
ADsP 키워드(Keyword)

ADsP 1과목 Overview

데이터 이해

- 데이터에 대한 종류, 습득 방법, 정보와의 관계
- 데이터베이스 정의, 특징, 종류, 관련 용어, SQL, 데이터베이스 설계 절차
- 기업 내부 데이터베이스 솔루션 (중요)
  OLTP, OLAP, CRM, SCM, DataWare House, Data Mart, ERP, BI, BA, 블록체인, KMS, RFID …
- 데이터웨어하우스 특징

데이터의 가치와 미래

- 빅데이터 정의, 4V, 7V, 빅데이터 출현 배경, 역할
- 빅데이터 가치 산정, 빅데이터가 가져온 본질적 변화
- 빅데이터 활용 기법, 빅데이터 활용을 위한 3대 요소
- 빅데이터 위기 요인, 통제방안
- 개인 정보 비식별화 기법, 분석 애플리케이션 사례

빅데이터 이해

- 데이터 사이언스의 정의, 다른 학문과의 차이점
- 데이터 사이언티스트의 역량
- 정보, 통찰력, 인문학 열풍
- 의사 결정 오류, 가치 패러다임 변환
- 데이터사이언스의 한계와 인문학

추가로 알아야 하는 내용

- SQL 명령의 종류, SELECT는 세부적으로
- 데이터의 크기 (K-M-G-T-P-E-Z-Y)
- 구글 자동 번역 시스템 - Volume과 관련 …
- IoT, 딥러닝 종류

1과목 키워드 살펴보기

데이터 정의	데이터 유형	지식의 차원	암묵지와 형식지 상호작용	DIKW 피라미드
	정성적, 정량적	암묵지, 형식지	공통화-표출화-연결화-내면화	데이터 - 정보 - 지식 - 지혜

데이터베이스 특징	DBMS 정의, 종류	데이터베이스 관련 용어
통합, 저장, 공용, 변화되는 데이터	RDBMS(관계), ODBMS(객체)	DD(데이터 사전), ERD, SQL, 메타데이터

데이터베이스 설계 절차	NoSQL 특징, 종류
요구조건 분석 - 개념적 - 논리적 - 물리적 설계	RDBMS보다 덜 제한적, 확장성, 세세한 통제, MongoDB, HBase, Redis

기업 내부 데이터베이스 솔루션 1	데이터웨어하우스의 4대 특징
OLTP, OLAP, CRM, SCM, Data Warehouse, Data Mart, ERP	데이터 통합, 시계열성, 주제 지향적, 비소멸(비휘발)성

1과목 키워드 살펴보기

기업 내부 데이터베이스 솔루션 2

BI, BA, 블록체인, KMS, RFID,

빅데이터 정의, 4V

Volume, Variety, Velocity + Value
투자비용 요소, 비즈니스 효과 요소

빅데이터 출현 배경

클라우딩 컴퓨팅 ➔ 경제적 효과 제공
양질 전환, 비정형 데이터, 처리기술

빅데이터 역할

석탄/철, 원유, 렌즈, 플랫폼
(공동 활용 목적, 페이스북, API..)

빅데이터 가치 산정 - 어려움

데이터를 언제, 어디서, 누가 활용
기존에 없던 가치 창출, 분석 기술 발달

빅데이터가 ... 본질적인 변화

사후처리, 전수조사,
양(Quantity), 상관관계

빅데이터 활용 기법

연관규칙학습, 유형분석(분류), 유전 알고리즘, 기계학습, 회귀 분석, 감정분석, 소셜 네트워크 분석(=사회 관계망 분석)

빅데이터 활용을 위한 3대 요소

자원(=데이터), 기술, 인력

1과목 키워드 살펴보기

빅데이터 위기요인, 통제방안

사생활 침해 ,책임 원칙의 훼손, 데이터의 오용
알고리즈미스트 등장!

개인정보 비식별화 기법

데이터 마스킹, 데이터 범주화, 가명, 잡음 첨가, 총계, 평균값 대체, 데이터 값 삭제

분석 애플리케이션 사례

에너지 - 트레이딩, 수요예측

데이터 사이언스의 정의, 다른 학문과 차이점

다양한 유형의 데이터, 분석+구현+전달=포괄적
총체적 접근법, 과학과 인문학의 교차로

데이터 사이언티스트의 역량

가트너 - 데이터관리, 분석 모델링, 비즈니스 분석, 소프트 스킬
하드 스킬, 소프트 스킬 (세부 항목도!)

정보 vs 통찰력

통찰력 : 모델링, 실험설계, 권고, 예측, 최적화

인문학 열풍

복잡한 세계, 비즈니스 중심 → 서비스, 시장창조

의사 결정 오류

로직 오류, 프로세스 오류

가치 패러다임 변화

Digitalization - Connection - Agency

데이터 사이언스의 한계와 인문학

분석은 가정에 근거 → 의구심, 가정과 현실 불일치에 대해 계속 고찰, 예측 위험 살피기

1과목 - 2021년, 2022년 추가

데이터의 크기	구글의 실시간 자동 번역시스템
K-M-G-T-Peta-Exa-Zeta-Yotta	관련 빅데이터의 특징 : Volume

SQL 함수 중 '그룹' '조건'

SELECT 컬럼 FROM 테이블 WHERE 조건식 GROUP BY 그룹화할 컬럼 HAVING 조건식 ORDER BY 정렬 컬럼

미래사회의 특성, 빅데이터의 역할

불확실성 - 통찰력, 리스크 - 대응력, 스마트 - 경쟁력, 융합 - 창조력

2021년 : 키워드 살펴보기에서 8 ~ 9개 풀기 가능 했음

SQL 함수 중 DML 고르기

DML - SELECT, UPDATE, INSERT, DELETE	DDL - CREATE, ALTER, DROP	DCL - GRANT, REVOKE, COMMIT, ROLLBACK

ADsP 2과목 Overview

데이터분석 기획의 이해

- 분석 기획 유형 4가지 (최적화, 솔루션, 통찰, 발견)
- 과제 단위, 마스터플랜 단위
- 분석 기획 시 고려사항
- 데이터 유형 분류
- 데이터 저장 방식
- 분석방법론의 구성요소
- 기업의 합리적 의사결정 장애요소
- 분석 방법론의 모델 3가지(폭포수, 나선형, 프로토타입)
- KDD 분석 방법론, CRISP-DM 분석 방법론
- 빅데이터 분석 방법론
- 분석 과제 도출 방법
- 분석 과제 주요 특성, 분석 프로젝트의 관리 영역

분석 마스터 플랜

- 분석 마스터플랜 우선순위 고려요소
- 적용범위/방식 고려요소
- 포트폴리오 사분면을 통한 우선순위 선정
- 분석 거버넌스 체계 수립
- 빅데이터 거버넌스 특징
- 관련 용어(Servitization, CoE, ISP, Sandbox …)

추가로 알아야 하는 내용

- 7V, SQL 구문

2과목 키워드 살펴보기

분석 기획 유형 4가지	과제단위(당면한 분석 주제 해결), 마스터플랜 단위(지속적 분석문화 내재화)
what(분석대상), how(분석방법) 최적화, 솔루션, 통찰, 발견	과제단위 : Speed & Test, Quick-Win, Problem Solving 마스터플랜 단위 : Accuracy & Deploy, Long Term View, Problem Definition

분석 기획 시 고려사항	데이터 유형 분류	데이터 저장 방식
가용한 데이터, 적절한 유스케이스 탐색, 장애요소들에 대한 사전 계획 수립	정형, 반정형, 비정형 데이터 분류	RDBMS : Oracle, MSSQL, MySQL NoSQL : MongoDB, HBase, Redis, Cassandra 분산파일시스템 : HDFS

분석방법론의 구성요소	기업의 합리적 의사결정 장애요소
상세한 절차, 방법, 도구와 기법 템플릿과 산출물	고정관념, 편향된 생각, 프레이밍 효과

2과목 키워드 살펴보기

분석 방법론의 모델 3가지

폭포수 모델 - 하향식 접근 방법
나선형 모델 - 점증적 개발
프로토타입 모델 - 상향식 접근 방법

KDD 분석 방법론 (5단계)

데이터셋 선택 - 데이터 전처리 - 데이터 변환 - 데이터 마이닝 - 데이터 마이닝 결과평가

잡음, 이상치, 결측치 식별/제거

변수 선택, 차원 축소, 데이터셋 변경 작업

CRISP-DM 분석 방법론 (6단계) - 영문도!

업무(Business) 이해 - 데이터 이해 - 데이터 준비 - 모델링 - 평가 - 전개

모델 평가

분석 결과 평가
모델링 과정 평가
모델 적용성 평가

CRISP-DM : Business Understanding - Data Understanding - Data Preparation - Modeling - Evaluation - Deployment

CRISP-DM 분석 방법론 - 업무 이해 순서

업무 목적 파악 - 상황 파악 - 데이터 마이닝 목표 설정 - 프로젝트 계획 수립

2과목 키워드 살펴보기

빅데이터 분석 방법론

분석 기획 — 비즈니스 이해 및 프로젝트 범위 설정 - 프로젝트 정의 및 수행 계획 수립 - 프로젝트 위험 계획 수립

데이터 준비 — 필요 데이터 정의 - 데이터 스토어 설계 - 데이터 수집 및 정합성 점검

데이터 분석 — 분석용 데이터 준비 - 텍스트 분석 - 탐색적 분석 - 모델링 - 모델 평가 및 검증

추가적 데이터 확보가 필요한 경우
반복적인 피드백을 수행하는 구간

위험 대응 방법

회피(Avoid)
전이(Transfer)
완화(Mitigate)
수용(Accept)

시스템 구현

평가 및 전개

2과목 키워드 살펴보기

분석 과제 도출 방법

하향식 접근 방법	상향식 접근 방법	디자인 싱킹(Thinking)
문제가 확실할 때 사용	문제 정의 자체가 어려운 경우 사용	중요 의사결정시 상향식(발산)과 하향식(수렴)을 반복적 사용

하향식 접근 방법 (Top-Down Approach) 의 데이터 분석기획 단계	상향식 접근 방법
Problem Discovery - Problem Definition - Solution Search - Feasibility Study 문제 탐색 - 문제 정의 - 해결 방안 탐색 - 타당성 검토	비지도 학습, 디자인 싱킹의 발산 단계, 반복적인 시행착오를 통해 수정하며 문제 도출

비즈니스 모델 기반 문제 탐색 : 비즈니스 모델 캔버스, 5가지 영역 : 업무, 제품, 고객, 지원 인프라, 규제와 감사
분석 기회 발굴의 범위 확장 : 거시적 관심 요인, 경쟁자 확대, 시장의 니즈 탐색, 역량의 재해석 관점
외부 참조 모델 기반 문제 탐색 : 유사/동종 사례 벤치마킹 (Quick & Easy 방식)
분석 유즈 케이스 : 향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용

2과목 키워드 살펴보기

분석 프로젝트 특징

분석 과제 주요 특성

Data Size, Data Complexity, Speed, Analytic Complexity, Accuracy & Precision

정확도, 복잡도는 트레이드 오프 관계
기준점을 사전에 정의해 두어야 함

Accuracy : 분석의 활용적 측면
Precision : 분석의 안정성 측면
트레이드 오프 관계

분석 프로젝트의 관리 영역

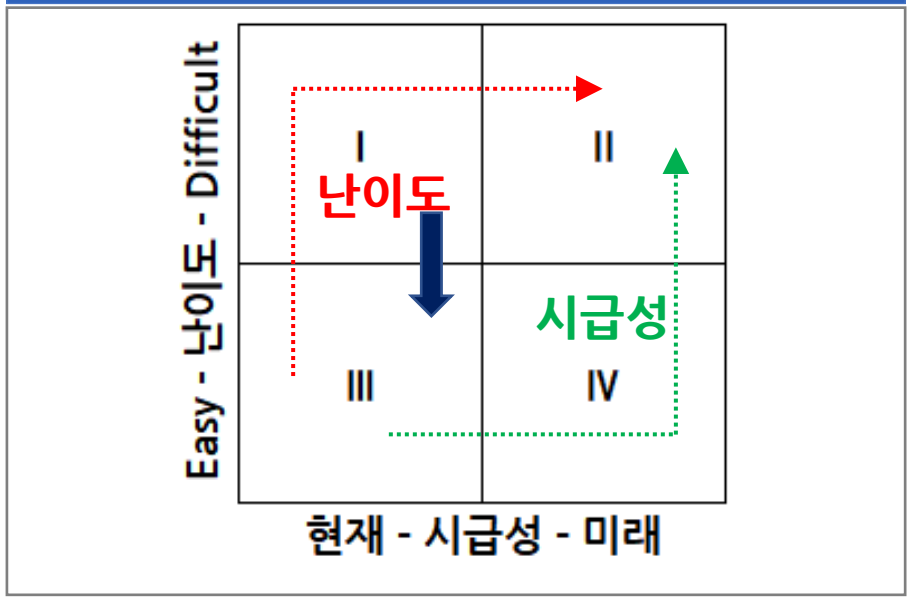
시간, 범위, 품질, 통합, 이해관계자, 자원, 원가, 리스크, 조달, 의사소통

2과목 키워드 살펴보기

분석 마스터플랜 수립

우선순위 고려 요소	적용 범위/방식 고려요소
전략적 중요도 ROI(투자자본 수익률) 실행 용이성	업무 내재화 적용 수준 분석 데이터 적용 수준 기술 적용 수준

포트폴리오 사분면을 통한 우선순위 선정



2과목 키워드 살펴보기

분석 거버넌스 체계 수립

- 기업에서 데이터가 어떻게 관리, 유지, 규제되는지에 대한 내부적인 관리 방식이나 프로세스
- 5가지 분석 거버넌스 체계 구성 요소 : 과제 기획/운영 프로세스, IT기술/프로그램, 분석 기획/관리 및 추진 조직, 분석교육, 데이터 거버넌스 (분석 비용 및 예산 없음)

1. 거버넌스 체계 개요	거버넌스, 분석 거버넌스, 데이터 거버넌스
2. 데이터 분석 준비도	분석 업무 파악, 이력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라
3. 분석 성숙도 모델	비즈니스 부문, 조직/역량 부문, IT 부문을 대상 도입단계, 활용 단계, 확산 단계, 최적화 단계로 구분하여 살펴봄 - CMMI (능력 성숙도 통합 모델) : ~ 성숙도를 평가하기 위한 모델
4. 분석 수준 진단 결과	
5. 분석 지원 인프라 방안 수립	
6. 데이터 거버넌스 체계 수립	데이터 거버넌스 체계요소 (세부적인 내용 꼭 알아 둘 것!) - 데이터 표준화, 데이터 관리체계, 데이터 저장소 관리, 표준화 활동 데이터 거버넌스 구성 요소 : 원칙, 조직, 프로세스
7. 데이터 조직 및 인력 방안 수립	데이터 분석을 위한 조직 구조 : 집중형, 기능중심, 분산 (각 특징 중요!)
8. 분석 과제 관리 프로세스 수립	과제 발굴 : 분석 아이디어 발굴, 분석 과제 후보 제안, 분석 과제 확정 과제 수행 : 팀 구성, 분석 과제 실행, 분석 과제 진행 관리, 결과 공유/개선
9. 분석 교육 및 변화 관리	모든 구성원이 데이터를 분석, 활용할 수 있도록 분석 문화를 정착, 변화시도

2과목 키워드 살펴보기

빅데이터 거버넌스 특징

- 분석 대상 및 목적 명확히 정의
- 품질 관리도 중요하지만, 데이터 수명주기 관리방안을 수립하지 않으면 … 가용성, 관리비용 증대 문제
- ERD는 계속해서 변경사항을 관리해야 함
- 산업 분야, 데이터 유형, 정보 거버넌스 요소별로 구분하여 작성
- 분석 조식 및 인력에 대해 지속적인 교육과 훈련을 실시
- 개인정보보호 및 보안에 대한 방법 마련

관련 용어

- Servitization : 제조업과 서비스업의 융합을 나타내는 용어
- CoE(Center of Excellence) : 조직 내 분석 전문조직을 말함
- ISP(정보전략계획) : 기업의 경영목표 달성에 필요한 전략적 주요 정보를 포착하고, 주요 정보를 지원하기 위해 전사적 관점의 정보 구조를 도출하며, 이를 수행하기 위한 전략 및 실행 계획을 수립하는 전사적인 종합추진 계획
- Sandbox : 보안모델, 외부 접근 및 영향을 차단하여 제한된 영역 내에서만 프로그램을 동작시키는 것

SQL 구문 - DDL, DML, DCL

DDL : CREATE, ALTER, DROP, TRUNCATE
DML : DELETE, INSERT, UPDATE, SELECT
DCL : GRANT, REVOKE, COMMIT, ROLLBACK

분석 과제 관리 프로세스

발굴 단계, 수행 단계의 종류, 분석과제 중 발생된 시사점과 분석 결과물은 풀(Pool)로 관리, 분석 과제로 확정된 분석과제를 풀로 관리 X

분석 마스터플랜과 ISP의 관계

분석 마스터 플랜
- 일반적인 ISP 방법론을 활용하되 데이터 분석 기획의 특성을 고려하여 수행,
- 기업에 필요한 데이터 분석 과제를 빠짐없이 도출한 후 과제의 우선순위를 결정하고 단기 및 중,장기로 나누어 계획을 수립함

ISP - 기업 및 공공기관에서는 시스템의 중장기 로드맵을 정의하기 위한 ISP를 수행한다 (중장기 마스터플랜을 수립하는 절차)

분석 성숙도 모델 단계

도입/활용/확산/최적화 중 어떤 것인가?

분석 준비도 영역

분석 업무파악, 인력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라 중 무엇?

빅데이터 7V

Volume, Variety, Velocity, Value
Veracity, Validity, Volatility

summary 함수

연속형, 범주형 변수의 해석
- 연속형 : 최솟값, 1사분위수, 중간값, 평균, 3사분위수, 최댓값
- 범주형 : 범주, 범주별 데이터 개수가 콜론으로 구분되어 표시

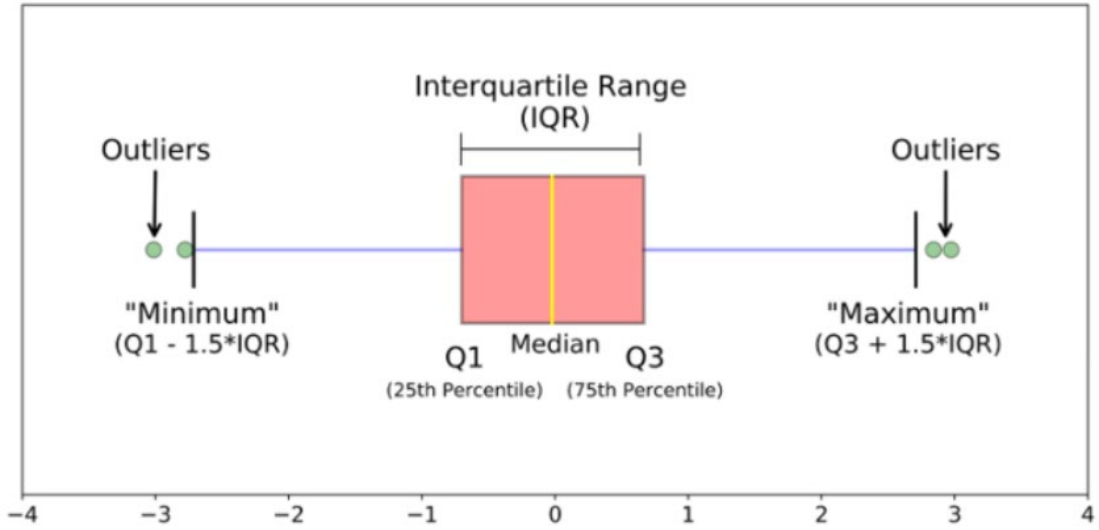
결측치 대치법

단순 대치법: 완전 응답 개체분석, 평균대치법, 단순확률 대치법
다중 대치법: 추정량 표준오차의 과소추정, 계산의 난해성 문제

이상값(outlier)

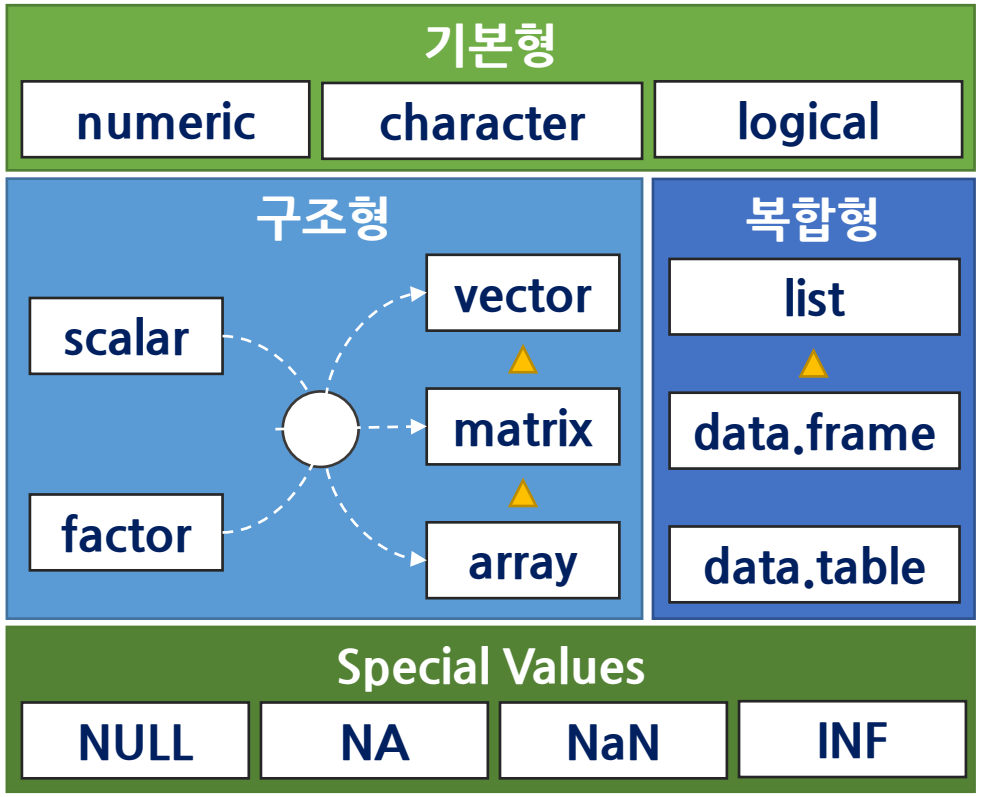
- 평균 - 3*표준편차, 평균 + 3*표준편차 밖의 값
- Q1 - IQR * 1.5, Q3 + IQR *1.5 밖의 값
- IQR = Q3 - Q1
- 분석 대상이 될 수 있어 무조건 삭제 X

boxplot



3-01. R의 데이터 형(=타입)

- R의 데이터 형은 기본형, 구조형, 복합형으로 나눌 수 있으며, Special Values 가 존재함



- NULL: 변수 값이 초기화되지 않음
- NA: Not Available, 데이터 값 없음(결측치)
- NaN: Not Available Number, 계산 불가능
- INF: Infinite, 무한대

데이터 형	특징
numeric	정수, 실수, 복소수, 수학적 연산 및 통계적 계산
character	문자, 단어로 구성, “” 또는 ‘’ 내에 표현됨
logical	TRUE, FALSE, 산술 연산 시 1, 0으로 사용됨

데이터 형	차원	원소	원소의 타입
scalar	단일	수치/문자/논리	단일
factor	1D	수치/문자	단일, 범주형
vector	1D	수치/문자/논리	단일
matrix	2D		단일
data.frame	2D		복합 가능
array	2D 이상		단일
list	2D 이상		복합 가능

3-03. summary 함수예

```
> summary(Hitters)
     AtBat             Hits         HmRun            NewLeague
 Min.   : 16.0   Min.   :  1   Min.   : 0.00   A:176
 1st Qu.:255.2   1st Qu.: 64   1st Qu.: 4.00   N:146
 Median :379.5   Median : 96   Median : 8.00
 Mean   :380.9   Mean   :101   Mean   :10.77
 3rd Qu.:512.0   3rd Qu.:137   3rd Qu.:16.00
 Max.   :687.0   Max.   :238   Max.   :40.00
```

범주형

```
     Salary
 Min.   :  67.5
 1st Qu.: 190.0
 Median : 425.0
 Mean   : 535.9
 3rd Qu.: 750.0
 Max.   :2460.0
 NA's   :59
```

결측치 59개

- AtBat, Hits, HuRun : 연속형 변수, Min, 1st Qu., Median, Mean, 3rd Qu., Max. 값이 표시되어 있음
- Salary : 연속형 변수 + 결측치, 결측치는 NA's : 59로 표시되어 있음
- NewLeague : 범주형, A 범주가 176개, N 범주가 146개 있음

3과목 2-1 통계학 개론 - 키워드 살펴보기

모집단 / 모수	표본 / 통계량	확률적 표본추출법의 종류
평균(μ), 분산(σ^2), 표준편차(σ)	평균($\bar{x}$), 분산(s^2), 표준편차(s)	단순 무작위 추출, 계통추출, 층화추출, 군집추출

척도의 종류	집중화 경향 측정	평균 vs 중앙값
명목척도, 서열(순위)척도, 등간(구간)척도, 비율척도	평균, 중앙값, 최빈값	평균은 양 꼬리값의 영향을 크게 받음

Negative-Skewed	Positive-Skewed	편차(bias)	범위(range)
평균 < 중앙값 < 최빈값	평균 > 중앙값 > 최빈값	평균 - 변량(=데이터)	최대값 - 최소값

분산 (s^2) Variance

$$s^2 = \frac{1}{n-1}\sum_{x=1}^{n}(x_i - \bar{x})^2$$

표준편차(s) Standard Deviation

$$\sigma = \sqrt{\frac{1}{n}\sum_{x=1}^{n}(x_i - \mu)^2} \qquad s = \sqrt{\frac{1}{n-1}\sum_{x=1}^{n}(x_i - \bar{x})^2}$$

변동 계수(CV)

$$\text{cv} = \frac{s}{\bar{x}}$$

편차, 분산, 표준편차 ➔ 데이터의 퍼짐 정도 측정

3과목 2-1 통계학 개론 - 키워드 살펴보기

확률(probability)

A가 발생할 확률
P(A) = A사건 / 표본공간
0 ~ 1의 값

사건의 종류

독립사건 : 두 사건 A, B가 독립이면 P(B|A)=P(B), P(A|B) = P(A), P(A∩B) = P(A)·P(B) 성립
배반사건 : 교집합이 공집합, P(A∩B) = 0, P(AUB) = P(A) + P(B)
종속사건 : 한 사건의 결과가 다른 사건에 영향을 주는 사건 P(A∩B) = P(A|B) · P(B)

조건부 확률

P(A|B) = P(A∩B) / P(B), 단 P(B) > 0

이산형 확률분포

베르누이분포, 이항분포, 기하분포, 포아송분포

이산적 확률변수 기댓값

$$E(X) = \sum x \cdot f(x)$$

연속형 확률분포

정규분포, 지수분포, 연속균일분포, 카이제곱분포, F분포, t 분포, z분포

연속적 확률변수 기댓값

$$E(X) = \int x \cdot f(x)$$

베르누이 분포: 실험 결과 0 또는 1의 결과
이항분포 : 베르누이 시행을 n회 반복
기하분포 : 베르누이 시행에서 처음 성공까지 시도한 횟수 X의 분포
포아송 분포 : 단위 시간/공간에서 어떤 사건이 몇 번 발생할 것인지 표현

3과목 2-1 통계학 개론 - 키워드 살펴보기

정규분포(normal distribution)	정규분포의 당위성
▪ 평균과 표준편차(σ)에 대해 모양이 결정되고 N(μ,σ^2)로 표기 ▪ N(0, 1) 를 표준 정규 분포, z 분포 ▪ z분포의 평균 주위로 표준편차의 1배 범위에 있을 확률 68%, 2배 범위 안 95%, 3배 범위 안 99.7% (3시그마규칙)	▪ 이항분포의 근사 : 시행횟수 N이 커질 때 정규분포 ▪ 중심 극한 정리 : 확률표본의 표본평균은 N이 충분히 크면 근사적으로 정규분포를 따르게 됨 ▪ 오차의 법칙 : 오차는 정규분포를 따름

t 분포	표본을 많이 뽑지 못하는 경우에 대한 대응책으로 예측범위가 넓은 분포를 사용하며, 이것이 t-분포임
Z 분포	정규분포를 표준화 한 것, 표준정규분포, N(0, 1), t 분포, Z분포는 평균 검정에 사용됨
χ^2 분포	분산의 특징을 확률분포로 만든 것으로 0이상의 값만 가질 수 있으며, 오른쪽 꼬리가 긴 비대칭 모양
F 분포	두 집단의 분산이 크기가 서로 같은지 또는 다른지 비교하는데 사용, 두 분산의 나눗셈을 확률분포로 나타낸 것

3과목 2-1 통계학 개론 - 키워드 살펴보기

모수적 추론

특정 분포를 가정하고 모수 추론

비모수적 추론

모집단에 대해 특정 분포 가정 하지 않음

추정의 종류

점추정(MLE, 최소제곱법), 구간 추정

가설 검정

모수에 대한 가설 검정 (귀무가설, 대립가설)

통계적 추론

Frequentist, Bayesian

좋은 추정량

일치성, 비편향성, 효율성

표준오차(Standard Error)

표준오차= $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$ (표본평균의 표준편차)

표본오차(Sampling Error)

표본오차 = 임계값 * $\frac{\sigma}{\sqrt{n}}$ 표준정규분포에서 표본오차 = $z * \frac{\sigma}{\sqrt{n}} \approx z * \frac{s}{\sqrt{n}}$

구간 추정 (P. 353, 확실한 이해 필요)

신뢰수준(1- α), 신뢰구간, 신뢰구간의 길이

신뢰구간

모수가 포함되리라고
기대되는 범위

신뢰구간의 길이 (Z 분포)

$$l = 2 * Z * \frac{\sigma}{\sqrt{n}}$$

- 신뢰도 95%, α=0.025 ➔ Z=1.96
- 신뢰도 99%, α=0.005 ➔ Z=2.58

신뢰수준, 신뢰구간

- 99% 신뢰수준에 대한 신뢰구간이 95% 신뢰수준에 대한 신뢰구간보다 길다
- 표본의 크기가 커지면 신뢰구간의 길이는 줄어든다

3과목 2-1 통계학 개론 - 키워드 살펴보기

귀무가설(H_0)	가설검정의 대상이 되는 가설, 연구자가 부정하고자 하는 가설, 변화 없는 것
대립가설 (H_1)	귀무가설이 기각될 때 채택되는 가설, 연구자가 연구를 통해 입증 / 증명되기를 기대하는 예상이나 주장
기각역	검정통계량(t-value)의 분포에서 유의수준의 크기에 해당하는 영역

제 1종 오류	제 2종 오류	유의수준(α)
α error, 귀무가설이 참인데 기각	β error, 귀무가설이 거짓인데 채택	1종 오류의 최대 허용 한계, 주로 0.05 사용

유의 확률 (p-value)

- 1종 오류를 범할 확률, p-value < α 일 때 귀무가설 기각, 대립가설 채택
- p-value가 0.05(5%) : 귀무가설을 기각했을 때 기각 결정이 잘못될 확률이 5%임
- 검정 통계량에 관한 확률, 극단적인 표본 값이 나올 확률

3과목 2-1 통계학 개론 - 키워드 살펴보기

모수적 검정	▪ One sample T test, Paired T test, Two sample T test, ANOVA test

평균 차이 검정	분산 간의 차이	t 검정 방법 (검정 결과 해석할 수 있어야 함), P. 367-369
t-test, z-test, ANOVA	F 분포	One sample t-test, Paired t-test, Two sample t-test

자유도(degree of freedom)	데이터 정규성 검정 종류
df = n - 1 two sample t-test : df = n-2	Q-Q plot, Histogram, Shapiro-Wilk test, Kolmogorov-Smirnov test, Anderson-Darling test

3과목 2-1 통계학 개론 - 키워드 살펴보기

비교 대상 집단 수에 따른 모수/비모수적 추론 방법

비교대상 집단 수	관계	비모수적 검정 (명목척도)	비모수적 검정 (서열척도/연속형)	모수적 검정
1	-	Run test, χ^2 적합성 검정	Wilcoxon Signed-Rank Test Sign Test Kolmogorov-Smirnov Test(연속형)	One sample t-test
2	독립	Crosstab χ^2 독립성/동질성 검정	Mann-Whitney U Test Wilcoxon rank-sum Test Kolmogorov-Smirnov Test(연속형)	Two sample t-test
	대응 자료	McNemar test	Wilcoxon Signed -Rank Test Sign test	Paired t-test
k(다변량)	독립	χ^2 독립성/동질성 검정	Kruskal-Wallis H Test	ANOVA Test
	대응 자료	Cochran test	Friedman test	

- 카이제곱검정(χ^2 검정)을 모수적 추론에서 사용할 때는 분산에 대한 검정에 사용되며,
  비모수적 추론 방법에서는 명목 척도 데이터인 경우 적합성, 동질성, 독립성 검정에서 사용한다

다음 중 비 모수적 추론이 아닌 것은? Run test, Sign test, Wilcoxon Rank test 등과 함께 카이제곱검정/분포가 보기로 나올때 카이제곱검정은 모수적 추론으로 보아야 함

3과목 2-2 기초 통계분석 - 키워드 살펴보기

독립변수	종속변수	오차(모집단) / 잔차
입력 값, 원인을 나타냄	분석의 대상, 결과물	예측값과 실제값 차이

회귀 분석		
종속변수가 연속형 변수	최소자승법	(측정값 - 함수값)2 의 합이 최소가 되는 직선의 그래프를 찾는 것
	회귀 모형 해석	단일, 다중 회귀 방정식(종속, 독립변수), 절편, 회귀 계수
	회귀 모형의 가정	선형성, {독립성, 정상성, 등분산성, 비상관성} ➔잔차 관련 Normal Q-Q, Scale-Location, Residuals vs Fitted 그래프 해석 3과목 QnA 보기
	회귀 모형 해석 R 코드 해석	표본 회귀선의 유의성 검정 : 회귀 계수 $\beta 1 = 0$ 일 때 귀무가설, $\beta 1 \neq 0$일 때 대립가설로 설정
		모형이 통계적으로 유의미 : F 통계량, p-value 확인 회귀계수들이 유의미 : 회귀계수의 t값, p-value 확인 모형의 설명력 : 결정계수(R^2) 확인 = SSR/SST, 1-(SSE/SST)

3과목 2-2 기초 통계분석 - 키워드 살펴보기

다중공선성

독립변수들끼리 상관관계 갖음
학습 방해 요인이므로
VIF 검사로 10 이상 제거

설명변수 선택방법

모든 가능한 조합, 후진제거법, 전진선택법, 단계별 선택법
step 함수 사용, direction = 'backward', 'forward', 'both'

변수 선택 판단 기준

AIC, BIC, Mallow's Cp : 값이 작을수록 좋음

Regularization
L은 1번에
R은 2번에

- Lasso - L1 norm, 회귀계수가 0이 되거나 0에 가깝게 됨 (변수선택)
- Ridge - L2 norm, 회귀계수가 0에 가까워질 뿐 0은 되지 않음
- 람다값이 클 수록 패널티를 강하게 준 것임

Normalization (MinMax normalization)

$X' = \frac{X - Xmin}{X_{max} - Xmin}$, 값의 범위를 [0, 1]로 변환

Standardization

$Z' = \frac{X - \mu}{\sigma}$, 값의 평균 0, 분산 1

상관계수

- -1 ~ 1의 값, 0은 무상관, 인과관계가 있다할 수 없음
- 피어슨 : 선형적인 크기만 측정 가능 (등간, 비율 척도)
- 스피어만 : 비선형적인 관계도 측정 가능 (서열 척도)
- 분석 결과 해석 할 수 있어야 함

과적합(overfitting)

- 학습데이터에 치중한 학습
- 학습데이터에 대한 성능 매우 높음
- 검증데이터(새로운 데이터)에 대한 성능이 매우 낮음
- Low bias, High variance
- 해결방법
  - 데이터를 더 수집
  - Regularization 사용
  - 앙상블 모델 사용

3과목 2-2 기초 통계분석 - 키워드 살펴보기

차원축소법 | 주성분분석(PCA), 요인분석, 판별분석, 군집분석, 정준상관분석, 다차원척도법

주성분 분석 (PCA)

- 주성분 분석은 **가장 분산이 큰 것을 제 1주성분으로 설정**한다
- 주성분 분석은 상관관계가 있는 변수들을 결합해 상관관계가 없는 변수로 분산을 극대화하는 변수로 선형결합을 해 변수를 축약하는 데 사용하는 방법이다
- 공분산 행렬은 변수의 특정단위를 그대로 반영한 것이고, 상관행렬은 모든 변수의 측정단위를 표준화한 것이다
- 공분산 행렬 또는 상관계수 행렬을 사용
- 모든 변수들을 잘 설명하는 주성분 찾음
- 결과 해석 할 수 있어야 함 (매우 중요!)
- prcomp(data, scale=TRUE) : 상관계수 행렬 사용
- princomp(data, cor=TRUE) : 상관계수 행렬 사용
- 주성분 결정 기준 : 누적 분산 비율(70~90),
  고윳값이 1보다 큰 성분, Scree Plot 해석

공분산, 상관계수

- 두 확률변수의 선형관계를 나타냄
- 공분산 - 측정단위에 영향 받음(민감함)
- 상관계수 - 측정 단위 영향 받지 않음

3과목 2-2 기초 통계분석 - 키워드 살펴보기

정상성(stationary)	정상 시계열 조건
시계열의 평균과 분산에 체계적인 변화 및 주기적 변동이 없다는 것	▪ 평균은 모든 시점(시간 t)에 대해 일정하다 $E(x_t) = \mu$ ▪ 분산은 모든 시점(시간 t)에 대해 일정하다 $Var(x_t) = \sigma^2$ ▪ 공분산은 시점(시간 t)에 의존하지 않고, 단지 시차에만 의존 $Cov(x_{t+h}, x_t) = \gamma_h$

정상 시계열 전환	차분	분해 시계열 분해 요인
▪ 평균이 일정하지 않은 경우 : 원계열에 차분 사용 ▪ 계절성을 갖는 비정상시계열 : 계절 차분 사용 ▪ 분산이 일정하지 않은 경우 : 원계열에 자연로그(변환) 사용	현 시점의 자료 값에서 전 시점의 자료 값을 빼 주는 것	추세요인(Up/Down), 계절요인 (고정된 주기), 순환요인 (알려지지 않은 주기), 불규칙요인 구분

시계열 데이터 분석 절차	1. 시간 그래프 그리기 ➔ 2. 추세와 계절성 제거하기 ➔ 3. 잔차 예측하기 ➔ 4. 잔차에 대한 모델 적합하기 ➔ 5. 예측된 잔차에 추세와 계절성을 더해 미래예측하기
AR 모형	백색 잡음의 현재 값과 자기 자신의 과거 값의 선형 가중 값으로 이루어진 정상 확률 모형 (정상 시계열)
MA 모형	현시점의 자료가 유한 개의 과거 백색잡음의 선형결합으로 표현 (정상 시계열)
ARIMA 모형	ARIMA 모형은 비정상시계열 모형이며 차분/변환을 통해 AR, MA, ARMA 모형으로 정상화

AR $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$

MA $y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$

3과목 3 정형 데이터 마이닝 - 키워드 살펴보기

데이터 마이닝

모든 사용가능한 원천 데이터를 기반으로 감춰진 지식, 기대하지 못했던 경향 또는 새로운 규칙 등을 발견하고 이를 실제 비즈니스 의사결정 등에 유용한 정보로 활용하는 일련의 작업

데이터 마이닝 기법

- 분류, 추정, 연관분석, 예측, 군집, 기술에 대한 정의/예시
- 분류 : 답이 있는 상태(기존의 분류, 정의된 집합에 배정)
- 군집 : 미리 정의된 기준,예시 없음, 유사성에 의해 그룹화되고 이질성에 의해 세분화
- 연관 : 카탈로그 배열 및 교차판매, 마케팅 계획
- 기술 : 데이터가 가진 특징 및 의미를 단순하게 설명

로지스틱 회귀

- 종속변수 범주형(=이산형)
- 최대우도법, 가중최소자승법, χ^2 test
- Sigmoid 함수 : log odds값을 연속형 0 ~ 1의 비선형 값으로 바꾸는 함수
- 승산비(odds ratio) = 관심있는 사건이 발생할 상대 비율

분류분석의 종류

로지스틱 회귀, 의사결정나무, 앙상블, 신경망 모형, kNN, 나이브베이즈, SVM, 유전 알고리즘

의사결정 나무

- 나무 구조로 나타내 전체 자료를 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석 방법
- 순수도가 높아지고 불확실성이 낮아지는 방향 분리
- 분류 : 지니, 엔트로피, 카이제곱 통계량의 p-value 작은 것
- 회귀 : 분산 감소량이 큰 것, F 통계량의 p-value 작은 것
- Gini(T) = 1 - ∑(각 범주별수/전체수)² $1-\sum_{i=1}^{k} P_i^2$
- Entropy(T) = $-\sum_{i=1}^{k} P_i \log_2 P_i$
- CART : 지니지수, 분산 감소량 C5.0 : 엔트로피 지수
- CHAID : 카이제곱 통계량의 p-value, ANOVA F-통계량 p-value

3과목 3 정형 데이터 마이닝 - 키워드 살펴보기

앙상블 모형

- 여러 개의 분류 모형에 의한 결과를 종합하여 분류의 정확도를 높이는 방법
- 약하게 학습 된 여러 모델들을 결합하여 사용
- 성능을 분산시키기 때문에 과적합(overfitting) 감소 효과가 있음

Voting

- 서로 다른 여러 개 알고리즘 분류기 사용, Hard voting(빈도수), Soft voting(확률)

Bagging

- 서로 다른 훈련 데이터 샘플로 훈련, 서로 같은 알고리즘 분류기 결합, 원 데이터에서 중복 허용
- 여러 모델이 병렬로 학습, 그 결과를 집계하는 방식

Boosting

- 여러 모델 순차적 학습, 분류가 잘못된 데이터에 가중치를 부여하여 표본 추출, 이상치에 약함
- AdaBoost, GradientBoost (XGBoost, Light GBM)

Random Forest

- 배깅(Bagging)에 랜덤 과정을 추가한 방법
- 여러 개 의사결정 나무를 사용해, 하나의 나무를 사용할 때보다 과적합 문제를 피할 수 있음

kNN

- 새로운 데이터에 대해 주어진 이웃의 개수(k) 만큼 가까운 멤버들과 비교하여 결과를 판단
- 스케일링 중요, k는 hyper parameter, lazy learning, 지도 학습

SVM

- 서로 다른 분류에 속한 데이터 간의 간격(margin)이 최대가 되는 선을 찾아 이를 기준으로 데이터를 분류하는 모델

3과목 3 정형 데이터 마이닝 - 키워드 살펴보기

인공신경망(ANN)	▪ 분류, 예측 모두 가능, 입력층, 은닉층, 출력층으로 구성
Bias, variance	▪ Overfitting : Low Bais High Variance (=유연성 크다, 복잡도 높다)
신경망 활성화 함수	▪ Sigmoid, softmax (3개 이상의 범주)
은닉 층 노드 수	▪ 많으면 - 과적합 문제 발생, 레이어가 많아지면 기울기 소실 문제 ▪ 적으면 - 과소적합 문제 발생, 복잡한 의사결정 경계를 만들 수 없음
기울기 소실	▪ 다층신경망에서 역전파 알고리즘이 입력층으로 갈 수록 Gradient가 점차적으로 작아져 0에 수렴하여, weight가 업데이트 되지 않는 현상
홀드 아웃	▪ 원천 데이터를 랜덤하게 두 분류로 분리하여 성능을 평가하는 방법 ▪ 하나는 모형 학습 및 구축을 위한 훈련용 자료, 다른 하나는 성과평가를 위한 검증용 자료로 사용하는 방법
교차 검증	▪ 데이터가 충분하지 않을 경우 Hold out은 많은 양의 분산 발생, 이에 해결책으로 교차검증 사용 ▪ 클래스 불균형 데이터에 적합하지 않음 ▪ K-fold Cross Validation : 전체데이터 shuffle, K개로 데이터 분할, K 번째의 하부 집합을 검증용 자료, K-1 개는 훈련용 자료로 사용해 K번 측정 후 결과를 평균 낸 값을 사용함
붓스트랩	관측치를 한 번 이상 훈련용 자료로 사용하는 복원추출법에 기반 평가 방법
데이터 불균형 해결 방법	▪ under sampling : 적은 class의 수에 맞추는 것 ▪ over sampling : 많은 class의 수에 맞추는 것

3과목 3 정형 데이터 마이닝 - 키워드 살펴보기

분류 모델 평가지표

오분류표

Confusion matrix		예측값 TRUE	예측값 FALSE	실Sen, 예Pre
실제값	TRUE	40 (TP)	60 (FN) Type II Error	Sensitivity TP / (TP+FN)
실제값	FALSE	60 (FP) Type I Error	40 (TN)	Specificity TN / (TN+FP)
		Precision TP / (TP+FP)	Negative Predictive Value TN / (TN + FN)	Accuracy (TP+TN) / (TP+TN+FP+FN)

FP Rate

- FP Rate = FP / (FP + TN), 1 - Specificity
- 실제가 N 인데 예측이 P로 된 비율 (Y가 아닌데 Y로 예측된 비율, 1종 오류 비율)

F1

F1 = 2 * (Precision * Recall) / (Precision + Recall)

F_2

- $F_\beta = (1+\beta^2) * \frac{precision * recall}{(\beta^2 * precision)+recall}$, $F_2 = (1+2^2) * \frac{precision * recall}{(2^2 * precision)+recall}$

3과목 3 정형 데이터 마이닝 - 키워드 살펴보기

ROC Curve

- X축 FP Rate(1-Specificity), Y축 민감도(Sensitivity)를 나타내 이 두 평가 값의 관계로 모형 평가
- ROC 그래프의 밑부분의 면적(AUC)이 넓을수록 좋은 분류 모형으로 평가함

이익 도표

- 얼마나 예측이 잘 이루어졌는지를 나타내기 위해 임의로 나눈 각 등급별로 반응검출율, 반응률, 향상도 등의 정보를 산출하여 나타내는 도표, 분류모형의 성능 평가 척도

계층적 군집 + 댄드로그램 해석

- 최단 연결법(군집 구하는 방법 이해), 완전 연결법, 평균 연결법, 중심 연결법, 와드 연결법(군집내 오차제곱합에 기초)
- 거리측정에 대한 정의 필요
- 이상치 민감
- 군집수를 사전에 설정할 필요 없음

계층적 군집의 거리

- 유클리드, 맨해튼 구하는 법
- 민코프스키, 마할라노비스, 코사인 거리 특징

비계층적 군집

- K-means
- 혼합분포 군집
- 중심밀도 군집

K-means 군집

비계층적, K를 정해줘야 함
이상값(outlier) 영향 받기 쉬움

K-means 절차

1. 초기 군집의 중심으로 K개의 객체를 임의로 선택한다
2. 각 자료를 가장 가까운 군집의 중심에 할당한다
3. 각 군집 내의 자료들의 평균을 계산하여 군집의 중심을 갱신한다
4. 군집 중심의 변화가 거의 없을 때까지 2, 3을 반복한다

3과목 3 정형 데이터 마이닝 - 키워드 살펴보기

DBSCAN 군집	비계층적, 밀도기반, k 값 없음, 임의적 모양 군집에 적합, outlier 영향 적음
혼합분포군집	데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단 모형에서 나왔다는 가정하에, 추정된 k개의 모형 중 어느 모형으로부터 나왔을 확률이 높은지에 따라 군집 분류를 수행
실루엣 계수	▪ 군집내 거리와 군집 간의 거리를 기준으로 군집 분할 성과를 측정하는 방식 ▪ 클러스터 안의 데이터들이 다른 클러스터와 비교해 얼마나 비슷한가를 나타내는 군집평가 ▪ 실루엣 지표가 1에 가까울수록 군집화가 잘 되었다고 판단
SOM Self Organizing Maps	▪ 인공신경망의 한 종류로, 차원축소와 군집화를 동시에 수행하는 기법 ▪ 비지도 학습, 고차원 데이터를 저차원으로 변환해서 보는데 유용 (2개의 층)
연관분석	항목들 간의 '조건-결과' 식으로 표현되는 유용한 패턴을 연관규칙이라 하며, 이러한 패턴 규칙을 발견해 내는 것을 연관분석이라 함, 장바구니 분석이라고도 하며 Apriori, FP Growth 알고리즘이 있음
연관규칙 측정지표	▪ 지지도 = $P(A \cap B)$: A와 B가 동시에 포함된 거래 수 / 전체 거래 수 ▪ 신뢰도 = $P(B\|A) = P(A \cap B) / P(A)$: A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수 ▪ 향상도 = $P(B\|A)/P(B) = P(A \cap B) / (P(A)*P(B))$
향상도 해석	▪ 1보다 큰 수 : 연관성 높음, 서로 양의 관계 ▪ 1 : 두 품목 사이에 상호 관계가 없음 ▪ 1보다 작은 수 : 서로 음의 상관 관계

문제를 많이 풀어 보시는 것을 "강추"합니다!