기대도수와의 차이가 커질수록 높아지는 값은?

① 불순도 ② 순수도
③ 지니점수 ④ 엔트로피

46 정보이론에서 순도가 증가하고 불확실성이 감소하는 것으로 발생 사건의 확률이 높아질수록 가치는 줄어드는 이것을 통칭하면?

① 정보손실 ② 정보공유
③ 정보획득 ④ 정보전파

47 임의로 크기가 동일한 여러 개의 표본자료들을 생성하는 것으로 랜덤 포레스트가 앙상블 학습하는데 기반이 되는 이것은?

① 복원추출 ② 배깅생성
③ 부트스트래핑 ④ 부스팅분류

48 의사결정나무의 장점으로 부적합한 설명은?

① 연속형, 범주형 변수 모두 적용이 가능하다.
② 데이터 변형에 민감하다.
③ DB마케팅, 시장조사, 기업 부도/예측 등에 활용한다.
④ 구조 복잡성에 관계없이 손쉽게 해석할 수 있다.

49 인공신경망의 주요 요소 설명으로 부적합한 것은?

① 노드는 신경계 시냅스에 비유된다.
② 은닉층은 입력층과 출력층 사이에서 데이터를 전파 학습한다.
③ 활성화함수는 임계값을 이용, 활성화 여부를 결정한다.
④ 가중치와 입력값이 활성화함수를 통해 전달된다.

50 신경망 학습에서 실제 출력과 목표 출력값과의 오차를 출력층에서 입력층으로 전달, 가중치와 편향을 계산, 업데이트하는 것은?

① 손실함수 ② 오차역전파
③ 연쇄법칙 ④ 매개변수 갱신

51 최적의 딥러닝 모델 구현을 위해 수동으로 딥러닝 모델에 설정하는 변수인 초매개변수(하이퍼파라미터) 종류와 거리가 먼 것은?

① 배치크기
② 훈련 반복 횟수
③ 가중치 초기화 방법
④ 편향 조정

52 RNN의 단점을 보완하기 위한 변형된 알고리즘인 LSTM은 오랫동안 데이터를 잘 기억하기 위하여 3가지 게이트를 가지고 있는데 이에 해당되지 않는 것은?

① 입력 게이트 ② 망각 게이트
③ 복원 게이트 ④ 출력 게이트

53 다차원 데이터를 저차원으로 바꾸고 바꾼 저차원 데이터를 다시 고차원 데이터로 바꾸면서 특징점을 찾아내는 대표적인 비지도학습 알고리즘은?

① GAN
② 오토인코더
③ RNN
④ CNN

54 SVM에서 초평면의 마진은 각 ( )을/를 지나는 초평면 사이의 거리를 의미할 때 괄호에 알맞는 것은?

① 오프셋
② 결정영역
③ 서포트 벡터
④ 커널

55 군집분석의 척도로 L1 거리로도 통칭되며 사각형 격자, 블록으로 이뤄진 지도에서 출발점에서 도착점까지 가로지르지 않고 도착하는 최단거리 개념은?

① 유클리드 거리
② 민코우스키 거리
③ 마할라노비스 거리
④ 맨하탄 거리

56 다음 아래와 같은 분석에 대해 사용 가능한 가장 적합한 통계량과 분석기법은 무엇인가?

방송사에서 방송중인 두 개의 프로그램에 대한 시청률에 대해 성별에 따른 차이 유무를 검증하기 위하여 100명의 표본을 선출하여 조사하였다.

① t, 단일평균분석
② χ^2, 교차분석
③ Z, 회귀분석
④ F, 이원분산분석

57 시계열자료에 대한 설명으로 잘못된 것은?

① 추세성분(Trend Component)은 관측 값이 지속적 증가 또는 감소하는 추세(Trend)를 포함한다.
② 계절성분(Seasonal Component)은 주기적 성분에 의한 변동을 가지는 형태(계절, 주, 월, 년 등)이다.
③ 자기상관성(Autocorrelation)은 시차값 사이 이동평균에 대한 값으로 분석하는 것을 말한다.
④ 백색잡음(White Noise)은 자기상관성이 없는 시계열 데이터를 지칭한다.

58 다음은 어떤 모델에서 문서분류에 대한 원리를 나열한 것이다. 빈칸에 들어갈 알맞은 말을 고르시오.

문서 doc가 주어졌을 때 범주 C_1와 C_2로 분류 시
$P(C_1|Doc) = \frac{P(Doc|C_1)P(C_1)}{P(Doc)}$, $P(C_2|Doc) = \frac{P(Doc|C_2)P(C_2)}{P(Doc)}$
(가) 모델은 $P(C_1|Doc)/P(Doc)$와 $P(C_2|Doc)/P(Doc)$를 비교해서 그 값이 (나) 쪽으로 범주를 할당한다는 개념이다.

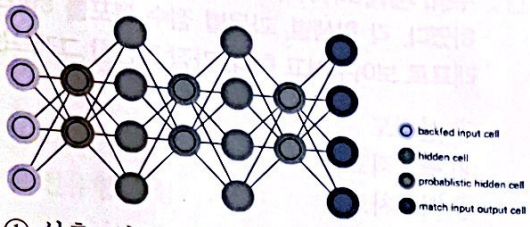
① (가) 합성곱신경망(CNN) (나) 작은
② (가) K-means (나) 동일한
③ (가) 나이브 베이즈 모델 (나) 큰
④ (가) 딥러닝 (나) 큰

59 다음의 설명 중 옳은 것은?

가. Voting은 서로 다른 알고리즘이 도출해 낸 결과물에 대하여 최종 투표하는 방식을 통해 최종 결과를 선택한다.
나. 부스팅(Boosting)은 가중치를 활용하여 연속적인(sequential) 약학습기를 생성하고 이를 통해 강학습기를 만드는 방법이다.
다. 부스팅(Boosting)은 순차적이기 때문에 병렬 처리에 어려움이 있고, 그렇기 때문에 다른 앙상블 대비 학습 시간이 오래 걸린다는 단점이 있다.
라. Bagging은 같은 알고리즘 내에서 다른 sample 조합을 사용한다.

① 가, 다, 라
② 가, 나, 라
③ 가, 나, 다, 라
④ 나, 다, 라

60 다음은 어떤 신경망 모델에 대한 다이어그램인지 고르시오.



① 심층 신뢰 신경망(DBN: Deep Belief Network)
② 순환 신경망(RNN: Recurrent Neural Network)
③ 합성곱 신경망(CNN: Convolutional Neural Network)
④ 심층 신경망(DNN: Deep Neural Network)

4과목 | 빅데이터 결과 해석

61 분류 평가지표로 맞지 않는 것은?

① 재현율 ② 정확도
③ 정상도 ④ 정밀도

62 모든 분류 임계값에서 분류 모델 성능을 보여주는 그래프에서 다음 곡선 아래 영역을 의미하는 용어로 1에 가까울수록 최적의 분류모델을 나타낸다고 할 때 이에 해당되는 것은?

① ROC ② MSE
③ F1 점수 ④ AUC

63 회귀에서 자주 사용되는 회귀지표로 예측한 값을 실제 값과 빼고 제곱한 값을 평균한 것은?

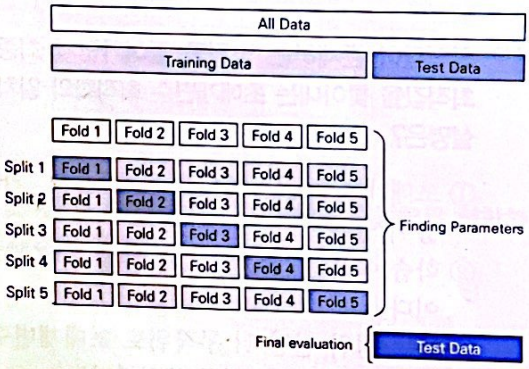
① MASE ② MSE
③ MAE ④ RMSE

64 다음 중 k-평균군집 분석의 분석절차 순서로 맞는 것은?

a. 군집중심으로 원하는 수(k)만큼 선택
b. 반복 과정으로 최종 군집 형성
c. 군집내 자료들의 평균 계산 뒤 중심점 갱신
d. 각 개체를 가장 가까운 중심에 할당

① d→ a→ c→ b
② d→ c→ a→ b
③ a→ d→ c→ b
④ a→ c→ d→ b

65 교차검증 K-Fold 검증에 대한 다음 예시로 부적합한 설명은?



① 훈련 데이터셋을 5개 Fold로 나눈다.
② 각 Fold마다 한 번씩 평가 데이터셋으로 사용, 나머지 Fold들을 훈련한다.
③ 테스트 횟수는 총 5회이다.
④ 5개 성능 결과가 나오면 이를 평균한 것이 해당 모델의 성능이라고 할 수 있다.

66 딥러닝 학습에서의 과적합을 예방하기 위한 방법으로 훈련할 때 은닉층의 뉴런 일부 연결을 삭제하여 신호를 전달하지 않게 하며 테스트 시에 모든 뉴런을 사용하는 기법은?

① 드롭아웃
② 가중치 초기화
③ 언더피팅
④ 앙상블

67 딥러닝 학습 동안 가중치 갱신 시에 가중치 값이 커지지 않도록 규제를 하는 기법으로 손실함수에 가중치의 절대값을 추가하는 규제기법명은?

① 정규화
② L1
③ L2
④ L3

68 손실함수를 최소화하는 매개변수를 찾는 방법 중 확률적 경사 하강법(SGD)와 관련이 없는 항목은?

① 손실함수의 기울기
② 학습률
③ 가중치
④ 편향

69 최적값이 존재하는 범위를 줄여가면서 최종적인 최적값을 찾아내는 초매개변수 최적화와 일치하는 설명은?

① 초매개변수에는 배치크기, 학습률, 가중치 등이 있다.
② 학습 에폭(epoch)을 크게 하는 것이 효과적이다.
③ 특정 범위 설정 뒤 무작위로 초매개변수 값을 샘플링하여 범위를 좁혀간다.
④ 최적화 이후 딥러닝 학습시간이 짧아진다.

70 부스팅과 배깅을 비교 시, 차이점과 관련된 학습 용어는?

① 독립성
② 순차성
③ 결합성
④ 변동성

71 두 변수 x와 변수 y값의 관계를 표현하기에 적합하지 않은 도구는?

① 스캐터 플롯
② 히트맵
③ 버블차트
④ 파이차트

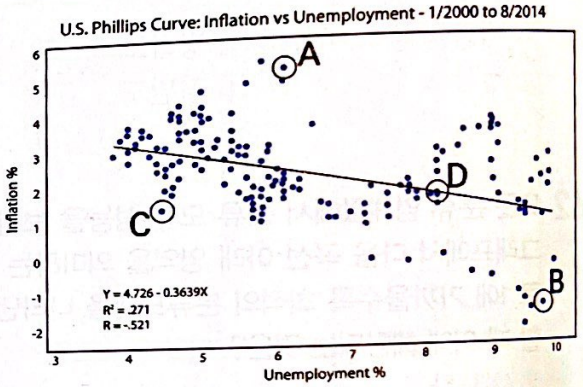
72 A반은 30명 학생이 있으며, 지난 주에 국어, 영어, 수학, 사회, 과목 5과목에 대해서 중간고사를 실시하였다. 각 학생별, 과목별 점수 분포를 하나의 그래프로 보여주려고 한다. 가장 적당한 그래프는?

① 산점도
② 평행좌표계
③ 버블차트
④ 도넛차트

73 데이터시각화 응용분야로 보기 어려운 것은?

① 인포그래픽
② 탐색적 데이터 분석(EDA)
③ 히트맵
④ 연관분석

74 다음 산점도는 인플레이션에 따른 실업률 변화를 보여준다. 맞게 설명한 것은?



① 인플레이션과 실업률은 양의 상관관계가 있다.
② A는 실업률이 높아서 이상값으로 분류될 수 있다.
③ D는 추세선 위에 존재하므로, 삭제를 해도 추세선의 모양이 변하지 않는다.
④ B와 C는 인플레이션이 비교적 낮은 도시로 분류할 수 있다.

75 다음 중 연속형 데이터와 이산형 데이터에 모두 적용할 수 없는 데이터의 시각적 속성은 어느 것인가?

① 위치 속성
② 색 속성
③ 크기 속성
④ 선유형 속성

76 다음 기업 월별 매출 그래프를 파이차트로 변환하려고 한다. 변환된 파이차트에서 3월이 차지하는 영역(조각)의 각도는 얼마인가?

① 25도
② 90도
③ 45도
④ 30도

77 다차원척도법에 대한 설명으로 맞지 않는 것은?

① 모든 변수를 비교하여 비슷한 대상을 그래프 상에서 가깝게 배치한다.
② 2차원 평면에 나타나는 경우 각 관측값이 (x, y) 좌표로 표시된다.
③ 원래의 차원보다 낮은 차원으로 위치시킬 수 있다.
④ 유사한 특징을 갖는 데이터들이 서로 뭉쳐져서 나타난다.

78 단계구분도는 구분되는 지역의 넓이가 각각 다르기 때문에, 넓은 지역의 값이 전체를 지배하는 것과 같이 보일 수 있다. 이러한 단점을 극복할 수 있는 방법으로 제시된 그래프는 다음 중 어느 것인가?

① 스타차트
② 카토그램
③ 히트맵
④ 인포그래픽

79 빅데이터 분석 방법론에서 완성된 분석 모형 모니터링과 가장 밀접한 관계가 있는 단계는?

① 분석기획 단계
② 데이터 준비 단계
③ 데이터 분석 단계
④ 평가 및 전개 단계

80 분석모형 리모델링에서 수행하는 활동으로 적합하지 않은 것은?

① 성능 모니터링
② 분석 알고리즘 개선
③ 매개변수 최적화
④ 신규분석과제 발굴

1 과목 | 빅데이터 분석 기획

01 다음 중 정량적 데이터와 정성적 데이터에 대한 설명으로 적절하지 않은 것은?

① 정량적 데이터의 유형은 비정형 데이터, 정성적 데이터는 정형, 반정형 데이터이다.
② 정량적 데이터는 수치나 기호 등으로, 정성적 데이터는 문자나 언어 등으로 구성되어 있다.
③ 정량적 데이터는 통계 분석에 용이한 반면 정성적 데이터의 경우 통계 분석 시 어려움이 있다.
④ 정량적 데이터는 주로 객관적 내용을 다루지만, 정성적 데이터는 주관적 내용을 다룬다.

02 다음 중 지식의 피라미드에 대한 예시로 적합하지 않은 것은?

① 데이터 : 텀블러의 온라인 가격은 1만원, 오프라인 가격은 1만5천원이다.
② 지식 : 텀블러를 저렴한 온라인 상점에서 구매하고, 커피도 온라인 상점에서 구매할 것이다.
③ 정보 : 텀블러를 온라인 상점에서 구매하는 것이 오프라인보다 더 저렴하다.
④ 지혜 : 텀블러가 온라인 상점에서 더 저렴하니 머그잔도 온라인 상점이 더 저렴할 것이다.

03 다음 중 OLTP와 OLAP에 대한 설명으로 적절하지 않은 것은?

① OLTP는 데이터 구조가 복잡하지만, OLAP는 단순하다.
② OLTP는 응답 시간이 수초 이내로 빠르지만, OLAP는 수 초에서 몇 분 사이로 느린 편이다.
③ OLTP는 데이터 액세스 빈도가 보통이지만, OLAP는 데이터 액세스 빈도가 높은 편이다.
④ OLTP는 현재 데이터를 담고 있지만, OLAP는 요약된 데이터를 다루고 있다.

04 다음 중 빅데이터의 특징에 대한 설명으로 적절하지 않은 것은?

① 정형 데이터 외 반정형 및 비정형 데이터로 유형이 확대되었다.
② 대용량 데이터의 신속하고 즉각적인 분석이 요구되고 있다.
③ 다른 데이터들과 연계 시 가치가 배로 증대된다.
④ 저품질의 다양한 데이터를 통해서 고수준 인사이트 도출이 가능하다.

05 다음 중 빅데이터의 기능과 효과에 대한 설명으로 적절하지 않은 것은?

① 빅데이터는 시장에 새롭게 진입하려는 잠재적 경쟁자에게 사업의 발판을 마련해준다.
② 빅데이터는 이를 활용하는 기존 사업자에게 경쟁 우위를 제공한다.
③ 빅데이터는 알고리즘 기반으로 의사결정을 지원하거나 이를 대신한다.
④ 빅데이터는 투명성을 높여 연구개발 및 관리 효율성을 제고한다.

06 다음 중 데이터 산업의 진화과정을 순서대로 알맞게 나열한 것은?

① 데이터 통합 → 데이터 분석 → 데이터 연결 → 데이터 권리 → 데이터 처리
② 데이터 처리 → 데이터 통합 → 데이터 분석 → 데이터 연결 → 데이터 권리
③ 데이터 권리 → 데이터 처리 → 데이터 통합 → 데이터 분석 → 데이터 연결
④ 데이터 연결 → 데이터 권리 → 데이터 처리 → 데이터 통합 → 데이터 분석

07 다음 중 분산형 조직구조에 대한 설명으로 적절하지 않은 것은?

① 분석 전문 인력을 현업 부서에 배치하여 분석 업무를 수행한다.
② 전사 차원에서 분석과제의 우선순위를 선정하고 수행한다.
③ 분석 수행의 일반적 구조이다.
④ 분석 결과를 현업에 빠르게 적용 가능하다.

08 다음 중 빅데이터 처리과정에서 요구되는 요소기술이 아닌 것은?

① 수집 기술 ② 저장 기술
③ 처리 기술 ④ 설계 기술

09 다음 중 기계학습의 종류에 대한 설명으로 적절하지 않은 것은?

① 강화학습은 선택 가능한 행동들 중 보상을 극대화하는 행동을 역순서로 선택하는 방법이다.
② 지도학습은 학습 데이터로부터 하나의 함수를 유추 해내기 위한 방법이다.
③ 비지도학습은 데이터가 어떻게 구성되었는지 알아내는 문제의 범주에 속한다.
④ 준지도학습은 목표 값이 표시된 데이터와 표시되지 않은 데이터 모두 학습에 사용한다.

10 다음 중 개인정보의 판단기준으로 적합하지 않은 것은?

① 생존하는 개인에 관한 정보여야 한다.
② 다른 정보와 결합하여 개인을 알아볼 수 있는 정보는 배제한다.
③ 정보의 내용 및 형태 등은 제한이 없다.
④ 개인을 알아볼 수 있는 정보여야 한다.

11 다음 중 2020년 데이터 3법의 주요 개정 내용에 대한 설명으로 적절하지 않은 것은?

① 데이터 이용 활성화를 위한 가명정보 개념 도입 및 데이터간 결합 근거를 마련하였다.
② 개인정보보호 관련 법률의 유사 · 중복 규정을 정비 및 거버넌스 체계를 효율화 하였다.
③ 데이터 활용 따른 개인정보처리자의 책임을 조직 대표자가 연대하여 책임지도록 강화하였다.
④ 다소 모호했던 개인정보의 판단기준을 명확하게 하였다.

12 다음 중 데이터 분석 기획의 특징에 대한 설명으로 적절한 것은?

① 분석 주제를 정의한 상태에서 분석 방법을 알고 있을 때 인사이트 발굴이 가능하다.
② 분석 주제를 정의하지 못하였지만 분석 방법을 알고 있다면 솔루션을 찾아낼 수 있다.
③ 분석 주제를 정의한 상태에서 분석 방법을 알고 있을 때 솔루션을 찾아낼 수 있다.
④ 분석 주제를 정의하지 못하였지만 분석 방법을 알고 있다면 인사이트를 발굴 할 수 있다.

13 다음 중 IT 프로젝트에서 과제 우선순위 평가기준으로 적합하지 않은 것은?

① 기술 완전성 ② 전략적 필요성
③ 시급성 ④ 투자 용이성

14 다음 중 분석 과제 우선순위 조정 시 고려사항에 대한 설명으로 적절하지 않은 것은?

① 분석 과제의 전체 범위를 한 번에 일괄적으로 적용하여 추진할 수 있다.
② 기존 시스템에 미치는 영향을 최소화하여 적용하는 방안이 가장 적절하다.
③ 분석 과제 중 일부만 PoC로 진행하고 평가 후 범위를 확대할 수 있다.
④ 기존 시스템과 별도로 시행하여 난이도 조율을 통한 우선순위를 조정할 수 있다.

15 다음 중 하향식 접근 방식의 해결방안 탐색 단계에 대한 설명으로 적절한 것은?

① 분석 역량을 확보하고 있고 분석 기법이나 시스템을 보유하고 있다면 고도화를 진행한다.
② 분석 역량을 확보하고 있고 분석 기법이나 시스템을 확보하지 못하였다면 아웃소싱한다.
③ 분석 역량을 확보하고 있고 분석 기법이나 시스템을 보유하고 있다면 개선하여 활용한다.
④ 분석 역량을 확보하지 못하였으나 분석 기법이나 시스템을 보유하고 있다면 아웃소싱한다.

16 다음 중 분석 방법론으로 활용 가능한 소프트웨어 개발생명주기에 대한 설명으로 옳은 것은?

① 폭포수 모형은 이해하기 쉽고 관리가 용이하며, 요구사항 도출이 쉽다.
② 원형 모형은 의사소통을 향상시키며, 폐기되는 프로토타입도 재활용 가능하다.
③ 나선형 모형은 계획수립, 개발, 위험분석, 고객평가 순으로 진행된다.
④ 반복적 모형은 시스템을 여러 번 나누어 릴리즈하는 방법이다.

17 다음 중 CRISP-DM 분석 방법론의 분석절차로 올바른 것은?

① 업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 → 전개
② 데이터셋 선택 → 데이터 전처리 → 데이터 변환 → 데이터마이닝 → 데이터마이닝 결과 평가
③ 추출 → 탐색 → 수정 → 모델링 → 평가
④ 분석 기획 → 데이터 준비 → 데이터 분석 → 시스템 구현 → 평가 및 전개

18 다음 중 분석 프로젝트 관리 시 중요한 속성들에 대한 설명으로 적절하지 않은 것은?

① Accuracy는 모형과 실제 값 사이의 차이를 측정하는 지표이다.
② 데이터의 크기는 현 시점을 기준으로 하며, 지속적 증가 여부는 고려하지 않는다.
③ Precision은 모형을 계속하여 반복했을 때 결과의 일관성을 측정하는 지표이다.
④ 분석 모형의 정확도와 복잡도는 Trade off 관계에 있다.

19 다음 중 데이터 수집을 위한 비용 요소에 대한 설명으로 적절하지 않은 것은?

① 데이터의 수집 주기는 실시간, 매시, 매일, 매주, 매달 단위로 할 수 있다.
② 데이터의 수집 방식은 자동 수집과 수동 수집으로 나뉜다.
③ 데이터의 종류는 관계형 데이터베이스나 파일에 있는 정형 데이터로 한정한다.
④ 데이터를 수집하기 위한 기술로는 ETL이나 크롤러 등이 있다.

20 다음 중 개인정보 비식별화를 위한 데이터 범주화 방법에 대한 설명으로 적절하지 않은 것은?

① 감추기는 명확한 값을 숨기기 위하여 데이터의 평균 또는 범주 값으로 변환하는 방식이다.
② 범위 방법은 수치 데이터를 임의의 수 기준 범위로 설정하는 기법이다.
③ 랜덤 라운딩은 수치 데이터를 임의의 수 기준으로 올림 또는 내림하는 기법이다.
④ 임의 잡음 추가는 개인 식별이 가능한 정보에 임의의 숫자 등 잡음을 추가하는 기법이다.

2 과목 | 빅데이터 탐색

21 데이터의 정제과정에 관련한 설명으로 올바른 것은?

① 수집된 데이터를 대상으로 초기 분석하여 원하는 결과를 얻어내는 과정이다.
② 정제과정을 거치지 않으면 데이터 구성의 일관성이 없어지므로 분석처리의 어려움이 발생한다.
③ 데이터로부터 원하는 결과나 분석을 얻기 위해서 분석도구나 기법에 상관없이 데이터의 객관성을 확보하는 처리가 필요하다.
④ 후처리 과정이란 도출된 결과를 보정하는 과정으로 정제된 데이터의 신뢰성확보에 필요하다.

22 질적 자료의 설명으로 옳은 것을 고르시오.

① 정량적 자료라고 하며 수치의 크기 자체의 의미를 부여하는 자료를 말한다.
② 서열 자료는 수치나 기호가 서열을 나타내는 자료이다.
③ 명목 자료는 측정대상이 범주나 종류에 대해 구분 되어지는 것을 수치 또는 기호로 분류 될 수 없는 자료이다.
④ 정성적 자료라고 하며 분류가 불가능한 비정형 자료이다.

23 다음은 결측값의 종류에 대한 설명이다. 틀린 설명을 고르시오.

① 완전 무작위 결측은 어떤 변수상에서 결측 데이터가 관측된 혹은 관측되지 않는 다른 변수와 아무런 연관이 없는 경우로 정의한다.
② 결측 데이터를 가진 모든 변수가 완전 무작위 결측(MCAR)이라면 소규모 데이터에서 단순 무작위 표본추출을 통해 처리 가능하다.
③ 무작위결측(MAR)은 변수상의 결측데이터가 관측된 다른 변수와 연관되어 있지만 그 자체가 비관측값들과는 연관되지 않은 경우이다.
④ 비 무작위 결측(NMAR)은 어떤 변수의 결측 데이터가 완전 무작위 결측(MCAR) 또는 무작위 결측(MAR)이 아닌 결측데이터로 정의하는 것이다.

24 다음 보기는 어떠한 대치법(Imputation)에 대한 설명인지 고르시오.

> 평균 대치법에서 추정량 표준오차의 과소 추정을 보완하는 대치법으로 Hot-deck 방법이라고도 한다. 확률추출에 의해서 전체 데이터 중 무작위로 대치하는 방법이다.

① 평균 대치법(Mean Imputation)
② 회귀 대치법(Regression Imputation)
③ 최근방 대치법(Nearest-Neighbor Imputation)
④ 단순확률 대치법(Single stochastic Imputation)

5 다음은 어떠한 변수선택의 설명인가?

- 영 모형에서 시작, 모든 독립변수 중 종속변수와 단순 상관계수의 절댓값이 가장 큰 변수를 분석모형에 포함시키는 것을 말한다.
- 부분 F 검정(F test)을 통해 유의성 검증을 시행, 유의한 경우는 가장 큰 F 통계량을 가지는 모형을 선택하고 유의하지 않은 경우는 변수선택 없이 과정을 중단한다.
- 한번 추가된 변수는 제거하지 않는 것이 원칙이다.

① 후진 선택법(Backward Selection)

② 단계적 선택법(Stepwise Selection)

③ 전진 선택법(Forward Selection)

④ 부분 선택법(Piecewise Selection)

차원축소 필요성에 대한 설명으로 틀린 것은?

① 데이터를 분석하는데 있어서 분석시간의 증가(시간복잡도: Time Complexity)와 저장변수 양의 증가(공간복잡도: Space Complexity)를 고려 시 동일한 품질을 나타낼 수 있다면 효율성 측면에서 데이터 종류의 수를 줄여야 한다.

) 차원이 작은 간단한 분석모델일수록 내부구조 이해가 용이하고 해석이 쉬워진다.

차원의 증가는 분석모델 파라메터의 증가 및 파라메터 간의 복잡한 관계의 증가로 분석결과의 오적합 발생의 가능성이 커진다. 이는 분석모형의 정확도(신뢰도) 저하를 발생시킬 수 있다.

작은 차원만으로 안정적인(robust) 결과를 도출해낼 수 있다면 많은 차원을 다루는 것보다 효율적이다.

27 주성분 분석(PCA: Principal Component Analysis)에 대한 설명으로 틀린 것을 모두 고르시오.

가) 분포된 데이터들의 특성을 설명할 수 있는 하나 또는 복수개의 특징(주성분: Principal Component)을 찾는 것을 의미한다.

나) 서로 연관성이 있는 고차원공간의 데이터를 선형적 관성이 없는 저차원(주성분)으로 변환하는 과정을 거친다(직교변환을 사용).

다) 기존의 기본변수들을 새로운 변수의 세트로 변환하여 차원을 줄이되 기존 변수들의 분포특성을 최대한 보존하여 이를 통한 분석결과의 신뢰성을 확보한다.

라) 차원 축소에 폭넓게 사용된다. 각 차원 간 사전 분포는 독립적인 정규분포를 따른다.

마) 차원의 축소는 본래의 변수들이 서로 독립일 때만 가능하다.

① 가, 마 ② 가, 나

③ 라, 마 ④ 다, 라

28 불균형 데이터에 대한 설명 중 옳은 것은?

① 데이터에서 각 클래스가 갖고 있는 데이터의 질에 차이가 큰 경우, 클래스 불균형이 있다고 말한다.

② 데이터 클래스 비율이 너무 차이가 나면 재현율(recall-rate)이 높아도 데이터 개수가 적은 클래스의 정확도(accuracy)가 급격히 작아지는 현상이 발생할 수 있다.

③ 클래스 균형은 다수의 클래스에 특별히 더 큰 관심이 있는 경우에 필요하다.

④ 클래스에 속한 데이터의 개수의 차이에 의해 발생하는 문제들을 불균형 데이터 문제 또는 비대칭 데이터 문제(Imbalanced Data Problem)이라고 한다.

29 이상치 발견의 통계적 기법 활용에 대한 방법으로 옳은 것은?

① 중앙값은 전체변수의 범위중에서 가운데값을 사용하므로 이상값이 존재하면 영향을 받는다.

② 데이터의 중심을 알기 위해서는 평균(mean), 중앙값(median), 최빈값(mode), 첨도(kurtosis) 를 사용할 수 있다.

③ 데이터의 분산도를 알기 위해서는 범위(range), 분산(variance), 왜도(skewness)를 사용할 수 있다.

④ 평균에는 집합 내 모든 데이터 값이 반영되기 때문에, 이상값이 있으면 값이 영향을 받는다.

30 피어슨 상관계수(Pearson Correlation Coefficient)에 대한 설명으로 옳은 것은?

① 두 변수 X 와 Y 간의 비선형 상관관계를 계량화한 수치이다.

② 두 변수 간의 연관 관계가 있는지 없는지를 밝혀주며 자료에 이상점이 있거나 표본크기가 작을 때 유용하다.

③ 피어슨 상관계수는 +1과 −1 사이의 값을 가지며, +1은 완벽한 양의 선형 상관관계, 0은 선형 상관관계 없음, −1은 완벽한 음의 선형 상관관계를 의미한다.

④ 데이터가 서열자료인 경우 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수로서, 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용해 상관계수를 구한다.

31 100명의 여자에 대한 신장과 체중을 비교한 자료이다. 체중의 개인차가 신장의 개인차보다 크다고 할 수 있는가?

	평균	표준편차
체중	52.3kg	2.54kg
신장	152.7 Cm	2.28cm

① 체중에 대한 개인차가 크다.

② 신장에 대한 개인차가 크다.

③ 체중에 대한 개인차와 신장에 대한 개인차는 동일하다.

④ 체중과 신장의 개인차는 알 수 없다.

32 다음 보기는 공간데이터 용어의 어떤 정의의 설명인가?

> 공간 객체간의 관계를 표현하며, 방위, 공간 객체간의 중첩, 포함, 교차, 분리 등과 같은 위치적 관계

① 비 공간 타입 ② 래스터 공간 타입

③ 벡터 공간 타입 ④ 위상적 공간 타입

33 정준분석의 설명 중 틀린 것은?

① 두 변수집단 간의 연관성(Association)을 각 변수집단에 속한 변수들의 선형결합(Linear Combination)의 상관계수를 이용하여 분석하는 방법이다.

② 정준상관계수(Canonical Correlation Coefficient)는 정준변수들 사이의 상관계수이다.

③ 두 집단에 속하는 변수들의 개수 중에서 변수의 개수가 적은 집단에 속하는 변수의 개수만큼의 정준변수 상이 만들어질 수 있다.

④ 정준분석의 경우 하나의 반응변수를 여러 개의 설명변수로 설명하고자 할 때, 가장 설명력이 높은 변수들의 선형결합을 찾아 이들 사이의 인과관계를 생각하는 방법이다.

34 다음은 표본추출오차에 관한 설명이다. 틀린 것은?

① 최대 대표는 모집단에서 추출된 표본이 너무 많이 추출되어 전수조사에 가까운 조사가 되는 현상이다.

② 표본추출 시 표본의 크기(Sample Size)보다는 대표성을 가지는 표본을 추출하는 것이 중요하다.

③ 과잉 대표는 중복선택 등의 원인으로 모집단이 반복·중복된 데이터만으로 규정되는 현상을 지칭한다.

④ 최소 대표는 실제모집단의 대표성을 나타낼 표본이 아닌 다른 데이터가 표본이 되는 현상이다.

35 다음 중 포아송분포를 적용할 수 있는 예가 아닌 것을 고르시오.

① 10시부터 11시사이에 은행지점창구에 도착한 고객의 수

② 하루 동안 걸려오는 전화수

③ 원고집필 시 원고지 한 장당 오타의 수

④ 금융상품 가입 상담 건수 10회중 실제 가입이 이루어진 수

36 스튜던트 t 분포에서 자유도에 대한 설명으로 맞는 것은?

① 자유도는 자료집단의 변수 중에서 자유롭게 선택될 수 있는 변수의 수를 말한다.

② 스튜던트 t 분포는 정규분포의 평균 측정 시 주로 사용하는 분포이다. 분포의 모양은 Z-분포와 유사하며 t-곡선의 대칭/비대칭 여부를 결정하는 것은 자유도이다.

③ 자유도가 클수록 정규분포보다 더욱 높은 종모양을 가지게 된다.

④ 자유도가 1보다 클 때만 스튜던트 t 분포에서 기대값은 1이다.

37 편향에 대한 설명으로 틀린 것은?

① 기대하는 추정량과 모수의 차이를 편향(bias)이라고 한다.

② 임의의 추정량의 편향을 $B(\hat{\theta})$이라고 하면 $B(\hat{\theta}) = E(\hat{\theta}) - \theta$로 정의할 수 있다.

③ 불편추정량(Unbiased Estimator)은 $B(\hat{\theta})=0$ 즉, 편향이 0이 되는 상황의 추정량 $\hat{\theta}$을 불편추정량이라고 한다.

④ 표본 평균은 이상치의 영향으로 값의 변화가 커지므로 대표적인 불편추정량이 아니다.

38 모평균에 대한 신뢰구간에 대한 각 상황 별 정리이다. 옳은 것을 모두 고르시오.

	구분	신뢰구간 $100(1-\alpha)\%$
(가)	모집단의 분산을 아는 경우	$\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
(나)	모집단의 분산을 모르는 경우 (표본크기가 작은 경우)	$\bar{X} - t_{\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}}$
(다)	모집단의 분산을 모르는 경우 (표본크기가 큰 경우)	$\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$

① 가 ② 가, 나

③ 가, 다 ④ 가, 나, 다

39 가설검정에 대한 설명으로 옳은 것은?

① 연구자에 의해 설정된 가설은 모집단 전체를 근거로 하여 채택여부를 결정짓게 되는데 이때 사용되는 통계량을 검정통계량이라 정의한다.

② 귀무가설(Null Hypothesis, H_0)은 연구자가 모수에 대해 새로운 통계적 입증을 이루어 내고자 하는 가설이다.

③ 검정통계량의 표본분포에 따라 채택여부를 결정짓는 일련의 통계적 분석과정을 가설검정이라 하며 일반적으로 몇 단계의 절차를 거쳐 검정이 수행된다.

④ 대립가설(Alternative Hypothesis, H_1) 현재 통념적으로 믿어지고 있는 모수에 대한 주장 또는 원래의 기준이 되는 가설이다.

40 두 독립표본(각 n, m 표본수) 사이의 평균차이의 검정을 하기 위한 검정 통계량 식으로 옳은 것은?

① 검정 통계량 $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$ 여기서

$S_p^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$ 으로 공통분산 σ^2의 합동표본분산이며 S_1^2, S_2^2는 각각의 표본의 표본분산을 말한다. 검정 통계량 T는 자유도 m+n−2인 t 분포를 따른다.

② 검정 통계량 $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$ 여기서

$S_p^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m+2}$ 으로 공통분산 σ^2의 합동표본분산이며 S_1^2, S_2^2는 각각의 표본의 표본분산을 말한다. 검정 통계량 T는 자유도 n+m+2인 t 분포를 따른다.

③ 검정 통계량 $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{n}{m}+\frac{1}{n}}}$ 여기서

$S_p^2 = \frac{(n-1)S_1^2+(n-1)S_2^2}{n+m-2}$ 으로 공통분산 σ^2의 합동표본분산이며 S_1^2, S_2^2는 각각의 표본의 표본분산을 말한다. 검정 통계량 T는 자유도 n+m−2인 t 분포를 따른다.

④ 검정 통계량 $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{m}{n}+\frac{1}{m}}}$ 여기서

$S_p^2 = \frac{(m-1)S_1^2+(m-1)S_2^2}{n+m-2}$ 으로 공통분산 σ^2의 합동표본분산이며 S_1^2, S_2^2는 각각의 표본의 표본분산을 말한다. 검정 통계량 T는 자유도 n+m−2인 t 분포를 따른다.

3 과목 | 빅데이터 모델링

41 지도학습 모델은 분류와 예측모델로 구분되는데 분류모델에 해당되지 않는 것은?

① 의사결정트리
② 인공신경망
③ 서포트 벡터 머신(SVM)
④ 다중회귀분석

42 로지스틱 회귀분석이 갖는 선형 회귀분석과 비교 시 차이점으로 맞는 설명은?

① 종속변수: 연속형 변수, 분포: 정규분포
② 종속변수: 범주형 변수, 분포: 정규분포
③ 종속변수: 범주형 변수, 분포: 이항분포
④ 종속변수: 연속형 변수, 분포: 이항분포

43 다중회귀분석 결과를 해석할 시 진행순서가 올바른 것은?

① 다중공선성 진단 → 모형의 적합도 평가 → 회귀계수 유의성 확인 → 수정된 결정계수 확인
② 수정된 결정계수 확인 → 모형의 적합도 평가 → 회귀계수 유의성 확인 → 다중공선성 진단
③ 모형의 적합도 평가 → 회귀계수 유의성 확인 → 수정된 결정계수 확인 →다중공선성 진단
④ 다중공선성 진단 → 회귀계수 유의성 확인 → 수정된 결정계수 확인 → 모형의 적합도 평가

44 정보이론에서 순도가 증가하고 불확실성이 감소하는 것을 의미하는 용어는?

① 재귀적 분기 ② 가치치기
③ 정보 획득 ④ 엔트로피 지수

45 의사결정나무의 대표적 알고리즘인 CART (Classification and Regression Tree)는 불순도 측도로 범주형 또는 이산형일 경우 (ㄱ)를, 연속형인 경우 분산의 감소량을 이용한 (ㄴ)를 활용한다. 빈칸에 맞는 말을 고른다면?

① ㄱ 엔트로피 지수, ㄴ 다지분리
② ㄱ 지니 지수, ㄴ 다지분리
③ ㄱ 지니 지수, ㄴ 이진분리
④ ㄱ 엔트로피 지수, ㄴ 이진분리

46 여러 부트스트랩 자료를 생성하여 학습하는 모델링으로 랜덤포레스트가 속한 알고리즘 기법은?

① 부스팅 ② 배깅
③ 앙상블 ④ 의사결정트리

47 한 회사의 직원 3명의 메일함에서 스팸 메일들을 분류할 때 앙상블 값이 맞는 경우는?

	김철수	나윤아	이성희
나이브 베이지안	1	0	0
KNN	0	0	1
SVM	1	0	1
의사결정나무	1	1	1

① 김철수, 앙상블 값 = 3/4
② 나윤아, 앙상블 값 = 1/4
③ 이성희, 앙상블 값 = 1
④ 정답이 없음

48 다음 빈칸에 들어갈 단어로 맞는 것을 고른다면?

(ㄱ)함수는 신경망이 출력한 값과 실제 값과의 오차에 대한 함수로 손실 함수값이 최소화되도록 하기 위해 가중치와 (ㄴ)을 찾는 것이 인공신경망의 학습이라고 하며 일반적인 손실 함수로는 (ㄷ)오차 또는 교차 엔트로피 오차를 활용한다.

① ㄱ 손실, ㄴ 교차점, ㄷ 평균
② ㄱ 오차, ㄴ 편향, ㄷ 평균제곱
③ ㄱ 손실, ㄴ 편향, ㄷ 평균제곱
④ ㄱ 오차, ㄴ 교차점, ㄷ 평균

49 일반적인 신경망 알고리즘 학습 프로세스 순서로 적합한 것은?

① 미니배치 – 가중치 매개변수 기울기 산출 – 매개변수 갱신
② 샘플선정 – 가중치 매개변수 기울기 산출 – 매개변수 갱신
③ 미니배치 – 매개변수 갱신 – 가중치 매개변수 기울기 산출
④ 샘플선정 – 매개변수 갱신 – 가중치 매개변수 기울기 산출

50 가중치 매개변수의 기울기를 미분을 통해 진행하는 것은 시간비용이 크므로 오차를 출력층에서 입력층으로 전달, 연쇄법칙을 활용하여 가중치와 편향을 계산, 업데이트하는 기법을 칭하는 것은?

① 퍼셉트론
② 활성화함수
③ 확률적 경사하강법
④ 오차역전파

51 다음 (1, 3), (4, 4)간의 유클리디안 거리 값을 계산, 선택한다면?

① $\sqrt{5}$ ② $\sqrt{10}$
③ $\sqrt{12}$ ④ $\sqrt{16}$

52 분류모델이 틀린 곳에 집중하여 새로운 분류규칙을 생성, 즉 weak classifier에 중점을 두는 지도학습 알고리즘은?

① 부스팅 ② 배깅
③ 랜덤포레스트 ④ 회귀분석

53 활성화함수 종류 중 0보다 크면 입력값을 그대로 출력 0 이하의 값만 0으로 출력하는 함수명은?

① Sigmoid(시그모이드)
② Relu(렐루)
③ Softmax(소프트맥스)
④ Perceptron(퍼셉트론)

54 랜덤포레스트는 여러 개의 의사결정 나무를 활용, 예측 결과를 ( ) 방식으로 예측 결정한다. 빈칸에 적합한 용어는?

① 투표 ② 평균
③ 분류 ④ 군집

55 기저귀와 맥주 간 support, confidence, lift 값은?

TID	Items
1	빵, 우유
2	빵, 기저귀, 맥주, 달걀
3	우유, 기저귀, 맥주, 콜라
4	빵, 우유, 기저귀, 맥주
5	빵, 우유, 기저귀, 콜라

① $\frac{3}{5}, \frac{2}{5}, \frac{3}{4}$
② $\frac{4}{5}, \frac{3}{5}, \frac{2}{3}$
③ $\frac{3}{5}, \frac{3}{4}, \frac{5}{4}$
④ $\frac{2}{5}, \frac{5}{4}, \frac{3}{5}$

56 다음은 범주형 분석방법에 대한 설명이다 옳지 않은 것은?

① 빈도분석은 질적 자료를 대상으로 빈도와 비율을 계산할 때 쓰인다.
② 로지스틱분석은 두 범주형 변수가 서로 상관이 있는지 독립인지를 판단하는 통계적 검정방법이다.
③ T 검정은 독립변수가 범주형(두개의 집단)이고 종속변수가 연속형인 경우 사용되는 검정 방법으로 두 집단간의 평균 비교 등에 사용된다.
④ 독립변수가 범주형(두개 이상 집단)이고 종속변수가 연속형인 경우 사용되는 검정 방법으로 분산분석이 사용된다.

57 다음은 시계열 자료의 정상성(Stationarity 定常性)에 대한 설명이다. 틀린 것을 고르시오.

① 정상성을 가진다는 의미는 시계열 데이터가 평균과 분산이 일정한 경우를 지칭한다.
② 시계열 데이터가 정상성을 가지면 분석이 용이한 형태로 볼 수 있다.
③ 시계열 데이터가 평균이 일정하지 않으면 차분(difference)을 통해 정상성을 가지도록 할 수 있다.
④ 시계열 데이터가 분산이 일정하지 않으면 평행이동을 통해 정상성을 가지도록 할 수 있다.

58 다음은 인공신경망과 딥러닝에 대한 설명이다. 틀린 것은?

① 딥러닝은 인공신경망의 단점(계산속도의 저하, 과적합 문제) 등이 극복되면서 부각된 기계학습이라고 할 수 있다.
② 딥러닝은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계학습 알고리즘의 집합으로 정의된다.
③ 소수의 동일레이어 내의 노드의 수직체계 개수를 다수로 늘려서 정확도를 높이는 것이 기존 인공신경망과 딥러닝의 차이이다.
④ 인공신경망과 딥러닝은 사람의 사고방식을 컴퓨터에게 가르치는 기계학습의 한 분야라고 이야기할 수 있다.

59 다음은 어떤 딥러닝에 대한 설명이다. 아래 설명에 해당되는 딥러닝 알고리즘은 무엇인가?

- 인공신경망을 구성하는 유닛 사이의 연결이 Directed cycle을 구성하는 신경망을 말한다.
- 앞먹임 신경망(Feed forward Neural Network)과 달리, 임의의 입력을 처리하기 위해 신경망 내부의 메모리를 활용할 수 있다.
- 필기체 인식(Handwriting recognition)과 같은 분야에 활용되고 있고, 높은 인식률을 나타낸다
- 기존의 뉴럴 네트워크와 다른 점은 '기억'을 갖고 있다는 점인데, 네트워크의 기억은 지금까지의 입력 데이터를 요약한 정보라고 볼 수 있다.

① 합성곱신경망(Convolutional Neural Network, CNN)
② 순환신경망(Recurrent Neural Network, RNN)
③ 심층신뢰신경망(Deep Belief Network, DBN)
④ 심층신경망(Deep Neural Network, DNN)

60 앙상블 기법에 대한 설명으로 틀린 것은 무엇인가?

① 약학습기(Weak Learner)는 무작위 선정이 아닌 성공확률이 높은 즉 오차율이 일정 이하(50% 이하)인 학습 규칙을 말한다.
② 강학습기(Strong Learner)는 약학습기로부터 만들어내는 강력한 학습 규칙을 의미한다.
③ 앙상블 기법은 서로 다른 학습 알고리즘을 경쟁시켜 각 알고리즘 간의 장점를 결합하여 학습하는 개념이다.
④ 한 개의 Single Leaner에 의한 분석보다는 더 나은 분석성능을 이끌어 낼 수 있으며 다양한 Weak Learner를 통해 Strong Learner를 만들어가는 과정이다.

61 분류모델을 평가하는 지표에 대한 설명으로 거리가 먼 것은?

① 정확도는 True인 데이터를 모델에서 True로 분류하는 정도를 말한다.
② 정밀도는 True로 분류한 대상 중에 실제로 True인 비율을 말한다.
③ 예측 성능을 측정하기 위해 예측값과 실제값을 비교한 표를 오차행렬이라고 한다.
④ AUC는 ROC 곡선 하단영역의 넓이를 구한 값으로 0~1 사이의 값을 갖는다.

62 다음 보기와 같이 실제값과 예측값이 존재할 때 평균제곱오차는 얼마인가?

실제값	예측값
10	11
13	12
17	19
21	23

① 10
② 2.5
③ −4
④ 6

63 군집분석 모델을 평가하기 위한 고려사항으로 거리가 먼 것은?

① 같은 군집내에 속한 요소가 군집의 중심으로부터 가깝게 분포할 때 좋은 모델이다.
② 군집과 이웃군집 사이의 거리가 멀수록 좋은 모델로 평가할 수 있다.
③ 같은 군집에 속한 요소들의 평균 거리를 실루엣 계수라 한다.
④ 군집의 수가 많아질수록 군집내 속한 요소들 간의 거리는 줄어든다.

64 '동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다'는 이론은 다음 중 무엇인가?

① 중심극한정리
② 평균과 표준편차
③ 베이즈 정리
④ 교차 검증

65 회귀분석의 잔차진단 방법 중 잔차의 분산이 특정 패턴이 없이 순서와 무관하게 일정한지 진단하는 방법은 다음 중 어느 것인가?

① 정규성 진단
② 등분산성 진단
③ 독립성 진단
④ 평균성 진단

66 모델의 과대적합 방지를 위한 기법에 해당되지 않는 것은?

① 드롭아웃
② L2 규제
③ 편향-분산 트레이드오프
④ 경사하강법

67 다음 중 군집분석의 타당성 지표로 적당하지 않은 것은?

① 군집간 거리
② 군집의 지름
③ 군집의 분산
④ 군집의 평균

68 두 종류 이상의 결과변수를 동시에 분석할 수 있는 방법으로 결과 변수 간의 유의성, 관련성을 설명할 수 있는 방법은 다음 중 어느 것인가?

① 앙상블 학습
② 결합분석 모형
③ 매개변수 최적화
④ 경사하강법

69 다음 중 분류모델에 대한 시각화 방법으로 적절하지 않은 것은?

① 히트맵
② SVM
③ KNN
④ 의사결정트리

70 다음 분석모형 해석에 관한 설명 중 맞지 않는 것은?

① 분석 후 적합한 모형을 도출하는데 사용되는 지표는 설명력, 오차율, 인자수, 잔차 등이 있다.
② 딥러닝에서의 적합 모형 해석은 분류문제인 경우 정확도나 오차율을 사용한다.
③ 연관분석 모델은 두 개 또는 그 이상의 품목들 사이의 상호 관련성으로 해석한다.
④ 군집분석 모델은 군집그룹에 속한 요소들 사이의 거리 평균을 사용한다.

71 분석모델별 활용되는 시각화 기법 연결로 잘못된 것은?

① SVM – 산점도
② KNN – 평행좌표계
③ 연관분석 – 파이차트
④ 군집분석 – 산점도

72 다음 중 데이터 시각화에 대한 설명으로 잘못된 것은?

① 데이터 시각화는 데이터의 특징을 직관적으로 제공한다.
② 데이터의 시각적 속성으로는 위치, 형태, 크기 등이 있다.
③ 데이터의 분포를 시각적으로 보여주는데 유용한 도구이다.
④ 비정형데이터는 구조화하기 어렵기 때문에 정형데이터로 변환하여 시각화해야 한다.

73 다음 보기의 개념을 가장 정확하게 설명하는 개념은 어느 것인가?

> 주로 뉴스 기사의 그래픽에 사용되며 복잡한 정보와 지식을 차트, 지도, 다이어그램, 일러스트레이션 등을 활용해 한눈에 파악할 수 있도록 시각적으로 표현한다.

① 인포그래픽
② 정보디자인
③ 시각적 분석
④ 데이터추상화

74 다음 보기에서 설명하는 용어는 무엇인가?

> 그래프나 차트에서 사용되는 기호나 선 등이 어떤 의미인지 설명하는 역할을 함

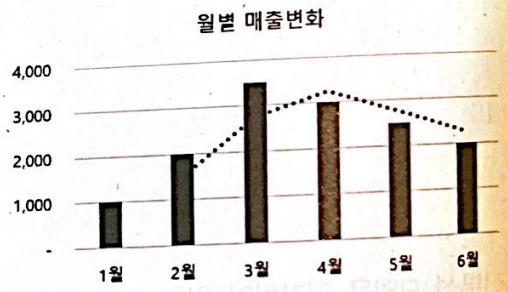
① 축
② 범례
③ 스케일
④ 스코프

75 다음 보기의 특징을 갖는 시각화 방법은 무엇인가?

> • 일정 기간에 걸쳐 진행되는 변화를 표현하기에 적합
> • 막대의 영역을 구분하여 두 개 이상의 변수를 동시에 표현

① 막대그래프
② 누적막대그래프
③ 히스토그램
④ 파이차트

76 기업의 월별 매출을 보여주는 막대그래프이다(추세선은 현재월과 전월의 이동평균값을 보여준다). 다음 중 그래프를 통해서 이해할 수 있는 정보로 적절하지 않은 것은?



① 1월 매출보다 2월 매출이 증가하였다.
② 3월 총매출은 3,000 이상이다.
③ 1월부터 6월까지 매출의 평균은 1,000이상이다.
④ 추세선이 막대보다 위에 있는 경우 다음달 매출이 줄어들 것을 예측할 수 있다.

77 다음 중 데이터 시각화 도구에 대한 설명으로 맞지 않는 것은?

① 트리맵은 사각형 영역을 사용하여 데이터 분포를 시각화하는데 적합하다.
② 산점도는 점들의 분포에 따라 집중도를 확인할 수 있다.
③ 도수분포표는 일정한 간격으로 구분된 구간에 대해 데이터의 분포를 표현하는데 적합하다.
④ 파이차트는 시간에 따른 데이터의 변화를 표현하는데 적합하다.

78 데이터 시각화에 대한 다음 설명 중 가장 거리가 먼 것은 어느 것인가?

① 데이터 시각화는 분석된 결과를 해석하는 대표적인 방법이다.
② 공간시각화를 위한 대표적인 도구는 카토그램이 있다.
③ 누적막대그래프는 이산형 데이터를 표현하는데 적합하다.
④ 데이터 시각화를 통해서 데이터의 결측치를 효율적으로 발견할 수 있다.

79 다음 보기에서 설명하는 CRISP-DM 데이터 분석 프로세스는 무엇인가?

- 분석 모델을 실제 운영 데이터에서 동작시킨다.
- 분석 모델 수정이 이루어진다.

① 데이터 준비 ② 모델링
③ 평가 ④ 전개

80 분석 프로젝트 성과 평가에서 이루어지는 활동으로 거리가 먼 것은?

① 분석 프로젝트의 성과 지표는 정량적, 정성적 지표를 동시에 고려한다.
② 목표치를 달성하기 위해서 분석 모델을 리모델링 한다.
③ 성과가 목표치보다 부족한 경우 분석과제의 개선사항을 검토한다.
④ 성과평가 결과를 관련 부서 및 조직과 공유한다.