# I. 데이터 이해

## 1. 데이터의 이해

**1) 데이터**

- 존재적 특성: 객관적 사실

- 당위적 특성: 추론, 예측, 전망, 추정을 위한 근거

**2) 특징**

- 정성적 데이터: 언어, 문자

- 정량적 데이터: 수치, 도형, 기호

**3) 지식**

- 암묵지: 학습과 경험을 통한 지식(내면화->공통화 필요)

- 형식지: 형식화된 지식(표준화-> 연결화)

## 2. 데이터와 정보

**1) 데이터의 활용 단계**

- 데이터: 객관적 사실

- 정보: 연관관계=>의미 도출

- 지식: 정보 구조화&분류+개인적 경험=내재화

- 지혜: 지식의 축적+아이디어=창의적 산물

## 3. 데이터베이스

**1) 정의**

- EU: 체계적으로 정리되고 전자식으로 개별접근 가능한 독립된 데이터

- 국내 저작권법: 소재를 체계적으로 구성한 편집물로 개별적으로 접근 가능한 것

- 국내 컴퓨터 용어사전: 동시에 복수 업무 지원이 가능한 데이터의 집합

**2) 특징**

- 통합된 데이터: 중복 없음

- 저장된 데이터: 저장매체에 저장

- 공용 데이터: 공동 이용

- 변화되는 데이터: 변화하면서도 현재의 정확한 데이터 유지

## 4. 빅데이터

**1) 정의**

Mckinsey: 일반적인 DBMS 이용 규모를 초과하는 데이터

IDC: 대규모 데이터로부터 저비용 고가치 추출 가능한 차세대 기술

Doug Laney: 3V(Volume, Variety, Velocity)

**2) 빅데이터 정의의 범주 및 효과**

- 데이터 변화: 규모, 형태, 속도

- 기술변화: 신데이터 처리, 저장, 분석 기술, 클라우드 컴퓨팅 활용

- 인재, 조직 변화: 데이터 중심 조직으로 변화

**3) 빅데이터 활용 기본 테크닉**

- 연관 규칙 학습, 유형분석, 유전 알고리즘, 기계학습, 회귀분석, 감정분석, 소셜네트워크 분석

**4) 위기요인**

- 사생활 침해 -> 동의에서 책임으로

- 책임 원칙 훼손 -> 결과 기반 책임 원칙 고수

- 데이터 오용 -> 알고리즘 접근 허용

**5) 빅데이터로 인한 변화**

- 사전처리 -> 사후처리

- 표본조사 -> 전수조사

- 데이터의 질 -> 데이터의 양

- 인과관계 -> 상관관계

**6) 빅데이터의 미래**

- 모든 것의 데이터화, 인공지능, 데이터 사이언티스트, 알고리즈미스트

## 5. 데이터 사이언스

**1) 빅데이터 회의론 원인**

- 부정적 학습 효과: 과거의 CRM(고객 관계 관리) -> 공포 마케팅, 투자대비 효과 미흡

- 부정적 성공 사례: 빅데이터가 불필요한 분석사례, 기존 CRM 분석 성과

**2) 실패 사례**

- 싸이월드 퇴보 원인: OLAP같은 분석인프라 있으나 활용 못함, 웹 로그와 같은 일차원적 분석에 집중, 관련 프레임워크 혹은 평가지표 없음, 전략적 통찰 없음

**3) 전략적 통찰이 없는 분석의 함정**

- 일차원적 분석의 반복: 해당 부서 업무 영역에는 효과적, 환경|고객 변화 대처의 어려움

- 가치기반 분석 단계: 일차원적 분석 -> 경험 -> 활용범위 확장 -> 사업성과

**4) 데이터 사이언스**

- 데이터로부터 정보 추출-> 분석 -> 구현 -> 전달

- 영역: 분석 분야, IT 분야, 비즈니스 분야

- 다양한 유형의 데이터(정형, 비정형)를 대상으로 분석

- 데이터 공학, 수학, 통계학, 컴퓨터공학, 시각화, 해커의 사고방식, 분야별 전문지식 종합

**5) 데이터 사이언티스트**

- 역할: 데이터 소스 발견, 대용량 데이터 구조화, 불완전 데이터 연결

- 요구사항: 창의적 사고, 호기심, 논리적 비판을 통한 통찰력, 전달력, 커뮤니케이션 능력

**6) 빅데이터와 데이터 사이언스의 미래**

- 외부 환경적 측면의 인문학 열풍

세계화: 단순 세계화->복잡한세계화(컨버전스->디버전스)

비즈니스 중심: 제품 생산 위주 -> 서비스(품질->서비스)

경제산업 논리: 생산->시장창조(공급자 중심 기술경쟁->무형자산의 경쟁)

- 가치 패러다임의 변화

1단계 디지털화: 가치 형상화, 표준화(문서의 디지털화)

2단계 연결: 정보의 효율적인 연결 및 제공(인터넷 모바일 기술의 발전)

3단계 에이전시: 필요정보 효과적 제공 및 관리(개인, 기기, 사물 하이퍼 연결)

- 데이터 사이언스의 한계: 분석에서 인간의 개입(가정->해석)

# II. 데이터 분석 기획

## 1. 데이터 분석 기획

**1) 분석 방향성 도출**

- 분석 기획: 분석과제 정의, 관리 방안 사전 계획

- 고려 사항: 데이터에 대한 고려(확보, 유형), 활용방안, 유즈케이스, 장애요소 대책

**2) 분석 방법론**

- KDD: 9개의 프로세스와 5단계의 절차를 통해데이터 마이닝(순차적)

데이터셋 선택 -> 데이터 전처리 -> 데이터 변환 -> 데이터 마이닝 -> 결과평과

- CRISP-DM: 4계층과 일정치 않은 방향의 6단계 프로세스로 구성된 데이터마이닝(비순차적)

업무 이해 <-> 데이터이해 -> 데이터준비 <-> 모델링 -> 평가 -> 업무이해 회귀 or 전개

- 빅데이터: 3계층과 5단계의 절차로 구성

분석기획 -> D준비 -> D분석 -> 시스템 구현 -> 평가 및 전개

**3) 분석 과제**

분석 과제 정의서: 소스 데이터, 분석 방법, 데이터 입수 및 분석의 난이도, 분석 수행 주기, 검증 오너십, 상세 분석 과정 등 정의

**4) 하향식 접근 방식 분석 과제  발굴**

- 주어진 문제를 단계적으로 수행하는 방식

- 문제 탐색

비즈니스 모델 기반의 5개 영역: 업무, 제품, 고객, 규제와 감사, 지원 인프라

경제적요인(STEEP): 사회, 기술, 경제, 환경, 정치

경쟁자 확대 관점: (직접 경쟁자, 제품, 서비스) + 대체재, 신규진입자

시장의 니즈 탐색 관점: 고객, 채널, 영향자

역량의 재해석 관점: 내부 역량 영역, 파트너&네트워크 영역

- 문제 정의: 비즈니스 문제 -> 데이터와 분석 문제로 변환

- 해결방안 탐색: 분석역량, 분석기법 및 시스템 활용해서 해결방안 탐색

- 타당성 검토: 경제적 타당성, 데이터 및 기술적 타당성

**5) 상향식 접근 방식 분석 과제 발굴**

- 원천 데이터 분석을 통해 통찰력과 지식을 얻는 방식

- 비지도 학습 방법으로 수행, 데이터 자체의 결합, 연관성, 유사성 중심 접근

- 프로토 타이핑 접근법: 시행착오를 통한 문제 해결(빅데이터 분석 환경에서 필요)

## 2. 분석 마스터플랜

**1) 마스터 플랜 수립**

- ISP: 중장기 마스터플랜 수립 절차

- 비즈니스 관점 분석 과제 -> 적용 우선순위 평가

- 4V(크기, 다양성, 속도 + 가치)를 고려해서 우선순위 기준 적용

- ROI관점의 특징: 투자비용(크기, 다양성, 속도), 비즈니스 효과 요소(가치)

- 우선순위: 전략적 중요도(전략적 필요성, 시급성), 실행 용이성(투자 용이성, 기술 용이성)

- 기업의 분석적용 수준 고려->시급성, 난이도 평가기준에 맞춰 우선순위 조정

- 단계적 로드맵, 추진 계획 수립 필요

**2) 분석 거버넌스 체계**

- 정의: 표준화된 관리체계 수립, 운영을 위한 프레임 워크, 저장소 구축

- 목적: 데이터 분석의 지속 적용과 확산(마스터 플랜 수립 시점)

- 데이터 거버넌스 구성요소: 원칙, 조직, 프로세스

- 데이터 거버넌스 체계

데이터 표준화: 용어 설정, 명명 규칙, 메타데이터 구축, 데이터사전 구축

데이터 관리 체계: 메타데이터 및 데이터 사전 관리원칙 수립

데이터 저장소 관리: 메타데이터 및 표준 데이터를 관리하기 위한 저장소 구성

표준화 활동: 모니터링을 통해 표준 준수 여부 점검, 변화 관리 및 교육

- 분석 조직

집중구조: 별도의 전담 조직에서 담당

기능구조: 별도의 조직 없이 해당 업무 부서에서 분석

분산구조: 분석 조직 인력을 현업 부서에 배치

**3) 분석 거버넌스 수준 진단**

- 분석 준비도: 업무, 데이터, 인력/조직, 문화, 기법, IT인프라

- 분석 성숙도

도입: 환경과 시스템 구축

활용: 실제 업무 적용

확산: 전사차원에서 관리 및 공유

최적화: 분석 진화-> 혁신 및 성과 향상

- 성숙도 진단 결과

준비형(준비도 하, 성숙도 하): 사전준비가 필요한 상태

정착형(준비도 하, 성숙도 상): 준비도는 낮으나 분석 기법을 제한적 사용(1차적 정착 필요)

도입형(준비도 상, 성숙도 하): 기법은 부족하지만 준비도가 높아서 바로 도입 가능

확산형(준비도 상, 성숙도 상): 6가지 분석 구성요소를 갖추고 부분적 도입(지속적 확산 필요)

# III - 1. 데이터 분석

## 1. 데이터 분석 기법

**1) 데이터 처리**

- 데이터 웨어하우스, 데이터마트를 통해 분석 데이터 활용

- 스테이징, 운영데이터저장소(ODS)에 저장된 신규데이터를 DW, DM에 결합

**2) 시각화**

- 복잡한 분석을 효율적으로 해석(차트, 트리, 다이어그램 맵, 워드 클라우드)

- 탐색적 분석시 필수

**3) 공간분석**

- 공간적 차원과 관련된 속성의 시각화

- 공간적 차원에 생성한 속성의 크기, 모양, 선굵기 등을 통해 인사이트 도출

**4) 탐색적 자료분석(EDA)**

- 다양한 차원의 값을 조합해가며 관계 도출, 빅데이터에 활용

- 주제: 저항성의 강조, 잔차 계산, 자료변수의 재표현, 그래프를 통한 현시성

**5) 통계분석**

- 어떠한 현상을 체계기반으로 숫자, 표, 그림 형태로 표현

- 기술통계: 표본의 정보를 정리, 요약 -> 숫자, 그래프 형태로 표현

- 추측통계: 표본통계량 -> 모수(모집단의 특성) 통계적 추론

**6) 데이터마이닝**

- 대용량의 정보 요약 -> 관계, 패턴, 규칙 탐색 -> 모형화 -> 지식 추출

- 방법론

DB 지식탐색: 데이터웨어하우스에서 데이터마트 생성 -> 속성 사전 분석 -> 지식 추출

기계학습: 컴퓨터 자체 학습 알고리즘(인공신경망, 의사결정나무, 클러스터링, 베이지안분류, SVM 등)

패턴인식: 사전지식 및 패턴 통계정보 추출 -> 자료 및 패턴 분류(장바구니 분석, 연관규칙)

**7) 시뮬레이션**

- 복잡한 상황 모델링 후 결과 예측 및 변경 활용

**8) 최적화**

- 목적함수와 제약조건 정의로 목표값 개선

# 2. R

**1) R 기능 기초**

- 함수 선언 방법: function(매개변수, 매개변수,...) {}

- setwd("작업디렉토리"): 작업환경 설정

- help(), ?함수, RsiteSearch(): 도움말

- history(), savehistory(file="파일명"), loadhistory(file="파일명"): 히스토리

- Ctrl+L: 콘솔 청소

- Ctrl+R: 한 줄 실행(드래그 하면 여러 줄 실행)

- #: 주석 처리

- install.packages("패키지명"): 패키지 자동 설치(매개변수 "패키지위치" 추가시 수동)

- R CMD BATCH 주소\batch.R: 배치(자동 구동)파일 실행

**2) 변수**

- 변수명 선언 후 입력된 값을 스스로 인식

- 변수에 값 할당: 대입연산자 활용( <-, <<-, =, ->, ->>)

- a<-1, a=1: 변수 a에 1 할당

- print(변수명), 변수명: 출력

- ls(): 메모리에 불필요 변수 여부 확인(삭제는 rm())

**3) 기본 코드**

- seq(from=시작점, to=끝지점, by=증가값, length.out=갯수 지정, length=간격): 수열생성

- rep(시작,time=횟수, each=각반복횟수): 반복 /  rep(c(내용), 횟수) : 문자열 반복

- substr("문자열",시작,끝): 해당 문자열 추출

**4) 벡터**

- c(), V[n]: 벡터 생성

- mean(): 평균

- sum(): 합계

- median(): 중간값

- sd(): 표준편차

- var(): 분산

- cov(): 공분산

- cor(): 상관계수

- split(vec, fac): 집단 분할(벡터, 팩터값 일치 필수)

**5) 행렬**

- v<-c(v, newItems): 데이터 추가

- f<-foctor(v, levels): 요인 생성

- rownames(mtrx)<-c("rowname1","rowname2"...): 행 이름 설정(colnames로 열 이름)

- apply(mtr, 1, func): 함수 적용

**6) 데이터 프레임**

- data.frame(v1,v2...): 데이터 프레임 생성

- dfm[[1]], dfm[["name"]], dfm$name 등: 데이터 프레임 내 데이터 선택

- rbind(), cbind(), merge(): 데이터 프레임 결합

- subset(): 데이터 프레임 추출

- split(dfm, fac): 데이터 프레임 내 집단 분할

- apply(dfm,func): 함수 적용(동질인 경우 가능), lapply, sapply

## 3-1. 데이터마트(데이터 변경 요약)

**1) 요약변수**

- 수집된 정보를 분석에 맞게 종합하는 변수(데이터마트의 기본 변수)

- 재활용성: 많은 모델이 공통 사용 가능

**2) 파생변수**

- 특정 함수에 의해 값을 만들어 의미를 부여

- 주관성: 논리적 타당성 필요

**3) reshape 패키지**

- 2개의 핵심 함수 melt() (//녹임) + cast() (//구조 생성)로 구성

**4) sqldf 패키지**

- R에서 sql 명령어를 사용가능하게 해주는 패키지

**5) plyr패키지**

- apply함수를 기반으로 데이터 및 출력변수 동시 치환 처리

- split-apply-combine방식: 분리 -> 처리 -> 결합

**6) data.table 패키지**

- R에서 가장 많이 사용하는 데이터 핸들링 패키지

- 목적: 대용량 데이터 탐색, 연산, 병합 / 기존 data.frame 방식보다 빠름

## 3-2. 데이터마트(데이터 가공 및 기초분석)

**1) 변수의 구간화**

- 변수들을 구간화해서 점수를 적용하는 방식

- binning: 연속형 변수->범주형 변수 변환 // 50개 이하 구간에서 점차 구간 축소

- 의사결정나무: 연속형 변수-> 범주형 변수 변환 // 모형 활용

**2) 결측값**

- 변수에 데이터가 비어있는 경우: Null, NA, ., 999999, Unknown, Not Answer등

- 단순 대치법: completes analysis(결측값 레코드 삭제)

평균 대치법: 실험 데이터의 평균으로 대치

조건부 평균 대치법: 회귀분석을 통해 대치

\*단순확률 대치법: 평균대치법 발전형(Hot-deck, nearest Neighbor)

- 다중 대치법: 단순 대치법 m번 실시->m개의 가상 자료를 만들어 대치

**3) 이상값**

- bad data: 오입력, 비적합 값 -> 삭제

- 이상값: 의도 외의 입력, 의도된 극단값-> 처리해서 활용

- 이상값 인식 오차: 평균~3표준편차 / 기하평균~2.5표준편차 / 1,3사분위 값~2.5배

- 이상값 처리: 절단(레코드 삭제), 조정(상하한 조정)

## 4. 통계분석

**1) 통계**

- 특정 집단 대상 조사 결과에 대한 요약된 형태의 표현

- 총조사: 대상 집단 모두 조사(큰 비용 및 시간 소모)

- 표본조사: 표본을 추출해서 조사

모집단: 조사 대상 집단 전체

원소: 모집단 구성 개체

표본: 모집단의 일부 원소

모수: 모집단을 나타내기 위한 정보값

- 표본 추출: 모집단을 대표할 수 있는 표본 추출

단순랜덤추출: 임의 추출

계통추출: 구간 구분 후 임의 인덱스 위치 추출

집락추출: 군집별 단순랜덤 추출

층화추출: 유사한 원소끼리 층을 나누어서 랜덤 추출

- 측정

명목척도: 집단 분류

순서척도: 서열관계 측정

구간척도: 속성의 양 측정

비율척도: 간격에 대한 비율에서 의미 도출

**2) 통계분석**

- 통계적추론: 수집자료 기반으로 모집단에 대한 의사 결정

모수추정: 모집단의 특성(평균, 분산 등)을 분석

가설검정: 특정 가설 설정 후 채택 여부 결정

예측: 미래의 불확실성 해결

- 기술통계: 주관 배제 후 객관성 확보(평균, 표준편차, 중위수, 최빈값, 그래프의 표현)

**3) 확률 및 확률 분포**

- 확률 변수: 특정값이 나타날 가능성

표본공간: 실험 결과물의 집합

사건: 관찰자가 관심있는 사건(표본공간의 부분집합)

원소: 나타날 수 있는 모든 결과

- 이산형 확률변수: 확률값을 셀 수 있는 경우(0제외)

베르누이분포, 이항분포, 기하분포, 다항분포, 포아송분포

- 연속형 확률변수:  확률값이 특정 구간 전체에 해당

균일분포, 정규분포, 지수분포, t분포, F분포, x제곱 분포

**4) 추정 및 가설검정**

- 추정: 표본으로부터 미지의 모수 추측

- 점추정: 표본 평균, 중위수, 최빈값 등을 사용해서 모수의 값 추정

- 점추정 조건: 불편성, 효율성, 일치성, 충족성

- 구간추정: 점추정의 정확성을 보완하기 위해 모수의 특정 구간 추정(신뢰수준 필수)

- 모분산이 알려진 경우 표준정규분포 사용, 그렇지 않으면 모분산 대신 표본분산 사용(t분포)

- 가설검정: 가설 설정 후 표본 관찰을 통해 가설 채택여부 결정

- 귀무가설: 비교대상값과 동일함을 전제하는 가설

- 대립가설: 정확한 증거를 전제하는 가설

1종오류: 옳은 귀무가설 기각

2종오류: 틀린 귀무가설 채택

오류해결: 1종 오류의 크기를 고정하고 2종오류의 최소 기각역 설정

**5) 비모수 검정**

- 모수적 검정: 특정 가정 하에 검정통계량 및 분포를 유도해 검정

- 비모수 검정: 모집단의 분포에 대한 제약 없이 검정(가정이 불가능할 경우 사용)

관측값의 순위나 차이의 부호 등을 이용해서 검정

- 순위 or 두 관측값 차이를 부호 이용 검정

- 비모수 검정의 예시: 부호검정, 윌콕슨의 순위합검정; 부호순위합검정, 만-위트니의 U검정, 런검정, 스피어만의 순위상관계수

## 5. 기초 통계 분석

**1) 기술 통계**

-자료의 특성을 표, 그림, 통계량 등을 사용해서 정리요약하는 기초적 통계

**2) 통계량에 의한 자료분석**

- 중심위치: 평균, 중앙값, 최빈값

- 산포의 척도: 분산, 표준편차, 범위, 사분위수범위, 변동계수, 표준오차

- 분포의 형태: 왜도, 첨도

**3) 그래프를 통한 자료 분석**

- 범주형자료: 막대그래프와 파이차트 등

- 연속형자료: 히스토그램, 줄기-잎 그림, 상자그림 등

- 시계열자료: 꺾은 선 그래프

**4) 연관성 분석**

- 종속변수: 다른변수의 영향을 받는 변수

- 독립변수: 영향을 주는 변수

- 산점도: 좌표평면 위에 점으로 표현

- 산점도에서 확인사항: 선형관계, 함수관계 성립 여부, 이상값 존재여부, 집단의 수

- 공분산: 두 확률변수 간의 방향성을 확인

**5) 상관 분석**

- 정의: 두 변수간의 관계를 알아보는 분석방법(상관계수 이용)

- 상관계수는 -1~1사이 값(0이면 무관한 상태)

- 피어스 상관계수: 등간척도 이상으로 측정된 두 변수의 상관관계(연속형 변수, 정규성 가정)

- 스피어만 순서상관계수: 순서 or 서열척도인 두 변수의 상관관계(순서형 변수, 비모수)

## 6. 회귀 분석

**1) 정의**

- 독립변수가 종속변수에 미치는 영향 추정이 가능한 통계기법

- 단순선형회귀분석: 독립변수가 하나인 경우

- 다중선형회귀분석: 독립변수가 둘 이상인 경우

**2) 선형 회귀 분석**

-선형회귀분석의 가정

선형성: 입력변수와 출력변수의 관계

등분산성: 오차의 분산이 입력변수와 무관하게 일정

독립성: 독립변인과 잔차는 관련 없음

비상관성: 오차들간 상관없음

정상성: 오차의 분포가 정규분포를 따름

- F검정: 회귀식(모형) 검증

- t검정: 회귀계수들의 의미 검증

- 모형의 설명력: 결정계수로 확인

- 모형의 적합성: 잔차와 종속변수의 산점도로 확인

**3) 다중 선형 회귀 분석**

- 다중공선성: 다중회귀 분석에서 설명변수들간 선형관계 존재시 정확한 회귀계수 추정 불가

- 다중 공선성 검사법

분산팽창요인(VIF): 10 이상 심각한 문제

상태지수: 10 이상은 문제, 30이상은 심각한 문제

- 해결: 선형관계가 강한 변수 제거, 주성분 회귀, 능형회귀로 활용

**4) 변수선택법**

-가능한 모든 독립변수들의 조합에 대한 회귀모형 분석 후 적합한 모형 선택

-전진선택법: 절편만 있는 상수모형으로 시작, 중요한 설명변수부터 차례로 모형에 추가

\*이해가 쉽고 많은 변수에서 활용 가능하지만 안정성 부족(작은 변동에도 결과가 달라짐)

\*변수 추가시 기존 변수들의 중요도에 영향을 줌

-후진소거법: 독립변수를 모두 포함 후 가장 적은 영향을 주는 변수부터 하나씩 제거

\*전체 변수들의 정보 이용 가능, 변수가 많은 경우 활용이 어려움, 안정성 부족

-단계별 방법: 전진선택법에 의해 변수 추가, 중요도 약화시 삭제

## 7. 시계열 분석

**1) 시계열 자료**

- 시간의 흐름에 따라 관찰된 값들

- 분석목적: 미래의 값 예측, 시계열 데이터의 특성 파악(경향, 주기, 계절성, 불규칙성)

- 분석방법: 회귀분석방법, Box-Jenkins방법, 지수평활법, 시계열 분해법 등

**2) 정상성**

- 일정한 평균: 모든 시점에 대해 일정한 평균을 가짐

평균이 일정하지 않은 시계열 -> 차분(현시점 자료에서 전 시점 자료 제거)으로 정상화

- 일정하지 않은 평균, 분산도 시점에 미의존: 변환을 통해 정상화

- 공분산도 특정 시점에서 t, s 미의존 상태로 일정함

**3) 정상시계열의 특징**

- 자기공분산 측정시 동일한 값(어떤 시점에서 평균과 분산, 시차의 길이를 갖는 자기공분산)

- 평균값 회귀 경향(평균값 주변에서 일정한 폭을 유지한 변동)

- 비정상시계열은 다른 시기로 일반화 불가

**4) 시계열 모형**

- 자기회귀 모형(AR 모형)

자기상관함수(ACF)는 빠르게 감소하고 부분자기함수(PACF)는 절단점이  존재

- 이동평균 모형(MA모형)

자기상관함수(ACF)는 절단점이 존재하고 부분자기함수(PACF)는 빠르게 감소

- 자기회귀누적이동평균 모형(ARIMA(p,d,q)

비정상 시계열 모형으로 차분, 변환을 통해 AR, MA, ARMA 모형으로 정상화

**5) 분해 시계열**

- 일반적인 요인을 시계열에서 분리해서 분석하는 방법(주로 회귀분석 사용)

- 추세요인: 자료 그림의 형태 오르내림 추세, 선형, 이차식 형태, 지수적 형태

- 계절요인: 요일반복, 월 변화, 분기 변화, 주기 변화

- 순환요인: 명백한 이유 없이 주기변환하는 자료

- 불규칙요인: 위 세가지 요인으로 설명 불가한 오차 요인

## 8. MDS/PCA

**1) 다차원 척도법(MDS)**

- 객체간 근접성을 시각화하는 통계기법

- 개체들의 변수 측정 후 유사성/비유사성 측정, 2~3차원 공간에 점으로 표현

- 개체간 거리계산은 유클리드 거리행렬 활용(루트 각값의차제곱++)

- 계량적 MDS:  데이터가 구간척도, 비율척도인 경우 활용

N개의 케이스에 대해 p개의 특성변수 존재시 거리를 계산한 비유사성 S를 공간상에 표현

- 비계량적 MDS:  데이터가 순서척도인 경우 활용

순서척도를 거리의 속성과 같도록 변환하여 거리를 생성해서 적용

- (s-)stress와 정합도 수준 M: 부적합도 기준으로 사용, 부적합도 최소화 목적 반복 수행

**2) 주성분 분석(PCA)**

- 상관관계 내 변수를 결합 상관관계 없는 변수로 분산 극대화, 선형결합으로 변수축약

- 요인 분석: 요인 추출을 위한 방법의 집합(주성분 분석은 원소) / 데이터 축소에 활용

- 주성분 분석: 보통 2개의 변수 생성(중요도에 따라 제1,2성분으로 분류), 목표 변수 고려

**3) 주성분 분석 활용**

- 주성분(상관성 높은 변수들)의 선형결합으로 고상관성 변수들을 요약

- 목적

주성분차원으로 변수 축소로 이해 및 관리 용이

다중공선성 존재시 상관도 낮은 변수 축소로 모형개발에 활용

고연관성 변수의 주성분 분석으로 군집화, 연산속도 개선

기계 센서데이터를 주성분분석으로 차원 축소 후 시계열 분석으로 고장징후 포착

# III - 2. 데이터 마이닝

## 1. 데이터마이닝

**1) 데이터마이닝 개요**

- 정의: 대용량 데이터에서 의미있는 패턴 파악, 예측으로 의사결정에 활용하는 방법

- 활용: 시각화분석, 분류, 예측, 군집화, 포캐스팅

**2) 작업 유형**

- 예측 목적

분류규칙: 분류모형을 만들어 레코드의 결과값 예측(회귀분석, 판별분석, 신경망, 의사결정나무)

- 설명 목적

연관규칙: 데이터간 관계 발견(동시 발생 매트릭스)

연속규칙: 연관규칙에 시간관련 정보 포함(동시발생 매트릭스)

데이터 군집화: 특성에따라 그룹화, 결과값 없음(K-Means Clustering)

**3) 학습법**

- 교사학습: 목적변수가 존재하는 분석

의사결정나무, 인공신경망, 일반화 선형 모형(회귀 분석, 로지스틱 회귀 분석), 사례기반 추론, k-Nearest Neighbor

- 비교사학습: 목적변수 없이 설명을 위한 분석

OLAP, 연관성 규칙발견, 군집분석, SOM

**4) 추진단계**

- 목적 설정: 명확한 목적 설정

- 데이터 준비: 다양한 데이터 준비, 정제를 통해 품질 보장

- 데이터 가공: 목적변수 정의, 모델링을 위한 데이터 형식으로 가공

- 기법 적용: 데이터 마이닝 기법 적용으로 정보 추출

- 검증: 추출 결과 검증 -> 업무 적용 -> 투자대비 성과(ROI) 등 전파

**5) 데이터 분할**

- 구축용: 50% 데이터를 모델링 훈련용으로 활용

- 검정용: 30% 데이터를 구축된 모형의 과대/과소 추정의 판정 목적 활용

- 시험용: 20% 데이터를 테스트or과거 데이터를 활용해서 모델 성능 평가에 활용

## 2. 의사결정나무 분석

**1) 분류분석 VS 예측분석**

- 공통점: 레코드의 특정 속성의 값 예측

- 차이점

분류분석: 레코드의 범주형 속성 예측

예측분석: 레코드의 연속형 속성 예측

- 분류 모델링: 신용평가모형, 사기방지모형, 이탈모형, 고객세분화

- 분류기법: 회귀분석, 로지스틱 회귀분석, 의사결정나무, CART, C5.0, 베이지안 분류, naive Bayesian, 인공신경망, 지지도벡터기계, k 최근접 이웃, 규칙기반의 분류와 사례기반추론

**2) 특징**

- 분류함수를 의사결정 규칙으로 이루어진 나무 모양으로 그리는 방법

- 의사결정 문제를 시각화 -> 의사결정 시점, 성과 파악

- 계산 결과가 직접 나타나서 분석이 간편함

- 분류 정확도 높음

- 쉬운 계산으로 대용량 데이터에서도 빠른 작성

- 비정상 잡음 데이터에 민감함 없이 분류 가능

- 한 변수와 상관성 높은 불필요 변수가 있어도 영향 없음

**3) 활용**

- 세분화: 특성에 따른 그룹화를 통해 그룹별 특성 발견

- 분류: 예측변수들에 근거해서 목표변수의 범주를 등급별 분류

- 예측: 자료간의 규칙을 발견해서 미래의 사건을 예측

- 차원축소 및 변수선택: 목표변수에 영향이 큰 예측변수 도출

- 교호작용효과의 파악: 예측 변수들의 결합으로 목표 변수에 작용하는 규칙 파악

- 범주의 병합, 연속형 변수의 이산화: 범주형 목표변수의 범주 병합, 연속형 목표변수 이산화

**4) 의사결정나무 분석**

- 분석단계

성장: 각 마디에서 최적의 분리규칙을 발견해서 나무를 성장, 정지규칙 만족 시 중단

가지치기: 오차에 대한 위험도가 높거나 부적절한 추론규칙을 가진 가지와 불필요 가지 제거

타당성 평가: 이익도표, 위험도표, 시험자료를 이용해서 평가

해석 및 예측: 구축된 나무 모형 해석해서 예측모형 설정 후 적용

- 대형모형: 과대적합 위험 // 소형모형: 과소적합 위험

마디에 속한 자료가 일정 수 이하일 때 분할을 정지하고 가지치기 실시해서 해결

- 불순도에 따른 분할 측도: 카이제곱 통계랑, 지니지수, 엔트로피 지수

**5) 의사결정나무 분석의 종류**

- CART: 연속형 목표변수를 예측하는 의사결정 나무

불순도의 측도가 범주형 목적변수일 경우에는 지니지수, 연속형 목적변수일 경우엔 분산을 이용한 이진분리를 사용

- C4.5 & C5.0: CART와는 다르게 다지분리 가능, 범주 수 만큼 분리가 일어남, 불순도의 측도로 엔트로피 지수를 사용

- CHAID: 가지치기 없이 적당한 크기까지 성장 후 중지, 범주형 변수를 입력변수로 사용, 불순도의 측도로 카이제곱 통계량 사용

## 3. 앙상블 분석

**1) 앙상블 기법**

- 여러개의 예측 모형 조합해서 최종 예측모형 생성

- 종류: 다중 모델조합 방법, classifier combination 방법

- 학습방법의 불안정성 해결 목적(의사결정나무: 불안정성 높음 / 1-Nearest Neighbor: 낮음)

**2) 배깅**

- 주어진 자료에서 여러개의 부트스트랩 자료 생성 후 각 예측모형 생성, 결합 모형 생성

- 훈련자료를 모집단으로 사용해서 평균 예측 모형 생성(분산을 줄이고 예측력은 향상 가능)

**3) 부스팅**

- 예측력이 약한 모형 결합해서 강한 예측 모형 생성

- 훈련 오차를 쉽고 빠르게 줄일 수 있다.

- 경우 예측오차가 향상되어 배깅에 비해 성능이 뛰어난 경우가 많음

**3) 랜덤 포레스트**

- 배깅, 부스팅에 비해 많은 무작위성을 주어 약한 학습기 생성 후 선형 결합하는 방법

- 단점: 이론적인 설명이나 결과에 대한 해석이 어려움

- 장점: 예측력이 매우 높음(입력 변수가 많을수록 예측력이 높아짐)

**4) 스태킹**

- 다양한 학습 모델을 통해 구성

## 4. 인공신경망 분석

**1) 인공신경망**

- 인간의 뇌를 기반으로 한 추론 모델

- 세포체: 뉴런 / 수상돌기: 입력 / 축색돌기: 출력 / 시냅스: 가중치

**2) 인공신경망 연구**

- 1943년 매컬럭과 피츠가 인간의 뇌를 디지털 네트워크 모형으로 간주하고 신경세포의 신호처리 과정을 모형화해서 단순 패턴분류 모형개발

- 헵: 신경세포 사이의 연결강도 조정해서 학습규칙 개발

- 로젠블럿: 퍼셉트론 인공세포를 개발했으나 비선형성 한계점 발생 -> XOR 문제

- 홉필드, 러멜하트, 맥클랜드: 역전파 알고리즘을 활용해서 비선형성을 극복한 다계층 퍼셉트론 개발

**3) 뉴런**

- 기본적인 정보처리 단위

- 입력 링크에서 가중치가 있는 여러 신호를 받아 활성화 수준 계산 후 하나의 출력신호 출력

- 전이함수(활성함수) 사용: 입력 가중치합 계산 -> 임계값 비교(가중치 합<임계값 ? -1 : 1)

**4) 신경망모형 구축시 고려사항**

- 입력변수: 복잡성으로 인해 입력자료 선택에 매우 민감

범주형 변수: 모든 범주에서 일정 빈도 이상의 값을 가지고 범주의 빈도가 일정 -> 가변수화 적용

연속형 변수: 입력변수 값들의 범위가 큰 차이가 없을 때-> 평균 기준 비대칭 분포시 안좋은 결과 도출,  변환 및 범주화 활용

- 가중치 초기값: 역전파 알고리즘의 경우 초기값 선택이 매우 중요

가중치 0이면 시그모이드 함수는 선형, 신경망 모형은 선형 모형이 됨

일반적으로 초기값은 0 근처의 랜덤값 선정, 선형모형은 가중치 증가에 따라 비선형으로 변화

- 학습모드

온라인 학습모드: 각 관측값을 순차적으로 신경망에 투입(가중치 추정값 매번 바뀜)

확률적 학습모드: 온라인 학습 모드와 같으나 신경망에 투입되는 관측값 순서가 랜덤

배치 학습 모드: 전체 훈련자료를 동시 투입

- 학습률: 상수값 사용 / 처음에는 큰값->반복->해가 가까울 수록 0에 수렴

- 은닉층, 은닉노드의 수: 많으면 가중치 많아져서 과대적합 / 적으면 과소적합 문제 발생

은닉층 수 결정: 은닉층이 하나인 신경망은 범용근사자가 되므로 가급적 1개로 설정

은닉노드 수 결정: 적절한 큰값으로 결정 -> 가중치 감소시키며 모수에 대한 벌점화 적용

- 과대적합 문제: 많은 가중치를 추정해야하므로 과대적합 문제 빈번

조기종료: 모형 적합 과정에서 검증 오차 증가하면 반복 중지

(선형모형의 능형회귀와 유사한)가중치 감소 벌점화 기법

- 신경망 모형은 비용함수, R은 비볼록 함수 / 여러개의 국소 최소값을 가짐

오차가 작은 것 or 평균을 구해서 최종 예측값 선정 / 배깅 적용으로 최종 예측치 선정

**5) 로지스틱 회귀분석**

- 반응변수가 범주형인 경우 적용되는 회귀분석모형

- 새로운 설명변수 or 예측변수 적용 시 반응변수의 각 범주에 속할 확률 추정(예측 모형)하여  추정 확률을 기준치로 분류하는 목적(분류모형)으로 활용

- 사후확률: 모형의 적합을 통해 추정된 확률

- glm(): 로지스틱 회귀분석 실행

- 표현: glm(종속변수~독립변수1+...+독립변수k, family=binomial, data=데이터셋명)

## 5. 군집분석

- 객체의 유사성 측정 -> 유사성 높은 대상집단 분류 -> 유사성과 상이성 규명

- 특성에 따라 여러개의 배타적인 집단으로 나누는 것

- 군집의 개수, 구조에 대한 가정 없이 데이터로부터 거리를 기준으로 군집화 유도

**1) 군집분석의 특징**

- 비교사학습법 -> 목표변수(종속변수) 정의 없이 학습 가능

- 목적에 따라 분석자가 군집 정의 가능

- 요인분석과의 차이: 유사한 변수가 아닌 객체를 묶어줌

- 판별분석과의 차이: 집단이 없는 상태에서 집단 구분

**2) 군집분석의 거리측정**

- 연속형 데이터: 유클리드 거리, 표준화 거리, 마할라노비스 거리, 체비셔프 거리, 맨하탄 거리, 캔버라 거리, 민코우스키 거리 등 활용

- 범주형 데이터: 자카드 거리 활용

**3) 계층적 군집분석: 군집을 점차 줄여나가는 방법**

- 최단 연결법: n\*n 거리행렬에서 가장 가까운 데이터를 묶어서 군집 형성

군집, 데이터간 최단거리 계산 후 거리행렬 수정 -> 새로운 군집 형성

- 최장 연결법: 군집, 데이터간 최장거리를 계산 후 거리행렬 수정

- 평균 연결법: 군집, 데이터간 평균거리를 계산 후 거리행렬 수정

- 와드 연결법: 군집 내 편차의 제곱합 고려, 군집화를 통해 정보의 손실 최소화

**4) 비계층적 군집분석: 모든 가능한 방법으로 최적화된 군집 형성**

- K-평균 군집분석

군집의 수, 초기값을 정해 값(seed) 중심으로 군집 형성 -> 각 데이터를 가장 가까운 see 군집으로 분류 -> 군집의 seed값 재 계산 -> seed값 없고 모든 개체 할당될 때 까지 반복

- K평균 군집분석의 특징: 연속형 변수 활용 가능, 초기 중심값 임의 선택 가능, 가급적 가까이 선택, 일렬 선택하지 않는게 좋음, 초기중심의 오차 제곱합 최소화 방향으로 군집이 형성되는 탐욕적 알고리즘-> 안정된 군집 보장하지만 최적이라는 보장 없음

- 장점: 단순한 알고리즘, 빠른 수행, 분석 방법 적용 용이, 계층적 대비 많은 데이터 활용 가능

- 단점: 군집 수, 가중치와의 거리 정의 어려움, 목적이 없으므로 결과 해석 어려움, 잡음 혹은 이상값 영향 큼, 볼록한 형태가 아닌 군집이 존재할 경우 성능이 떨어짐

**5) 혼합 분포 군집**

- 모형 기반 군집 방법, 가중치를 자료로부터 추정하는 방법

- 추정된 k개의 모형 중 출신 모형 추정에 따라 군집 분류

- EM 알고리즘을 사용해 혼합 모형에서의 모수와 가중치를 추정

- 군집의 크기가 작으면 추정이 어려움

- K-평균군집처럼 자료에 민감

**6) SOM**

- SOM 알고리즘(코호넨 맵): 코호넨에 의해 개발

- 비지도 신경망: 고차원 데이터->저차원 뉴런 정렬->지도로 형상화(입력변수의 위치관계 보존)

- 특징: 쉬운 시각적 이해, 패턴 발견이나 이미지 분석 용이, 단 하나의 전방패스 사용으로 빠름

## 6. 연관 분석

**1) 연관성 분석**

- 기업 DB에서 사건들 간의 규칙을 발견하기 위한 분석

- 장바구니 분석: 장바구니에 함께 들어있는 객체 분석

- 순차 분석: 구매 후 추가 구매 품목 분석

**2) 연관성분석의 측도**

- 산업 특성에 따라 지지도, 신뢰도, 향상도 값을 고려해서 규칙 선택

- 지지도: 전체 거래 중 항목 A와 B를 동시에 포함하는 거래의 비율

- 신뢰도: 항목 A를 포함한 거래 중 B가 포함될 확률

- 향상도: B 단독구매 -> A+B 동시구매시 B 구매 확률 증가 비율

**3) 연관분석 특징**

- 절차: 최소지지도 선정(-5%시작) -> 최소 지지도 넘는 품목 분류 -> 2가지 품목 집합 생성 -> 반복 수행으로 빈발품목 집합 선정

- 장점

탐색적 기법: 연관성 분석(조건 반응으로 표현) 결과의 쉬운 이해 가능

비목적성 분석기법: 방향, 목적이 없을 경우 목적변수가 없으므로 유용하게 활용

데이터 형태의 용이성: 데이터 변환 없이 이용

계산의 용이성: 간단한 계산

- 단점: 많은 계산 과정, 적절한 품목 결정, 품목의 비율 차이

- 두 항목의 높은 지지도, 신뢰도가 높은 연관이 없을 수도 있음(향상도 함께 고려)

**4) Apriori**

- 연관 규칙 알고리즘 중 우선시되는 알고리즘

- 원리: 항목집합이 빈발 -> 모든 부분집합도 빈발

- 지지도의 anti-monotone 성질: 어떤 항목집합의 지지도는 그 부분집합들의 지지도 이하