1과목

데이터의 유형

▶ 정성적 데이터: 언어·문자로 표현 (ex. 회사 매출이 증가함. ←이런 식으로 기록)
↳ 언어나 문자로 표현하면 저장·검색·분석에 당연히 많은 비용이 소모
↳ 정성스럽게 길게도 썼네~ 이런 식으로 암기

▶ 정량적 데이터: 수치·도형·기호로 표현 (ex. 나이, 몸무게..)
↳ 간단하게 정형화되어있어 비용소모가 적다.
↳ 정량(양) 으로 표현해서 정량적 데이터

지식 경영의 핵심 이슈

▶ 암묵지: 메뉴얼화 되어있지 않고 개인에게 체화되어 있어 겉으로 드러나지 않음
↳ 김치 담그기, 자전거 타기 등등
↳ 개인에게 내면화 → 조직에게 공통화
▶ 형식지: 문서나 매뉴얼화된 지식
↳ 교과서, 비디오 등등
↳ 언어·숫자·기호로 표출화 → 개인의 지식으로 연결화

DIKW

▶ D(Data) 데이터: 가공하기 전의 순수한 수치나 기호
▶ I(Information) 정보: 패턴을 인식하고 의미를 부여한 데이터
▶ K(Knowledge) 지식: 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물
▶ W(Wisdom) 지혜: 창의적 아이디어

데이터베이스의 특징

▶ 통합된 데이터 : 동일한 내용이 중복되어 있지 않음

▶ 저장된 데이터 : 컴퓨터가 접근할 수 있는 저장 매체에 저장됨

▶ 공용 데이터 : 여러 사용자가 데이터를 공동으로 이용

▶ 변화되는 데이터 : 새로운 데이터의 삽입, 기존 데이터 삭제, 갱신으로 항상 변화해도 현재의 정확한 데이터를 유지해야함

▶데이터베이스 설계 절차: 요구사항 분석→개념적 설계→논리적 설계→물리적 설계

기업 내부 데이터베이스

▶ OLTP(On-Line Transaction Processing)
↳데이터베이스의 데이터를 수시로 갱신하는 프로세싱
↳온라인 거래처리로 다양한 과정의 연산이 하나의 단위 프로세싱으로 실행되도록 하는 단순 자동화 시스템

▶ OLAP(On-Line Analytical Processing)
↳ 다차원의 데이터를 대화식으로 분석

▶ CRM(Customer Relationships Management: 고객 관계 관리)
↳ 고객과 관련된 내·외부 자료를 분석해 고객 중심 자원을 극대화하고 이를 토대로 효율적인 마케팅에 활용
↳ 설명에 고객 관리 단어가 있으면 CRM이라 생각

▶ SCM(Supply Chain Management: 공급망 관리)
↳ 공급망 단계를 최적화해 고객을 만족시킴
↳ 공급망·최적화 단어가 있으면 SCM

▶ ERP(Enterprise Resource Planning)
↳ 기업 전체를 경영자원의 효과적 이용이라는 관점에서 통합적으로 관리하고 경영의 효율화를 기하기 위한 시스템
↳ 기업 경영자원·효율화 단어가 있으면 ERP

▶ BI(Business Intelligence)
↳ 기업이 보유하고 있는 데이터를 정리하고 분석해 의사결정에 활용
↳ 리포트 중심의 도구

▶ BA(Business Analytics)
↳ 데이터 통계를 기반으로 성과와 비즈니스 통찰력에 초점을 둔 방법
↳ 통계·수학에 초점

빅데이터

▶ 빅데이터를 나타내는 4V : Volume(양) Variety(다양성) Velocity(속도) Value(가치)
▶ 클라우드 컴퓨팅 기술은 빅데이터 분석에 경제성을 제공해주었다.

빅데이터에 거는 기대 비유

▶ 산업혁명의 석탄, 철: 혁명적 변화를 가져올 것

▶ 21세기 원유: 산업전반에 생산성을 향상시키고 기존에 없던 새로운 범주의 산업 형성

▶ 렌즈: 렌즈를 통해 현미경이 생물학 발전에 미친 영향 만큼이나 데이터가 산업 발전에 영향을 줄 것 (ex. 구글의 Ngram Viewer)

▶ 플랫폼: 공동 활용의 목적으로 구축된 유무형의 구조물로서 역할

본질적인 변화

▶ 사전처리 → 사후처리 : 기술이 발전해서 그냥 닥치는 대로 데이터를 모아도 된다.
▶ 표본조사 → 전수조사 : 마찬가지로 닥치는 대로 데이터를 모으면 된다.
▶ 질 → 양 : 이것도 마찬가지..
▶ 인과관계 → 상관관계 : 엄청나게 모은 데이터들을 분석해 서로 상관이 있는지 본다.

가치 선정이 어려운 이유

▶ 데이터를 재사용하거나 재조합해 활용하면서 특정 데이터를 언제·어디서·누가 활용할지 알 수 없게 되어 가치 산정하기 힘듦
▶ 기술이 발전하면서 '기존에 없던 가치'를 창출해서 가치 산정이 어려움
▶ 현재는 가치가 없어도 나중에 기술이 발전하면 가치가 있는 데이터가 될 수도 있기 때문에 현재 어떤 데이터가 쓸모없는지 확정짓기 힘듦

빅데이터 기본 테크닉

▶ 연관규칙학습 : 커피를 구매하는 사람이 탄산음료를 더 많이 사는가?

▶ 유형분석 : 이 사용자는 어떤 특성을 가진 집단에 속하는가?

▶ 유전자 알고리즘 : 최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송하는가?
↳최적해를 구하는 알고리즘(최적화·최적해 라는 단어가 있으면 유전자 알고리즘)

▶ 회귀분석 : 구매자의 나이가 구매 차량 타입에 어떤 영향을 끼치는가?

▶ 감정분석 : 고객의 평가나 리뷰를 통한 분석

빅데이터 시대 위기 요인

▶ 사생활 침해: SNS에 올린 데이터로 사생활 침해 할 수 있을 것
↳익명화 기술 발전이 필요할 것
↳ 개인정보 사용을 제공자의 동의에서 사용자의 책임으로

▶ 책임 원칙 훼손: 범죄 예측 프로그램을 돌려서 범죄도 저지르지 않은 사람을 체포하면 문제가 될 것
↳ 명확히 나타난 결과에 대해서만 책임을 물어야함 (결과기반책임원칙)

▶ 데이터 오용: 데이터 분석 결과가 항상 옳은 것은 아님
↳ 스티브 잡스가 사람들에게 아이디어를 물었다면 아이폰은 나오지 못했을 것이라 함
↳ 불이익을 당한 사람들을 대변할 전문가(알고리즈미스트)가 필요해짐

빅데이터 활용 3요소

▶ 데이터 기술 인력
↳빅데이터니까 데이터는 당연히 있어야하고 다양한 분석 기술이 있으니 기술도 있어야하고 데이터 사이언티스트나 알고리즈미스트 같은 직업이 나타나므로 인력도 필요하다.

데이터 사이언스

▶ 데이터와 관련된 모든 분야의 전문지식을 종합한 학문
↳정형·비정형을 막론하고 데이터 분석 (총체적 접근법)
↳데이터로부터 의미있는 정보를 추출해 내는 학문

▶ 분석적 영역: 수학, 확률모델, 분석학 등등 과 같은 이론적인 지식
▶ IT: 프로그래밍, 데이터 엔지니어링, 고성능 컴퓨팅 과 같은 프로그래밍적 지식
▶ 비즈니스 분석: 커뮤니케이션, 시각화, 프레젠테이션 과 같은 비즈니스적 능력

데이터 사이언티스트 요구 역량 : 대부분의 전문가들이 호기심이 필요하다고 언급

▶하드 스킬(Hard Skill): 빅데이터에 대한 이론적 지식, 분석 기술
↳(가트너가 제시한 역량에는 포함되어있지 않음)

▶소프트 스킬(Soft Skill): 통찰력(정확성 보다는) 있는 분석능력, 설득력 있는 전달(스토리텔링·시각화), 협력 능력과 같은 인문학적 능력
↳인문학적 능력의 중요성이 왜 나타났는가?
↳단순(컨버전스)세계에서 복잡(디비전스)세계로의 변화
비즈니스의 중심이 생산에서 서비스 & 시장 창조로 이동

DBMS(Data Base Management System)

▶데이터베이스를 공유하고 관리할 수 있는 환경을 제공하는 소프트웨어

▶관계형 DBMS: 테이블(표)로 데이터 정리한다고 생각
▶객체지향 DBMS: 정보를 객체(이미지나 영상)으로 정리한다고 생각

SQL: 데이터 베이스에 접근할 수 있는 데이터 베이스 하부언어
↳집계함수 중 COUNT()만 수치형, 문자형 둘 다 사용가능

개인정보 비식별 기술

▶데이터 마스킹: 홍길동 → 홍**
▶가명처리: 홍길동 → 임꺽정
▶총계처리: 갑:165cm 을:170cm 병:175 → 학생들 키 합 :510cm 평균키 :170cm
▶범주화: 홍길동 35세 → 홍길동 30~40세

하둡: 여러 개의 컴퓨터를 하나인 것처럼 묶어 대용량 데이터를 처리하는 기술

데이터 유형

▶정형 데이터: 관계형 데이터베이스, CSV 등 형식이 정해져 있는 데이터

▶반정형 데이터: 눈으로 봤을 때 무슨 정보인지 모르는 데이터(센서데이터처럼), 한번 변환이 있어야함, 형태(스키마,메타데이터)가 있음

▶비정형 데이터: 소셜데이터, 영상, 이미지 와 같이 형태가 정해져있지 않은 데이터

데이터 웨어하우스

▶사용자의 의사결정에 도움을 주기 위해 정보를 기반으로 제공하는 하나의 통합적이고 시간성을 가지는 비휘발성 데이터의 집합

▶전사적 차원에서 일관적인 형식 유지

2과목

분석 기획에서의 데이터 사이언스 역량

▶ 분석 역량: 문제 영역에 대한 전문성, 수학·통계학적 지식
▶ 균형 잡힌 시각, 데이터·프로그래밍 기술 역량

분석 대상과 방법

방법 \ 대상	Known	Un-Known
Known	최적화 (Optimization)	통찰 (Insight)
Un-Known	솔루션 (Solution)	발견 (Discovery)

▶ 방법과 대상을 둘 다 알면 최적화, 방법을 모르면 방법을 찾아야 하므로 솔루션 둘 다 모르면 발견 대상을 모르면 통찰

분석 기획 방안

목표 시점 별 분석 기획 방안	
과제중심적 접근	장기적인 마스터 플랜
Speed & Test	Accuracy & Deploy
Quick & Win	Long Term View
문제 해결	문제 정의

▶ 과제 중심적 접근은 말그대로 과제를 정확도 보다 빠르게 해결하는데 중점을 둠

분석 기획시 고려사항

▶ 가용 데이터 : 분석의 기본이 되는 데이터 확보 및 파악

▶ 적절한 유스 케이스 : 기존에 잘 구현되어 활용되고 있는 유사 분석 시나리오 및 솔루션 최대한 활용

▶ 장애요소들에 대한 사전계획 수립 : 분석 수행 시 일어나는 문제에 대해 고려가 필요함 이행 저해요소 관리, 일회성 분석에 그치지 않고 조직의 역량으로 내재화

데이터 기반 의사결정의 필요성

▶ 경험과 감에 따른 직관적인 의사결정 → 데이터 기반의 의사결정
▶ 합리적 의사 결정을 가로막는 장애요소: 프레이밍 효과, 고정 관념, 편향된 생각
↳동일 상황임에도 개인의 판단·결정이 달라짐

분석 방법론 (구성요소 : 상세한 절차·방법·도구와 기법·템플릿과 산출물)

▶ 폭포수 모델: 단계를 순차적으로 진행, 이전 단계가 완료되어야 다음 단계로 진행 가능, 하향식(Top Down)으로 진행

▶ 나선형 모델: 여러번의 개발과정을 거쳐 점진적으로 프로젝트 완성
관리체계를 효과적으로 갖추지 못한 경우 복잡도가 상승한다.

▶ 프로토타입 모델: 일부분을 우선 개발하고 개선 작업을 거치는 것 중요한 기능들이 포함되어 있는 시스템의 초기 모델

KDD 분석 방법론

▶ 데이터셋 선택 → 데이터 전처리 → 데이터 변환 → 데이터 마이닝 → 분석 결과 평가

▶ 전처리 과정에서 이상값, 잡음을 식별하고 데이터 변환 과정에서 분석 목적에 맞는 변수 선택 및 차원 축소 과정을 거친다.

CRISP-DM 분석 방법론

▶ 업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 & 분석

▶ CRISP-DM 에서의 데이터 준비 과정은 KDD 분석에서의 데이터 변환과정과 유사
모델링 단계에서 모델 평가는 수행하지만 모델 적용성 평가는 그 다음 단계에서 진행함

빅데이터 분석 방법론

▶ 분석 기획 → 데이터 준비 → 데이터 분석 → 시스템 구현 → 평가 & 전개
↳ 추가적인 데이터 확보가 필요한 경우 데이터 준비 단계로 다시 진행

▶ 분석 기획 단계는 범위 설정 → 프로젝트 정의 & 계획 → 위험 식별 & 대응으로 이루어짐

지도 학습 vs 비지도 학습

지도 학습	비지도 학습
명확한 목적하에 데이터 분석을 실시	데이터 자체의 결합, 연관성을 중심으로 데이터의 상태를 표현하는 것
자료가 입력 변수와 출력변수로 주어짐, 예측 모형을 얻을 때 사용	데이터 마이닝에서 자료가 출력변수 없이 입력변수만 주어지는 경우

하향식 접근 방식: 문제가 주어지고 이에 대한 해법을 찾기 위해 과정이 진행(지도 학습)
분석적으로 사물을 인식하려는 'Why' 관점
문제 탐색 → 문제 정의 → 해결방안 → 탐색 타당성 검토 순으로 진행

▶ 문제 탐색: 빠짐없이 문제를 도출하고 식별하는 것이 중요하다.
문제를 해결함으로써 발생하는 가치에 중점을 두는 것이 중요하다.
거시적 관점: STEEP(사회·기술·경제·환경·정치)
경쟁자 확대 관점: 대체자·경쟁자·신규 진입자
시장의 니즈 탐색 관점: 고객·채널·영향자
비즈니스 모델 기반 : 업무·제품·고객·규제와 감사·지원 인프라
↳「지원 인프라」「업무」 중에 「고객」이 「제품」을 「규제와 감사」 했다. 로 외우기

▶ 타당성 검토: 대안을 과제화 하기 위해서 다각적인 타당성 분석이 수행되어야함
경제적 타당성은 비용대비 편익 분석 관점의 접근이 필요
데이터 타당성은 데이터 존재 여부, 분석 역량이 필요
기술적 타당성은 역량 확보 방안을 사전에 수립

상향식 접근 방식 : 비지도 학습 방법에 의해 데이터 분석을 함
↳문제의 정의 자체가 어려운 경우 데이터를 기반으로 문제를 탐색
사물을 있는 그대로 인식하는 'What' 관점

분석과제 정의서: 분석별로 필요한 소스데이터, 분석방법, 데이터 입수 및 분석 난이도, 상세 분석 등을 정의함

분석 과제 관리를 위한 5가지 주요 요인

▶ 데이터 크기 & 데이터 복잡성 & 속도 & 분석 복잡성 & 정확성·정밀도

▶ 분석 복잡성에서 정확도와 복잡도는 trade off 관계가 존재한다. (정확도를 생각해 분석을 실행하면 복잡해질 것) & 정확성과 정밀도도 trade off 관계가 되는 경우가 많다.

분석 프로젝트 관리 방안 10가지

▶범위 시간 원가 품질 통합 조달 자원 리스크 의사소통 이해관계자
↳외운다면 [범통이] [조리품] [시원 의자] 로 외우기
범위통합이해관계자 조달리스크품질 시간원가 의사소통자원

분석 마스터 플랜 수립

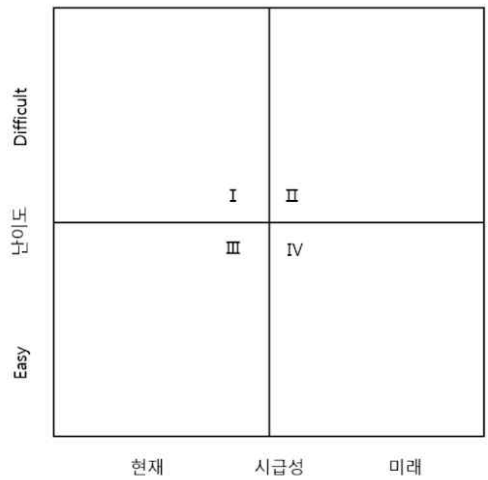
▶ 전략적 중요도, 비즈니스 성과 및 ROI, 실행 용이성 등 기준을 고려해 적용 우선순위 결정
↳전략적 중요도는 전략적 필요성과 시급성으로 이루어짐
실행 용이성은 투자 용이성과 기술 용이성으로 이루어짐

▶ 업무 내재화 적용 수준, 분석 데이터 적용 수준, 기술 적용 수준을 고려해 Analytics 구현 로드맵 수립

ISP: 정보기술 또는 정보 시스템을 전략적으로 활용하기 위하여 중장기 마스터 플랜을 수립하는 절차

ROI 관점 빅데이터 특징

▶ 투자 비용 요소(난이도) : 3V= Volume, Variety, Velocity
▶ 비즈니스 효과(시급성) : Value



사분면 분석

▶우선순위 : 시급성 기준으로는 Ⅲ→Ⅳ→Ⅱ (반시계 방향)
난이도 기준으로는 Ⅲ→Ⅰ→Ⅱ (시계 방향)

분석 마스터 플랜 세부 이행 계획 수립

▶ 폭포수 방식도 있으나 반복적인 정련과정을 통하여 프로젝트의 완성도를 높이는 방식을 주로 사용

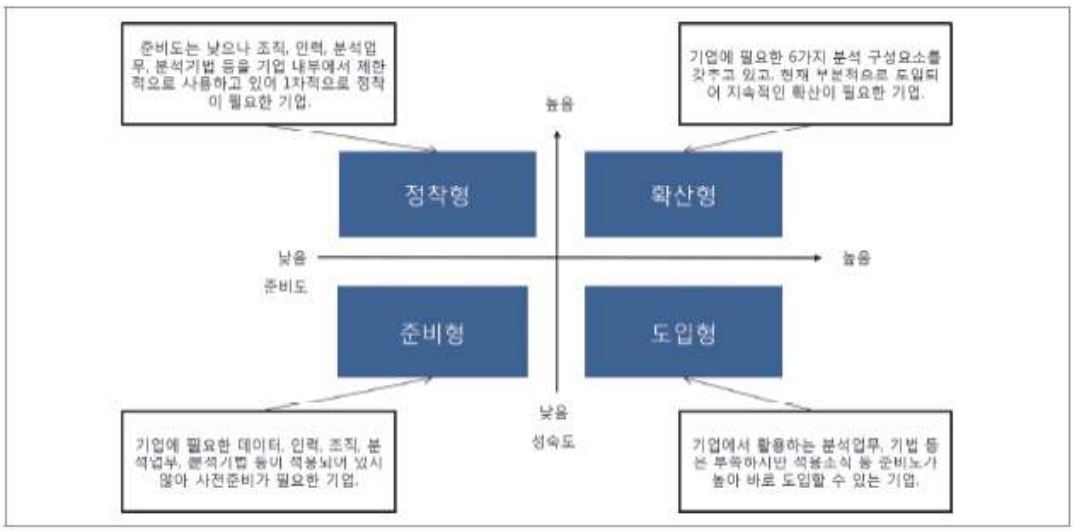
▶ 반복적인 분석체계라고 모든 단계를 반복하는 것이 아닌 모델링 단계를 중점으로 반복적으로 진행하고 세부적인 일정계획도 수립해야함

분석 거버넌스 구성 요소 : 조직(Organization) 프로세스(Process) 시스템(System) 데이터(Data) 분석 관련 교육 및 마인드 육성 체계(Human Resource)

데이터 분석 수준 진단

▶ 분석 준비도 : 분석 업무·분석 인력·분석 기법·분석 데이터·분석 문화·분석 인프라

▶ 분석 성숙도 : 도입 → 활용 → 확산 → 최적화



출처: http://www.dbguide.net/

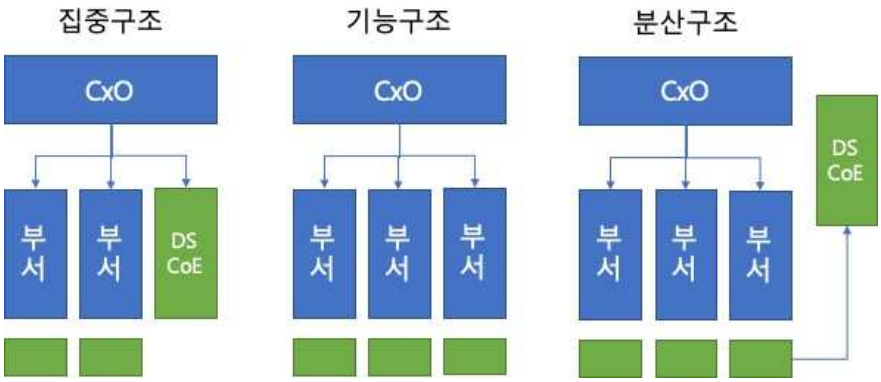
준비도가 x축 성숙도가 y축일 때 왼쪽 위를 시작으로 시계방향으로 정 확 도 준 으로 외우기

데이터 거버넌스: 전사 차원의 모든 데이터에 대하여 표준화된 관리체계를 수립하고 운영을위한 프레임 워크 및 저장소를 구축하는 것

▶마스터 데이터, 메타 데이터, 데이터 사전은 데이터 거버넌스의 중요한 관리 대상이다.
↳마스터 데이터는 사람 이름, 생일과 같이 변하지 않아 처리 운영에 기본이 되는 자료
▶구성 요소 : 원칙 조직 프로세스

▶데이터 거버넌스 체계에는 데이터 표준화, 데이터 관리 체계, 데이터 저장소 관리, 표준화 활동이 있다.
↳뭔가를 구축한다는 내용이 있으면 데이터 표준화
데이터 관리라는 내용이 있으면 데이터 관리 체계
데이터 저장소 내용이 있으면 데이터 저장소 관리

분석을 위한 3가지 조직 구조



출처: https://wikidocs.net/

▶ 집중 구조는 말 그대로 분석 부서 하나가 따로 있는 것, 현업 업무부서와 분석 업무와 이원화 가능성이 높다. (이원화라는 말이 있으면 집중구조)
▶ 기능 구조는 별도의 분석 조직(DSCoE)이 없음
▶ 분산 구조는 분석 조직 인력들을 현업 부서로 직접 배치, 신속한 Action이 가능함.

분석 과제 관리 프로세스

▶ 과제 발굴 : 분석 idea 발굴 → 분석 과제 후보 제안 → 분석 과제 확정 ↴

▶ 과제 수행 : 결과 공유 & 개선 ← 분석 과제 진행 관리 ← 분석과제 실행 ← 팀 구성

※ 분석 과제 후보 제안, 결과 공유 단계에서만 Pool이라는 것에 관리함

CMMI(능력 성숙도 통합 모델): 1~5단계로 구성된 성숙도 모델

3과목

탐색적 자료분석(EDA): 데이터가 가지고있는 특성을 파악하기위해 시각화하여 분석하는 방식, 시각화하면 이상점을 식별하기 쉽다.

▶ EDA의 4가지 주제: 저항성의 강조·잔차 계산·자료변수의 재표현·그래프를 통한 현시성

기술통계: 모집단으로부터 표본을 추출하고 표본이 가지고 있는 정보를 하나의 숫자 또는 그래프의 형태로 표현 (우리가 흔히 생각하는 통계)

추측통계: 모집단으로부터 추출된 표본의 표본통계량으로 모집단을 통계적으로 추론

데이터마이닝 모델링

▶ 대표적인 고급 데이터 분석법이다.
↳'시뮬레이션'도 고급 분석법 중 하나
▶ 지나치게 통계적 가설이나 유의성에 집착하지 말고 훈련 및 테스트 성능에 큰 편차가 없고 예상 성능을 만족하면 중단한다. 반드시 다양한 옵션을 주어야 하는 것이 아니다.

R (이론보다는 문제를 보고 이론을 배우는 편이 좋음)

▶ 오픈소스 프로그램이며 많은 패키지가 수시로 업데이트되고 질의를 위한 커뮤니티가 매우 활발

▶ 문제가 발생할 경우 다양한 의견들을 들을 수 있으나 신속하게 유지보수가 되는 것은 아님
↳왜냐하면 전문가가 후딱 고치는 것이 아니라 사용자들끼리 얘기하는 것이기 때문

▶ S 언어 기반 (GNU S라고도 함)

R기초

▶ 벡터 생성 c: c(1,2,3)은 [1 2 3]이고 숫자형 벡터이지만 c(1,2,'a')처럼 문자형이 원소로 하나라도 껴있으면 문자형 벡터가 됨. c(1:5)는 1부터 5까지라는 뜻 :를 ~로 생각하면 될 듯
↳즉 숫자가 아니므로 +나 -와 같은 연산이 안될 것
↳True/False 는 1/0을 나타내고 pi는 원주율을 나타내므로
c(3.14 pi True)는 문자형 벡터가 아님

▶ 패키지 설치 및 로드 : install.packages("패키지명") → library(패키지명)

▶ 행렬을 as.vector함수에 입력하면 열방향으로 1열부터 원소를 나열하는 벡터 생성
예를 들어 $\begin{matrix}1&3\\2&4\end{matrix}$ 행렬이 있다면 as.vector 함수에 넣으면 [1 2 3 4]가 됨(열방향으로! 행방향 아님)

▶ 만약 [1 2] + [3 4 5 6 7]를 하게되면 상식적으로는 계산이 불가하지만 R에서는 가능함
[3+1 4+2 5+1 6+2 7+1] 와 같이 연산됨(경고메세지와 함께 출력) 부족한 벡터성분을 짧은 벡터의 원소([1 2])를 재활용해가면서 사용.

▶ summary함수는 4분위수, 최소, 최대, 중앙값, 평균을 출력함

▶ 데이터프레임: 2차원 목록(list) 데이터 구조, 각 열이 다른 데이터 타입을 가질 수 있다.
↳데이터 테이블: 데이터 프레임과 유사하지만 보다 월등히 빠른 그루핑과 ordering, 짧은 문장 지원측 면에서 더 매력적

▶na.rm=T 는 na이가 not answer(결측값) rm이 remove(삭제) T가 True라 생각하면 결측값을 삭제하라는 뜻으로 생각.
↳mean(x,na.rm=T)는 결측값을 제외한 x의 평균이라는 뜻

데이터 마트

▶ 데이터 웨어하우스와 사용자 사이의 중간층에 위치

▶ CRM관련 업무 중에서 핵심

요약변수

▶ 수집된 정보를 분석에 맞게 종합한 변수, 재활용성이 높다.

파생변수

▶ 특정조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여한 변수

▶ 주관적일 수 있어 논리적 타당성을 갖추어야한다.

sqldf: R에서 sql명령어를 사용 가능하게 해주는 패키지

plyr: apply 함수를 기반으로 가장 필수적인 데이터 처리기능 제공

결측값(na) 처리 방법

▶ completes analysis: 결측값이 존재하는 레코드 삭제
↳삭제하면 활용할 수 있는 변수의 수가 작아져 효율적이지 못함

▶ 평균 대치법: 결측값을 데이터의 평균으로 대치

▶ 다중 대치법: 대치 → 분석 → 결합 단계로 진행

이상값

▶ 잘못 입력한 경우, 의도하지 않게 입력되었고 분석 목적에 맞지 않는 경우
↳이런 경우는 웬만하면 삭제하는 것이 좋음

▶ 꼭 제거해야하는 것은 아니므로 적절한 판단 필요

이상값 인식 방법 3가지

▶ ESD: 평균으로부터 3 표준편차 떨어진 값

▶ 기하평균-2.5*표준편차 < data < 기하평균+2.5*표준편차를 벗어나는 값

▶ 사분위수 이용하기: Q1-1.5*(IQR)<data<Q3+1.5*(IQR)를 벗어나는 값
↳Q1: 데이터의 하위 25%에 해당하는 값(순서로 봤을 때)
Q3: 데이터의 상위 25%에 해당하는 값
IQR=Q3-Q1
보통 이러한 범위를 넘어가면 제거하지 않고 범위의 최대 최소값으로 바꾸어 활용

▶ 사기탐지, 의료, 부정사용방지 등등에 쓰일 수 있다.

통계자료 획득

▶ 전수조사: 전체를 다 조사하는 것 → 시간과 비용이 많이 소요.

▶ 표본조사: 일부만 추출하여 모집단(표본이 포함된 전체 집단)을 분석

표본 추출 방법

▶ 단순랜덤 추출법: 말 그대로 랜덤하게 표본을 뽑음

▶ 계통추출법: 번호를 랜덤하게 부여한 후 특정한 간격별로 추출

▶ 집락추출법: 군집을 나눈 후 군집별로 단순랜덤 추출

▶ 층화추출법: 계층을 고루 대표할 수 있도록 표본 추출

표본 오차 & 표본 편의

▶ 표본오차는 모집단을 대표할 수 있는 표본 단위들이 조사대상으로 추출되지 못함으로서 발생하는 오차

▶ 표본편의는 모수를 작게 또는 크게 할 때 추정하는 것과 같이 표본추출방법에서 기인하는 오차를 의미, 확률화로 최소화하거나 없엘 수 있다.

척도 구분

▶ 명목척도: 어느 집단에 속하는지 분류 (성명/성별 등등)

▶ 순서척도: 서열관계가 있을 때 (만족도/학년/등수 등등)

▶ 구간척도: 속성의 양을 측정하는 것으로 구간이나 구간 사이의 간격이 의미가 있는 자료 (온도, 지수 등등), 절대적 원점은 없다.

▶ 비율척도: 절대적 기준인 0이 존재하고 사칙연산이 가능 (무게/ 나이 등등)

확률변수

▶ 이산형: 0이 아닌 확률값을 갖는 셀 수 있는 실수값
↳이산형 확률 변수로는 베르누이·이항·기하·다항·포아송 분포가 있음

▶ 연속형: 확률이 함수형태로 주어져 있다고 생각하면 됨.
↳균일분포·정규분포·지수분포·t-분포·카이제곱 분포·F-분포 가있다.
t-분포는 평균이 동일한지 알고자할 때 사용
카이제곱 분포는 모분산에 대한 가설 검정에 사용
F-분포는 분산의 동일성 검정에 사용

조건부 확률

▶ $P(A|B) = \frac{P(A \cap B)}{P(B)}$ A와 B가 독립(배반)사건 이라면 $P(A \cap B) = P(A) * P(B)$ 이다.
↳뒤에 있는 B가 분모로 감

추정: 표본으로부터 모수(모집단의 특징)을 추측하는 것

▶ 점추정: 모수가 특정한 값일 것 이라고 추정하는 것

▶ 구간추정: 모수가 특정한 구간에 있을 것이라고 선언하는 것
추정량의 분포에 대한 전제가 주어져야 하고 구해진 구간 안에 모수가 있을 신뢰구간이 주어져야한다.

가설검정

▶ 귀무가설: 비교하는 값과 차이가 없다를 기본개념으로 하는 가설
↳ex.) '우리 반 평균키는 대한민국 남성 평균키보다 크다'라는 주장에 가설검정을 한다면 귀무가설은 '우리 반 평균키는 대한민국 남성 평균키와 차이가 없다.(같다.)'를 가설로 두는 것 (일반적으로 기각되는 것이 목표)

▶ 대립가설: 뚜렷한 증거가 있을 때 주장하는 가설 (채택되는 것이 목표)

▶ p값(value): 귀무가설이 사실일 때, 관측된 검정통계량이 대립가설을 지지하는 방향으로 나올 확률 ↳관찰된 표본으로부터 구한 통계량
우리가 내린 판정이 잘못되었을 실제 확률을 의미

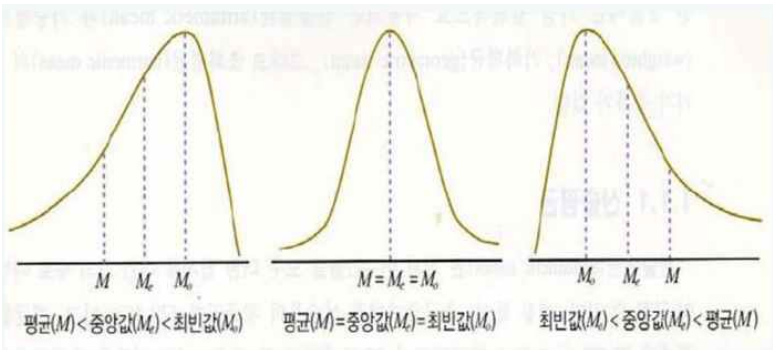
▶ 유의수준: 귀무가설이 옳은데도 이를 기각(채택 안하는)하는 확률의 크기

▶ 제1종 오류: 귀무가설이 옳은데 귀무가설을 기각하게 되는 오류
▶ 제2종 오류: 귀무가설이 틀린데 귀무가설을 채택하는 오류

비모수적 방법

▶ 자료가 추출된 모집단의 분포에 아무 제약 않고 검정 실시
▶ 분포의 형태에 대해 설정
▶ 절대적인 크기에 의존하지 않는 관측값들의 순위나 두 관측값 차이의 부호 등을 이용
▶ 부호검정, 윌콕슨의 순위합검정, 윌콕슨이 부호순위합검정, 만-위트니의 U검정, 스피어만 순위상관계수 (순위, 부호가 들어가면 비모수적 방법이라고 기억, + U검정)

왜도: 분포의 비대칭 정도를 나타내는 측도(평균 중앙값 최빈값 관계 그림으로 기억)



왜도<0 왜도=0 왜도>0

상관분석 : 데이터 안의 두 변수간의 관계를 알아보기 위한 것

▶ 상관계수의 절댓값이 0이면 전혀 상관이 없는 것이고 0.3보다 작으면 거의 상관이 없는 것 절댓값이 0.3과 0.7사이라면 약한 상관관계가 있는 것 1과 0.7사이라면 강한 상관이 있는 것

▶ cor() 혹은 rcorr() 함수로 상관계수를 구할 수 있다.

피어슨 상관계수	스피어만 상관계수
등간척도인 변수	서열척도인 변수
연속형 변수, 정규성 가정	순서형 변수, 비모수적 방법
두 변수간의 선형관계 크기 측정	비선형 관계도 측정가능

↳스피어만 상관계수의 특징은 'ㅅ'이 많이 들어가 있다.

▶ R로 상관분석을 했을 때, p-value값이 0.05이하인 경우 귀무가설을 기각하고 대립가설을 채택할 수 있다. 즉 변수 간 상관관계가 있다고 볼 수 있다.

회귀분석: 하나 또는 그 이상의 독립변수들이 종속변수에 미치는 영향을 추정하는 통계법
(함수의 개념과 비슷)

▶ p-값이 0.05보다 작으면 추정된 회귀식은 통계적으로 유의수준 5%이하에서 유의미하다고 봄

▶ 결정계수는 R-squared값으로 확인하면 되고 1에 가까울수록 설명력이 높다고 판단할 수 있다. 총 변동 중에서 회귀모형에 의하여 설명되는 변동이 차지하는 비율이다.

▶ 선형회귀분석의 가정

선형성	입력변수와 출력변수의 관계가 선형이다.
등분산성	오차의 분산이 입력변수와 무관하게 일정하다.
독립성	입력변수와 오차는 관련이 없다.
비상관성	오차들끼리 상관이 없다.
정상성	오차의 분포가 정규분포를 따른다. (Q-Q plot, Shapiro-Wilks test로 확인)

[정성(선)스럽게 비등한 독립] 으로 외우기 → 선형성 말고는 오차에 대한 내용이 주를 이룸.
↳정상성성(선)형성스럽게 비상관성등분산성한 독립성

▶ 데이터의 정규성은 Q-Q plot, Shapiro-Wilks test, 히스토그램을 사용해 확인 가능

단계적 변수선택

▶ 전진선택법: 절편만 있는 상수모형으로부터 시작해 중요하다고 생각되는 변수를 차례로 추가

▶ 후진제거법: 모든 변수를 포함한 모형에서 출발해 가장 적은 영향을 주는 변수부터 하나씩 제거

▶ 단계선택법: 전진선택법으로 변수를 추가하는데 기존 변수가 영향을 받아 중요도가 약화되면 변수를 다시 제거하는 등 단계별로 추가,제거 여부를 검토하는 방법

▶ 최적회귀방정식은 모든 후보 모형들에 대해 AIC, BIC를 계산하고 그 값이 최소가 되는 모형 선택 (AIC, BIC가 뭔지는 굳이 알 필요 없음)

시계열 자료: 시간의 흐름에 따라 관찰된 값을 뜻함

정상시계열 (일반적으로 분산이 시점에 의존하지 않음을 나타냄)

▶ 모든 시점에 대해 일정한 평균과 분산을 가진다.
▶ 특정한 시차의 길이를 갖는 자기공분산을 측정하더라도 동일한 값을 갖는다.

비정상시계열을 정상시계열로 전환하는 방법

▶ 평균이 일정하지 않은 경우: 원시계열에 차분(현 시점에서 바로 전 시점의 자료값을 뺌)
▶ 계절성을 갖는 경우: 계절차분 사용
▶ 분산이 일정하지 않은 경우: 자연로그를 취함('변환' 이라고 함)

자기회귀 모형 (AR)

▶ 시계열 모델 중 자기 자신의 과거 값을 사용하여 설명하는 모형

▶ 백색 잡음의 현재값과 자기 자신의 과거값의 선형 가중합으로 이루어진 정상 확률 모형

분해시계열: 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법
↳ 경향(추세)요인, 계절요인, 순환요인, 불규칙요인으로 이루어짐
경향은 말 그대로 자료가 오르거나 내리는 추세를 의미
계절은 고정된 주기에 따라 자료가 변하는 경우
순환은 경제적이나 자연적인 이유 없이 알려지지 않은 주기를 갖고 변화
불규칙은 위 3가지로 설명할 수 없을 때

다차원척도법(MDS)

▶ 객체간 근접성을 시각화하는 통계기법
▶ 개체들을 2차원 또는 3차원 공간상에 점으로 표현하여 개체들 사이의 집단화를 시각적으로 표현하는 분석방법
▶ 계량적 MDS → 비율척도, 구간척도 데이터 활용
▶ 비계량적 MDS → 순서척도 데이터 활용

주성분분석(PCA)

▶ 여러 변수들을 상관관계를 이용해 소수의 주성분으로 차원을 축소하는 것

▶ Scree plot을 이용하는 경우에는 그래프의 기울기가 완만해져 거의 0에 다다르는 지점에서 주성분의 개수를 구한다.

▶ 대략 85%의 분산설명력을 갖게끔 주성분의 수를 결정한다.

관측치 수 = 자유도(df) + 1

데이터마이닝

▶ 대용량 데이터에서 의미있는 데이터 패턴을 파악, 예측하여 의사결정에 활용하는 방법

▶ 가설이나 가정없이 다양한 수리 알고리즘을 이용

▶ 데이터 마이닝 도구가 다양하고 체계화되어 환경에 적합한 제품을 선택하여 활용 가능

▶ 알고리즘에 대한 깊은 이해가 없어도 분석에 큰 어려움이 없다.

▶ 분석 결과의 품질을 위해서 풍부한 경험을 가진 전문가가 하면 좋다.

지도학습	비지도학습
인공신경망 의사결정나무 회귀분석 로지스틱회귀분석 사례기반추론	OLAP 연관성 규칙발견 군집분석 SOM

↳지도학습에는 [인공 의사 회귀]로 외우기
군집분석 & SOM은 비지도학습인 것 기억

데이터마이닝 추진 단계

▶ 목적설정 → 데이터 준비 → 가공 → 기법 적용 → 검증

데이터 분할

▶ 구축용(train data): 데이터마이닝 모델을 만드는데 활용하며 보통 50%사용
▶ 검정용(validation data): 구축된 모형의 과대추정 또는 과소추정을 미세 조정을 하는데 활용 보통 30% 사용
▶ 시험용(test data): 모델의 성능을 검증하는데 활용 보통 20% 사용

▶ 홀드아웃 방법: 주어진 데이터를 학습용과 시험용 데이터로 분리하여 사용하는 방법
▶ 교차확인 방법: 주어진 데이터를 k개의 집단으로 구분하여 k-1개의 집단을 학습용으로, 나머지는 검증용으로 설정해 학습하는 방법

성과 분석 (중요)

예측 \ 실제	참	거짓
참	1	2
거짓	3	4

▶ 1번은 실제로도 참인데 예측도 참이라고 한 경우
2번은 실제로는 거짓인데 예측을 참이라고 한 경우
3번은 실제로 참인데 예측을 거짓이라고 한 경우
4번은 실제로도 거짓인데 예측도 거짓이라고 한 경우

▶ Accuracy= (1+4)/(1+2+3+4)
↳실제와 맞게 예측한 확률이라고 볼 수 있다.
▶ 특이도(Specificity)= 4/(2+4)
↳실제로 거짓인 사건을 예측도 거짓으로 한 확률이다.
▶ 정확도(Precision)= 1/(1+2)
↳예측을 참이라고 했는데 실제로도 참일 확률이다.
▶ 재현율(Recall)=민감도(Sensitivity)= 1/(1+3)
↳실제로 참인 경우 중에 예측을 참으로 한 확률

※쉽게 생각하면 정확도는 예측력이라고 생각해라. 분모에는 예측을 참이라고 한 경우를 합한 (1+2)가 들어간다.
※재현율과 특이도는 분모에 실제를 기준으로 넣으면 된다. 특이도의 경우에는 실제로 거짓인 경우를 합한 (2+4) 재현율은 실제가 참인 경우를 합한 (1+3)이 들어간다.

▶ F1 Score= $2*\frac{Precision*Recall}{Precision+Recall}$

▶ ROC Curve: 가로축을 1-특이도 세로축을 민감도 값으로 두어 시각화한 그래프
그래프의 면적이 클수록(1에 가까울수록) 모형의 성능이 좋다고 평가

분류분석

▶ 데이터가 어떤 그룹에 속하는지 예측하는데 사용되는 기법

▶ 지도학습에 해당

▶ 인공신경망, 의사결정나무, 회귀분석 등등 위에서 다뤘던 지도학습의 대부분이 분류분석에 속함

▶ 향상도 곡선: 분류 분석의 모형평가 방법으로 랜덤 모델과 비교하여 해당 모델의 성과가 얼마나 향상되었는지를 각 등급별로 파악. (단답형문제)

로지스틱 회귀분석

▶ 반응변수가 범주형인 경우에 적용되는 회귀분석 모형

▶ exp(β)는 나머지 변수(x1,....xk)가 주어질 때 x1이 한단위 증가할 때마다 성공의 오즈가 몇 배 증가하는지를 나타내는 값 ('오즈' 기억하고 그냥 같이 증가하는구나 알고 있으면 됨)
↳오즈=p/(1-p)= 확률/(1-확률) : 성공할 확률이 실패할 확률의 몇 배인지를 나타냄

▶ β>0이면 S자모양 β<0이면 역 S자 모양이 됨

의사결정나무

▶ 연속적으로 발생하는 의사결정 문제를 시각화해 의사결정이 이뤄지는 시점과 성과를 한눈에 볼 수 있게 한다. 해석이 간편하다.
▶ 예측력에 치중할 때도 있고 해석력에 치중할 때도 있다.
▶ 대용량 데이터에서도 빠르다.
▶ 비정상 잡음 데이터에 대해서도 민감함이 없다.
▶ 상관성이 높은 다른 불필요한 변수가 있어도 크게 영향을 받지 않는다.
↳조건이 맞지 않는 변수는 그냥 버리면 되니 크게 영향을 안받을 것

▶ 새로운 자료에 대한 과대적합이 발생할 수 있다.
↳과대적합: 너무 자세하게 만들어서 다른 자료에 적용할 때 성능이 떨어짐
과소적합: 모형이 너무 단순해서 성능이 떨어짐

▶ 아래로 내려갈수록 각 마디에서의 불순도는 감소한다.

형성과정

▶ 성장 (분리규칙·정지규칙) → 가지치기 → 타당성평가(이익도표·위험도표) → 해석 및 예측
↳과대적합되어 현실문제에 적응할 수 있는 적절한 규칙이 나오지 않는 현상 방지

불순도 측도

▶ 카이제곱 통계량 & 지니지수 & 엔트로피 지수 (책 예시로 이해하는게 쉬움)

의사결정나무 알고리즘

▶ CART: 가장 많이 활용되는 알고리즘, 출력변수가 범주형일 경우 지니지수, 연속형일 경우 이진분리 사용

▶ C4.5 /C5.0: 범주형 입력변수에 대해서는 범주의 수만큼 분리가 일어남, 측도로는 엔트로피지수 사용

▶ CHAID: 가지치기를 없이 적당한 크기에서 나무 성장을 중지, 입력변수는 반드시 범주형 사용, 측도로는 카이제곱 통계량 사용(알고리즘 이름이 카이드니까 카이제곱 사용)

앙상블 분석: 여러개의 예측모형들을 만든 후 조합

▶ 배깅: 여러개의 부스트랩 자료를 생성한 후 각 자료에 예측모형을 만든 후 결합
가지치기를 하지않고 최대로 성장한 의사결정나무 활용

▶ 부스팅: 배깅과 다른 점은 각 자료에 동일한 가중치를 주는 것이 아닌 분류가 잘못된 데이터에 더 큰 가중을 준다.

▶ 랜덤포레스트: 배깅에 랜덤과정을 추가한 방법

▶ 샘플에 한번도 선택되지 않는 원데이터가 발생할 수 있는데 전체 샘플의 36.8%가 해당

인공신경망

▶ 인간의 뇌를 기반으로한 추론모델
▶ 역전파 알고리즘을 활용해 비선형성 극복한 모형 등장
  ↳연결강도를 갱신하기 위하여 예측된 결과와 실제값의 차이인 error로 가중치 조절
▶ 활성화함수를 사용해 출력을 결정(입력변수의 속성에 따라 선택하는 것은 아님)
  ↳시그모이드 함수라는 것이 있는데 0~1의 확률값을 갖으며 로지스틱 회귀분석과 유사

▶ softmax 함수: 출력값이 여러개로 주어지고 목표치가 다범주인 경우 (단답형에 가끔 나옴)

▶ 은닉노드의 수는 적절히 큰 값으로 놓고 가중치를 감소시키며 적용하는 것이 좋음
  ↳ 은닉층의 뉴런개수는 자동으로 설정되는 것이 아닌 직접 설정해야함.
  너무 많이 설정하면 과대적합 너무 적으면 과소적합 일어남 (기울기 소실문제)

군집분석

▶ 군집의 개수나 구조에 대한 가정없이 데이터들 사이의 거리를 기준으로 군집화 유도
  ↳인공신경망은 개수를 직접 설정해줘야 했음

▶ 각 개체의 유사성을 측정하여 분류하고 서로 다른 군집의 객체들과의 상이성을 규명하는 분석방법

실루엣: 군집분석의 품질을 정량적으로 평가하는 대표적인 지표, 군집 내 데이터간 거리가 짧을 수록 군집 간 거리가 멀수록 값이 커짐
(단답형으로 가끔 나옴, 분류분석은 향상도 곡선)

연속형 변수 거리

▶ 유클리디안 거리: 우리가 흔히 아는 좌표계에서의 거리
▶ 표준화 거리: 표준화하게되면 척도, 분산의 차이로 인한 왜곡을 피할 수 있다.(개념 기억)
▶ 마할라노비스 거리: 통계적 개념 포함, 변수의 표준화와 상관성을 동시에 고려(개념 기억)
▶ 맨하탄 거리, 민코우스키 거리: 책에 나와있는 식 확인

범주형 변수 거리 (개념기억)

▶ 자카드 유사도: Boolean 속성으로 이루어진 두 객체간의 유사도 측정에 사용된다.
▶ 코사인 유사도: 두 단위 벡터의 내적을 이용, 내각의 크기로 유사도를 측정

계층적 군집 분석

▶ 최단연결법: 최단거리를 이용해 군집형성, 고립된 군집을 찾는데 중점, 사슬모양의 군집이 생길 수 있음
▶ 최장연결법: 최장거리를 이용해 군집형성, 내부 응집성에 중점을 둠
▶ 중심연결법: 중심간의 거리를 이용해 군집형성
▶ 평균연결법: 계산량이 많지만 모든 데이터를 포함하는 하나의 군집 형성
▶ 와드연결법: 군집내의 오차제곱합에 기초하여 군집 형성

비계층적 군집 분석 (K 평균 군집분석)

▶ 원하는 군집의 개수와 초기값들을 정해 seed를 중심으로 군집을 형성
↳ 집단 내 제곱합 그래프를 참고해 군집수를 정할 수 있음(어려움)
▶ 한번 군집이 형성되어도 개체들은 다른 군집으로 이동할 수 있음
▶ 잡음이나 이상값에 영향을 많이 받음
▶ 주어진 목적이 없어 결과 해석이 어렵다. (극복하기위해 PAM 방법 사용)
▶ 볼록한 형태가 아닌 군집이 존재할 경우 성능이 떨어짐
▶ 내부 구조에 대한 사전정보가 없어도 의미있는 자료구조를 찾을 수 있다는 장점이 있다.

혼합분포군집

▶ EM 알고리즘이 사용된다. (이름만 기억)

SOM(자기조직화지도)

▶ 입력변수의 위치 관계를 그대로 보존한다는 특징이 있다.
▶ 역전파 알고리즘을 사용하는 인공신경망과 달리 단 하나의 전방 패스를 사용
▶ 연결강도는 입력패턴과 가장 유사한 경쟁층 뉴런이 승자가 된다.
▶ BMU: SOM에서 선택된 프로토타입 벡터 (이름 기억)
▶ 입력 변수의 개수와 동일하게 뉴런 수가 존재
▶ 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬

연관분석: A→B일 때 (커피를 사는 사람은(A) 탄산음료도 산다(B))

▶ 지지도: $\frac{P(A \cap B)}{\text{전체}}$
▶ 신뢰도: $\frac{P(A \cap B)}{P(A)}$ (A확률이 분모)
▶ 향상도: $\frac{P(A \cap B)}{P(A)*P(B)}$ (독립/배반사건이면 1이 될 것)
↳향상도가 1보다 크면 해당 규칙이 결과를 예측하는데 있어 우수하다는 것을 의미