ADSP

1과목 - 데이터의 이해

1.1 데이터와 정보

- 데이터의 유형
  - 정성적 데이터
    - 저장, 검색, 분석에 많은 비용이 소모되는 언어, 문자 형태의 데이터
    - 예: 회사 매출의 증가 등
  - 정량적 데이터
    - 정형화된 데이터
    - 수치, 도형, 기호 등의 형태를 가진 데이터
    - 양이 크게 증가하더라도 DBMS에 저장, 검색, 분석하여 활용하기가 용이
    - 예: 나이, 몸무게, 주가, 매출액, 영업이익 등

- 지식경영의 핵심 이슈
  - 암묵지
    - 학습과 경험을 통해 개인에게 체화되어 있지만 겉으로 드러나지 않는 지식
    - 사회적으로 중요하지만 공유되기 어려움
    - 예: 김장하기, 자전거 타기
    - 상호작용: 공통화, 내면화
      - 개인에게 축적된 내면화 → 조직의 지식으로 공통화

    - 형식지
        - 문서나 메뉴얼처럼 형상화된 지식
        - 전달과 공유가 용이함
        - 예: 교과서, 비디오, DB
        - 상호작용: 표출화, 연결화
            - 문서로 표출화 → 개인의 지식으로 연결화
            - 표출화: 개인에게 내재된 경험을 객관적인 데이터로 문서나 매체에 저장, 가공, 분석하는 과정

- DIKW 피라미드
    - 데이터(Data)
        - 존재형식을 불문하고, 타 데이터와의 상관관계가 없는 가공 전 순수한 수치나 기호
        - 예: A마트는 연필이 100원, B마트는 200원이다.
    - 정보(Information)
        - 데이터의 가공 및 상관관계간 이해를 통해 패턴을 인식하고 그 의미를 부여한 데이터
        - 예: A마트의 연필이 더 싸다
    - 지식(Knowledge)
        - 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물
        - 예: 상대적으로 저렴한 A마트에서 연필을 사야겠다.
    - 지혜(Wisdom)
        - 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어
        - 예: A마트의 다른 상품들도 B마트보다 쌀 것이라고 판단한다.

1.2 데이터베이스 정의와 특징

- 데이터베이스의 정의
  - EU - 체계적이나 조직적으로 정리되고 전자식 또는 기타 수단으로 개별적으로 접근할 수 있는 독립된 저작물, 데이터 또는 기타 소재의 수집물
  - 국내 저작권법 - 소재를 체계적으로 배열·구성한 편집물로 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것
  - 국내 컴퓨터용어사전 - 동시의 복수의 적용업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서 데이터를 받아들이고 저장, 공급하기 위해 일정한 구조에 따라 편성된 데이터의 집합

- 데이터베이스의 특징
  - 통합된 데이터(integrated data)
    - 동일한 내용의 데이터가 중복되어 있지 않다는 것
    - 데이터 중복은 관리상의 복잡한 부작용을 초래
  - 저장된 데이터(stored data)
    - 자기 디스크나 자기 테이프 등과 같이 컴퓨터가 접근할 수 있는 저장 매체에 저장되는 것
    - 데이터베이스는 기본적으로 컴퓨터 기술을 바탕으로 함
  - 공용 데이터(shared data)
    - 여러 사용자가 서로 다른 목적으로 데이터를 공동으로 이용한다는 것
    - 대용량화되고 구조가 복잡한 것이 보통
  - 변화되는 데이터(changeable data)
    - 데이터베이스에 저장된 내용은 곧 데이터베이스의 현 시점에서의 상태를 나타냄
    - 이 상태는 새로운 데이터의 삽입, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 정확한 데이터를 유지해야 함

1.3 데이터베이스의 활용

- OLTP(On-Line Transaction Processing)
  - 호스트 컴퓨터가 데이터베이스를 액세스하고 바로 처리 결과를 돌려보내는 형태
  - 데이터를 수시로 갱신
  - 주문입력시스템, 재고관리시스템 등

- OLAP(On-Line Analytical Processing)
  - 정보 위주 분석 처리
  - 다양한 비즈니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근해 의사 결정에 활용할 수 있는 정보를 얻게 해주는 기술

- CRM(Customer Relationship Management)
  - 고객 관계 관리
  - 고객 중심 자원 극대화 → 고객 특성에 맞는 마케팅 활동

- SCM(Supply Chain Management)
  - 공급망 관리
  - 기업이 외부 공급 업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화시키기 위한 것
  - 자재 구매, 생산, 제고, 유통, 판매, 고객 데이터로 구성

2.1 빅데이터의 이해

- 빅데이터의 정의
  1. 관점에 따른 정의
     - Mckinsey(2011)
       - 데이터 규모에 중점

        - 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터
    - IDC(2011)
        - 분석 비용 및 기술에 초점
        - 다양한 종류의 대규모의 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처
    - 3V(가트너그룹(Gartner Group)과 더그레니(Doug Laney))
        - Volume: 데이터의 규모 측면
        - Variety: 데이터의 유형과 소스 측면
        - Velocity: 데이터의 수집과 처리 측면
2. 빅데이터 정의의 범주 및 효과
    a. 데이터 변화
        - 규모(Volume)
        - 형태(Variety)
        - 속도(Velocity)
    b. 기술 변화
        - 데이터 처리, 저장, 분석기술 및 아키텍처
        - 클라우드 컴퓨팅 활용
    c. 인재, 조직 변화
        - Data Scientist와 같은 새로운 인재 필요
        - 데이터 중심 조직

- 출현 배경과 변화
    - 산업계: 고객 데이터 축척, 보유를 통해 데이터에 숨어있는 가치를 발굴
    - 학계: 거대 데이터를 다루는 학문 분야가 늘어나면서 필요한 기술 아키텍처 및 통계 도구의 발전
    - 기술발전: 관련 기술의 발달(저장 기술, 인터넷 보급, 클라우드 컴퓨팅, 모바일 혁명)

- 빅데이터에 거는 기대의 비유적 표현
  - 산업혁명의 석탄과 철: 생산성을 획기적으로 끌어올려 혁명적 변화를 가져올 것
  - 21세기의 원유: 산업 전반의 생산성을 향상시키고, 새로운 범주의 산업을 만들어낼 것
  - 렌즈: 현미경이 생물학 발전에 미쳤던 영향 만큼 산업 발전에 영향을 미칠 것
  - 플랫폼: 공동 활용의 목적으로 구축된 유무형의 구조물 → 다양한 서드파티 비즈니스에 활용되며 플랫폼 역할을 할 것

- 빅데이터가 만들어 내는 본질적인 변화
  - 사전처리 → 사후처리
  - 표본조사 → 전수조사
  - 질 → 양
  - 인과관계 → 상관관계

2.2 빅데이터의 가치와 영향

- 빅데이터의 가치 산정이 어려운 이유
  - 데이터 활용 방식: 재사용, 재조합, 다목적용 개발
  - 새로운 가치 창출
  - 분석 기술 발전

- 빅데이터가 미치는 영향
  - 기업
    - 혁신, 경쟁력제고, 생산성 향상
    - 빅데이터를 활용해 소비자의 행동을 분석하고, 시장 변동을 예측해 비즈니스 모델을 혁신하거나 신사업 발굴

    - 정부
        - 환경 탐색, 상황 분석, 미래 대응
        - 기상, 인구이동, 각종 통계, 법제 데이터 등을 수집해 사회 변화를 추정, 정보 추출
    - 개인
        - 목적에 따른 활용
        - 개인은 빅데이터를 서비스하는 기업의 출현으로 비용이 지속적으로 하락하여 정치인이나 대중가수 등이 인지도 향상에 빅데이터를 활용

⇒ 생활 전반의 스마트화

2.3 비즈니스 모델

- 빅데이터 활용 사례
    - 관점에 따른 정의
        - 구글: 사용자의 로그 데이터를 활용한 검색엔진 개발, 기존 페이지랭크 알고리즘 혁신
        - 월마트: 고객의 구매패턴을 분석해 상품 진열에 활용
    - 정부
        - 실시간 교통정보 수집, 기후 정보, 소방 서비스 등을 위한 실시간 모니터링 등을 실시하여 국가 안전 확보에 활용
    - 개인
        - 정치인: 선거 승리를 위해 사회관계망 분석을 활용해 유세 지역 선거
        - 가수: 팬들의 음악 청취 기록을 분석해 공연 시 노래 순서 선정

- 빅데이터 활용 기본 테크닉
    - 연관 규칙 학습
        - 상관관계가 있는지

        - 커피 구매하는사람이 탄산음료도 구매하는가?
    - 유형 분석/군집 분석
        - 분류 또는 그룹화
        - 어느 집단에 속하는가?
    - 유전자 알고리즘
        - 생명의 진화를 모방해 최적해를 구하는 알고리즘
        - 어떤 미지의 함수 Y=f(x)를 최적화하는 해 x를 찾기 위해 진화를 모방한 탐색 알고리즘
        - 최고 시청률을 얻으려면 어느 시간대에 방송해야하는가?
    - 기계학습
        - 훈련 데이터로부터 학습한 알려진 특성을 활용해 예측
        - 기존 시청 기록을 바탕으로 어떤 걸 보고싶어할지?
    - 회귀분석
        - 독립변수와 종속변수 간의 관계 파악
        - 구매자의 나이가 구매 물품 타입에 어떤 영향을 미치는가?
    - 감정분석
        - 사람의 감정 분석
        - 고객 평가는 어떤가?
    - 소셜네트워크분석 (=사회관계망 분석)
        - 특정인과 다른 사람의 관계, 영향력 있는 사람 파악
        - 고객 간 관계망 구성은 어떤가?

2.4 위기 요인과 통제 방안

- 위기 요인에 따른 통제 방안
    - 사생활 침해 → 동의에서 책임으로
    - 책임 원칙 훼손 → 결과 기반 책임 원칙 고수

- 데이터 오용 → 알고리즘 접근 허용

2.5 미래의 빅데이터

- 빅데이터 활용의 3요소
  - 데이터: 모든 것의 데이터화(datafication)
  - 기술: 진화하는 알고리즘, 인공지능
  - 인력: 데이터 사이언티스트, 알고리즈미스트

3.1 빅데이터 분석과 전략 인사이트

- 빅데이터 회의론의 원인
  - 부정적 학습효과 → 과거의 고객관계관리(CRM): 공포 마케팅, 투자대비 효과미흡
  - 부적절한 성공사례 → 빅데이터가 필요 없는 분석사례, 기존 CRM의 분석 성과를 빅데이터 분석 성과로 과대포장

  ⇒ 단순히 빅데이터에 포커스를 두지 말고, 분석을 통해 가치를 만드는 것에 집중해야 함

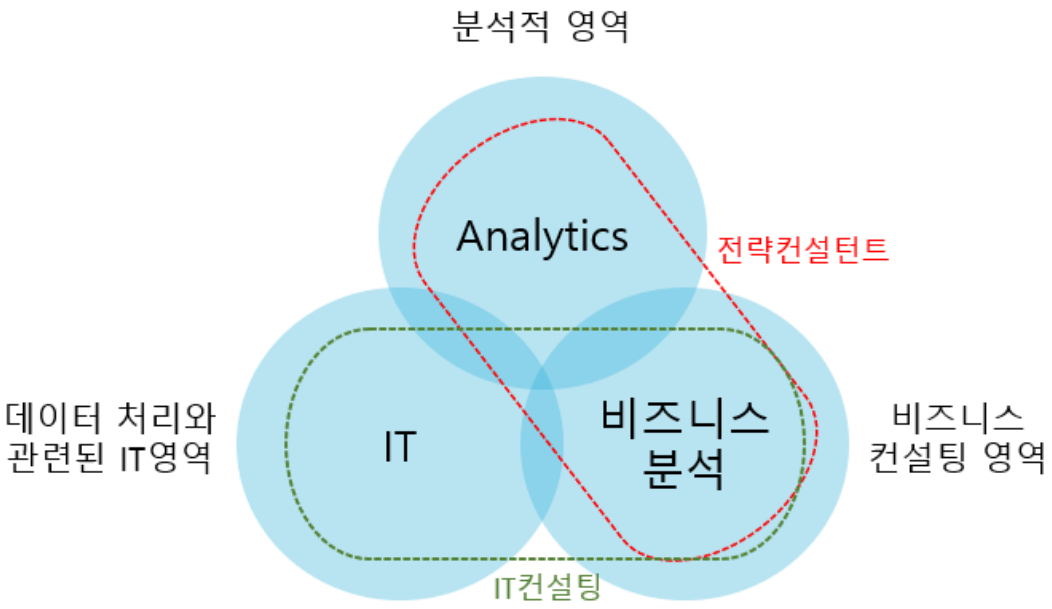
- 일차원적인 분석 vs 전략 도출을 위한 가치 기반 분석
  - 산업별 일차원적 분석 애플리케이션
    - 금융 서비스: 신용점수 산정, 사기 탐지, 가격 책정, 프로그램 트레이딩, 클레임 분석, 고객 수익성 분석
    - 병원: 가격 책정, 고객 로열티, 수익 관리
    - 에너지: 트레이딩, 공급, 수요 예측
    - 정부: 사기 탐지, 사례 관리, 범죄 방지, 수익 최적화
  - 전략 도출 가치 기반 분석

- 전략적 통찰력의 창출에 포커스 → 해당 사업에 중요한 기회를 발굴, 주요 경영진의 지원
- 분석의 활용 범위를 더 넓고 전략적으로 변화 시키고, 전략적 인사이트를 주는 가치기반의 분석 단계로 나아가야 함

3.2 전략 인사이트 도출을 위한 필요 역량

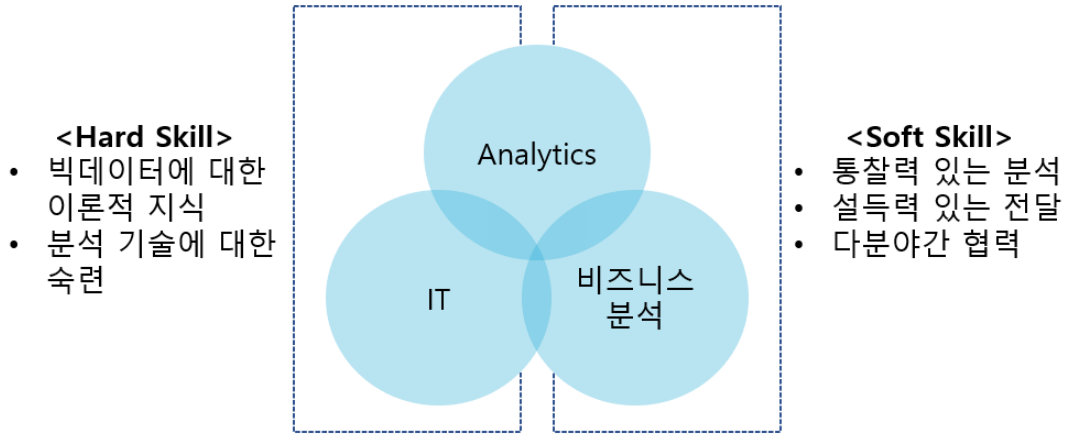
- 데이터 사이언스의 의미
  - 데이터 공학, 수학, 통계학, 컴퓨터 공학, 시각화, 해커의 사고방식, 해당 분야의 전문 지식을 종합한 학문

- 데이터 사이언스의 구성요소
  - 데이터 사이언스의 영역



- 분석적 영역: 수학, 확률모델, 머신러닝, 분석학, 패턴 인식과 학습, 불확실성 모델링
- 데이터 처리와 관련된 IT영역: 시그널 프로세싱, 프로그래밍, 데이터 엔지니어링, 데이터 웨어하우징, 고성능 컴퓨팅, 분산 컴퓨팅

        - 비즈니스 컨설팅 영역: 커뮤니케이션, 프레젠테이션, 스토리텔링, 데이터 시각화
    - 데이터 사이언티스트의 요구 역량



- 데이터 사이언스: 과학과 인문의 교차로
    - 분석의 기술보다 더 중요한 것은 소프트 스킬
    - 전략적 통찰을 주는 분석은 단순 통계 및 데이터 처리 능력보다 스토리텔링, 커뮤니케이션, 창의력, 열정, 직관력, 비판적 시각, 대화 능력 등의 인문학적 요소가 필요함

- 전략적 통찰력과 인문학의 부활
    - 외부 환경적 측면에서 본 인문학 열풍의 이유
        - 컨버전스 → 디버전스
            - 단순 세계화에서 복잡한 세계화로의 변화
            - 규모의 경제, 세계화, 표준화, 이성화 → 복잡한 세계, 다양성, 관계, 연결성, 창조성
        - 생산 → 서비스
            - 비즈니스 중심이 제품생산에서 서비스로 이동
            - 고장 나지 않는 제품의 생산 → 뛰어난 서비스로 응대
        - 생산 → 시장창조
            - 공급자 중심의 기술경쟁에서 무형자산의 경쟁으로 변화

- 생산 기술 중심, 기술 중심 대규모 투자 → 현재의 패러다임에 근거한 시장 창조, 현지 사회와 문화에 관한 지식

3.3 빅데이터 그리고 데이터 사이언스의 미래

- 데이터 사이언스의 의미
  - 빅데이터 분석은 선거 결과에 결정적인 영향을 미칠 수도 있고, 기업의 비용 절감, 시간 절약, 매출 증대, 고객 서비스 향상, 신규 비지니스 창출, 내부 의사결정 지원 등에 있어 상당한 가치를 발휘하고 있음

- 빅데이터 회의론을 넘어 가치 패러다임의 변화
  - 과거: Digitalization
    - 가치 창출 원천: 아날로그 세상을 어떻게 효과적으로 디지털화하는가
  - 현재: Connection
    - 디지털화된 정보와 대상들은 서로 연결 시작
    - 성공요인: 연결을 더 효과적이고 효율적으로 제공하는가
  - 미래: Agency
    - 복잡한 연결을 얼마나 효과적이고 믿을 수 있게 관리하는가

- 데이터 사이언스의 한계와 인문학
  - 데이터 사이언스의 한계
    - 분석 과정에서는 가정 등 인간의 해석이 개입되는 단계를 반드시 거침
    - 분석결과가 의미하는 바는 사람에 따라 전혀 다른 해석과 결론을 내릴 수 있음
    - 아무리 정량적인 분석이라도 모든 분석은 가정에 근거
  - 데이터 사이언스와 인문학
    - 인문학을 이용하여 빅데이터와 데이터 사이언스가 데이터에 묻혀있는 잠재력을 풀어내고, 새로운 기회를 찾고, 누구도 보지 못한 창조의 밑그림을 그릴 수 있는 힘을 발휘하게 될 것

+++

- DBMS의 종류
  - 관계형 DBMS
    - 데이터를 컬럼과 로우를 이루는 테이블로 정리
    - 테이블 → 엔티티 타입
    - 로우(튜플, 레코드) → 엔티티 타입의 인스턴스
    - 컬럼 → 인스턴스의 속성값
  - 객체지향 DBMS
    - 정보를 객체 형태로 표현
  - 네트워크 DBMS
    - 그래프 기반 모델
    - 노드 → 레코드
    - 간선 → 레코드 사이의 관계
  - 계층형 DBMS
    - 트리 구조 기반

- 개인정보 비식별 기술
  - 데이터 마스킹: 길이/유형/형식 등의 속성을 유지한 채 새로운 데이터를 익명으로 생성
  - 가명처리: 개인정보 주체의 이름을 다른 이름으로 변경
  - 총계처리: 데이터의 총합 값을 보여 개별 데이터의 값을 보이지 않도록 함
  - 데이터 삭제: 공유/개방 목적에 따라 필요없는 값 삭제
  - 데이터 범주화: 데이터 값을 범주값으로 변환
  - 난수화: 사생활 침해를 막기위해 개인정보를 무작위 처리하는 등 데이터가 본래 목적 외에 가공되고 처리되는 것을 방지하는 기술

- 데이터 무결성
  - 데이터베이스 내 데이터의 일관성, 유효성, 신뢰성을 보장하기 위해 데이터 변경에 제한을 두어 데이터의 정확성을 보증하는 것
  - 데이터 무결성의 유형
    - 개체 무결성 (Entity integrity)
    - 참조 무결성 (Referential integrity)
    - 범위 무결성 (Domain integrity)

- 데이터 레이크
  - 수 많은 정보 속의 의미있는 내용을 찾기 위해 방식에 상관 없이 데이터를 저장하는 시스템
  - Apache Hadoop, Teradata Integrated Big Data Platform 1700 등

- 빅데이터 분석 기술
  - 하둡(Hadoop)
    - 여러개의 컴퓨터를 하나인 것처럼 묶어 대용량 데이터를 처리하는 플랫폼 기술
    - 맵리듀스(Map Reduce)로 분산파일 시스템(HDFS)에 저장된 대용량 데이터를 대상으로 SQL을 이용해 사용자 질의를 실시간으로 처리
  - Apache Spark
    - 실시간 분산형 컴퓨팅 플랫폼
    - 스칼라로 작성되어 있지만 스칼라, 자바, R, 파이썬, API 지원
    - In-Memory 방식으로 처리 → 하둡보다 처리속도가 빠름
  - Smart Factory
    - 공장 내 설비와 기계에 IoT 설치로 생산성 극대화
  - Machine Learning
    - 인공지능 연구 분야 중 하나
    - 인간의 학습 능력과 같은 기능을 컴퓨터에서 실현하고자 하는 기법
  - Deep Leanrning

        - 컴퓨터가 많은 데이터를 이용해 사람처럼 학습할 수 있게 하기 위해 인공 신경망(ANN) 등의 기술 기반으로 구축한 기계 학습 기술 중 하나

- 데이터 단위
    - 바이트: 1Byte
    - 킬로바이트: 1KB = 1024B
    - 메가바이트: 1MB = 1024KB
    - 기가바이트: 1GB = 1024MB
    - 테라바이트: 1TB = 1024GB
    - 페타바이트: 1PB = 1024TB
    - 엑사바이트: 1EB = 1024PB
    - 제타바이트: 1ZB = 1024EB
    - 요타바이트: 1YB = 1024ZB

- 비즈니스 모델
    - B2B
        - 기업과 기업 사이의 거래를 기반으로 한 비즈니스 모델
        - 기업 사용 장비, 재료, 공사입찰 등
    - B2C
        - 기업과 고객 사이의 거래를 기반으로 한 비즈니스 모델
        - 이동통신사, 여행회사, 신용카드회사, 옥션 등

- 블록체인
    - 거래 정보를 하나의 덩어리를 보고 이를 차례로 연결한 거래장부
    - 거래에 참여하는 모든 사용자에게 거래 내역을 보내주며 거래때마다 이를 대조해 데이터 위조를 막는 방식 사용

- 데이터 유형

- 정형 데이터
  - DB로 정제된 데이터
  - 형태가 있으며(고정 필드) 연산이 가능
  - 주로 RDBMS에 저장
  - 데이터 수집 난이도가 낮고 처리가 쉬운 편
  - 관계형 DB, 스프레드시트, CSV 등
- 반정형 데이터
  - 센서 중심으로 스트리밍되는 머신데이터
  - 형태가 있으며(스키마, 메타데이터) 연산이 불가능
  - 주로 파일로 저장
  - 데이터 수집 난이도가 중간이고 데이터 처리 기술(파싱) 필요
  - XML, JSON, HTML 등
- 비정형 데이터
  - 보고서, 소셜미디어 데이터
  - 형태가 없으며 연산이 불가능
  - 주로 NoSQL에 저장
  - 데이터 수집 난이도가 높으며 수집 데이터 처리가 어려움
  - 소셜데이터, 영상, 이미지, PDF 등

2과목 - 데이터 분석 기획

1.1 분석 기획 방향성 도출

- 분석 기획의 특징

- 분석 기획: 실제 분석을 수행하기에 앞서 분석을 수행할 과제를 정의하고, 의도했던 결과를 도출할 수 있도록 이를 적절하게 관리할 수 있는 방안을 사전에 기획하는 일련의 작업
- 데이터 사이언티스트의 역량
  - 수학/통계학적 지식
  - 정보기술(IT기술, 해킹기술, 통신기술 등)
  - 비즈니스에 대한 이해와 전문성

- 분석 대상과 방법
  - 분석은 분석의 대상(what)과 분석의 방법(how)에 따라 4가지로 분류

분석 주제 유형		분석의 대상(What)	
		Known	Un-known
분석의 방법 (How)	Known	Optimization	Insight
	Un-known	Solution	Discovery

  - 하향식 접근방법: 최적화 → 해결책
  - 상향식 접근방법: 발견 → 통찰

- 목표 시점별 분석기획 방안

당면한 분석 주제의 해결 (과제단위)		지속적 분석 문화 내재화 (마스터플랜단위)
Speed & Test	1차 목표	Accuracy & Deploy
QuickWin	과제의유형	LongTerm View
Problem Solving	접근방식	Problem Definition

- 분석 기획시 고려사항
  - 분석의 기본인 가용데이터(Available data)에 대한 고려가 필요

    - 분석을 통해 가치가 창출될 수 있는 적절한 활용방안과 유즈케이스(Proper Business Use Case) 탐색이 필요
    - 분석 수행시 발생하는 장애요소들에 대한 사전계획 수립이 필요

1.2 분석 방법론

- 분석 방법론의 개요
    - 기업의 합리적 의사결정을 가로막는 장애 요소
        - 고정 관념(Stereotype)
        - 편향된 생각(Bias)
        - 프레이밍 효과(Framing Effect): 표현방식에 따라 동일한 상황에도 판단이나 선택이 달라질 수 있는 현상
    - 방법론의 적용 업무의 특성에 따른 모델
        - 폭포수 모델(Waterfall Model)
            - 단계를 순차 진행하는 방법
            - 기존 IT의 SW개발방식
        - 프로토타입 모델(Prototype Model)
            - 폭포수 모델의 단점 보완
            - 고객의 요구를 완전하게 이해하고 있지 못하거나 완벽한 요구 분석의 어려움을 해결하기 위해 일부분을 우선 개발하여 제공하고 요구 분석 및 성능 평가를 통한 개선 작업을 시행하는 모델
            - 처음 시도하는 프로젝트에 적용이 용이하지만, 반복에 대한 관리 체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있음
        - 나선형 모델(Spiral Model)
            - 반복을 통해 점증적으로 개발

- **KDD 분석 방법론**

Data

⬇️ 데이터셋 선택(selection)

Target Data

⬇️ 데이터 전처리(preprocessing)

Preprocessed Data

⬇️ 데이터 변환(transformation)

Transformed Data

⬇️ 데이터 마이닝(data mining)

Patterns

⬇️ 결과 평가(interpretation/evaluation)

Knowledge

- 데이터셋 선택
  - 분석 대상의 비즈니스 도메인에 대한 이해와 프로젝트 목표 설정이 필수
  - 데이터 마이닝에 필요한 목표데이터를 구성하여 분석에 활요
- 데이터 전처리
  - 잡음과 이상치, 결측치를 식별하고 제거하거나 재처리하여 데이터 셋을 정제하는 단계
  - 추가로 요구되는 데이터 셋이 필요한 경우 데이터 선택 프로세스 재실행
- 데이터 변환
  - 정제된 데이터에 목적에 맞는 변수를 생성, 선택
  - 데이터 차원 축소
  - 학습용 데이터와 검증용 데이터로 데이터를 분리하는 단계
- 데이터 마이닝
  - 분석 목적에 맞는 기법을 선택하고 데이터마이닝 작업을 실행하는 단계
  - 필요에 따라 데이터 전처리와 변환 프로세스 추가 실행
- 결과 평가
  - 데이터마이닝 결과에 대한 해석과 평가, 분석목적과의 일치성 확인
  - 발견한 지식을 업무에 활용하기 위한 방안 마련

- **CRISP-DM 분석 방법론**
  - CRISP-DM의 4레벨 구조
    - 단계(Phase): 최상위레벨
    - 일반화 태스크(Generic Tasks): 데이터마이닝의 단일 프로세스를 완전하게 수행하는 단위
    - 세분화 태스크(Specialized Tasks): 구체적인 수행 레벨
    - 프로세스 실행(Process Instances): 데이터마이닝을 위한 구체적인 실행
  - CRISP-DM의 프로세스
    - 업무 이해(business understanding)
      - 목적과 요구사항 이해
      - 도메인지식 → 데이터 분석을 위한 문제 정의
      - 초기 프로젝트 계획 수리
    - 데이터 이해(data understanding)
      - 데이터 수집 및 데이터 속성 이해
      - 데이터 품질 문제점 식별
      - 숨겨진 인사이트 발견
    - 데이터 준비(data preparation)
      - 분석 기법에 적합한 데이터 편성
      - 데이터셋 선택, 데이터 정제
    - 모델링(modeling)
      - 모델링 기법과 알고리즘 선택
      - 파라미터 최적화
      - 모델 평가
    - 평가(evaluation)
      - 모델링 결과가 프로젝트 목적에 부합하는지 평가
      - 마이닝 결과를 최종 수용할 지 판단
      - 분석 결과 평가, 모델링 과정 평가, 모델 적용성 평가

        - 전개(deployment)
            - 완성 모델을 실무에 적용하기 위한 계획 수립
            - 모니터링, 모델 유지보수 계획 마련

- 빅데이터 분석 방법론
    - 빅데이터 분석의 계층적 프로세스
        - 단계(Phase)
            - Process Group을 통해 완성된 단계별 산출물 생성
            - 각 단계는 기준선으로 설정되어 관리
            - 버전관리로 통제
        - 태스크(Task)
            - 단계를 구성하는 단위 활동
            - 물리적/논리적 단위로 품질 검토의 항목
        - 스텝(Step)
            - WBS의 워크패키지
            - 입력자료, 처리 및 도구, 출력자료로 구성된 단위프로세스(Unit Process)
    - 빅데이터 분석 방법론의 5단계
        1. 분석 기획
            - 비지니스 이해
            - 프로젝트 범위 설정
            - 프로젝트 정의
            - 프로젝트 수행 계획 수립
            - 데이터 분석 위험 식별
            - 위험 대응 계획 수립
        2. 데이터 준비
            - 필요 데이터 정의
            - 데이터 획득 방안 수립

- 데이터 스토어 설계
- 데이터 수립 및 저장
- 데이터 정합성 점검

3. 데이터 분석
    - 비즈니스 룰 확인: 프로젝트 목표 정확히 인식하고 데이터의 범위 확인
    - 분석용 데이터 셋 준비
    - 텍스트 데이터 확인 및 추출
    - 텍스트 데이터 분석: 모델 구축 및 용어사전 확보, 시각화 도구를 이용
    - 탐색적 데이터 분석: 기초통계량 산출, 통계적 특성 이해
    - 데이터 시각화: 탐색적 데이터 분석을 위한 도구로 활용, 시스템 구현을 위한 프로토타입으로 활용
    - 데이터 분할: 모델 과적합과 일반화를 위해 데이터 분할, 검증 횟수, 생성 모델 개수 설정
    - 데이터 모델링: 분류, 예측, 군집 등의 모델 생성, 프로젝트 목적에 맞는 통합 모델 수행
    - 모델 적용 및 운영 방안: 알고리즘 설명서 작성(의사코드 수준의 상세한 작성 필요), 모니터링 방안 수립
    - 모델 평가: 알고리즘 파악, 모델 검증 데이터 활용
    - 모델 검증: 모델의 실적용성 검증, 모델링 검증 보고서 작성, 모델 품질 최종 검증
4. 시스템 구현
    - 시스템 분석 및 설계: 응용시스템 구축
    - 시스템 구현
    - 시스템 테스트: 품질 관리 차원에서 진행함으로써 적용된 시스템의 객관성과 완전성 확보
    - 시스템 운영 계획: 교육 실시, 시스템 운영계획 수립
5. 평가 및 전개
    - 모델 발전 계획 수립
    - 프로젝트 성과 평가: 성과 평가서 작성

- 프로젝트 종료: 모든 산출물 및 프로세스를 지식 자산화

1.3 분석 과제 발굴

- 분석과제 발굴 방법론
  - 하향식 접근 방식: 분석 과제가 주어지고 이에 대한 해법을 찾기 위해 각 과정이 체계적으로 단계화되어 수행하는 방식
  - 상향식 접근 방식: 문제의 정의 자체가 어려운 경우 데이터를 기반으로 문제를 지속적으로 개선하는 방식

- **하향식 접근 방식**(Top Down Approach)
  1. 문제 탐색(Problem Discovery)
     - 비즈니스 모델 기반 문제 탐색
       - 업무(operation), 제품(product), 고객(customer), 규제와 감사(regulation & audit), 지원 인프라(IT & human resource) 등 5가지 영역으로 기업 비즈니스 분석
     - 분석 기회 발굴의 범위 확장
       1. 거시적 관점: 사회, 기술, 경제, 환경, 정치
       2. 경쟁자 확대: 대체제, 경쟁자, 신규 진입자
       3. 시장 니즈 탐색: 고객, 채널, 영향자들
       4. 역량의 재해석: 내부 역량, 파트너 네트워크
     - 외부 사례 기반 문제 탐색: 유사, 동종 사례를 벤치마킹을 통해 분석 기회 발굴
     - 분석 유즈 케이스(Analytics Use Case) 정의
  2. 문제 정의(Problem Definition)
     - 비즈니스 문제를 데이터 분석 문제로 변환하여 정의
  3. 해결방안 탐색(Solution Search)
     - 분석 역량(Who), 분석 기법 및 시스템(How)으로 해결 방안 탐색

        - 수행 옵션 도출
    4. 타당성 검토(Feasibility Study)
        - 경제적 타당성, 데이터 및 기술적 타당성 검토: 분석역량
        - 과제 선정

- **상향식 접근 방식**(Bottom Up Approach)
    - 기업이 보유하고 있는 다양한 원천 데이터로부터 분석을 통해 통찰력과 지식을 얻는 접근방법
    - 다양한 원천 데이터를 대상으로 분석을 수행해 가치있는 모든 문제를 도출하는 일련의 과정

- 상향식 접근법의 특징
    - 하향식 접근법은 논리적 단계별 접근법으로, 최근의 복잡하고 다양한 환경에서 발생하는 문제를 해결하기 어렵기 때문에 디자인적 사고(Design Thinking) 접근법을 통해 WHY → WHAT 관점으로 존재하는 데이터 그 자체를 객관적으로 관찰하여 문제를 해결하려는 접근법을 사용
    - 상향식 접근법은 비지도 학습 방법으로 수행
    - 시행착오를 통한 문제 해결: **프로토타이핑 접근법**
        - 일단 분석을 시도해보고 결과를 확인해가면서 반복적으로 해석해나가는 방법
        - 신속하게 해결책이나 모형을 제시해 문제를 좀 더 명확하게 인식하고 필요한 데이터를 식별하여 구체화할 수 있게 하는 유용한 상향식 접근 방법
        - 프로세스: 가설의 생성(Hypotheses) → 디자인에 대한 실험(Design Experiments) → 테스트(Test) → 통찰 도출 및 가설 확인(Insight)

- **디자인 사고**(Design Thinking)
    - 상향식 접근 방식의 발산 단계와 하향식 접근 방식의 수렴 단계를 반복 수행
    - 상호 보완적인 동적 환경을 통해 분석 가치를 높일 수 있는 최적의 의사결정 방식
    - 디자인 사고의 프로세스
        - 감정이입(Empathize)

        - 문제 정의(Define)
        - 아이디어 도출(Ideate)
        - 구현(Prototype)
        - 테스트(Test)

- 지도학습과 비지도학습
    - 비지도학습(Unsupervised Learning)
        - 데이터 자체의 결합, 연관성, 유사성 등을 중심으로 데이터의 상태를 표현하는 것
        - 목표 값을 사전에 정의하지 않고 데이터 자체만을 가지고 그룹을 도출
        - 장바구니 분석, 군집 분석, 기술 통계 및 프로파일링 등
    - 지도학습(Supervised Learning)
        - 명확한 목적 하에 데이터 분석을 실시하는 것
        - 도출되는 값에 대해 사전에 인지하고 결과를 예측
        - 분류, 추측, 예측, 최적화를 통해 사용자 주도하에 분석을 실시하고 지식을 도출하는 것이 목적
        - 분류 분석

- 분석과제 정의
    - 분석과제 정의서를 통해 분석별 필요 소스 데이터, 분석법, 데이터 입수 및 분석의 난이도, 분석 수행주기, 검증 오너십, 상세 분석 과정 등을 정의

1.4 분석 프로젝트 관리 방안

- 분석과제 관리를 위한 5가지 주요 영역
    - 분석프로젝트는 범위, 일정, 품질, 리스크, 의사소통 등 영역별 관리가 수행되어야 할뿐 아니라 데이터에 기반한 분석 기법을 적용한다는 특성 때문에 5가지 주요 속성을 고려해 추가적인 관리가 필요하다.

    - 5 Analytic Dimensions
        - 데이터 복잡성(Data Complexity)
        - 데이터 크기(Data Size)
        - 속도(Speed)
        - 정확도와 정밀도(Accuracy & Precision)
        - 분석 복잡성(Analytic Complexity)

- 분석 프로젝트의 특성
    - 분석가의 목표: 개별적인 분석업무 수행 뿐만 아니라 전반적인 프로젝트 관리 또한 중요
    - 분석가의 입장: 데이터 영역과 비즈니스 영역의 현황을 이해하고, 프로젝트의 목표인 분석의 정확도 달성과 결과에 대한 가치 이해를 전달하는 조정자로서의 분석가 역할이 중요
    - 분석 프로젝트는 도출된 결과의 재해석을 통한 지속적인 반복 및 정교화가 수행되는 경우가 대부분이므로, 프로토타이핑 방식의 어자일(Agile) 프로젝트 관리 방식에 대한 고려도 필요

2.1 마스터 플랜 수립 프레임워크

- 마스터 플랜 수립 프레임워크
    - 분석 과제를 대상으로 다양한 기준을 고려해 적용 우선순위를 설정하고, 데이터 분석 구현을 위한 로드맵을 수립한다.
    - 우선순위 고려요소
        1. 전략적 중요도
        2. 비즈니스 성과/ROI
        3. 실행 용이성

        → 적용 우선순위 설정
    - 적용범위/방식 고려요소

1. 업무 내재화 적용 수준
2. 분석 데이터 적용 수준
3. 기술 적용 수준

→ Analytics 구현 로드맵 수립

- ISP(Information Strategy Planning)
  - 기업 및 공공기관에서 시스템의 중장기 마스터 플랜을 정의하기 위한 정보전략계획
  - 조직 내·외부 환경을 분석해 기회나 문제점 도출
  - 사용자 요구사항을 분석하여 시스템 구축 우선순위 결정

- 분석 마스터 플랜
  - 일반적인 ISP 방법론을 활용하되 데이터 분석 기획의 특성을 고려해 수행
  - 기업에서 필요한 데이터 분석 과제를 빠짐없이 도출한 후 과제 우선순위를 결정하고 단기 및 중장기로 나누어 계획 수립

- 우선순위 평가에 활용하기 위한 ROI 관점에서 빅데이터의 핵심 특징
  - 3V: 난이도 - 비용, 분석수준
    1. 크기(Volume): 데이터 규모/양
    2. 다양성(Variety): 데이터 종류/유형
    3. 속도(Velocity): 데이터 생성속도, 처리속도)

    ⇒ 투자 비용 요소(Investment)
  - 4V: 시급성 - 전략적 중요도, 목표가치(KPI)

    4. 가치(Value): 분석결과 활용 및 실행을 통한 비즈니스 가치

    ⇒ 비즈니스 효과(Return)

2.2 분석 거버넌스 체계 수립

- 분석 거버넌스 체계 구성요소
  - 분석 기획 및 관리 수행 조직(Organization)
  - 과제 기획 및 운영 프로세스(Process)
  - 분석관련시스템(System)
  - 데이터(Data)
  - 분석 교육/마인드 육성 체계(Human Resource)

- 데이터 분석 수준진단
  - **분석 준비도**(Readiness)
    - 목표: 기업의 데이터 분석 도입 수준 파악
    - 구성
      - 분석업무 파악: 발생한 사실 분석, 예측 분석, 시뮬레이션 분석, 최적화 분석, 분석 업무 정기적 개선
      - 분석인력, 조직: 분석 전문가 직무 존재, 교육 훈련 프로그램, 관리자들의 기본적 분석 능력, 분석업무 총괄 조직 존재, 경영진의 분석 업무 이해 능력
      - 분석기법: 업무별 적합한 분석기법, 도입 방법론, 라이브러리, 효과성 평가, 정기적 개선
      - 분석데이터: 데이터 충분성, 신뢰성, 적시성, 비구조적 데이터 관리, 외부 데이터 활용 체계, 기준 데이터 관리(MDM)
      - 분석문화: 사실에 근거한 의사결정, 데이터 중시 정도, 활용 상황, 데이터 기반 의사결정, 데이터 공유/협업 문화
      - 분석인프라: 운영시스템 데이터 통합, 데이터유통체계, 분석 서버 및 스토리지, 분석환경
  - **분석 성숙도**(Maturity)
    - 도입 → 활용 → 확산 → 최적화
    - 비즈니스, 조직역량, IT부문
  - 조직의 성숙도 평가 도구: CMMI(Capability Maturity Model Integration)
    - 소프트웨어와 시스템 공학의 역량 숙성도를 측정하기 위한 1~5단계로 구성된 성숙도 평가 모델

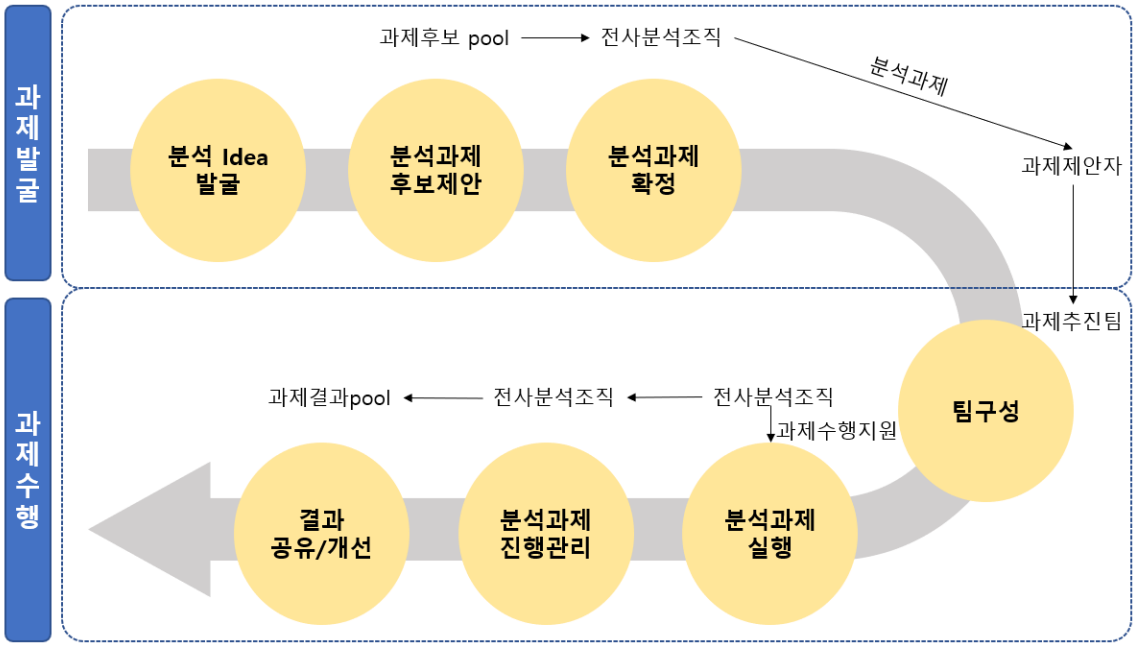
        - 소프트웨어 품질보증과 시스템 엔지니어링 분야의 품질보증 기술을 통합하여 개발
    - 분석 수준 진단 결과: 정착형, 확산형, 준비형, 도입형

- 데이터 거버넌스 체계 수립
    - 데이터 거버넌스
        - 전사 차원의 모든 데이터에 대해 정책 및 지침, 표준화, 운영 조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크(Framework) 및 저장소(Repository)를 구축하는 것
        - 데이터 거버넌스의 중요 관리 대상: 마스터 데이터(Master Data), 메타데이터(Meta Data), 데이터 사전(Data Dictionary)
    - 데이터 거버넌스의 구성요소
        - 원칙(Principle): 보안, 품질기준, 변경관리
        - 조직(Organization): 데이터 관리자, 데이터베이스 관리자, 데이터 아키텍트
        - 프로세스(Process): 작업 절차, 모니터링 활동, 측정 활동
    - 데이터 거버넌스 체계
        - 데이터 표준화
            - 데이터 표준 용어 설정, 명명 규칙 수립, 메타데이터 구축, 데이터사전 구축
            - 데이터 표준용어: 표준 단어사전, 표준 도메인사전, 표준 코드 등으로 구성되며 사전 간 상호 검증이 가능하도록 점검 프로세스 포함
            - 명명 규칙: 언어별로 작성되어 매핑 상태 유지
        - 데이터 관리 체계
            - 메타데이터와 데이터사전의 관리 원칙 수립
            - 데이터의 생명 주기 관리 방안을 수립하지 않으면 데이터 가용성 및 관리 비용 증대 문제에 직면하게 될 수 있음
        - 데이터 저장소 관리(Repository)
            - 메타 데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소 구성
            - 저장소는 데이터 관리 체계 지원을 위한 워크플로우 및 관리용 응용 소프트웨어를 지원하고 관리 대상 시스템과 인터페이스를 통한 통제가 이루어져야 함

        - 데이터 구조 변경에 따른 사전영향평가도 수행되어야 효율적인 활용 가능
    - 표준화 활동
        - 표준 준수 여부 주기적 점검 및 모니터링 실시
        - 변화 관리 및 주기적 교육 진행
        - 지속적인 데이터 표준화 개선 활동을 통한 실용성 증대

- 데이터 분석을 위한 3가지 조직 구조
    - 집중구조
        - 별도의 분석 전담 조직에서 분석업무 담당
        - 분석 조직이 전략적 중요도에 따라 우선순위를 정해 진행 가능
        - 현업 업무부서의 분석 업무와 이중화/이원화 가능성 높음
    - 기능구조
        - 일반적인 분석 수행 구조
        - 업무부서에서 분석 수행
        - 전사적 핵심 분석이 어려우며, 과거 실적에 국한된 분석 수행 가능성 높음
    - 분산구조
        - 분석조직 인력을 현업부서에 배치
        - 분석결과에 따른 신속한 액션 가능
        - 베스트 프랙티스 공유 가능
        - 업무과다 이원화 가능성 → 역할분담을 명확히 해야함

- 분석과제 관리 프로세스



3과목 - 데이터 분석

1.1 데이터 분석 기법의 이해

- 데이터 처리 과정
  - 데이터 분석을 위해서는 데이터 웨어하우스(DW)나 데이터마트(DM)를 통해 분석 데이터를 구성
  - 신규데이터나 DW에 없는 데이터는 기존 운영시스템(legacy)에서 직접 가져오거나 운영데이터저장소(ODS)에서 정제된 데이터를 가져와서 DW의 데이터와 결합하여 활용

- 시각화 기법
  - 가장 낮은 수준의 분석이지만 잘 사용하면 복잡한 분석보다 더 효율적
  - 대용량 데이터를 다룰 때와 탐색적 분석을 할 때 시각화는 필수

- 공간분석
  - 공간적 차원과 관련된 속성들을 시각화하는 분석
  - 지도 위에 관련된 속성들을 생성하고 크기, 모양, 선 굵기 등을 구분하여 인사이트를 얻음

- 탐색적 자료분석(EDA)
  - 다양한 차원과 값을 조합해가며 특이점이나 의미있는 사실을 도출하고 분석의 최종 목적을 달성해가는 과정
  - EDA의 4가지 주제
    - 저항성의 강조
    - 잔차 계산
    - 자료변수의 재표현
    - 그래프를 통한 현시성

- 통계분석
  - 어떤 현상을 종합적으로 한눈에 알아보기 쉽게 일정한 체계에 따라 숫자와 표, 그림의 형태로 나타내는 것
  - 기술 통계: 모집단으로부터 표본을 추출하고 표본이 가지고 있는 정보를 쉽게 파악할 수 있도록 숫자 또는 그래프 형태로 표현하는 절차
  - 추측 통계(추론 통계): 모집단으로부터 추출된 표본의 표본통계량으로부터 모집단의 특성인 모수에 관해 통계적으로 추론하는 절차

- **데이터 마이닝**
  - 대용량의 자료로부터 정보를 요약하고 미래에 대한 예측을 목표로 자료에 존재하는 관계, 패턴, 규칙 등을 탐색하고 이를 모형화함으로써 이전에 알지 못한 유용한 지식을 추출하는 분석 방법
  - 방법론
    - 기계학습: 인공신경망, 의사결정나무, 클러스터링, SVM

        - 패턴인식: 연관규칙, 장바구니 분석

- 모델링 성능 평가 기준
    - 데이터마이닝: 정확도, 정밀도, 디텍트레이트, 리프트
    - 시뮬레이션: Throughput, Average Waiting Time, Average Queue Length, Time in System

2.1 R 소개

- R의 탄생
    - R: 오픈소스 프로그램으로 통계·데이터마이닝과 그래프를 위한 언어
    - 다양한 최신 통계분석과 마이닝 기능을 제공하며, 5000개에 이르는 패키지가 수시로 업데이트 됨

- 통계분석 도구의 비교

구분	SAS	SPSS	오픈소스 R
프로그램비용	유료, 고가	유료, 고가	오픈소스
설치용량	대용량	대용량	모듈화로 간단
다양한 모듈 지원 및 비용	별도구매	별도구매	오픈소스
최근 알고리즘 및 기술 반영	느림	다소느림	매우빠름
학습자료 입수의 편의성	유료 도서 위주	유료 도서 위주	공개 논문 및 자료 많음
질의를 위한 공개 커뮤니티	NA	NA	매우활발

- R의 특징
    - 오픈소스 프로그램
    - 뛰어난 그래픽 및 성능
    - 시스템 데이터 저장 방식

    - 모든 운영체제에서 사용 가능(윈도우, 맥, 리눅스)
    - 표준 플랫폼(S 언어 기반)
    - 객체 지향언어이면서 함수형 언어

2.2 R기초

- 편리한 기능
    - R 작업환경 설정: R 단축아이콘 우측 클릭 → 속성 → 바로가기 → 시작위치에 현재 작업위치 입력 → 저장
    - 프로그램에서 작업환경 설정: setwd("작업디렉토리")
    - 도움말: help(함수), ?함수, RSiteSearch("함수명")
    - 히스토리: history(), savehistory(file="파일명"), loadhistory(file="파일명")
    - 콘솔청소: Ctrl+L

- 스크립트 사용하기
    - 한줄 실행: Ctrl+R
    - 여러줄 실행: 드래그 후 Ctrl+R
    - 주석처리: #

- 패키지
    - 패키지: R함수, 데이터 및 컴파일 코드의 모임
    - 패키지 자동설치: install.packages("패키지명")
    - 패키지 수동설치: install.packages("패키지명", "패키지 위치")

- 배치 실행
    - 매일 실행 되어야하는 프로그램을 시스템에서 프로세스를 자동으로 구동하는 작업

    - 배치파일 실행 명령: 윈도우창 - batch.R 실행파일이 있는 위치에서 R CMD BATCH batch.R
    - Path 지정: 내 컴퓨터에 오른쪽 마우스 클릭 → 속성 → 고급시스템설정 → 환경변수 → path 클릭 → R프로그램의 실행파일 위치를 찾아서 추가 → 저장

- 변수 다루기
    - R에서는 변수명만 선언하고 값을 할당하면 자료 형태를 스스로 인식하고 선언
    - 화면에 프린트하고자 할 때, print()를 사용해도 되지만 변수 값만 표현해도 출력 가능
    - 변수에 값을 할당할 때는 대입연산자 (<-, <<-, =, ->, ->>)를 사용할 수 있으나 <-를 추천
    - 메모리에 불필요한 변수가 있는지 확인하기 위해서는 ls(), 삭제는 rm()

- 기본적인 통계량 게산
    - 평균: mean()
    - 중간값: median()
    - 표준편차: sd()
    - 분산: var()
    - 공분산: cov()
    - 상관계수: cor()

- 함수의 생성 및 활용
    - R은 함수형 언어이기 때문에 프로그래머가 직접 활용 가능한 함수를 생성하여 활용 가능
    - 함수는 function(매개변수1, 매개변수2, ...)로 선언
    - 표현식이 2줄 이상인 경우 {}로 묶어 함수의 범위를 설정
    - 표현식은 변수 할당, 조건문(if)과 반복문(for, while, repeat), 전달값(return)으로 구성

- 연산자 우선순위

연산자 우선순위	뜻	예시
[[[	인덱스	a[1]
$	요소 뽑아내기, 슬롯 뽑아내기	a$coef
^	지수	5^2
- +	단항 마이너스와 플러스 부호	-3, +5
:	수열 생성	1:10
%any%	특수 연산자	%/% 나눗셈 몫 %% 나눗셈 나머지 %*% 행렬 곱
* /	곱하기, 나누기	3*5
+ -	더하기, 빼기	3+5
== != <> <= >=	비교	3==5
!	논리부정	
&	논리 "and", 단축 "and"	TRUE & TRUE
|	논리 "or", 단축 "or"	TRUE | TRUE
~	식(formula)	lm(log(brain)~log(body),data=Animals)
-> ->>	오른쪽 대입	3->a
=	오른쪽을 왼쪽으로 대입	a=3
<- <<-	오른쪽을 왼쪽으로 대입	a<-3
?	도움말	?lm

2.3 입력과 출력

- 데이터 입력과 출력
  - R에서는 텍스트 데이터 뿐만 아니라 데이터베이스와 다양한 통계 프로그램에서 작성된 데이터를 불러들여 적절한 데이터 분석을 수행할 수 있음

    - R에서는 부동소수점 표현시 7자리 수를 기본으로 세팅되어 있으며, option()함수, digit="숫자"를 지정해서 자릿수를 변경할 수 있음
    - 문자열을 파일로 저장하고자 할 때: cat("저장할 문자열", file="파일명")
    - 역슬래시(\)는 인식하지 못하므로 파일 경로는 슬래시(/) 또는 이중 역슬래시(\\)로 지정

- 외부 파일 입력과 출력
    - 고정자리 변수 파일: read.fwf("파일명", width=c(w1, w2, ...))
    - 구분자 변수 파일: read.table("파일명", sep="구분자")
    - csv 파일 읽기: read.csv("파일명", header=T) # 1행이 변수인 경우: header=T
    - csv 파일 출력: write.csv(데이터 프레임, "파일명")

- 웹 페이지(web page)에서 데이터 읽어 오기
    - 파일 다운로드: read.csv(http://www.~~/data.csv)
    - ftp 파일 다운로드: read.csv(ftp://ftp.~~~/data.csv)
    - html에서 테이블: readHTMLTable("url")

2.4 데이터 구조와 데이터 프레임

- 데이터 구조의 정의
    - 벡터(vector)
        - 원소 자료형: 동질적
        - 원소를 위치로 인덱싱 가능
        - 인덱싱으로 여러 개 원소로 구성된 하위 데이터 생성 가능
        - 원소들에 이름 부여 가능
    - 리스트(lists)
        - 원소 자료형: 이질적

        - 원소를 위치로 인덱싱 가능
        - 인덱싱으로 여러 개 원소로 구성된 하위 데이터 생성 가능
        - 원소들에 이름 부여 가능
    - 데이터 프레임(data frames)
        - 원소 자료형: 이질적
        - 원소를 위치로 인덱싱 가능
        - 인덱싱으로 여러 개 원소로 구성된 하위 데이터 생성 가능
        - 원소들에 이름 부여 가능
    - 단일값(scalars): 원소가 하나인 벡터로 인식/처리
    - 행렬(matrices): 차원을 가진 벡터로 인식
    - 배열(arrays): 3차원 또는 n차원까지 확장된 형태
    - 요인(factors): 고유값(value)이 요인의 수준(level)으로 구성된 벡터(범주형 변수, 집단 분류)

- 리스트 다루기
    - 리스트 원소 선택: L[[n]], L[["name"]], L$name

- 행렬 다루기
    - 행렬 설정: dim(vec)<-c(2,3)
    - 행과 열 이름 붙이기
        - rownames(mtrx)<-c("lowname1", "lowname2", ...)
        - colnames(mtrx)<-c("colname1", "colname2", ...)

- 데이터 구조 변환
    - 벡터→리스트: as.list(vec)
    - 벡터→행렬
        - 1열짜리: cbind(vec) 또는 as.matrix(vec)
        - 1행짜리: rbind(vec)

        - nxm행렬: matrix(vec, n, m)
    - 벡터→데이터프레임
        - 1열짜리: as.data.frame(vec)
        - 1행짜리: as.data.frame(rbind(vec))
    - 리스트→벡터: unlist(lst)
    - 리스트→행렬
        - 1열짜리: as.matrix(lst)
        - 1행짜리: as.matrix(rbind(lst))
        - nxm행렬: matrix(lst, n, m)
    - 리스트→데이터프레임
        - 원소들이 데이터의 열이면: as.data.frame(lst)
        - 원소들이 데이터의 행이면: rbind(obs[[1]], obs[[2]])
    - 행렬→벡터: as.vector(mat)
    - 행렬→리스트: as.list(mat)
    - 행렬→데이터프레임: as.data.frame(mat)
    - 데이터프레임→벡터
        - 1열짜리: dfm[[1]] 또는 fm[, 1]
        - 1행짜리: dfm[1,]
    - 데이터프레임→리스트: as.list(dfm)
    - 데이터프레임→행렬: as.matrix(dfm)

- 집단으로 분할하기
    - 벡터: split(vec, fac) - 벡터값과 팩터값의 길이가 같아야 한다.
    - 데이터프레임: split(dfm, fac)

- 함수 적용하기
    - 벡터, 행렬: apply(mtr, 1, func), apply(mtr, 2, func) → 1이면 행, 2면 열 자료 적용

    - 리스트: lapply(lst, func), sapply(lst, func)
    - 데이터프레임: lapply(dfm, func), sapply(dfm, func), apply(dfm, 1 or 2, func)

- 집단별로 함수 적용하기
    - tapply(vec, fac, func)
    - by(dfm, fac, func)

- 병렬 벡터들과 리스트들에 함수 적용하기
    - mapply(func, vec1, vec2, vec3, ...)
    - mapply(func, lst1, lst2, lst3, ...)

- 문자열 다루기
    - 문자열 길이: nchar("문자열")
    - 벡터의 길이: length(vec)
    - 문자열 연결하기: paste("단어", "문장", scalar)
    - 하위 문자열 추출하기: substr("문자열", 시작번호, 끝번호)
    - 구분자로 문자열 추출하기: strsplit("문자열", 구분자)
    - 문자열 대체하기: sub("대상문자열", 변경문자열", s), gsub("대상문자열", "변경문자열", s)

- 날짜 다루기
    - 문자열→날짜: as.Date("2014-12-25"), as.Date("12/25/2014, format="%m/%d/%Y")
    - 날짜→문자열: format(Sys,Date(), format="%m/%d/%Y")
    - format 인자값
        - %b: 축약 월이름(Jan)
        - %B: 전체 월이름(January)
        - %d: 두자리 숫자 일(31)

        - %m: 두자리 숫자 월(12)
        - %y: 두자리 숫자 년(14)
        - %Y: 네자리 숫자 년(2014)

3.1 데이터 변경 및 요약

- 데이터 마트
    - 데이터웨어하우스와 사용자 사이의 중간층에 위치한 것
    - 하나의 주제 또는 하나의 부서 중심의 데이터 웨어하우스

- 요약변수와 파생변수
    - 요약변수
        - 수집된 정보를 분석에 맞게 종합한 변수
        - 데이터 마트에서 가장 기본적인 변수
        - 많은 모델이 공통으로 사용할 수 있어 재활용성 높음
        - 예시
            - 기간별 구매 금액, 횟수, 여부
            - 위클리 쇼퍼
            - 상품별 구매 금액, 횟수, 여부
            - 상품별 구매 순서
            - 유통 채널별 구매 금액
            - 단어 빈도
            - 초기 행동변수
            - 트랜드 변수
            - 결측값과 이상값 처리
            - 연속형 변수의 구간화

- 파생변수
  - 사용자(분석가)가 특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여한 변수
  - 매우 주관적일 수 있으므로 논리적 타당성을 갖출 필요가 있다.
  - 예시
    - 근무시간 구매자 수
    - 주 구매 매장 변수
    - 주 활동 지역 변수
    - 주 구매 상품 변수
    - 구매상품 다양성 변수
    - 선호하는 가격대 변수
    - 시즌 선호 고객 변수
    - 라이프 스테이지 변수
    - 라이프스타일 변수
    - 휴면가망 변수
    - 최대가치 변수
    - 최적 통화시간

- **reshape** 패키지
  - 2개의 핵심적인 함수로 구성
    - melt(): 쉬운 casting을 위해 데이터를 적당한 형태로 만들어주는 함수
    - cast(): 데이터를 원하는 형태로 계산 또는 변형시켜주는 함수
  - 변수를 조합해 변수명을 만들고 변수들을 시간, 상품 등의 차원에 결합해 다양한 요약변수와 파생변수를 쉽게 생성하여 데이터 마트를 구성할 수 있게 해주는 패키지

- **sqldf** 패키지
  - R에서 sql 명령어를 사용할 수 있게 해주는 패키지
  - SAS의 proc sql과 같은 기능

- head([df]) → sqldf("select * from [df] limit 6")
- subset([df], [col] %in% c("BF", "HF")) → sqldf("select * from [df] where [col] in('BF', 'HF')")
- merge([df1], [df2]) → sqldf("select * from [df1], [df2]")

- **plyr** 패키지
  - apply함수를 기반으로 데이터와 출력변수를 동시에 배열로 치환하여 처리하는 패키지
  - split-apply-combine 방식으로 데이터를 분리하고 처리한 다음, 다시 결합하는 등 필수적인 데이터 처리 기능 제공

	array	data frame	list	nothing
array	aaply	adply	alply	a_ply
data frame	daply	ddply	dlply	d_ply
list	laply	ldply	llply	l_ply
n replicates	raply	rdply	rlply	r_ply
function arguments	maply	mdply	mlply	m_ply

- data.table
  - R에서 가장 많이 사용하는 데이터 핸들리 패키지 중 하나
  - 대용량 데이터의 탐색, 연산, 병합에 유용
  - 기존 data.frame 방식보다 월등히 빠른 속도
  - 특정 column을 key값으로 색인을 지정한 후 데이터를 처리
  - 빠른 grouping과 ordering, 짧은 문장 지원 측면에서 데이터 프레임보다 유용

3.2 데이터 가공

- 변수의 구간화
  - 신용평가모형, 고객 세분화 등의 시스템으로 모형을 적용하기 위해 각 변수들을 구간화하여 점수를 적용하는 방식이 활용됨
  - 변수의 구간화를 위한 rule
    - 10진수 단위로 구간화
    - 구간을 5개로 나누는 것이 보통
    - 7개 이상의 구간을 잘 만들지 않음

- 변수 구간화 방법
  - binning: 연속형 변수를 범주형 변수로 변환하기 위해 50개 이하의 구간에 동일한 수의 데이터를 할당하여 의미를 파악하면서 구간을 축소하는 방법
  - 의사결정나무: 모형을 통해 연속성 변수를 범주형 변수로 변환하는 방법

3.3 기초 분석 및 데이터 관리

- 결측값 처리
  - 변수에 데이터가 비어있는 경우: NA, ., 99999999, Unknown, Not Answer 등으로 표현
  - 단순 대치법(single imputation)
    - completes analysis
      - 결측값의 레코드를 삭제
      - 부분적 관측자료만 사용하므로 통계적 추론 타당성 문제가 있음
    - 평균대치법: 관측 및 실험을 통해 얻어진 데이터의 평균으로 대치
      - 비조건부 평균 대치법: 관측 데이터의 평균으로 대치
      - 조건부 평균 대치법: 회귀분석을 통해 데이터를 대치
    - 단순확률 대치법: 평균대치법에서 추정량 표준 오차의 과소 추정문제를 보완한 방법으로 Hot-deck 방법, nearest neighbor 방법이 있음

    - **다중 대치법**(multiple imputation)
        - 단순 대치법을 m번 실시하여, m개의 가상적 자료를 만들어 대치하는 방법
        - 순서: 대치 → 분석 → 결합

- R의 결측값 처리 관련 함수
    - complet.cases(): 데이터 내 레코드에 결측값이 있으면 FALSE, 없으면 TRUE 반환
    - is.na(): 결측값이 NA인지 여부 반환
    - DMwR 패키지
        - centralImputation(): NA값을 중앙값(center value)으로 대치
            - 숫자→중위수
            - factor→최빈값
        - knnImputation(): NA값을 k 최근 이웃 분류 알고리즘을 사용하여 대치
            - k개 주변 이웃까지의 거리를 고려하여 가중 평균한 값을 사용
    - Amelia 패키지 - amelia(): time-series-cross-sectional data set(여러 국가에서 매년 측정한 자료)에서 활용

- 이상값 처리
    - 이상값
        - 의도하지 않은 현상으로 입력된 값 / 의도된 극단값 → 활용할 수 있음
        - 잘못 입력된 값 / 의도하지 않은 현상으로 입력된 분석 목적에 부합되지 않는 값 → bad data이므로 제거
    - 이상값의 인식
        - ESD(Extreme Studentized Deviation): 평균으로부터 3표준편차 떨어진 값
        - 기하평균-2.5*표준편차 < data < 기하평균+2.5*표준편차
        - Q1-1.5(Q3-Q1) < data < Q3+1.5(Q3-Q1)를 벗어나는 데이터
    - 이상값의 처리

        - 절단(trimming): 이상값이 포함된 레코드를 삭제
        - 조정(winsorizing): 이상값을 상한 또는 하한 값으로 조정

4.1 통계분석의 이해

- 통계
    - 통계: 특정집단을 대상으로 수행한 조사나 실험을 통해 나온 결과에 대한 요약된 형태의 표현
    - 통계자료의 획득 방법: 총조사(census)와 표본조사(sampling)
    - 표본 추출 방법
        - 단순랜덤추출(simple random sampling): 각 샘플에 번호를 부여해 임의의 n개 추출
        - 계통추출법(systematic sampling): K개씩 n개의 구간을 나누고 첫 구간에서 임의로 하나를 선택한 후 K개씩 띄어서 n개의 표본 선택
        - 집락추출법(cluster sampling): 군집별로 단순랜덤추출법 수행 후 모든 자료를 활용하거나 샘플링
        - 층화추출법(stratified random sampling): 유사한 원소끼리 층을 나누어 각 층에서 랜덤추출
    - 자료의 측정 방법
        - 질적척도: 범주형자료로 크기 차이가 계산되지 않음
            - 명목척도: 측정 대상이 어느 집단에 속하는지 분류
            - 순서척도: 서열관계 관측 척도
        - 양적척도: 수치형자료로 크기 차이 계산 가능
            - 구간척도: 속성의 양을 측정하는 것으로 구간이나 구간 간격이 의미있음
            - 비율척도: 간격에 대한 비율이 의미를 가지는 자료로 절대적 기준인 0이 존재하고 사칙연산이 가능

- 표본조사의 유의점

    - 표본편의(Sampling Bias)
        - 표본추출 과정에서 특정 대상이 다른 대상에 비해 우선적으로 추출될 때 생기는 오차
        - 확률화(Randomization)에 의해서 최소화하거나 없앨 수 있음
    - 표본값으로 모집단의 모수를 측정할 때 표본오차의 비표본오차가 발생할 수 있음
    - 응답오차
    - 유도질문

- 통계분석
    - 기술통계(descriptive statistic): 평균, 표준편차, 중위수, 최빈값, 그래프
    - 통계적 추론(statistical inference): 모수추정, 가설검정, 예측

- 확률 및 확률 분포
    - 확률변수(random variable): 특정 값이 나타날 가능성이 확률적으로 주어지는 변수
    - **이산형** 확률분포(discrete distribution)
        - 베르누이 분포: 결과가 2개만 나오는 경우
        - 이항분포: 베르누이 시행을 n번 반복했을 때 k번 성공할 확률
        - 기하분포: p 확률의 베르누이 시행에서 첫번째 성공까지 x번 실패할 확률
        - 다항분포: 세가지 이상 결과를 가지는 반복 시행에서 발생하는 확률 분포
        - 포아송분포: 시간과 공간 내 발생하는 사건의 발생횟수에 대한 확률분포
    - **연속형** 확률분포(continuous distribution)
        - 균일분포: 모든 확률변수가 균일 확률을 가지는 확률분포
        - 정규분포: 평균이 μ이고 표준편차가 σ인 x의 확률밀도함수
            - 표준정규분포: 평균이 0이고 표준편차가 1인 정규분포
        - 지수분포: 어떤 시간이 발생할 때까지의 경과시간에 대한 연속확률분포
        - t분포: 두 집단 간 평균의 동일성 검정에 사용
        - F분포: 두 집단 간 분산의 동일성 검정에 사용

        - x^2분포(카이제곱분포): 모집단의 모분산 가설 검정에 사용, 두 집단 간의 동일성 검정에 사용

- 추정
    - 표본으로부터 미지의 모수를 추측하는 것
    - 점추정(point estimation)
        - '모수가 특정한 값일 것'이라고 추정하는 것
        - 평균, 표준편차, 중앙값 등을 추정
        - 점추정 조건: 불편성(unbiasedness), 효율성(efficiency), 일치성(consistency), 충족성(sufficient)
    - 구간추정(interval estimation)
        - 점추정을 보완하기 위해 모수가 특정 구간에 있을 것이라고 추정하는 것
        - 모분산을 알거나 대표본의 경우 표준정규분포 활용
        - 모분산을 모르거나 소표본의 경우 t분포 활용

- 가설검정
    - 모집단에 대한 가설을 설정한 뒤, 그 가설의 채택 여부를 결정하는 방법
    - 귀무가설(null hypothesis, H_0) vs 귀무가설(alternative hypothesis, H_1)
    - 1종 오류(Type Ⅰ Error, α): 귀무가설H_0가 옳은데도 귀무가설을 기각하게 되는 오류
        - p-value: 귀무가설이 사실인데도 불구하고 사실이라고 판정할 때, 판정이 잘못되었을 실제 확률
    - 2종 오류(Type Ⅱ Error, β): 귀무가설 H_0가 옳지 않은데도 귀무가설을 채택하게 되는 오류
    - 1종 오류의 크기를 0.1, 0.05, 0.01로 고정시키고 2종 오류가 최소가 되도록 기각역을 설정
    - 가설검정시 많이 활용되는 분포: t-분포, 카이제곱분포, F분포

- 모수적 검정

    - 모집단의 분포에 대한 가정을 하고 검정통계량과 통계량 분포를 유도해 검정 실시
    - 가설 설정 방법: 분포의 모수에 대해 가설 설정
    - 검정 방법: 표본평균, 표본분산 등을 이용

- 비모수 검정
    - 모집단의 분포에 대한 아무 제약을 가하지 않고 검정을 실시
    - 관측 자료 수가 많지 않거나 자료가 개체간 서열관계를 나타내는 경우에 이용
    - 가설 설정 방법: '분포의 형태가 동일하다', '분포의 형태가 동일하지 않다' 라는 식으로 분포의 형태에 대해 설정
    - 검정 방법: 관측값의 절대적 크기에 의존하지 않는 관측값들의 순위나 두 관측값 차이의 부호를 이용해 검정
        - 예: 부호검정(sign test), 윌콕슨의 순위합검정(rank sum test), 윌콕슨의 부호순위합검정(Wilcoxon signed rank test), 만-위트니의 U 검정, 런 검정(run test), 스피어만의 순위상관계수

4.2 기초 통계 분석

- 기술 통계(descriptive statistic)
    - 자료의 특성을 표, 그림, 통계량 등을 사용해 쉽게 파악할 수 있도록 정리/요약 하는 것
    - 통계량에 의한 자료 정리
        - 중심 위치의 측도: 평균, 중앙값, 최빈값
        - 산포의 측도: 분산, 표준편차, 범위, 사분위수범위, 변동계수, 표준오차
        - 분포의 형태: 왜도, 첨도
    - 그래프를 통한 자료 정리
        - 범주형 자료: 막대그래프, 파이차트, 모자이크 플랏 등
        - 연속형 자료: 히스토그램, 줄기-잎 그림, 상자그림 등

- 인과관계의 이해
  - 용어
    - 종속변수(반응변수 y)
    - 독립변수(설명변수 x)
    - 산점도(scatter plot)
      - 산점도에서 확인할 수 있는 것
        - 두 변수 사이의 선형관계가 성립하는가?
        - 두 변수 사이의 함수관계가 성립하는가?
        - 이상값의 존재 여부와 몇 개의 집단으로 구분되는지 확인
  - 공분산(covariance)
    - 두 변수간의 상관 정도를 상관계수를 통해 확인할 수 있음
    - $Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$

- 상관분석(correlation analysis)
  - 정의와 특성
    - 상관분석: 두 변수 간 관계를 상관계수를 이용해 알아보는 분석 방법
    - 상관계수가 1에 가까울수록 강한 양의 상관관계, -1에 가까울수록 강한 음의 상관관계를 가짐
    - 상관계수가 0인 경우 데이터 간 상관이 없음
  - 유형
    - 피어슨
      - 등간척도 이상으로 측정된 두 변수의 상관관계 측정
      - 두 변수 간의 선형관계 크기를 측정
      - 특징: 연속형 변수, 정규성 가정
      - 상관계수: 피어슨 γ(적률상관계수)
    - 스피어만
      - 순서, 서열 척도인 두 변수들 간의 상관관계 측정
      - 두 변수 간의 비선형적인 관계도 측정 가능하다.

        - 특징: 순서형 변수, 비모수적 방법
        - 상관계수: 순위상관계수 ρ(로우)
    - R코드: cor(x, y, method=c("pearson", "kendall", "spearman"))

4.3 회귀분석

- 회귀분석
    - 하나 또는 그 이상의 독립변수들이 종속변수에 미치는 영향을 추정할 수 있는 통계기법
    - $y_i = \beta_0 + \beta_i x_i + \epsilon_i$
      $i = 1, 2, ..., n,\ \epsilon_i$~$N(0, \sigma^2),\ y$: 종속변수, x: 독립변수
    - 독립변수가 1개 → 단순선형회귀분석
      독립변수가 2개 이상 →다중선형회귀분석
    - 최소제곱법
        - 측정값을 기초로 제곱합을 만들고 그것의 최소인 값을 구하여 처리하는 방법
        - 잔차 제곱이 가장 작은 선을 선택
        - 최소제곱: 제곱오차를 최소로 하는 회귀계수 추정량

- 회귀분석의 검정
    - 회귀식(모형)에 대한 검증: F-검정
    - 회귀계수들에 대한 검증: t-검정
    - 모형의 설명력은 결정계수(R^2)로 알 수 있음
        - 결정계수: 총 변동 중 회귀모형에 의해 설명되는 변동이 차지하는 비율
        - $R^2 = \frac{\text{회귀제곱합}}{\text{전체제곱합}} = \frac{SSR}{SST},\ 0 \le R^2 \le 1$
    - 단순회귀분석의 결정계수는 상관계수값의 제곱과 같음
    - 회귀식에 대한 검정은 독립변수의 기울기(회귀계수)가 0이라는 가정을 귀무가설, 0이 아니라는 가정을 대립가설로 둠

- 선형회귀분석
  - 가정
    - 선형성: 입력변수와 출력변수의 관계가 선형
    - 독립성: 잔차와 독립변인은 관련이 없음
      - 검정: Durbin-Waston 통계량 사용
    - 등분산성: 독립변인의 모든 값에 대한 오차들의 분산이 일정
    - 비상관성: 관측치들의 잔차들끼리 상관이 없어야 함
    - 정상성: 잔차항이 정규분포를 이뤄야 함
      - 검정: Q-Q plot, Kolmogolov-Smirnov 검정, Shaprio-Wilk 검정, 히스토그램
  - 다중선형회귀분석의 다중공선성(multicollinearity)
    - 다중회귀분석에서 설명변수들 사이에 선형관계가 존재하면 회귀계수의 정확한 추정이 곤란
  - 다중공선성 검사 방법
    - 분산팽창요인(VIF): 10보다 크면 심각한 문제
    - 상태지수: 10이상이면 문제가 있다고 보고, 30보다 크면 심각, 선형관계가 강한 변수는 제거

- 회귀분석의 종류
  - 단순회귀: $Y = \beta_0 + \beta_1 X_1 + \epsilon$
  - 다중회귀: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$
  - 로지스틱회귀: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$
  - 다형회귀: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_{11}^2 + \beta_{22} X_{22}^2 + \beta_{12} X_1 X_2 + \epsilon$ (k=2, 이차함수)
  - 곡선회귀: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ (2차 곡선)
  - 비선형회귀: $Y = \alpha e^{-\beta X} + \epsilon$

- 변수선택법(variable selection)

- 모든 가능한 조합: 모든 가능한 독립 변수들의 조합에 대한 회귀모형을 분석해 가장 적합한 모형 선택
- 전진선택법(forward selection)
  - 절편만 있는 상수모형으로부터 시작해 중요하다고 생각되는 설명변수부터 차례로 모형에 추가
  - 이해가 쉬움
  - 많은 변수에서 활용 가능
  - 변수 값의 작은 변동에 결과가 달라져 안정성이 부족
- 후진제거법(backward selection)
  - 독립변수 후보 모두를 포함한 모형에서 가장 적은 영향을 주는 변수부터 하나씩 제거
  - 전체 변수들의 정보를 이용 가능
  - 변수가 많은 경우 활용이 어려움
  - 안정성 부족
- 단계별방법(stepwise method)
  - 전진선택법에 의해 변수를 추가하면서 새롭게 추가된 변수에 기인해 기존 변수가 그 중요도가 약화되면 해당 변수를 제거하는 등 단계별로 추가/삭제되는 변수를 검토해 더 이상 없을 때 중단

4.4 시계열 분석

- 시계열 자료(time series)
  - 시계열 자료: 시간의 흐름에 따라 관찰된 값들
  - 시계열 데이터의 분석목적: 미래 값의 예측, 특성 파악(경향, 주기, 계절성, 불규칙성 등)
  - 정상성 (3가지를 모두 만족)
    - 평균이 일정(모든 시점에서 일정한 평균을 가짐)
      - 평균이 일정하지 않은 경우 차분(difference)을 통해 정상화

        - 분산도 일정
            - 분산이 일정하지 않을 경우 변환(transformation)을 통해 정상화
        - 공분산도 특정 시점에서 t, s에 의존하지 않고 일정

        ⇒ 정상 시계열

- 시계열 데이터 분석 절차
    - 시간그래프 그리기
    - 추세와 계절성 제거
    - 잔차 예측
    - 잔차에 대한 모델 적합
    - 예측된 잔차에 추세와 계절성을 더해 미래 예측

- 시계열 모형
    - 자기회귀모형(AR, Autoregressive model)
        - p시점 전의 자료가 현재 자료에 영향을 주는 모형
        - $Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + ... + \Phi_p Z_{t-p} + \alpha_t$
        - ACF는 빠르게 감소, PACF는 절단점이 존재 → AR(절단점-1)로 계산
    - 이동평균모형(MA, Moving Average model)
        - 자기 자신의 과거 값을 사용하여 설명
        - 백색 잡음의 현재값과 자기 자신의 과거값의 선형 가중합으로 이루어진 정상 확률 모형
        - $Z_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2} - ... - \theta_p \alpha_{t-p}$
        - ACF는 절단점이 존재, PACF는 빠르게 감소
    - 자기회귀누적이동평균모형(ARIMA(p, d, q))
        - $Z_t = \alpha_1 - \theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2}$
        - d(차분)=0이면 정상성 만족
        - p=0이면 d번 차분한 MA(q) 모델
        - q=0이면 d번 차분한 AR(p) 모델

- 분해 시계열
  - 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법
  - 추세요인(Trend factor): 형태가 오르거나 내리는 추세, 선형, 이차식, 지수 형태
  - 계절요인(Seasonal factor): 요일, 월, 사분기 별로 변화하여 고정된 주기에 따라 자료가 변화
  - 순환요인(Cyclical factor): 명백한 경제적, 자연적 이유 없이 알려지지 않은 주기로 자료가 변화
  - 불규칙 요인(Random factor): 위 세가지 요인으로 설명할 수 없는 회귀분석에서 오차에 해당하는 요인

4.5 다차원 척도법과 주성분분석

- 다차원 척도법
  - 군집 분석과 같이 개체들을 대상으로 변수들을 측정한 후 개체들 사이의 유사성/비유사성을 측정하여 개체들을 2차원 또는 3차원 공간상에서 점으로 표현하는 분석 방법
  - 목적
    - 개체들의 비유사성을 이용하여 2차원 공간상에 점으로 표시하고 개체들 사이의 집단화를 시각적으로 표현
  - 방법
    - 개체들의 거리 계산은 유클리드 거리행렬을 이용
      - $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}$
    - STRESS
      - 개체들을 공간상에 표현하기 위한 방법으로 STRESS나 S-STRESS를 부적합도 기준으로 사용
      - 최적모형의 적합은 부적합도를 최소로 하는 방법으로 일정 수준 이하로 될 때까지 반복해서 수행

    - 종류
        - 계량적 MDS(Metric MDS)
            - 데이터가 구간척도나 비율척도인 경우 활용(전통적인 다차원척도법)
            - N개의 케이스에 대해 p개의 특성변수가 있는 경우, 각 개체들 간의 유클리드 거리행렬을 계산하고 개체들 간 비유사성 S(거리제곱 행렬의 선형항수)를 공간상에 표현
        - 비계량적 mds(Nonmetric MDS)
            - 데이터가 순서척도인 경우 활용
            - 개체들 간 거리가 순서로 주어진 경우에는 순서척도를 거리 속성과 같도록 변환(monotone transformation)하여 거리를 생성한 후 적용

- 주성분 분석
    - 상관관계가 있는 변수들을 결합해 상관관계가 없는 변수로 분산을 극대화하는 변수
    - 선형결합으로 변수를 축약, 축소하는 기법
    - 목적
        - 여러 변수들을 소수의 주성분으로 축소하여 데이터를 쉽게 이해하고 관리
        - 주성분분석을 통해 차원을 축소하여 군집분석에서 군집화결과와 연산 속도 개선
        - 회귀분석에서 다중공선성 최소화
    - 주성분분석 vs 요인분석
        - 요인분석(Factor Analysis): 등간척도(비율척도)로 두 개 이상의 변수들이 잠재되어 있는 공통 인자를 찾아내는 기법
        - 공통점
            - 모두 데이터를 축소하는데 활용
            - 몇 개의 새로운 변수들로 축소
        - 차이점
            - 생성된 변수의 수와 이름
                - 요인분석: 몇 개로 지정할 수 없으나 이름을 붙일 수 있음
                - 주성분분석: 제1주성분, 제2주성분,...을 생성(보통 2개)

        - 생성된 변수들 간 관계
            - 요인분석: 생성된 변수들이 기본적으로 대등한 관계
            - 주성분분석: 제1주성분, 제2주성분 순으로 중요함
        - 목표변수와의 관계
            - 요인분석: 목표변수를 고려하지 않고 주어진 변수들 간 비슷한 성격들을 묶음
            - 주성분분석: 목표변수를 고려하여 주성분 변수 생성
    - 주성분의 선택법
        - 누적기여율(cumulative proportion)이 85% 이상이면 주성분의 수로 결정 가능
            - 제2주성분까지의 누적기여율=제1주성분의 기여율+제2주성분의 기여율
        - scree plot에서 고유값(eigen value)이 수평을 유지하기 전 단계로 주성분의 수를 선택
        - 전체 변이 공헌도방법: 전체 변이의 70~90% 정도가 되도록 주성분의 수를 선택
        - 평균 고유값 방법: 고유값들의 평균을 구한 후 고유값이 평균 이상이 되는 주성분을 선택

5.1 데이터 마이닝의 개요

- 데이터 마이닝
    - 대용량 데이터에서 의미있는 패턴을 파악하거나 예측하여 의사결정에 활용하는 방법
    - 통계분석과 차이점
        - 통계분석: 가설이나 가정에 따른 분석, 검증
        - 데이터마이닝: 다양한 수리 알고리즘을 이용해 데이터베이스의 데이터로부터 의미있는 정보를 추출
    - 활용 분야: 분류, 예측, 군집화, 시각화 등

    - 방법론: 의사결정나무, 로지스틱 회귀분석, 최근접 이웃법, 군집분석, 연관규칙분석 등

- 분석 방법
    - 지도학습(Supervised Learning)
        - 의사결정나무(Decision Tree)
        - 인공신경망(Artificial Neural Network)
        - 로지스틱 회귀분석(Logistic Regression)
        - 최근접 이웃법(K-Nearest Neighbor)
        - 사례기본 추론(Case-Based Reasoning)
    - 비지도학습(Unsupervised Learning)
        - OLAP(On-Line Analytic Processing)
        - 연관규칙분석(Association Rule Analysis)
        - 군집분석(k-Means Clustering)
        - SOM(Self Organizing Map)

- 데이터 마이닝 추진단계
    1. 목적설정: 데이터 마이닝을 위한 명확한 목적 설정
    2. 데이터 준비: 모델링을 위한 다양한 데이터를 준비, 데이터 정제를 통해 품질을 보장
    3. 데이터 가공: 목적변수 정의, 모델링을 위한 데이터 형식으로 가공
    4. 기법 적용: 데이터 마이닝 기법을 적용하여 정보를 추출
    5. 검증: 마이닝으로 추출한 결과를 검정하고 업무에 적용해 기대효과를 전파

- 데이터 분할
    - 구축용(training data): 50%의 데이터를 모델링을 위한 훈련용으로 활용
    - 검정용(validation data): 30%의 데이터를 구축된 모형의 과대/과소 추정의 판정 목적으로 활용

- 시험용(test data): 20%의 데이터를 테스트데이터나 과거 데이터를 활용해 모델의 성능 평가에 활용

- 모델의 성능 평가
  - 예: 은행 대출 문제 - 연이율 20%, 100만원을 100명에게 대출

실제분류

예측분류	모형1	우량	불량
	우량	65	10
	불량	5	20

실제분류

예측분류	모형1	우량	불량
	우량	75	0
	불량	15	10

  - 기대수익
    - 모형1=(65*20)-(10*100)=300만원
    - 모형2=(75*20)=1500만원
  - 기대손실
    - 모형1=(5*20)+(10*100)=1100만원
    - 모형2=(15*20)=300만원
  - 결과: 기대수익, 기대손실비용 면에서 볼 때 모형2가 우수함

5.2 분류분석

- 분류분석과 예측분석
  - 공통점: 레코드의 특정 속성의 값을 미리 알아 맞히는 것
  - 차이점
    - 분류: 레코드(튜플)의 범주형 속성의 값을 알아맞히는 것
    - 예측: 레코드(튜플)의 연속성 속성의 값을 알아맞히는 것
  - 예
    - 분류

        - 학생들의 국어, 영어 등의 점수를 통해 내신등급을 예측
        - 카드회사에서 회원들의 가입 정보를 통해 1년 후 신용등급 예측
    - 예측
        - 학생들의 여러 정보를 입력해 수능점수를 예측
        - 카드회사에서 회원들의 가입 정보를 통해 연 매출 예측
- 분류 모델링: 신용평가모형, 사기방지모형, 이탈모형, 고객세분화
- 분류 기법
    - 로지스틱 회귀분석(logistic regression)
    - 의사결정나무(decision tree), CART(Classification And Regression Tree), C5.0
    - 베이지안 분류(Bayesian classification), 나이브 베이지안 분류(naive Bayesian)
    - 인공신경망(artificial neural network)
    - 지지도백터기계(support vector machine)
    - k 최근접 이웃(k-nearest neighborhood)
    - 규칙기반의 분류와 사례기반추론(case based reasoning)

- **의사결정나무**
    - 분류함수를 의사결정 규칙으로 이뤄진 나무 모양으로 그리는 방법
    - 의사결정 문제를 시각화해 의사결정이 이뤄지는 시점과 성과를 한 눈에 볼 수 있게 함
    - 주어진 입력값에 대해 출력값을 예측하는 모형으로 분류나무와 회귀나무 모형이 있음
    - 훈련을 한 뒤 예측이 이루어지므로 교사학습법(지도학습법)이다.
    - 특징
        - 계산결과가 의사결정나무에 직접 나타나게 돼 분석이 간편함
        - 분류정확도가 좋음
        - 계산이 복잡하지 않아 대용량데이터에서도 빠르게 만들 수 있음
        - 비정상 잡음 데이터에 대해서도 민감함 없이 분류

        - 한 변수와 상관성이 높은 다른 불필요한 변수가 있어도 크게 영향 받지 않음
    - 활용
        - 세분화(segmentation): 데이터를 비슷한 특성을 갖는 몇 개의 그룹으로 분할해 그룹별 특성 발견
        - 분류(classification): 관측개체를 여러 예측변수들에 근거해 목표변수의 범주를 몇 개의 등급으로 분류하고자 하는 경우
        - 예측(prediction): 자료에서 규칙을 찾아내고 이를 이용해 미래의 사건을 예측하고자 하는 경우
        - 차원축소 및 변수선택(reduction, variable selection): 매우 많은 수의 예측변수 중 목표변수에 영향을 미치는 변수들을 골라내고자 하는 경우
        - 교호작용효과의 파악(interaction effect identification): 여러 개의 예측변수들을 결합해 목표 변수에 작용하여 파악하고자 하는 경우
        - 범주의 병합 또는 연속성 변수의 이산화(binning): 범주형 목표변수의 범주를 소수의 몇 개로 병합하거나 연속성 목표변수를 몇 개의 등급으로 이산화하고자 하는 경우
    - 분석 과정
        - 성장 → 가지치기 → 타당성평가 → 해석 및 예측
        - 가지치기(Pruning): 너무 큰 나무 모형은 자료를 과대적합하고 너무 작은 나무 모형은 과소적합할 위험이 있어 마디에 속한 자료가 일정 수 이하일 경우 분할을 정지하고 가지치기 실시
        - 불순도에 따른 분할 측도
            - 카이제곱 통계량
            - 지니지수
                - 노드의 불순도를 나타내는 값
                - 지니지수 값이 클수록 이질적이며 순수도가 낮음
                - 지니지수=$1-(p_1)^2-(p_2)^2$
            - 엔트로피 지수
                - 무질서 정도에 대한 측도
                - 엔트로피 지수 값이 클수록 순수도가 낮음
                - 엔트로피 지수=$-(p_1log_2p_1+p_2log_2p_2)$

- 의사결정나무 분석의 종류
  - CART(Classification And Regression Tree)
    - 연속형 목표변수를 예측하는 의사결정나무
    - 목적변수가 범주형인 경우 지니지수, 연속형인 경우 분산을 이용해 이진분리 사용
    - 개별 입력변수 뿐만 아니라 입력변수들의 선형결합들 중 최적의 분리를 찾을 수 있음
  - C4.5와 C5.0
    - 다지분리(multiple split)가 가능하고 범주형 입력변수의 범주 수만큼 분리 가능
    - 불순도의 측도로 엔트로피 지수 사용
  - CHAID(CHi-squared Automatic Interaction Detection)
    - 가지치기를 하지 않고 적당한 크기에서 나무 모형의 성장을 중지하며 입력변수가 반드시 범주형 변수여야 함
    - 불순도의 측도로 카이제곱 통계량 사용

- **앙상블 기법**
  - 주어진 자료로부터 여러 개의 예측모형들을 만든 후 조합하여 하나의 최종예측모형을 만드는 방법
  - 다중 모델 조합(combining multiple models), classifier combination 방법이 있음
  - 학습 방법의 불안전성을 해결하기 위해 고안된 기법
  - 가장 불안정성을 가지는 기법은 의사결정나무
  - 가장 안정성을 가지는 기법은 1-nearest neighbor
  - 기법의 종류
    - 배깅(bagging: bootstrap aggregating)
      - 여러 개의 붓스트랩 자료를 생성하고 각 붓스트랩 자료의 예측모형 결과를 결합하여 결과를 산정
        - 붓스트랩(bootstrap)

        - 주어진 자료에서 단순랜덤 복원추출 방법을 활용해 동일한 크기의 표본을 여러개 생성하는 샘플링 방법
        - 샘플에 한 번도 선택되지 않는 원데이터가 발생할 확률 = 전체 샘플의 36.8%
    - 배깅은 훈련자료를 모집단으로 생각하고 평균 예측모형을 구한 것과 같이 분산을 줄이고 예측력을 향상시킬 수 있음
    - 반복추출방법을 사용하므로 같은 데이터가 한 표본에 여러 번 추출될 수도 있고, 한 번도 추출되지 않을 수도 있다.
  - 부스팅(boosting)
    - 예측력이 약한 모형(weak learner)들을 결합하여 강한 예측모형을 만드는 방법
    - 훈련오차를 빨리 그리고 쉽게 줄일 수 있고, 예측오차의 향상으로 배깅에 비해 뛰어난 예측력을 보임
  - 랜덤 포레스트(Random forest)
    - 의사결정나무의 특징인 분산이 크다는 점을 고려하여 배깅과 부스팅보다 더 많은 무작위성을 주어 약한 학습기들을 생성한 후 이를 선형 결합하여 최종 학습기를 만드는 방법
    - 이론적 설명이나 해석이 어렵다는 단점이 있지만 예측력이 매우 높은 장점이 있음
    - 입력변수가 많은 경우 더 좋은 예측력을 보임

- 성과분석
  - 오분류표를 통한 모델 평가

		Condition(실제)		
		Positive	Negative	
Prediction (예측)	Positive	True Positive TP = 20	False Positive FP = 180	Positive predictive value =TP/(TP+FP) =20/(20+180) =10%
	Negative	False Negative FN = 10	True Negative TN = 1820	Negative Predictive value =TN/(FN+TN) =18020/(10+1820) =99.5%
		Sensitivity 민감도 =TP/(TP+FN) =20/(20+10) =67%	Specificity 특이도 =TN/(FP+TN) =1820/(180+1820) =91%	

- 정분류율=$\frac{TP+TN}{TP+TN+FP+FN}$: 전체 중 정답의 비율
- 특이도=$\frac{TN}{TN+FP}$: 실제로 거짓인 값 중 정답의 비율
- 민감도=$\frac{TP}{TP+FN}$: 실제로 참인 값 중 정답의 비율
- 재현율=민감도
- 정확도=$\frac{TP}{TP+FP}$: 참으로 예측한 값 중 정답의 비율.
- F1=$2 \times \frac{\text{정확도} \times \text{재현율}}{\text{정확도} + \text{재현율}}$
  - F1: 정확도와 재현율은 한 지표의 값이 높아지면 다른 지표의 값이 낮아질 가능성이 높은 관계를 지니고 있어 이러한 효과를 보정하여 하나의 지표로 만들어낸 것

◦ 향상도곡선

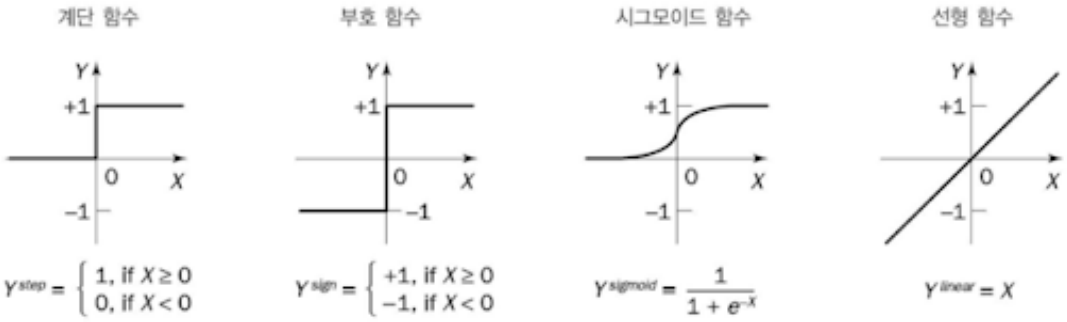
- 랜덤모델과 비교하여 해당 모델의 성과가 얼마나 향상되었는지를 각 등급별로 파악하는 그래프
- 상위등급에서 매우 크고 하위 등급으로 갈수록 감소하게 되면 일반적으로 모형의 예측력이 적절하다고 판단

◦ ROC(Receiver Operation Characteristic)

- 민감도와 1-특이도를 활용하여 모형을 평가
- AUROC(ROC 커브 밑부분의 넓이)
  - AUROC=(AR+1)/2
  - 0.9~1.0: Excellent
  - 0.8~0.9: Good

        - 0.7~0.8: Fair

- **인공신경망**
    - 신경망의 연구
        - 인공신경망은 뇌를 기반으로 한 추론모델
        - 1943년 매킬럭(McCulloch)과 피츠(Pitts): 인간의 뇌를 수많은 신경세포가 연결되어 하나의 디지털 네트워크 모형으로 간주하고 신경세포의 신호처리 과정을 모형화하여 단순 패턴분류 모형을 개발
        - 헵(Hebb): 신경세포(뉴런) 사이의 연결강도(weight)를 조정하여 학습규칙을 개발
        - 로젠블럿(Rosenblatt, 1955): 퍼셉트론(Perceptron)이라는 인공세포 개발, 비선형성의 한계점 발생 - XOR(Exclusive OR) 문제
        - 홉필드(Hopfild), 러멜하트(Rumelhart), 맥클랜드(McClelland): 역전파 알고리즘(Backpropagation)을 활용하여 비선형성을 극복한 다계층 퍼셉트론으로 새로운 인공신경망 모형 등장
    - 뉴런
        - 인공신경망은 뉴런이라는 아주 단순하지만 복잡하게 연결된 프로세스로 이루어져있음
        - 뉴런은 가중치가 있는 링크들로 연결되어있으며, 뉴런은 여러개의 입력신호를 받아 하나의 출력신호를 생성
        - 뉴런은 전이함수(활성화함수, activation function)을 사용
            - 뉴런은 입력 신호의 가중치의 합을 계산하여 임계값과 비교
            - 가중치 합이 임계값보다 작으면 -1, 같거나 크면 +1 출력



- 신경망 모형 구축시 고려사항
  - 입력변수
    - 신경망 모형은 복잡성으로 인해 입력자료의 선택에 매우 민감
    - 범주형 변수
      - 각 범주의 빈도가 일정수준 이상이고 각 범주의 빈도가 일정할 때 활용
      - ex) 가변수화하여 적용(성별[남, 녀] → 남성[1,0], 여성[1,0])
    - 연속형 변수
      - 입력 값의 범위가 변수들 간에 큰 차이가 없을 때 활용
      - 분포가 대칭이 아니면 좋지 않은 결과를 도출
      - 변환 또는 범주화 활용
  - 가중치 초기값
    - 역전파 알고리즘의 경우, 초기값에 따라 결과가 많이0 달라져 초기값 선택이 매우 중요
    - 가중치가 0이면 시그모이드 함수는 선형이 되고, 신경망 모형도 선형이 됨
    - 초기값은 0 근처의 랜덤 값으로 선정하고 초기에는 선형모형에서 가중치가 증가하면서 비선형으로 변경됨
  - 예측값 선정
    - 비용함수 R(θ)는 비블록함수이고 여러 개의 국소 최소값들(local minima)을 가짐
    - 랜덤하게 선택된 여러 개의 초기값에 대한 신경망을 적합한 후 얻은 해들을 비교하여 가장 오차가 작은 것을 선택하여 최종 예측값을 얻거나 평균(또는 최빈값)을 구하여 최종 예측값으로 선정
    - 훈련 자료에 대해 배깅(bagging)을 적용하여 최종 예측치를 선정
  - 학습률
    - 상수값을 사용
    - 처음에는 큰 값으로 정하고 반복이 진행되어 해가 가까울수록 0에 수렴
  - 은닉층(hidden layer), 은닉 노드(hidden node)의 수
    - 은닉층과 은닉노드가 많으면 가중치가 많아져서 과대적합 문제 발생
    - 은닉층과 은닉노드가 적으면 과소적합 문제 발생

        - 은닉층 수 결정: 은닉층이 하나인 신경망은 범용근사자(universal approximator)이므로 가급적이면 하나로 선정
        - 은닉노드 수 결정: 적절히 큰 값으로 결정하고 가중치를 감소하면서 모수에 대한 벌점화 적용
    - 과대적합문제
        - 신경망은 많은 가중치를 추정해야 하므로 과대적합 문제가 발생
        - 해결방법
            - 조기종료(모형이 적합하는 과정에서 검증오차가 증가하기 시작하면 반복을 중지)
            - 선형모형의 능형회귀와 유사한 가중치 감소라는 벌점화 기법 활용

- 로지스틱 회귀분석
    - 반응변수가 범주형인 경우에 적용되는 회귀분석모형
    - 새로운 설명변수(또는 예측변수)가 주어질 때 반응변수의 각 범주(또는 집단)에 속할 확률이 얼마인지를 추정(예측모형)하여, 추정 확률을 기준치에 따라 분류하는 목적(분류모형)으로 활용
    - 사후확률(Posterior probability)
        - 모형의 적합을 통해 추정된 확률
        - $log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$

        $\pi(x) = P(Y=1|x),\ x=(x_1,...,x_k)$
    - $exp(\beta_1)$: 나머지 변수 $(x_1,...,x_k)$가 주어질 때 x_1이 한 단위 증가할 때마다 성공(Y=1)의 오즈가 몇 배 증가하는지를 나타내는 값
    - 오즈(odds): 성공할 확률이 실패할 확률의 몇 배인지를 나타내는 확률
        - $Odds(A) = \frac{P(A)}{1-P(A)}$
        - 오즈비(odds ratio)=$\frac{Odds(A)}{Odds(B)}$
    - glm() 함수를 활용하여 로지스틱 회귀분석을 실행함
        - 표현: glm(종속변수 ~ 독립변수1+...+독립변수k, family=binomial, data=데이터셋명)

        - 로지스틱 회귀분석의 결과, β의 추정값이 5.14이면, 독립변수의 단위가 증가함에 따라 종속변수가 0에서 1로 바뀔 오즈(Odds)가 exp(5.140)≈170배 증가한다는 의미

- 일반화 가중치(generalized weight)
    - R에서 neuralnet 함수의 실행 결과로 도출
    - 각 공변량의 영향을 표현
    - 로지스틱 회귀모형에서의 회귀 계수와 유사하게 해석됨

5.3 군집분석

- 군집분석
    - 각 객체(대상)의 유사성을 측정하여 유사성이 높은 대상집단을 분류하고, 군집에 속한 객체들의 유사성과 서로 다른 군집에 속한 객체간의 상이성을 규명하는 분석 방법
    - 특성에 따라 고객을 여러 개의 배타적인 집단으로 나누는 것으로 군집의 개수, 구조에 대한 가정 없이 데이터로부터 거리 기준으로 군집화 유도

- 군집분석의 특징
    - 비교사학습법(unsupervised learning)에 해당하여 타겟변수(종속변수)의 정의가 없이 학습 가능
    - 데이터를 분석의 목적에 따라 적절한 군집으로 분석자가 정의 가능
    - 요인분석과의 차이: 유사한 변수를 함께 묶어주는 목적이 아니라 각 데이터(객체)를 묶어줌
    - 판별분석과의 차이: 판별분석은 사전에 집단이 나누어져 있어야 하지만 군집분석은 집단이 없는 상태에서 집단을 구분

- 거리 측정 방법
    - **연속형 변수** → dist 함수에서 지원

        - 유클리드 거리
            - 공통으로 점수를 매긴 항목의 크기를 통해 판단하는 측도
            - 데이터 간 유사성을 측정할 때 많이 사용
            - 통계적 개념 내포되어있지 않음
            - $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$
        - 표준화 거리
            - 해당 변수의 표준편차로 척도 변환 후 유클리드 거리 계산
            - 척도 차이, 분산 차이로 인한 왜곡 방지 가능
        - 마할라노비스 거리
            - 변수의 표준화와 변수 간 상관성을 동시에 고려한 통계적 거리
        - 체비셔프 거리
        - 맨하탄 거리
            - 맨하탄 도시에서 건물에서 건물로 가기 위한 최단 거리를 구하기 위해 고안
            - $|x_1 - y_1| + |x_2 - y_2|$
        - 캔버라 거리
        - 민코우스키 거리
            - 맨하탄 거리와 유클리드 거리를 한번에 표현한 공식
    - **범주형 변수** → dist 함수에서 미지원
        - 자카드 거리
            - Boolean 속성으로 이루어진 두 객체 간 유사도 측정에 사용
        - 코사인 거리
        - 코사인 유사도
            - 두 개체의 백터 내적의 코사인 값을 이용해 측정된 벡터간의 유사도

- **계층적 군집분석**
    - n개의 군집으로 시작해 점차 군집의 개수를 줄여나가는 방법
    - 최단연결법(single linkage, nearest neighbor)

        - n*n 거리행렬에서 거리가 가장 가까운 데이터를 묶어서 군집을 형성
        - 군집과 군집 또는 데이터와의 거리를 계산시 최단거리(min)를 거리로 계산하여 거리행렬 수정
        - 수정된 거리행렬에서 거리가 가까운 데이터 또는 군집을 새로운 군집으로 형성
    - 최장연결법(complete linkage, farthest neighbor)
        - 군집과 군집 또는 데이터와의 거리를 계산시 최장거리(max)를 거리로 계산하여 거리행렬 수정
    - 평균연결법(average linkage)
        - 군집과 군집 또는 데이터와의 거리를 계산시 평균거리(mean)를 거리로 계산하여 거리행렬 수정
    - 와드연결법(ward linkage)
        - 군집내 편차들의 제곱합을 고려한 방법으로 군집 간 정보의 손실을 최소화하기 위해 군집화를 진행

- **비계층적 군집분석**
    - n개의 개체를 g개의 군집으로 나눌 수 있는 모든 가능한 방법을 점검해 최적화한 군집을 형성하는 것
    - K-평균 군집분석(k-means clustering)
        - 프로세스
            1. 원하는 군집의 개수와 초기 값(seed)을 정해 seed 중심으로 군집을 형성
                - 집단 내 제곱합 그래프를 이용해 군집 수 결정
            2. 각 데이터를 거리가 가장 가까운 seed가 있는 군집으로 분류
            3. 각 군집의 seed 값을 다시 계산
            4. 모든 개체가 군집으로 할당될 때까지 위 과정들을 반복
        - 장점
            - 주어진 데이터의 내부구조에 대한 사전정보 없이 의미있는 자료구조를 찾을 수 있음
            - 다양한 형태의 데이터에 적용이 가능
            - 분석방법 적용이 용이함

        - 단점
            - 가중치와 거리 정의가 어려움
            - 초기 군집수를 결정하기 어려움
            - 사전에 주어진 목적이 없으므로 결과 해석이 어려움

- **혼합 분포 군집**(mixture distribution clustering)
    - 모형 기반(model-based)의 군집방법
    - 데이터가 k개의 모수적 모형(흔히 정규분포 또는 다변량 정규분포를 가정함)의 가중함으로 표현되는 모집단 모형으로부터 나왔다는 가정하에서 모수와 함께 가중치를 자료로부터 추정하는 방법을 사용
    - K개의 각 모형은 군집을 의미하며, 각 데이터는 추정된 k개의 모형 중 어느 모형으로부터 나왔을 확률이 높은 지에 따라 군집의 분류가 이루어짐
    - 흔히 혼합모형에서의 모수와 가중치의 최대가능도 추정에는 EM 알고리즘이 사용
    - 혼합분포군집모형의 특징
        - K-평균 군집의 절차와 유사하지만 확률분포를 도입하여 군집을 수행
        - 군집을 몇 개의 모수로 표현할 수 있으며, 서로 다른 크기나 모양의 군집을 찾을 수 있음
        - EM알고리즘을 이용한 모수 추정에서 데이터가 커지면 수렴에 시간이 걸림
        - 군집의 크기가 너무 작으면 추정의 정도가 떨어지거나 어려움
        - K-평균 군집과 같이 이상치 자료에 민감하므로 사전에 조치가 필요

- SOM(Self-Organizing Map)
    - SOM(자가조직화지도) 알고리즘은 코호넨(Kohonen)에 의해 제시, 개발되었으며 코호넨 맵(Kohonen Maps)이라고도 알려짐
    - SOM은 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화
        - 이러한 형상화는 입력 변수의 위치 관계를 그대로 보존한다는 특징이 있다.
        - 실제 공간의 입력변수가 가까이 있으면 지도상에도 가까운 위치에 있게 됨
    - SOM의 특징

        - 고차원의 데이터를 저차원의 지도 형태로 형상화하기 때문에 시각적으로 이해가 쉬움
        - 입력변수의 위치 관계를 그대로 보존하기 때문에 실제 데이터가 유사하면 지도상에서 가깝게 표현되며, 이러한 특징 때문에 패턴 발견, 이미지 분석 등에서 뛰어난 성능을 보임
        - 역전파(Back Propagation) 알고리즘 등을 이용하는 인공신경망과 달리 단 하나의 전방 패스(feed-forward flow)를 사용함으로써 속도가 매우 빠르므로 실시간 학습처리를 할 수 있는 모형

- 밀도 기반 군집
    - 어느 점을 기준으로 주어진 반경 내에 최소 개수만큼의 데이터를 가질 수 있도록 함으로써 특정 밀도함수 혹은 밀도에 의해 군집을 형성해나가는 기법
    - 임의적인 모양의 군집 탐색에 가장 효과적인 방법
    - DBSCAN, OPTICS, DENCLUE 등

- 실루엣
    - 군집분석의 품질을 정량적으로 평가하는 대표적인 지표
    - 군집 내 데이터의 응집도와 군집간 분리도를 계산하여 군집 내 데이터의 거리가 짧을수록, 군집 간 거리가 멀수록 값이 커지며 완벽한 분리일 경우 1의 값을 가지는 지표

5.4 연관분석

- **연관분석**
    - 기업의 데이터베이스에서 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간 규칙을 발견하기 위한 분석
    - 장바구니 분석(Market Basket Analysis)
        - 장바구니에 무엇이 같이 들어있는지에 대해 분석

        - ex) 주말을 위해 목요일에 기저귀를 사러 온 30대 직장인 고객은 맥주도 함께 사감
    - 순차분석(Sequence Analysis)
        - 구매 이력을 분석해 A 품목을 산 수 추가로 B품목을 사는지 분석
        - ex) 휴대폰을 새로 구매한 고객은 한달 내에 휴대폰 케이스를 구매

- 연관분석의 형태
    - 조건과 반응의 형태(if-then)
    - (item set A) → (item set B)

      ⇒ if A then B: 만일 A가 일어나면 B도 일어난다

- 연관분석의 측도
    - 지지도(support)
        - 전체 거래 중 항목 A와 항목 B를 동시에 포함하는 거래의 비율로 정의
        - 지지도=$P(A \cap B) = \frac{A\text{와 } B\text{가 동시에 포함된 거래수}}{\text{전체거래수}}$
    - 신뢰도(confidence)
        - 항목 A를 포함한 거래 중에서 항목 A와 항목 B가 같이 포함될 확률
        - 연관성의 정도를 파악할 수 있음
        - 신뢰도=$\frac{P(A \cap B)}{P(A)} = \frac{A\text{와 } B\text{가 동시에 포함될 확률}}{A\text{가 포함될 확률}}$
    - 향상도(lift)
        - A가 주어지지 않았을 때의 품목 B의 확률에 비해 A가 주어졌을 때의 품목 B의 확률의 증가 비율
        - 연관규칙 A→B는 품목 A와 품목 B의 구매가 서로 관련이 없는 경우에 향상도가 1이 됨
        - 향상도=$\frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{A\text{와 } B\text{가 동시에 포함될 확률}}{A\text{가 포함될 확률} \times B\text{가 포함될 확률}}$

- 연관분석의 특징
    - 절차

1. 최소 지지도 선정(보통 5%)
2. 최소 지지도를 넘는 품목 분류
3. 2가지 품목 집합 생성
4. 반복 수행으로 빈발품목 집합 선정

- 장점
  - 탐색적인 기법
    - 조건 반응으로 표현되는 연관성분석 결과를 쉽게 이해할 수 있음
  - 강력한 비목적성 분석기법
    - 분석방향이나 목적이 특별히 없는 경우 목적변수가 없으므로 유용하게 활용됨
  - 사용이 편리한 분석 데이터의 형태
    - 거래내용에 대한 데이터를 변환 없이 그 자체로 이용
  - 계산의 용이성
    - 분석을 위한 계산이 상당히 간단
- 단점
  - 상당한 수의 계산과정
    - 품목 수가 증가하면 분석에 필요한 계산은 기하급수적으로 늘어남
  - 적절한 품목의 결정
    - 너무 세분화한 품목을 갖고 연관성 규칙을 찾으면 수많은 연관성 규칙들이 발견되겠지만, 실제로 발생 비율 면에서 의미 없는 분석이 될 수도 있음
  - 품목의 비율 차이
    - 사용될 모든 품목을 자체가 전체자료에서 동일한 빈도를 갖는 경우, 연관성 분석은 가장 좋은 결과를 얻음
    - 그러나 거래량이 적은 품목은 당연히 포함된 거래수가 적을 것이고 규칙 발견 과정중에서 제외되기 쉬움

- 평가기준 적용시 주의점
  - 두 항목의 신뢰도가 높다고 해서 두 항목이 높은 연관관계가 있는 것은 아님(지지도를 함께 고려)

        - 만일 두 항목의 신뢰도가 높게 나왔어도 전체 항목 중 두 항목의 동시 구매율인 지지도가 낮게 나온다면 두 항목간 연관성을 신뢰하기에는 부족한 점이 있음
        - 구매율 자체가 낮은 항목이기에 일반적인 상관관계로 보기엔 어려움
    - 지지도와 신뢰도가 모두 높게 나왔더라도 꼭 두 항목이 높은 연관관계가 있는 것은 아님(향상도를 함께 고려)
        - 일반적으로 빈번하게 구매되는 항목들에 대해서는 지지도와 신뢰도가 높게 나올 수 있음
    - A,B 두 항목의 신뢰도(Confidence(A→B))가 높게 나왔을 때, 전체 거래에서 B의 자체 구매율보다 A의 자체 구매율이 더 높아야 의미있는 정보임

- Apriori 알고리즘
    - 어떤 항목집합이 빈발한다면, 그 항목집합의 모든 부분집합도 빈발
    - 최소지지도보다 큰 빈발항목집합에서 높은 측도(신뢰도, 향상도) 값을 갖는 연관규칙을 구하는 방법
    - 예
        - {우유, 빵, 과자}가 빈발 항목 집합이면, 부분집합인 {우유, 빵}, {빵, 과자}도 빈발항목집합 지지도의 anti-monotone 성질: 어떤 항목집합의 지지도는 그 부분집합들의 지지도를 넘을 수 있음