1과목

1) SELECT FROM WHERE age ( ) 20 AND 39 -> Between

2) 데이터 사이언티스트가 갖춰야 할 역량은 빅데이터의 처리 및 분석에 필요한 이론적 지식과 기술적 숙련에 관련된 능력인 ( ) Skill과 데이터 속에 숨겨진 가치를 발견하고 새로운 발전 기회를 만들어 내기 위한 능력인 ( ) Skill로 나누어진다 -> 하드, 소프트

3) ( )는 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 그 의미를 부여한 것이며, 지식을 도출하기 위한 재료가 된다. -> 정보

4) 기업의 의사결정 과정을 지원하기 위한 주제 중심적으로 통합적이며 시간성을 가지는 비휘발성 데이터의 집합을 ( )라 한다. -> 데이터웨어하우스

5) 별도로 정제되지 않은 자연스러운 상태의 아주 큰 데이터 세트인 ( )을 기업들이 구현하는 것은 2017년 새롭게 등장한 트랜드가 아니다. -> 데이터 레이크

6) ( ) 데이터는 지역별 매출액, 영업이익률, 판매량과 같이 수치로 명확하게 표현되는 데이터로, 그 양이 크게 증가하더라도 이를 DBMS에 저장, 검색, 분석하여 활용하기가 용이하다. -> 정량적 데이터

7) ( )은 기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화시키기 위한 것으로, 자재구매, 생산, 제고, 유통, 판매, 고객 데이터로 구성된다. -> SCM(Supply Chian Management)

8) 가) 페이스북은 자신들의 소셜 그래프 자산을 외부 개발자들에게 공개하고 서드파티 개발자들이 페이스북 위에서 작동하는 앱을 만들기 시작하면서 ( ) 역할을 하기 시작했다.
나) 하둡은 대규모 분산 병렬 처리의 업계 표준으로 맵리듀스 시스템과 분산 파일 시스템인 HDFS로 구성된 ( ) 기술이며, 아마존은 S3와 B2C 환경을 제공함으로써 ( )를 위한 클라우드 서비스를 최초로 실현하였다. -> 플랫폼

9) 가) 생명의 진화를 모강하여 최적해를 구하는 알고리즘으로 존 홀랜드가 1975년에 개발하였다.
나) 최대의 시청률을 얻으려면 어느 시간대에 방송해야 하는가?와 같은 문제해결에 사용된다.
다) 어떤 미지의 함수 Y=f(x)를 최적화하는 해x를 찾기 위해, 진화를 모방한 탐색 알고리즘이라고 말할 수 있다. -> 유전자 알고리즘

2과목

1) 분석 방법론의 “시스템 구현”단계에서 시스템으로 구현된 모델은 검증을 위하여 단위테스트, 통합테스트, 시스템테스트 등을 실시한다. 이 중 ( ) 테스트는 품질관리 차원에서 진행함으로써 적용된 시스템의 객관성과 안정성을 확보한다. -> 시스템

2) 저장소는 데이터 관리 체계 지원을 위한 워크플로우 및 관리용 응용소프트웨어를 지원하고 관리 대상 시트템과의 인터페이스를 통한 통제가 이루어져야한다. 또한 데이터 구조 변경에 따른 ( )도

수행되어야 효율적인 활용이 가능하다. -> 사전 영향 평가

3) 문제 탐색을 통해 식별된 비즈니스 문제를 변환하는 단계로써, 문제 탐색 단계가 무엇을 어떤 목적으로 수행해야 하는가에 대한 관점이었다면, ( ) 단계는 이를 달성하기 위해 필요한 데이터 및 기법(How)를 도출하기 위한 데이터 분석의 문제로의 변환을 수행하게 된다. -> 문제 정의

4) 분석 모델을 가동중인 운영 시스템에 적용하기 위해서는 모델에 대한 상세한 “알고리즘 설명서” 작성이 필요하다. “알고리즘 설명서”는 “시스템 구현”단계에서 중요한 입력자료로 활용되므로 필요시 ( ) 수준의 상세한 작성이 필요하다. -> 의사 코드

5) 분석과제 관리 푸로세스는 크게 과제 발굴과 ( )로 나누어진다. 조직이나 개인이 도출한 분석 아이디어를 발굴하고 이를 과제화하여 분석과제 풀(pool)로 관리하면서 분석과제가 확정되면 ( ), ( ), 분석과제 결과 공유/개선의 분석과제 프로세스를 수행하게 된다. -> 과제수행/탐구성, 분석과제 실행, 분석과제 진행관리

6) 저장소는 데이터 관리 체계 지원을 위한 ( ) 및 관리용 응용프로세스를 지원하고 관리 대상 시스템과의 인터페이스를 통한 통제가 이루어져야 한다. 또한 데이터 구조 변경에 따른 ( )도 수행되어야 효율적인 활용이 가능하다. -> 워크플로우/사전영향 평가

7) 비즈니스 모델 캔버스는 9가지 블록을 단순화하여 ( ), ( ), 고객단위로 문제를 발굴하고 이를 관리하는 규제와 감사, ( ) 영역으로 나눠 분석기회를 도출한다. -> 업무/제품/지원인프라

8) KDD 분석 방법론에서 잡음, 이상치, 결측치를 식별하여 분석용 데이터셋을 선택하고 분석에 필요한 변수 등을 선정하는 단계와 유사한 CRISP-DM 방법론의 단계는? -> 데이터 준비

9) 합리적 의사결정을 반대하는 요소로써 표현방식 및 발표자에 따라 동일한 사실에도 판단을 달리하는 현상을 무엇이라 하는가? -> 프레이밍 효과

10) 문제가 주어지고 이에 대한 해법을 찾기 위하여 각 과정이 체계적으로 단계화 되어 수행하는 분석과제 발굴 방식은? -> 하향식 접근 방식

11) 빅데이터 분석 프로세스에서 빅데이터 분석단계중 어떤 것에 대한 설명인가? / 분석용 데이터를 이용한 가설 설정을 통하여 통계모델을 만들거나 기계학습을 이용한 데이터의 분류, 예측, 군집 등의 기능을 수행하는 모델을 만드는 과정 -> 모델링

12) ( ) 모델은 반복을 통하여 점증적으로 개발하는 방법으로 처음 시도하는 프로젝트에 적용이 용이하지만, 반복에 대한 관리체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있다. -> 나선형

13) 소프트웨어와 시스템 공학의 역량 숙성도를 측정하기 위한 모델로 소프트웨어 품질보증과 시스템 엔지니어링 분야의 품질보증 기술을 통합하여 개발된 평가모델로 1~5단계로 구성된 성숙도 모델은?

->CMMI / 능력 성숙도 통합모델

14) 기업 및 공공기관에서는 시스템의 중장기 로드맵을 정의하기 위한 ( )을 수행한다. ( )은 정보기술 또는 정보시스템을 전략적으로 활용하기 위하여 조직 내외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터플랜을 수립하는 절차이다. -> ISP (Information Strategy Planning)

3과목

1) 공간적 차원과 관련된 속성들을 시각화에 추가하여 지도 위에 관련 속성들을 생성하고 크기, 모양, 선굵기 등으로 구분하여 인사이트를 얻는 분석방법은? -> 공간분석

2) 아래 R 코드의 출력 결과는? -> 4 1

▶ f <- function(x,a) return((x-a)^2)

▶ f(1:2,3)

3) R에서 다음의 명령을 수행했을 때 출력되는 결과는? -> NA

```
x <- c(1,2,3,NA)
mean(x)
```

4) R에서 다음의 명령을 수행했을 때 출력되는 결과는? -> 50

```
x<-1:100
sum(x>50)
```

5) A반과 B반 학생들이 동일한 과목을 들었다고 하자. A반과 B반 학생 모두를 대상으로 과목 별 성적의 평균을 구하려고 할 때, A반 학생 데이터와 B반 학생 데이터를 Class 라는 변수를 기준으로 합치려고 한다. R로 프로그래밍 하시오. -> merge(A,B,by="class")

6) 데이터 프레임명은 test라고 할 때, 경영학과 학생들의 데이터만 조회하고자 한다. R로 프로그래밍하시오. -> subset(test,subset=(학과==경영학과))

7) SQL을 활용하거나 SAS에서 PROC SQL로 작업하던 사용자들에게 R프로그램에서 지원해주는 패키지는 무엇인가? -> sqldf()

8) 평균으로부터 t Standard Deviation 이상 떨어져있는 값을 이상값으로 판단하고 t는 3으로 설정하는 이상값 검색 알고리즘은? -> ESD(Extreme Studentized Deviation)

9) 최적방정식을 선택하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수를 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하는 방법은? -> 후진제거법

10) 아래는 College 데이터의 Grad.Rate 변수의 기초통계량을 계산한 결과이다. Colledge 데이터의

Grad.Rate 변수의 몇 %가 78보다 큰 값을 가지는가? -> 25%
>summary(College$Grad.Rate)
min 10.00 / 1st Qu. 53.00 / Median 65.00 / Mean 65.46 / 3rd Qu. 78.00 / Max. 118.00

11) 아래 주성분분석의 결과에서 두 개의 주성분을 사용할 때 설명가능한 전체 분산의 비율은? -> 68.4%

12) 아래 회귀분석 모형의 추정에 대한 설명에서 ( )은? / 단순회귀분석 모형을 yi = B0+B1xi+εi로 표현할 수 있다. 주어진 자료를 가장 잘 설명하는 회귀계수의 추정치는 보통 제곱오차를 최소로 하는 값을 구한다. 이와같이 구해진 회귀계수 추정량을 ( )라고 한다. -> 최소제곱

13) 아래의 표본추출방방법은 무엇인가? / 번호를 부여한 샘플을 나열하여 k개씩 n개의 구간을 나누고 첫 구간에서 하나를 임의로 선택한 후에 k개씩 띄어서 표본을 선택하고 매번 k번째 항목을 추출하는 표본 추출 방법 -> 계통 추출 방법 (Systematic Sampling)

14) 아래의 설명은 어떤 오류에 관한 설명인가? / 귀무가설(H0)이 옳은데 귀무가설을 받아들이지 않고 기각하게 되는 오류 -> 제 1종 오류

15) 아래는 단순 로지스틱 회귀모형이다. “exp()의 의미는 x1,x2,...,xk가 주어질 때, x1이 한 단위 증가할 때마다 성공(y=1)의 ( )가 몇 배 증가하는 지를 나타내는 값이다.”( )는 무엇인가? -> 오즈비

16) 시계열의 수준과 분산에 체계적인 변화가 없고 엄밀하게 주기적 변동이 없다는 것으로 미래는 확률적으로 과거와 동일하다는 것을 의미하는 시계열 용어는? -> 정상성

17) 시계열 모형의 여러 종류 중 아래에서 설명하는 것은 무엇인가? -> 자기회귀모형 (AR모형)
 가) 시계열 모형 중 자기 자신의 과거 값을 사용하여 설명하는 모형임.
 나) 백색 잡음의 현재값과 자기 자신의 과거값의 선형 가중합으로 이루어진 정상 확률 모형
 다) 모형에 사용하는 시계열 자료의 시점에 따라 1차, 2차, ... p차 등을 사용하나 정상시계열 모형에서는 주로 1,2차를 사용함.

18) 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법은? -> 분해시계열

19) 최적회귀방정식을 선택하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을때까지 설명변수를 제거하는 방법은? -> 후진제거법(Backward Elimination)

20) 아래의 설명이 나타내는 척도는 무엇인가? / 자료의 위치를 나타내는 척도의 하나로 관측치를 크기순으로 배열했을 때 전체의 중앙에 위치한 수치이다. 평균에 비해 이상치에 의한 영향이 적기 때문에 자료의 분포가 심하게 비대칭인 경우 중심을 파악할 때 합리적인 방법이다. -> 중앙값

21) 모형평가방법 중 주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차 검정을 실시하는 방법으로 하나는 모형의 학습 및 구축을 위한 훈련용 자료로, 다른 하나는 성과 평가를 위한 검증용 자료로 사용하는 방법은 무엇인가? -> 홀드아웃방법

22) 분류문제를 예측하기 위한 모형을 개발하여 그 결과를 분석하고자 할 때, 특이도를 산출하는 방식을 나타내시오-> TN/TN+FP

23) 베이즈 정리와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전정보와 데이터로부터 추출된 정보를 결합하고 베이즈 정리를 이용하여 어떤 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘은 무엇인가? -> 나이브 베이지안 분류

24) 표준화 지수 함수라고 불리며, 출력값 z가 여러개로 주어지고, 목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공하여 출력노드에 주로 사용되는 함수는? -> softmax 함수

25) 두 개체 간의 거리에 기반하여 군집을 형성해가는 계층적 군집방법에서 사용되는 측도 중 두 개체의 벡터 내적을 기반하여 아래의 수식으로 계산할 수 있는 유사성 측도는 무엇인가? -> 코사인 유사도

26) 아래는 오분류표를 나타낸 것이다. F1값을 구하시오. -> 2*(정확도*민감도/정확도+민감도)
<성과분석>

TP	FP
FN	TN

정확도=TP/TP+FP, 민감도=TP/TP+FN

27) 혼합분포군집(Mixture Distribution Clustering)은 모형 기반의 군집 방법으로서 데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정 하에서 분석을 하는 방법이다. k개의 각 모형은 군집을 의미하며 이 혼합 모형의 모수와 가중치의 최대가능도 추정에 사용되는 알고리즘은 무엇인가? -> EM(Expectation=Maximization)알고리즘

28) 군집분석의 품질을 정량적으로 평가하는 대표적인 지표로 군집 내의 데이터 응집도와 군집간 분리도를 계산하여 군집 내의 데이터가 짧을수록, 군집 간 거리가 멀수록 값이 커지며 완벽한 분리일 경우 1의 값을 가지는 지표는? -> 실루엣

29) SOM(Self-Organizing Maps)에서는 각 학습 단계마다 입력층의 데이터 집합으로부터 하나의 표본 벡터를 임의로 선택하고 경쟁층의 프로토타입 벡터와의 거리를 계산하고 가장 가까운 프로토타입 벡터를 선택하는데 이때 선택된 프로토타입 벡터를 나타내는 용어는? -> BMU (Best-Matching Unit)

30) 랜덤 모델과 비교하여 해당 모델의 성과가 얼마나 좋아졌는지를 각 등급별로 파악하는 그래프로 상위등급에서 매우 크고 하위 등급으로 갈수록 감소하게 되면 일반적으로 모형의 예측력이 적절하다고 판단하게 된다. 모형 평가에 사용되는 이 그래프는? -> 향상도 곡선

31) 맨하탄 거리를 이용하여 군집분석을 하고자 한다. 맨하탄 거리를 이용하여 A와 B의 거리를 구하시오. -> |2-1|+|4-5|=2

사람	(키, 몸무게)
A	(1, 5)
B	(2, 4)

모의고사/기출 주관식

1) 개인의 사생활 침해를 방지하고 통계 응답자의 비밀사항은 보호하면서 통계자료의 유용성을 최대한 확보할 수 있는 데이터변환 방법은 무엇인가? -> 마스킹

2) 데이터 ( )이란 데이터베이스 내의 데이터에 대한 정확성, 일관성, 유효성, 신뢰성을 보장하기 위해 데이터 변경 혹은 수정 시 여러 가지 제한을 두어 데이터의 정확성을 보증하는 것을 말한다. -> 무결성

3) 데이터 분석 기획을 위해서 데이터 분석 수준진단이 필요하다. 분석 준비도와 분석 성숙도를 통해 데이터 분석 수준을 진단하게 되는데, 분석준비도 6개의 영역 중 2가지를 적으시오.
-> 분석 업무, 분석 인력/조직, 분석 기법, 분석 데이터, 분석 문화, 분석 인프라

4) 비즈니스 모델 캔버스는 9가지 블록을 단순화하여 ( ), ( ), 고객단위로 문제를 발굴하고 이를 관리하는 규제와 감사, ( )영역으로 나눠 분석 기회를 도출한다.
-> 업무, 제품, 지원 인프라

5) 이것은 데이터 안의 두 변수 간의 관계를 알아보기 위해 사용하는 값이다. 두 변수간의 공분산으로는 음과 양의 관계를 파악할 수 있으나 관계 정도를 확인하기는 힘들다. 그래서 각 변수의 공분산을 표준편차의 곱으로 나누어 -1에서 1사이 값으로 표준화하여 두 변수 간의 관계 정도를 확인할 수 있도록 수치화 한 이것을 활용한다. 이것은 무엇인가?->상관계수

6) 아래의 R 코드의 출력 결과는?
```
> f <- function(x,a) return((x-a)^2)
> f(1:2,3)
```
-> 4 1

7) 우리는 모집단을 조사하기 위해 추출한 모집단의 일부 원소를 이용한다. 통계자료의 획득 방법 중 모집단을 조사하기 위해 추출한 집단을 무엇이라 하는가? -> 샘플

8) 다음 중 아래 거래 전표에서 연관 규칙 “C-A”의 신뢰도를 구하시오. ->
C와 A 동시 포함된 확률/C가 포함된 확률

9) 동시에 구매될 가능성이 큰 상품군을 찾아내는 연관성 측정에 시간이라는 개념을 포함시켜 순차적인 구매 가능성이 큰 상품군을 찾아내는 데이터 마이닝 기법은? -> 순차 분석

10) 주성분분석을 통해 얻은 R 프로그램의 결과가 아래와 같이 나왔다. 3개의 변수를 활용할 경우 전체 데이터의 몇 %를 설명할 수 있는지 쓰시오.
-> Comp.3의 Cumulative Proportion 확인
---

1) 인터넷상의 서버에서 데이터 저장, 처리, 네트워크, 콘텐츠 사용 등 서로 다른 물리적인 위치에 존재하는 컴퓨팅 자원을 가상화 기술을 통해 IT 관련 서비스를 한번에 제공하는 혁신적인 컴퓨팅 기술은? -> 클라우드 컴퓨팅

2) 데이터 사이언스란 데이터로부터 의미있는 정보를 추출하는 학문이다. 통계학이 정형화된 실험 데이터를 분석 대상으로 하는 것에 비해, 데이터 사이언스는 정형 또는 ( )을 막론하고 인터넷, 휴대전화, 감시용 카메라 등에서 생성되는 숫자와 문자, 영상 정보 등 다양한 유형의 데이터를 대상으로 한다. -> 비정형

3) 풀어야 할 문제에 대한 상세한 설명 및 해당 문제를 해결했을 때 발생하는 효과를 명시함으로써 향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용하도록 하는 것은 무엇인가? -> 분석 유즈 케이스

4) R 명령의 결과를 쓰시오. 0/0
-> NaN (Not a Number)

5) 100명의 키를 cm로 측정한 데이터의 분산이 225였다. 동일한 100명의 키를 m로 측정한다면 데이터의 분산은 얼마인가? -> 0.0225

6) 고객은 늘 구매하지 않는다. 경쟁사의 고객 빼앗기에 따른 고객의 변심 또는 고객의 니즈나 취향이 변해 더 이상 상품과 서비스를 사용하지 않고 경쟁사와 거래하는 고객을 무엇이라고 하는가? -> 이탈고객

7) 의사결정나무 중 연속형 타깃변수(또는 목표변수)를 예측하는 의사결정나무를 무엇이라고 하는가? -> 회귀나무

8) 데이터 마이닝 모델링 분석 기법 중 random input에 따른 forest of tree를 이용한 분류방법으로 랜덤한 forest에는 많은 트리들이 생성된다. 새로운 오브젝트를 분류하기 위해 forest에 있는 트리에 각각 투입해 각각의 트리들이 voting함으로써 분류하는 방식의 R패키지는 무엇인가? -> 랜덤 포레스트

9) 다수 모델들의 예측을 관리하고 조합하는 기술을 메타학습(Meta Learning)이라고 한다. 여러 분류기들의 예측을 조합함으로써 분류 정확성을 향상시키는 기법은? -> 앙상블 기법

10) 데이터셋 x는 두 개의 변수와 5개의 관측치를 가지며 아래는 데이터와 관측치 간의 유클리드 거리를 나타낸다. 최단연결법을 사용하여 계층적 군집화를 할 때 첫 단계에서 발생되는 군집과 관측치

a와의 거리를 구하시오. -> 3.2

```
> dist(x)
   a   b   c   d
b 3.2
c 3.6 5.4
d 3.2 2.8 3.0
e 5.0 3.0 5.1 2.2
```

---

1) 다음 설명에 맞는 데이터 유형은 무엇인가? -> 정보
- 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 의미를 부여한 데이터
- 지식을 도출할 때 사용하는 데이터

2) 기업 내부 데이터베이스 중 기업 전체가 경영자원을 효과적으로 이용하기 위해 통합적으로 관리하고 경영의 효율화를 기하기 위한 수단으로 정보의 통합을 위해 기업의 모든 자원을 최적으로 관리하기 위한 기업 경영 정보 시스템은? -> ERP

3) 데이터 거버넌스란 전사차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운용조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축하는 것을 말한다. 특히 ( ), ( ), ( )는 데이터 거버넌스의 중요한 관리 대상이다. -> 마스터데이터, 메타데이터, 데이터사전

4) ( )은/는 전략적 중요도가 핵심이며, 이는 현재의 관점에서 전략적 가치를 둘 것인지, 미래의 중장기적 관점에 전략적인 가치를 둘 것인지를 고려하고, 분석 과제의 목표가치(KPI)를 함께 고려하여 ( )의 여부를 판단할 수 있다. (공통으로 들어갈 단어) -> 시급성

5) ~혼합분포 군집 방법으로 군집분석을 시행한 결과물(공분산 형태의 BIC)이다. 그래프보고 최적의 군집수는 몇 개인지 쓰시오.-> 4개 (해설: BIC 값이 가장 큰 지점을 찾았을 때 4가 제일 크므로 최적의 군집의 수는 4개)

6) 다수 모델의 예측을 관리하고 조합하는 기술을 메타학습이라 한다. 여러 분류기들의 예측을 조합함으로써 분류 정확성을 향상시키는 기법은? -> 앙상블 기법 (해설: 여러개의 예측 모형을 만든 후 조합하여 하나의 최종 모형을 만드는 방법)

7) 아래 데이터의 Terminal 변수는 약 몇 %가 92보다 큰 값을 가지는가? -> 3rd Qu. 92.0 -> 25%

8) 이산형 확률분포 중 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률 분포는 무엇인가? -> 포아송분포

9) 오분류표에서 실제/예측 True와 실제/예측 False가 100으로 동일하다고 한다. 민감도가 0.8이라고 할 때 정확도는 얼마인가? -> 0.8 (해설:민감도가 0.8이라면 TP=4FN, 100으로 동일하다면 TP+FP=100, FN+TN=100, TP+FN=100, FP+TN=100이다. 이중 TP+FN=100에서 FN=20, TP=80이니 정확도를 구하면 TP/TP+FP=80/100=0.8)

10) 아래에서 언급한 것은 무엇인가? -> 머신러닝 or 기계학습
- 데이터의 패턴을 발견하고 데이터 모델의 매개 변수를 자동으로 학습한다.
- 자체 알고리즘을 사용하여 시간이 경과함에 따라서 경험을 축적하면서 작업 성능이 향상된다.
---

1) 인공지능의 한 분야로, 컴퓨터가 스스로 많은 데이터를 분석해서 패턴과 규칙을 찾아내고, 학습된 패턴과 규칙을 활용하여 분류나 예측을 하는 것을 무엇이라고 하는가? -> 머신러닝 또는 기계학습

2) 조직 내 구성원들이 축적하고 있는 노하우 등 암묵적 지식을 형식지로 표출화 될 수 있도록 지원하는 등, 조직의 경쟁력 향상을 위해 지식자원을 체계화하고 원활하게 공유가 될 수 있도록 지원하는 시스템을 무엇이라고 하는가? -> KMS (지식경영시스템)

3) 기업 및 공공기관에서는 시스템의 중장기 로드맵을 정의하기 위한 ( )을 수행한다. ( )은 정보기술 또는 정보시스템을 전략적으로 활용하기 위하여 조직 내 외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터플랜을 수립하는 절차이다. -> ISP (정보전략계획)

4) 데이터 분석 도입의 수준을 파악하기 위한 분석 준비도의 6가지 구성요소 중 하나로서 운영시스템 데이터 통합, 빅데이터 분석 환경, 통계 분석 환경 등을 진단하는 구성요소는 무엇인가?
-> IT 인프라

5) 베이즈 정리와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전 정보와 데이터로부터 추출된 정보를 결합하고 베이즈 정리를 이용하여 특정 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘은 무엇인가? -> 나이브 베이즈 분류

6) 계층적 군집분석에서 두 군집을 병합하는 방법 중, 군집과 군집, 또는 데이터와의 거리계산 시 최단거리를 계산하여 거리가 가까운 데이터, 또는 군집을 새로운 군집으로 형성하는 방법을 무엇이라 하는가? -> 최단연결법(단일연결법)

7) 텍스트 마이닝에서 어근에 차이가 있더라도 관련이 있는 단어들을 동일한 어간으로 매핑이 될 수 있도록 정해진 규칙에 따라 단어에서 어간을 분리하여 공통 어간을 가지는 단어를 묶는 작업을 무엇이라 하는가? -> 스테밍 또는 어간 추출

8) 시계열 분석을 위해서는 정상성을 만족해야 한다. 따라서 주어진 자료가 정상성을 만족하는지 판단하는 과정이 필요하다. 자료가 추세를 보이는 경우에는 현 시점의 자료값에서 전 시점의 자료를 빼는 방법을 통해 비정상시계열을 정상시계열로 바꾸어 준다. 이 방법은 무엇인가? -> 차분

9) 문제오류 10) 원 데이터 집합으로부터 크기가 같은 표본을 여러 번 단순 임의 복원추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블하는 방법을 무엇이라 하는가? -> 배깅
---

1) 이것은 인터넷에 연결된 기기가 사람의 개입없이 상호간에 알아서 정보를 주고받아 처리한다. 구글의 Google Glass, 나이키의 Feul band 등이 있다. -> 사물인터넷

2) 기업 내부에서 활용되는 데이터베이스 활용에 대한 설명이다. ( )은 기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화시키기 위한 것으로, 자재 구매, 생산, 제고, 유통, 판매, 고객 데이터로 구성된다. -> SCM(Supply Chain Management)

3) 합리적 의사결정을 방해하는 요소로 표현방식 및 발표자에 따라 동일한 사실(Fact)에도 판단을 달리하는 현상을 이르는 말은? -> 프레이밍 효과

4) 분석적 기업으로 도약을 위해서는 가장 먼저 조직의 분석(Analytics) 도입 여부 및 활용 수준에 대한 명확한 진단이 요구된다. 특히 분석 수준 진단 방법 중 조직의 분석 및 활용을 위한 역량 수준을 파악하기 위해 “도입-> ( ) -> 확산 -> 최적화”의 분석 성숙도 단계 포지셔닝을 파악한다. -> 활용

5) 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법은 무엇인가?
-> 분해시계열

6) 다음 내용이 설명하고 있는 것을 적으시오 -> AR 모형 (해설; p시점 전의 자료가 현재자료에 영향)
- 시계열 모델 중 자기 자신의 과거 값을 사용하여 설명하는 모형임.
- 백색잡음의 현재값과 자기 자신의 과거값의 선형 가중합으로 이루어진 정상확률 모형
- 모형에 사용하는 시계열 자료의 시점에 따라 1차, 2차, … , pck 등을 사용하나 정상시계열 모형에서는 주로 1,2차를 사용함

7) 의사결정나무에서 끝마디가 너무 많으면 모형에 ( )인 상태로 현실문제에 적용될 수 있는 적절한 규칙이 나오지 않게 된다. 따라서 분류된 관측치의 비율 또는 MSE(Mean Square Error)등을 고려하여 적절한 수준의 가지치기 규칙을 제공해야 한다. -> 과대 적합 (해설: 너무 큰 모형은 자료 과대적합, 너무 작으면 과소적합 위험 있음)

8) 데이터 마이닝 기법 중 동물의 뇌신경계를 모방하여 분류(또는 예측)을 위해 만들어진 모형은? -> 인공 신경망

9) 분류 분석 모형을 사용하여 분류된 관측치가 각 등급별로 얼마나 포함되는지를 나타내는 도표는?
-> 이익도표

10) ( )은 데이터웨어하우스 환경에서 정의된 접근 계층으로, 데이터 웨어하우스에서 데이터를 꺼내 사용자에게 제공하는 역할을 한다. 보통 특정한 조직 혹은 팀에서 사용하는 것을 목적으로 한다.
-> 데이터마트

---

1) 데이터분석과 관련된 기술로 기업의 의사결정 과정을 지원하기 위한 주제중심적이고 통합적이며 시간성을 가지는 비휘발성 데이터의 집합을 ( )라고 한다. -> 데이터웨어하우스

2) ( )이란 데이터로부터 의미있는 정보를 추출해내는 학문으로, 통계학과와 달리 정형 또는 비정형을 막론하고 다양한 유형의 데이터를 분석 대상으로 한다. 또한 분석에 초점으로 두는 데이터 마이닝과는 달리 ( )는 분석뿐만 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함하는 포괄적인 개념이다. (공통으로 들어갈 단어) -> 데이터 사이언스

3) 분석은 분석의 대상(What)과 분석의 방법(How)에 따라 아래와 같이 분류한다. 다음 중 아래의 빈칸에 들어갈 용어로 가장 적절한 것은? -> Insight (통찰)

		분석의 대상 (What)	
		Known	Un-Known
분석의 방식 (How)	Known	Optimization	
	Un-Known	Solution	Discovery

4) 여러 분석 방법론 중 하나로 ( )은 반복을 통하여 점진적으로 개발하는 방법으로서 처음 시도하는 프로젝트에 적용이 용이해지면 관리체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있다. -> 나선형 모델

5) 오분류표를 활용하여 모형을 평가하는 지표 중 실제값이 False인 관측치 중 예측치가 적중한 정도를 나타내는 지표는? -> 특이도

6) 계층적 군집분석 결과를 아래와 같이 덴드로그램으로 시각화하였다고 할 때 Tree의 높이가 60일 경우 나타나는 군집의 수를 쓰시오 -> 3개 (문제 보면 알수 있으)

7) 가설검정 결과에서 귀무가설이 옳은데도 귀무가설을 기각하게 되는 오류는? -> 제 1종 오류

8) 로지스틱 회귀분석에서는 이산형 종속변수가 1일 확률을 모형화한다. 설명변수가 한 단위 증가할 때 종속변수가 1인 확률과 0인 확률 비의 증가율을 나타내는 것은? -> 오즈 (Odds) (해설: 성공할 확률이 실패할 확률의 몇 배인지를 나타냄)

9) 신경망 모형에서 출력값 z가 여러 개로 주어지고 목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공하여 출력노드에 주로 사용되는 함수는? -> 소프트맥스 함수 (해설: 세 개 이상으로 분류하는 다중 클래스 분류에서 사용되는 활성화 함수)

10) 신경망 모형의 학습을 위한 역전파 과정에서 오차를 더 줄일 수 있는 가중치가 존재함에도 기울기가 0이 되어버려 더 이상 학습이 진행되지 않는 문제를 나타내는 용어는? -> 기울기 소실

---

1) 데이터 가공 및 상관관계의 이해를 통해 패턴을 인식하고 그 의미를 부여하는 데이터를 무엇이라고 하는가? -> 정보

2) ( )는 인터넷을 기반으로 모든 사물을 연결해 사람과 사물, 사물과 사물 간의 정보를 상호 소통하는 지능형 기술 및 서비스이며, 사물에서 생성되는 Data를 활용한 분석을 통해 마케팅 등에 활용할 수 있다. -> 사물인터넷

3) ( )는 식별된 비즈니스 문제를 데이터의 문제로 변환하여 정의하는 단계 -> 문제 정의

4) ( )는 전시차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운용조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축하는 것을 말한다. 특히 마스터 데이터, 메타 데이터, 데이터 사전은 ( )의 중요한 관리 대상이다. (공통으로 들어갈 단어)
-> 데이터 거버넌스

5) ( )은 배깅에 랜덤 과정을 추가한 방법이다. 원 자료로부터 붓스트랩 샘플을 추출하고, 각 붓스트랩 샘플에 대해 트리를 형성해 나가는 과정은 배깅과 유사하나, 각 노드마다 모두 예측변수 안에서 최적의 분할을 선택하는 방법 대신 예측변수를 임의로 추출하고 추출된 변수 내에서 최적의 분할을 만들어 나가는 방법을 사용한다. -> 랜덤 포레스트 (해설: 의사결정나무 여러개로 구성)

6) 앙상블 기법 중 붓스트랩 표본을 구성하는 재표본 과정에서 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법은? -> 부스팅

7) 인공신경망에서 동일 입력층에 대해 원하는 값이 출력되도록 개개인의 가중치를 조정하는 방법은 무엇인가? -> 역전파 알고리즘 (해설: 출력값에 대한 입력값의 기울기를 출력층 layer에서부터 계산하여 거꾸로 전파시키는 것)

8) 군집분석의 품질을 정량적으로 평가하는 대표적인 지표로 군집 내의 데이터 응집도와 군집간 분리도를 계산하여 군집 내의 데이터의 거리가 짧을수록, 군집 간 거리가 멀수록 값이 커지며 완벽한 분리일 경우 1의 값을 가지는 지표는? -> 실루엣

9) 통계분석 개념 중 모집단의 특성을 단일한 값으로 추정하는 방법은 무엇인가? -> 점 추정 (해설: 추정하고자하는 하나의 모수에 대하여 모집단에서 임의로 추출된 n개 표본의 확률변수로 하나의 통계량을 만들고 주어진 표본으로부터 그 값을 계산하여 하나의 수치를 제시하려고 함)

10) 불순도를 측정하는 지표로 노드의 불순도를 나타내는 값이다. 클수록 이질적이며 순수도가 낮다고 볼 수 있으며, CART에서 목적변수가 범주형일 경우 사용하는 이 지표는 무엇인가? -> 지니 지수

---

1) 문자, 기호, 음성, 화상, 영상 등 상호 연관된 다수의 컨텐츠를 정보 처리 및 정보통신 기기에 의하여 체계적으로 수집, 축적하며 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체
-> 데이터베이스

2) 아래에서 설명하고 있는 빅데이터 활용 기본 테크닉은 무엇인가?
- 생명의 진화를 모방하여 최적해(Optimal Solution)을 구하는 알고리즘으로 존 홀랜드가 1975년에 개발하였다.
- “최대의 시청률을 얻으려면 어떤 시간대에 방송해야 하는가?”와 같은 문제를 해결할 때 사용된다.
- 어떤 미지의 함수 y=f(x)를 최적화하는 해 x를 찾기 위해 진화를 모방한 탐색 알고리즘이라고 말할 수 있다. -> 유전자 알고리즘

3) 문제가 주어지고 이에 대한 해법을 찾기 위하여 각 과정이 체계적으로 단계화되어 수행하는 분석 과제 발굴 방식을 무엇이라고 하는가? -> 하향식 접근 방식

4) 아래에서 설명하는 데이터 분석 조직 구조는? -> 집중구조
- 전사 분석업무를 별도의 분석 전담 조직에서 담당
- 전략적 중요도에 따라 분석조직이 우선순위를 정해서 진행 가능
- 현업 업무부서의 분석업무와 이중화/이원화 가능성 높음

5) 여러 대상 간의 관계에 관한 수치적 자료를 이용해 유사성에 대한 측청치를 상대적 거리로 시각화하는 방법은? -> 다차원 척도법 (해설: 개체들 사이의 유사성/비유사성을 측정하여 개체들을 2차원 공간상에 점으로 표현하는 번석법)

6) 최적방정식을 선택하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수를 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하는 방법은? -> 후진제거법

7) 모형 평가방법 중 주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차 검정을 실시하는 방법으로 하나는 모형의 학습 및 구축을 위한 훈련용 자료로, 다른 하나는 성과 평가를 위한 검증용 자료로 사용하는 방법은? -> 홀드아웃 방법 (보통 70:30 또는 80:20으로 분리)

8) P(A)=0.3, P(B)=0.4이다. 두 사건 A와 B가 독립일 경우 P(B|A)는 얼마인가? -> 0.4 (독립일 때, P(B|A)=P(B)

9) 이산형 확률분포 중 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률분포는 무엇인가? -> 포아송 분포

10) ( )은 계층적 군집분석 방법 중 하나로 군집과 군집, 또는 데이터와의 거리계산 시 최단거리르 계산하여 거리가 가까운 데이터, 또는 군집을 새로운 군집으로 형성하는 방법이다. 이 방법은 사슬 구조의 군집이 생길 수 있다. -> 최단연결법