<ADsP 요약정리 및 오답노트>

-1과목-

(객관식)
데이터 마스킹 : 데이터의 속성은 유지한 채, 익명으로 생성

Cinematch -> 넷플릭스에서 개발한 알고리즘

데이터마이닝 vs 머신러닝(딥러닝) 구분하기 다른거임

트레이딩, 공급, 수요예측 -> 에너지 산업

CRM -> 고객관계관리 데이터베이스 (기업내부)

ERP -> 기업 전체를 통합적으로 관리하고 경영의 효율화 목적

빅데이터 가치측정 어려윤 이유 : 1) 데이터 재사용,재조합,다목적용 개발
2) 새로운 가치 창출
3) 분석 기술 발전
cf. 전문인력증가는 가치측정과 관련 없음

사생활 침해를 막기 위한 개인정보 무작위 처리 (본래 목적 외에 사용 방지기술) -> 난수화

유형분석 -> 특성에따라 분류할때 사용한다.

핀테크(금융)분야에서 빅데이터 활용의 핵심분야 -> 신용평가

K-NN 분석기법 - 딥러닝과 관련 x
LSTM, Autoencoder, RNN -> 딥러닝과 관련 o

Caffe, Tnsorflow, Theano -> 딥러닝 관련된 오픈소스
Anaconda -> 머신러닝 관련된 오픈소스

책임 원칙의 훼손 -> 일어나지도 않은 일을 예측해서 행동함 (범행 전에 체포, 신용불량 전에 대출금지)

멀티미디어 등 복잡한 데이터베이스 관리 -> 객체지향 DBMS
cf. 일반적으로 사용되는 테이블 기반 -> 관계형 DBMS

데이터 시각화 -> 비즈니스 컨설팅 영역 (IT영역 X)

빅데이터 분석 기법과 방법론 확대 -> 인문학 열풍과 무관하다

데이터 사이언티스트 필요역량 -> 하드스킬(전문지식 및 기술) + 소프트스킬(스토리텔링, 통찰력, 커뮤니케이션)
cf. 네트워크 최적화는 필요역량이 아니다.

데이터사이언스 -> 주로 통찰력있는 분석에 초점을 둔다 (정확성에 초점 x)

(단답형)
SQL문에서 -> WHERE ~~ BETWEEN ~~~~ AND ~~~~

데이터사이언티스트 필요역량 -> 하드스킬, 소프트스킬

지식을 도출하기 위한 재료 -> 정보

기업의 의사결정 지원, 통합적이며 시간성을 가지는 비휘발성 데이터집합 -> 데이터 웨어하우스

정제되지 않은 자연스러운 상태의 아주 큰 데이터세트 -> 데이터 레이크

수치로 명확하게 표현되는 데이터 -> 정량적 데이터

기업이 외부공급업체(제휴업체)와 통합된 정보시스템 구축 -> SCM

페이스북 위에서 작동하는 앱 만들기 시작, 하둡 -> 플랫폼

생명의 진화를 모방한 알고리즘, 최적화하는 해답을 얻기 위한 알고리즘 -> 유전자 알고리즘

-2과목-

(객관식)
성공적인 분석을 위해 고려해야하는 요소 -> 관련 데이터 파악, 원점에서 솔루션 탐색, 이행 저해 요소 관리 (비즈니스 케이스 확보 x)

정형 데이터 -> ERP, CRM, SCM 등 정보시스템
반정형 데이터 -> 로그데이터, 모바일데이터, 센싱데이터(센서)
비정형 데이터 -> 영상, 음성 문자, 이메일

CRISP-DM 방법론에서 모델링 단계에서 수행하는 것 -> 모델링 기법 선택, 모델 테스트 계

획 설계, 모델 작성, 모델 평가 (모델 적용성 평가 X)

빅데이터 분석기획 단계 순서 : 프로젝트범위 설정 - 프로젝트 정의 - 프로젝트 수행계획 수립 - 데이터 분석 위험 식별 (마지막에 위험식별 기억!)

피드백이 반복적으로 많이 발생하는 단계 : 데이터 준비단계 -> 데이터 분석단계 (추가적인 데이터 확보 필요)

비즈니스 모델 캔버스 채널 기능 -> A/S 기능은 제공하지 않는다!!

비즈니스 모델 캔버스 문제탐색단계에서 사용하는 도구 -> 업무, 제품, 고객, 규제와 감사, 지원인프라

데이터 타당성을 평가할 때 -> 문제발생 포인트에 대한 확보는 중요하지 않다.

분석 프로젝트 일정계획 수립시 -> 융통성 있는 일정관리가 필요함 (철저한 통제와 관리 X)

분석 준비도를 평가할 때는 외부데이터 활용 체계를 평가한다 (내부데이터 X)

별도의 분석 전담 조직, 분석조직이 우선순위 정해서 진행, 현업과 분석업무의 이중화/이원화 가능성 높음 -> 집중구조

가치 -> 비즈니스 효과
크기, 다양성, 속도 -> 투자비용요소

분석 마스터 플랜 데이터 분석체계 -> 단계별 추진 내용 정렬 (반복적인 정련과정 X)

집중구조 -> 별도의 분석 전담조직 구성, 업무의 이원화 이슈
기능구조 -> 별도의 분석 전담조직 구성 X, 해당 부처에서 직접 분석을 수행, 국한된 분석수행 이슈
분산구조 -> 분석조직의 인력을 현업부사에 배치, 신속한 실무적용이 가능

빅데이터 거버넌스 -> 회사 내 모든 데이터 활용, 철저한 변경관리 필요, 요소별로 구분하여 작성, 수명주기관리 중요

분석과제 중에 발생한 결과물 -> 풀(pool)로 관리하고 공유
cf. 확정된 분석과제는 풀(pool)로 관리하지 않는다.

빅데이터 분석의 특징 -> 모든형태의 데이터 분석대상, 실시간에 가까운 분석, 데이터마트를 통해 모델생성
cf. 고급 분석기법 활용은 기존의 데이터분석에서도 가능했음.

Self Service Analytics 에서는 분산처리 지원 X

(단답형)
이중 '시스템' 테스트는 시스템의 객관성과 안정성을 확보한다.

데이터 거버넌스 체계, 데이터 구조 변경에 따른 -> 사전 영향 평가

데이터 및 기법(How)을 도출하기 위한 단계, 문제로의 변환 -> 문제 정의

알고리즘 설명서 -> 의사 코드

분석과제 관리 프로세스 -> 과제 발굴, 과제 수행

데이터 거버넌스 체계, 저장소, 데이터관리 체계 지원 -> 워크플로우
데이터 거버넌스 체계, 데이터 구조 변경에 따른 -> 사전 영향 평가

비즈니스 모델 캔버스 -> 업무, 제품, 고객 단위로 문제 발굴하고 이를 관리하는 규제와 감사, 지원인프라 영역으로 분석기회 도출

CRISP-DM 방법론에서 데이터셋 선택, 분석에 필요한 변수 선정 단계 -> 데이터 준비

동일한 사실에도 판단을 달리 하는 현상 -> 프레이밍 효과

문제가 주어지고 해법을 찾기 위한 분석과제 발굴 방식 -> 하향식 접근 방식

분석용 데이터를 이용해 특정 기능을 수행하는 모델을 만드는 과정 -> 모델링

반복을 통하여 점증적으로 개발하는 방법, 처음 시도에 용이함 -> 프로토타입 모델

품질보증 기술을 통합하여 개발된 평가모델 -> 능력 성숙도 통합 모델

기업 및 공공기관에서 시스템의 중장기 로드맵을 정의하기 위한 것 -> ISP (Information Strategy Planning)

-3과목 1장-

(단답형)
공간적 차원과 관련된 속성들을 지도 위에 생성하여 인사이트를 얻는 분석방법 -> 공간 분석

-3과목 2장-

as.vector 함수를 쓰면 열벡터를 생성한다

na.rm = T 는 결측값을 제외한다는 뜻이다.

"+"(2,3) 을 입력하면 숫자 5가 나온다.

R에서
int -> 무한대
NA -> 결측값
NaN -> 숫자가 아닌 값 (0 / 0)
dim -> 행렬의 차원

R에서 표준편차 구하는 함수는 sd() 이다. (stdev() 아니다)

substr()함수는 앞에 글자 몇개만 보이게 할때 쓴다. (Monday -> Mo)

두개의 테이블을 하나로 변경할때는 merge()함수를 쓴다.

모든행의 합 구하기 -> apply(m1, 1, sum)
모든열의 합 구하기 -> apply(m1, 2, sum)

생산지별로 도시의 평균 -> tapply()함수 사용

R에서는 NA가 포함된 벡터를 mean()함수 써버리면 결과 -> NA

x = 1:100
sum(x>50) -> 50 이다. (논리값 TRUE를 1로 계산)

어떤 변수를 기준으로 데이터를 합칠때 -> merge(A, B, by="class")

특정 데이터 조회할때 -> subset(test, subset = (학과 == 경영학과))

SQL 활용할때 R에서 지원해주는 패키지 이름 -> sqldf()

-3과목 3장-

(객관식)
결합해서 다양한 요약변수와 파생변수를 쉽게 생성할 수 있도록 해주는 패키지 -> reshape

파생변수는 매우 주관적일 수 있으므로 논리적 타당성을 갖춰야 한다

cast(md, id+variable~time)

관측치가 기록된 값을 결측치로 처리하면 안된다.

다중대치법 순서
대치 -> 분석 -> 결합

군집분석으로 이상치를 판정할 수 없음

(단답형)
평균으로부터 3표준편차 떨어져 있는 값을 이상치로 판단하는 알고리즘 -> ESD (Extreme Studentized Deviation)

-3과목 4장-
(객관식)
그래프 나올때는 x축 y축 잘 보기

비표본오차는 조사대상이 증가할수록 오차가 커진다

표본편의는 확률화에 의해 최소화하거나 없앨 수 있다 (모형추론 x)

구간척도는 절대적인 0이 없음 (절대영점은 비율척도만 가지고 있음)

p-value 는 귀무가설이 사실인데 사실이 아니라고 판정할 때 실제 확률을 나타낸다.

비모수검정은 관측값들의 순위나 차이의 부호를 이용해 검정한다. (평균x 분산 x)

줄기잎그림은 계산량이 많지 않다.

스피어만 상관계수는 순서형 변수를 사용하며 비선형적인 상관관계도 나타낼 수 있다 (비모수적 상관관계)

다중회귀모형은 F통계량을 이용해서 유의미한지 판단한다.

상관분석을 통해 분산은 알 수 없다.

상관계수가 0.27은 0에 가까우므로 선형관계를 가진다고 보기 어렵다 (0.5는 넘어야 선형관계로 취급)

상관계수 행렬로 인과관계를 설명할 수는 없다

모형의 적합도는 p-value를 통해 판단한다.

회귀분석의 가정 -> 선형성, 독립성, 등분산성, 비상관성
산점도가 나팔모양이면 등분산 가정이 무너진다 (이분산성)

회귀계수가 양수이면 양의 관계, 음수이면 음의 관계를 가진다.

결정계수
입력변수가 증가하면 결정계수도 증가한다
다중회귀분석에서는 수정된 결정계수를 사용하는 것이 적절하다
수정된 결정계수는 유의하지 않은 독립변수가 포함되었을때 값이 감소한다
결정계수는 총변동 중에서 설명되는 변동이 차지하는 비율이다.

Durbin Waston test 는 자기상관이 있는가에 대한 검정이다.

t값 = 추정치 / 표준오차 이다

시계열분석 정상성의 특징
1. 모든시점에서 일정한 평균을 가진다
2. 분산도 시점에 의존하지 않는다
3. 공분산은 단지 시차에만 의존하고 어느 시점 t,s에는 의존하지 않는다

시계열을 구성하는 4가지 요인
추세(경향)요인, 계절요인, 순환요인, 불규칙요인

고유값이 수평을 유지하기 전단계로 주성분의 개수를 정한다.

시계열데이터 분석 절차 순서
시간 그래프 그리기 -> 추세와 계절성 제거 -> 잔차 예측 -> 잔차에 대한 모델 적합 -> 예측된 잔차에 추세와 계절성 더하여 미래 예측 (예측하고 적합시키기)

교차분석은 두 문항 모두 범주형 변수일때 사용 가능하다

(단답형)

유의하지 않은 변수가 없을때까지 설명변수를 제거하는 방법 -> 후진제거법 (반대는 전진선택법)

제곱오차를 최소로 하는 값 ~~ 회귀계수의 추정량은 -> 최소제곱

구간을 나눠서 k개씩 띄어서 표본선택하고 매번 k번째 항목을 추출하는 표본추출 -> 계통추출방법

귀무가설이 옳은데 귀무가설을 기각하게 되는 오류 -> 제 1종오류

로지스틱 회귀모형 ~~ 몇배 증가하는지 나타내는 값 -> 오즈

시점에 상관없이 시계열의 특성이 일정한 것을 의미하는 용어 -> 정상시계열

시계열 모델 중 자기 자신의 과거 값을 사용하여 설명하는 모형 -> 이동평균모형

시계열에서 분리해서 분석하는 방법 -> 분해시계열

결정계수 구하는 공식
SSR/SST 또는 1 - SSE/SST (이게 더 자주쓰는듯)

유의하지 않은 변수가 없을 때까지 설명변수를 제거하는 방법 -> 후진제거법

전체의 중앙에 위치한 수치 -> 중앙값

-3과목 5장-

(객관식)
이질적인 모집단을 세분화 하는 기능 -> 군집분석

데이터마이닝 단계 중 목적변수를 정리하고 데이터를 적용할 수 있도록 준비하는 단계 -> 데이터 가공

검증용 자료로 사용되는 관측치 비율 -> 36.8%

의사결정나무 분석에서 뿌리마디 아래로 내려갈수록 불순도는 감소한다.

의사결정나무에서 지니계수 계산법
1 - (바뀐값 / 원래값)^2 - (바뀐값 / 원래값)^2

배깅은 부트스트랩을 사용하기 때문에 한 표본에 여러번 추출 가능(반복추출)

앙상블 기법 -> 배깅, 부스팅, 랜덤포레스트

실제값이 TRUE 인것 중에 예측치가 맞는 정도 -> 재현율(민감도)

정확도와 재현율을 보정하기위해 만들어낸 지표 -> F1

F1값 구하는법
2 / ((1 / 재현율) + (1/정밀도))

정밀도 -> 예측해서 TRUE 인것 중에서 실제값이 TRUE인 비율
재현율(민감도) -> 실제값이 TRUE인것 중에 예측한것이 TRUE인 비율
특이도 -> 실제값이 FALSE 인것 중에서 예측한것이 FALSE인 비율

일반화 가중치는 로지스틱 회귀모형에서 회귀계수와 유사하게 해석된다

사후확률을 제공하는 함수 -> softmax 함수

로지스틱회귀모형에서 회귀계수가 음수일때 -> 역S자 그래프

dist함수에서 지원하는 거리 측도에는 cosine(코싸인) 거리가 없다

군집의 개수를 줄여나가는 방법(계층적 군집 방법) -> 덴드로그램 사용

변수의 표준화와 상관성을 동시에 고려한 통계적 거리 -> 마할라노비스 거리

k-means 군집의 단점
1. 볼록한 형태가 아닌 군집이 존재하면 성능이 떨어짐
2. 사전에 목적이 없어서 결과해석이 어려움
3. 잡으이나 이상값에 영향을 많이 받음
cf. 군집이 형성되도 군집내 객체들은 다른 군집으로 이동 가능함!!

k-means 군집
1. 군집개수 초기에 정하고 seed 중심으로 군집 형성
2. 가장 가까운 seed가 있는 군집으로 분류
3. 개체의 적용에 따른 seed의 변화를 관찰한다
4. 100% 완전히 개체가 seed에 할당 되면 seed의 조정을 멈춘다.

군집분석의 유사도 측도로 피어슨 상관계수 사용 x

k-means 클러스터링 단점 극복을 위해서 -> PAM 사용

K-평균 군집에서 군집수 정하는데 활용하는 그래프 -> 집단 내 제곱합 그래프

시차연관분석은 원인과 결과의 형태로 해석할 수 있다.

K-평균 군집의 단점보완을 위해 평균대신 사용하는 것 -> 중앙값

향상도 구하는 공식
동시에 일어난 확률 / (A가 일어난 확률) X (B가 일어난 확률)

신뢰도 구하는 공식
동시에 일어난 확률 / A가 일어난 확률

지지도 구하는 공식
동시에 일어난 경우의수 / 전체경우의수

분류분석의 모델평가방법, 모델의 성과가 얼마나 향상되었는지 파악하는 그래프 -> 향상도 곡선

실제값이 FALSE인 것중 예측치가 FALSE인 정도를 나타낸 지표 -> 특이도

베이즈 정리를 이용한 분류 알고리즘 -> 나이브 베이지안 분류

사후확률 제공 함수 -> softmax 함수

벡터 내적을 기반으로 한 유사성 측도 -> 코사인 유사도

혼합모형의 모수와 가중치의 최대가능도 추정에 사용되는 알고리즘 -> EM 알고리즘

군집 내 응집도와 분리도를 계산하여 환벽한 분리일 경우 1을 가지는 지표 -> 실루엣

선택된 프로토타입 벡터를 나타내는 용어 -> BMU

랜덤모델과 비교하여 모델의 성과가 얼마나 좋아졌는지 등급별로 파악하는 그래프 -> 향상도 곡선

맨하탄 거리 : 절댓값(X1- X2) + 절댓값(Y1 - Y2)