

# 제 1 과목 - 데이터 이해 요약

## 데이터 유형

### 1. 데이터의 정의

- 데이터는 개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실(fact)
- 추론, 예측, 전망, 추정을 위한 근거(basis)로 기능하는 특성을 갖음
- 다른 객체와의 상호관계 속에서 가치를 찾음

### 2. 데이터의 유형

#### 1) 정성적 데이터(qualitative data)

- 자료의 성질, 특징을 자세히 풀어 쓰는 방식
- 언어, 문자로 기술 (예 : 설문조사의 주관식 응답, SNS 에 올린 글, 기상특보)
- 비정형 데이터 형태로 저장, 분석에 시간과 비용이 필요함

#### 2) 정량적 데이터(quantitative data)

- 수치, 기호, 도형으로 표시 (예 : 지역별 온도, 풍속, 강우량)
- 데이터 양이 증가하더라도 저장, 분석이 용이

## 암묵지 vs 형식지

### 1. 암묵지 vs 형식지

가장 널리 알려진 지식의 차원은 polanyi 에 의해 구분된 "암묵지와 형식지"

#### 1) 암묵지

- 학습과 체험을 통해 개인에게 습득 (현장 작업과 같은 경험을 통해 습득)
- 시행착오와 오랜 경험을 통해 개인에게 습득된 무형 지식
- 예) 김장김치 담그기, 자전거 타기
- 공유되기 어려움

#### 2) 형식지

- 교과서, 매뉴얼, 비디오, DB 등으로 형상화 된 지식을 의미
- 예) 회계 재무 관련 대차대조표에 요구되는 지식의 매뉴얼
- 외부로 표출되어 여러사람이 공유할 수 있는 지식

### 2. 지식경영이란?

- 개인의 암묵지와 집단에서의 형식지가 나선형의 형태로 회전하면서 생성, 발전, 전환되는 지식의 발전을 기반으로 한 기업의 경영

### 3. 암묵지, 형식지의 4 단계 지식전환 모드

- 1 단계 : 공통화(암-암) : 암묵적 지식을 다른 사람에게 알려주는 것
- 2 단계 : 표출화(암-형) : 암묵적 지식 노하우를 책이나 교본 등 형식지로 만드는 것
- 3 단계 : 연결화(형-형) : 책이나 교본(형식지)에 자신이 알고 있는 새로운 지식(형식지)를 추가하는 것
- 4 단계 : 내면화(형-암) : 만들어진 책이나 교본(형식지)를 보고 다른 직원이 암묵적지식(노하우)를 습득

## 데이터와 정보의 관계

### 1. 계층구조

- Data -> Information -> Knowledge -> Wisdom
- 데이터를 가공 처리하여 얻을 수 있는 것 : 정보 / 지식 / 지혜

#### 1) 데이터(Data)

타 데이터와의 상관관계가 없는 가공하기 전의 순수한 수치나 기호

ex. A 마트는 100 원에, B 마트는 200 원에 연필을 판매한다

#### 2) 정보(Information)

데이터의 가공 및 상관/연관 관계속에서 의미가 도출된 것

ex. A 마트의 연필이 더 싸다

#### 3) 지식(Knowledge)

상호연결된 정보패턴을 이해하여 이를 토대로 예측한 결과물

ex. 상대적으로 저렴한 A 마트에서 연필을 사야겠다

#### 4) 지혜(Wisdom)

근본 원리에 대한 깊은 이해를 바탕으로 도출되는 아이디어

ex. A 마트의 다른 상품들도 B 마트보다 쌀 것이라고 판단한다

## 데이터베이스

### 1. 특징

- 통합데이터(Integrated) : 데이터베이스에 같은 내용의 데이터가 중복되어 있지 않다는 것을 의미
- 저장데이터(Stored) : 자기디스크나 자기테이프 등과 같이 컴퓨터가 접근할 수 있는 저장매체에 저장되는 것을 의미
- 공유데이터(Shared) : 여러 사용자에게 서로 다른 목적으로 데이터베이스의 데이터를 공동으로 이용되는 것을 의미
- 변화되는 데이터(Changed) : 새로운 데이터의 추가, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 정확한 데이터를 유지해야한다

### 2. DBMS, RDBMS, ODBMS

#### 1) DBMS

- 사용자와 데이터베이스 사이에서 사용자의 요구에 따라 정보를 처리해주고 데이터베이스를 관리해주는 소프트웨어

#### 2) RDBMS

- 관계형 데이터베이스 관리 시스템
- 정형화된 테이블로 구성된 데이터 항목들의 집합체

- MySQL(오픈소스 RDBMS), PL/SQL(상용 RDBMS)
- SQL : RDBMS 의 데이터를 관리하기 위해 설계된 특수 목적의 프로그래밍 언어

### 3) ODBMS

- 객체지향 데이터베이스 관리 시스템
- 객체들을 생성하여 계층에서 체계적으로 정리하고, 다시 계층들을 하위계층이 상위계층으로부터 속성과 방법들을 물려받을 수 있는 DBMS
- 복잡한 데이터 구조를 표현 및 관리하는 DBMS

### 3. 데이터베이스 설계 절차

- 1) 요구조건 분석 : 데이터베이스 사용자, 사용목적, 사용범위, 제약조건 등을 정리, 명세서 작성
- 2) 개념적 설계 : DBMS 독립적인 E-R 모델 작성, 정보를 추상적 개념으로 표현하는 과정
- 3) 논리적 설계 : 자료를 컴퓨터가 이해할 수 있도록 특정 DBMS 의 논리적 자료 구조로 변환
- 4) 물리적 설계 : 논리적 구조로 표현된 데이터를 물리적 구조의 데이터로 변환하는 과정

### 4. NoSQL

#### 1) 개념

- 관계형 데이터베이스보다 덜 제한적인 일관성 모델을 이용하는 데이터의 저장 및 검색을 위한 매커니즘 제공, 디자인 단순화, 수평적 확장성, 세세한 통제 등을 포함
- 기존의 RDBMS 가 갖고있는 특성 뿐만 아니라 다른 특성들을 부가적으로 지원

#### 2) 저장방식 도구 : MongoDB, Apache HBase, Redis

##### - Mongo DB

데이터 교환 시 비산(BSON : Binary JSON) 문서 형태로 저장  
여러 서버에 분산 저장 및 확장이 용이하며  
방대한 데이터 처리가 빠름

C++로 작성됨

##### - Apache HBase

하둡 플랫폼을 위한 공개 비관계형 분산 데이터 베이스  
구글의 빅테이블(BigTable)을 본보기로 삼음  
자바로 작성됨

##### - Redis(Remote Dictionary Server)

"카-값" 구조의 비정형데이터를 저장하고 관리하기 위한 오픈소스기반의 비관계형 데이터베이스 관리 시스템

## 시대별 기업내부의 데이터베이스 솔루션

### 1. 1980년대 : OLTP, OLAP

#### 1) OLTP(On-Line Transaction Processing, 온라인 거래 처리)

- 예시 : 상품주문, 회원정보 수정
- 주 컴퓨터와 통신회선으로 접속되어 있는 복수의 사용자 단말에서 발생한 트랜잭션을 주 컴퓨터에서 처리하여 그 결과를 사용자에게 되돌려주는 처리형태

#### 2) OLAP(On-Line Analytical Processing, 온라인 분석 처리)

- 예시 : 10년간 A사의 직급별 임금 상승률
- 다차원으로 이루어진 데이터로부터 통계적인 요약정보를 제공할 수 있는 기술
- 다차원의 데이터를 대화식으로 분석하기 위한 SW

### 2. 2000년대 : CRM, SCM

#### 1) CRM(Customer Relationship Management)

- 고객별 구매 이력 데이터베이스를 분석하여 고객에 대한 이해를 돕고 이를 바탕으로 각종 마케팅 전략을 통해 보다 높은 이익을 창출할 수 있는 솔루션

#### 2) SCM(Supply Chain Management)

- 제조, 물류, 유통업체 등 유통공급망에 참여하는 모든 업체들이 협력을 바탕으로 정보기술(Information Technology)를 활용, 재고를 최적화하기 위한 솔루션
- 기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최소화 시키기 위한 것
- 자재구매 데이터, 생산, 재고 데이터, 유통/판매 데이터, 고객데이터로 구성됨

## 분야별 기업 내부 데이터베이스 솔루션 - 제조부문

### 1. Data warehouse

#### 1) 개념

- 기업 내의 의사결정 지원 애플리케이션을 위한 정보를 제공하는 하나의 통합된 데이터 저장 공간
- ETL : 추출(Extract), 변환(transform), 적재(load)

주기적으로 내부 및 외부 데이터베이스로부터 정보를 추출하고 정해진 규약에 따라 정보를 변환한 후에 정보를 적재함

- 데이터들은 시간적 흐름에 따라 변화하는 값 유지

#### 2) 특징

- 데이터의 통합 : 데이터들은 전사적 차원에서 일관된 형식으로 정의됨
- 데이터의 시계열성 : 관리되는 데이터들은 시간의 흐름에 따라 변화하는 값을 저장함
- 데이터 주제 지향적, 비소멸성(비휘발성) : 특정 주제에 따라 데이터들이 분류, 저장, 관리 됨

### 2. Data Mart

- 전사적으로 구축된 데이터 웨어하우스로부터 특정 주제, 부서 중심으로 구축된 소규모 단일 주제의 데이터웨어 하우스

- 재무, 생산, 운영과 같이 특정 조직의 특정 업무 분야에 초점을 두고 있음

### 3. ERP(Enterprise Resource Planning)

- 제조업을 포함한 다양한 비즈니스 분야에서 생산, 구매, 재고, 주문, 공급자와의 거래, 고객서비스 제공 등 주요 프로세스 관리를 돕는 여러 모듈로 구성된 통합 애플리케이션 소프트웨어 패키지

### 4. BI(Business Intelligence)

- 기업의 Data warehouse 에 저장된 데이터에 접근해 경영의사결정에 필요한 정보를 획득하고 이를 경영활동에 활용하는 것
- 데이터를 통합/분석하여 기업 활동에 연관된 의사결정을 돕는 프로세스를 말함
- 가트너는 '여러 곳에 산재하여 있는 데이터를 수집하여 체계적이고 일목요연하게 정리함으로써 사용자가 필요로 하는 정보를 정확한 시간에 제공할 수 있는 환경'으로 정의함
- 하나의 특정 비즈니스 질문에 답변하도록 설계

### 5. ad hoc report

- BI와 빅데이터 분석의 차이점을 표현한 키워드
- Optimization, forecast, insight : 빅데이터 분석 관련 키워드

### 6. BA(Business Analytics)

- 경영 의사결정을 위한 통계적이고 수학적인 분석에 초점을 둔 기법
- 성과에 대한 이해와 비즈니스 통찰력에 초점을 둔 분석방법
- 사전에 예측하고 최적화하기 위한 것으로 BI 보다 진보된 형태

## 분야별 기업 내부 데이터베이스 솔루션 - 금융부문

### 1. 블록체인(Block chain)

- 기존 금융회사의 중앙 집중형 서버에 거래 기록을 보관하는 방식에서 벗어나 거래에 참여하는 모든 사용자에게 거래 내용을 보내주며 거래 때마다 이를 대조하는 데이터 위조 방지 기술

## 분야별 기업 내부 데이터베이스 솔루션 - 유통부문

### 1. KMS(Knowledge Management System)

- 지식관리시스템의 약자, 조직 내의 지식을 체계적으로 관리하는 시스템을 의미

### 2. RFID

- 무선주파수(RF, Radio Frequency)를 이용하여 대상을 식별할 수 있는 기술
- RF 태그에 사용목적으로 알맞은 정보를 저장하여 적용대상에 부착한 후 판독기에 해당되는 RFID 리더를 통해 정보를 인식

## 빅데이터란?

### 1. 빅데이터 정의

- 빅데이터는 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터이다
- 빅데이터는 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처이다
- 데이터의 양(Volume) 데이터 유형과 소스측면의 다양성(Variety), 데이터 수집과 처리 측면에서 속도(Velocity)가 급격히 증가하면서 나타난 현상이다

### 2. 빅데이터

- 4V(ROI, Return on investment, 투자자본수익률 관점에서 보는 빅데이터)

Volume : 데이터의 크기, 생성되는 모든 데이터를 수집

Variety : 데이터의 다양성, 정형화된 데이터를 넘어 텍스트, 오디오, 비디오 등 모든 유형의 데이터를 대상으로 함

Velocity : 데이터의 속도, 사용자가 원하는 시간 내 데이터 분석 결과 제공, 업데이트 속도 빠름

Value : '비즈니스 효과 요소' (Volume, Variety, Velocity 는 '투자비용 요소')

### 3. 빅데이터의 출현 배경

- 산업계에서 일어난 변화를 보면 빅데이터의 현상은 양질 전환법칙으로 설명할 수 있다
- 학계에서도 빅데이터를 다루는 현상들이 늘어남. 대표적 사례로 인간게놈프로젝트
- 디지털화, 저장기술, 인터넷 보급, 모바일 혁명, 클라우드 컴퓨팅 등 관련 기술 발전과 관련이 있다
- \* 클라우드 컴퓨팅 : 빅데이터 분석에 경제적 효과를 제공해준 결정적 기술
- 소셜 미디어, 영상 등 비정형 데이터의 확산
- 데이터 처리 기술 발전

### 4. 빅데이터는 "석탄/철, 원유, 렌즈, 플랫폼" 이다!

- 석탄, 철 : 서비스 분야의 생산성을 획기적으로 끌어올려 혁명적 변화를 가져옴
- 원유 : 생산성 향상
- 렌즈 : 구글 'Ngram Viewer'를 통해 책을 디지털화
- 플랫폼

비즈니스 측면에서는 '공동 활용의 목적으로 구축된 유/무형의 구조물'을 의미함

페이스북은 SNS 서비스로 시작했지만, 자신들의 소셜그래프 자산을 외부 개발자들에게 공개하고 서드-파티 개발자들이 페이스북 위에서 작동하는 앱을 만들기 시작

각종 사용자 데이터나 M2M 센서 등에서 수집된 데이터를 가공, 처리, 저장해두고 이 데이터에 접근할 수 있도록 API를 공개하였다

#### 5. 빅데이터의 가치선정이 어려운 이유

- 데이터의 활용 방식 : 재사용이나 재조합, 다목적용 데이터 개발 등이 일반화되면서 특정 데이터를 언제, 어디서, 누가 활용할지 알 수 없다

- 새로운 가치 창출 : 데이터가 기존에 없던 가치를 창출함에 따라 그 가치를 측정하기 어렵다

- 분석기술의 발달 : 분석기술의 발달로 지금은 가치없는 데이터도 새로운 분석기술의 등장으로 거대한 가치를 만들어내는 재료가 될 가능성이 있다

#### 6. 빅데이터가 만들어내는 본질적인 변화

1) 사전처리 : 표준화된 문서포맷

표본조사 > 질(Quality) > 인과관계

2) 사후처리 : 데이터를 모은 뒤 그 안에서 숨은 정보를 찾아냄

전수조사 > 양(Quantity) > 상관관계

## 빅데이터 활용기법

#### 1. 연관규칙학습(Association rule Learning)

변수간 주목할만한 상관관계가 있는지 찾아내는 방법

우유구매자가 기저귀도 같이 구매하는가?

커피를 사는 사람들이 탄산음료도 많이 구매하는가?

#### 2. 유형분석(Classification tree Analysis)

사용자는 어떤 특성을 가진 집단에 속하는가?와 같은 문제 해결에 사용

문서를 분류하거나 조직을 그룹으로 나눌 때, 온라인 수강생들을 특성에 따라 분류할때 사용함

#### 3. 유전 알고리즘(Generic Algorithms)

최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 점진적으로 진화시켜나가는 방법

최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?

응급실에서 의사를 어떻게 배치하는 것이 가장 효율적인가

#### 4. 기계학습

훈련 데이터로부터 패턴을 학습해 '예측'하는 일에 활용

기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화 중 어떤것을 가장 보고 싶어할까? (넷플릭스 추천 시스템)

#### 5. 회귀분석

선형함수로 나타낼수 있는 수치 데이터 분석

사용자의 만족도가 충성도에 어떤 영향을 미치는가?

#### 6. 감정분석

특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석함

소셜 미디어에 나타난 의견을 바탕으로 고객이 원하는 것을 찾아낼 때 활용함

호텔에서 고객의 논평을 받아 서비스를 개선하기 위해 활용

#### 7. 소셜 네트워크 분석

사회관계망분석(SNA)

영향력 있는 사람을 찾아낼 수 있으면, 사람들 간 소셜관계를 파악할 수 있다

## 빅데이터 위기요인과 통제방안

빅데이터 위기요인의 종류에는 사생활 침해, 책임원칙의 훼손, 데이터 오용이 있다

#### 1. 사생활 침해

1) 위기요인

우리를 둘러싼 정보 수집 센서들의 수가 점점 늘어나고 있고, 특정 데이터가 본래 목적 외에 가공 처리돼 2차 3차적 목적으로 활용될 가능성이 증가

익명화(Anonymization) : 사생활 침해를 방지하기 위해 데이터에 포함된 개인 식별 정보를 삭제하거나 알아볼 수 없는 형태로 변환하는 것

2) 통제방안

동의제에서 책임제로 전환

개인정보의 활용에 대한 개인이 매번 동의하는 것은 경제적으로도 매우 비효율적임

사생활침해 문제를 개인정보 제공자의 동의를 통해 해결하기 보다는 개인정보사용자에게 책임을 지움으로써 개인정보 사용 주체가 보다 적극적인 보호장치를 강구하게 하는 효과가 발생할 것으로 기대됨

## 2. 책임원칙의 훼손

### 1) 위기요인

빅데이터 기반분석과 예측기술이 발전하면서 정확도가 증가한 만큼, 분석 대상이 되는 사람들은 예측 알고리즘의 희생양이 될 가능성이 증가함

그러나 잠재적 위험 사항에 대해서도 책임을 추궁하는 사회로 변질될 가능성이 높아 민주주의 사회원칙을 크게 훼손할 수 있다

예시 : 범죄 예측 프로그램을 통해 범죄 전 체포

### 2) 통제방안

기존의 책임원칙을 강화할 수 밖에 없다

## 3. 데이터 오용

### 1) 위기요인

빅데이터는 일어난 일에 대한 데이터에 의존함

그것을 바탕으로 미래를 예측하는 것은 적지않은 정확도를 가질 수 있지만, 항상 맞을 수는 없음

주어진 데이터에 잘못된 인사이트를 얻어 비즈니스에 직접 손실을 불러올 수 있음

### 2) 통제방안

데이터 알고리즘에 대한 접근권 허용 및 객관적 인증방안을 도입 필요성 제기

\* 알고리즘미스트 : 데이터 분석 알고리즘으로 부당한 피해를 보는 사람을 방지하기 위해 / 피해를 입은 사람을 구제하는 전문가

## 개인정보 비식별화 기법

### 1. 데이터 마스킹(Masking)

다양한 유형의 데이터 관리 시스템에 저장된 정보를 보호하는 데 사용되는 프로세스 (카드 뒤 4 자리 숨기기, 주민번호 뒤 6 자리 숨기기)

### 2. 데이터 범주화

변수가 가질 수 있는 가능한 값들을 몇 개의 구간으로 범주화

홍길동, 35 세 -> 홍씨, 30~40 세

### 3. 가명

개인식별 정보를 삭제, 알아볼 수 없는 형태로 변환

홍길동, 국제대 재학 -> 임격정, 한성대 재학

### 4. 값을 첨가

자료 값에 값을 추가하거나 곱해 원래 자료에 약간의 변형을 가하여 공개

### 5. 총계 처리 / 평균값 대체

데이터의 총합 값을 보임으로 개별 데이터의 값이 보이지 않도록 함

### 6. 데이터값 삭제

데이터 셋의 값 중 필요 없는 값 또는 개인 식별에 중요한 값 삭제

## 빅데이터 분석

### 1) 개념

- 빅데이터 분석은 'Big'이 핵심이 아님

유형의 다양성과 관련이 있음

음성, 텍스트, 이미지, 비디오 등과 같이 다양한 정보 원천의 활용이 핵심

- 전략적 통찰이 없는 분석의 함정

한국의 경영 문화는 여전히 분석을 국소적인 문제 해결 용도로 사용하는 단계

기업의 핵심가치와 관련해 전략적 통찰력을 가져다 주는 데이터 분석을 내재화하는 것이 어려움

- 일차적인 분석 vs 전략도출을 위한 가치기반 분석

일차적인 분석을 통해서도 부서나 업무영역에서 상당한 효과를 얻을 수 있음 / 일차적 분석 경험이 증가하고 분석의 활용 범위를 더 넓고 전략적으로 변화시켜야함

### 2) 사례

- 금융 서비스 : 신용점수 산정, 사기 탐지, 고객 수익성 분석

- 소매업 : 재고보충, 수요예측

- 제조업 : 맞춤형 상품 개발, 신상품 개발

- 에너지 : 트레이딩, 공급, 수요예측

- 온라인 : 웹 매트릭스, 사이트 설계, 고객 추천

## 데이터 사이언스의 정의

### 1. 정의

- 데이터로부터 의미있는 정보를 추출해내는 학문

- 정형, 반정형, 비정형의 다양한 유형의 데이터를 대상으로 함
- 분석 뿐만 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함한 포괄적 개념
- 데이터 공학, 수학, 통계학, 컴퓨터공학, 시각화, 해커의 사고방식, 해당 분야의 전문 지식을 종합한 학문 -> 총체적(holistic) 접근법을 사용함
- 과학과 인문학의 교차로에 서 있다고 할 수 있음 -> 스토리 텔링, 커뮤니케이션, 창의력, 직관력 필요

## 2. 데이터 사이언스의 핵심 구성요소

- IT(Data Management)
- 분석
- 비즈니스 컨설팅

## 3. 다른 학문과의 차이점

	데이터 사이언스	통계학	데이터 마이닝
분석 대상	정형, 비정형, 반정형 등 다양한 데이터 유형	정형화된 데이터	
분석 방법	분석 + 시각화 + 전달을 포함한 포괄적 개념		분석에 초점
학문 접근	종합적 학문 또는 총체적 접근법		

## 4. 데이터 사이언티스트의 역량

- 가트너(Gartner)가 본 데이터 사이언티스트의 역량
- 데이터 관리, 분석 모델링, 비즈니스 분석, 소프트 스킬 => 공통점은 호기심에서 시작
- 데이터 해커, 애널리스트, 커뮤니케이션, 신뢰받는 어드바이저 등의 조합이라 할 수 있다
- 하드 스킬과 소프트 스킬 능력을 동시에 갖추고 있어야 한다
- 데이터 처리 기술 이외에 사고 방식, 비즈니스 이슈에 대한 감각, 고객들에 대한 공감능력이 필요

## 5. 데이터 사이언티스트가 갖춰야하는 스킬

### 1) 하드 스킬

- Machine Learning, Modeling, Data Technical Skill
- 빅데이터에 대한 이론적 지식 : 관련 기법에 대한 이해와 방법론 습득
- 분석기술에 대한 숙련 : 최적의 분석 설계 및 노하우 축적

### 2) 소프트스킬

- 통찰력 있는 분석 : 창의적 사고, 호기심, 논리적 비판
- 설득력 있는 전달 : Storytelling, Visualization
- 다분야 간 협력 : Communication

## 6. 데이터 사이언티스트가 효과적 분석모델 개발을 위해 고려해야 하는 상황

- 분석 모델이 예측할 수 없는 위험을 살피기 위해 현실 세계를 돌아보고 분석을 경험과 세상에 대한 통찰력과 함께 활용한다
- 가정들과 현실의 불일치에 대해 끊임없이 고찰하고 모델의 능력에 대해 항상 의구심을 갖는다
- 분석의 객관성에 의문을 제기하고 분석 모델에 포함된 가정과 해석의 개입 등의 한계를 고려한다
- 모델범위의 바깥요인은 판단하지 않는다

## 7. 데이터 사이언티스트에 요구되는 인문학적 사고의 특성과 역할

	과거	현재	미래
정보 (information)	무슨 일이 일어났는가? 예) 리포팅(보고서)	무슨 일이 일어나고 있는가? 예) 경고	무슨 일이 일어날 것인가? 예) 추출
통찰력 (insight)	어떻게 왜 일어났는가? 예) 모델링, 실험설계	차선 행동은 무엇인가? 예) 권고	최악, 최선의 상황은? 예) 예측, 최적화

## 최근의 사회경제적 환경의 변화

### 1. 최근의 사회경제적 환경의 변화(인문학 열풍의 이유)

- 단순세계에서 복잡한 세계로의 변화 : 다양성과 각 사회의 정체성, 연결성, 창조성 키워드 대두
- 비즈니스의 중심이 제품생산에서 서비스로 이동 : 고객에게 얼마나 뛰어난 서비스를 제공 여부가 관건
- 경제와 산업의 논리가 생산에서 시장창조로 바뀜 : 무형자산이 중요

### 2. 데이터 기반 분석의 상관관계, 통계적 분석의 인과관계

- 신속한 의사결정을 원하는 비즈니스에서는 실시간 '상관관계' 분석에서 도출된 인사이트를 바탕으로 수익을 창출할 수 있는 기회가 점점 늘어나고 있음
- '상관관계'를 통해 특정 현상의 발생가능성이 포착되고, 그에 상응하는 행동을 하도록 추천되는 일이 늘어날 것
- 데이터 기반의 '상관관계' 분석이 주는 인사이트가 '인과관계'에 의한 미래 예측을 점점 더 압도해 나가는 시대가 도래하고 있음

### 3. 의사결정 오류

- 1) 논리(노직) 오류 : 부정확한 가정을 하고 테스트를 하지 않는 것
- 2) 프로세스 오류 : 결정에서 분석과 통찰력을 고려하지 않은 것, 데이터 수집이나 분석이 너무 늦어 사용할 수 없게 되는 것, 대안을 진지하게 고려하지 않은 것

### 4. 가치 패러다임의 변화

- Digitalization, Connection, Agency

### 5. 데이터 사이언스의 한계와 인문학

- 모든 분석은 가정에 근거함 -> 잘못된 분석은 안 좋은 결과를 가져올 수 있음
- 모델의 능력에 대해 항상 의구심을 갖고, 가정과 현실의 불일치에 대해 계속 고찰하고, 분석 모델이 예측할 수 없는 위험을 살펴야함

## 2 과목. 데이터 분석 기획

## 분석 기획이란?

- 분석을 수행할 과제의 정의 및 의도했던 결과를 도출할 수 있도록 이를 적절하게 관리할 수 있는 방안을 사전에 계획하는 작업
- 어떤 목표(what)을 달성하기 위해 어떤 데이터를 가지고 어떤 방식(how)을 수행할 지에 대한 일련의 계획을 수립하는 작업
- 성공적인 분석 결과 도출을 위한 중요사전작업
- 해당 문제 영역에 대한 전문성 역량 및 통계학적 지식을 활용한 분석 역량과 분석도구인 데이터 및 프로그래밍 기술 역량에 대한 균형 잡힌 시각을 가지고 방향성 및 계획을 수립해야함

## 분석 주제 유형 4 가지

분석의 대상(what), 분석의 방법(how)에 따라 4 가지로 구분

		분석대상 (what)	
분석방법 (how)		Known	Un-Known
	Known	최적화(Optimization)	통찰(Insight)
	Un-Known	솔루션(Solution)	발견(Discovery)

- Optimization : 분석 대상 및 분석 방법을 이해하고 현 문제를 최적화의 형태로 수행함
- Solution : 분석과제는 수행되고, 분석방법을 알지 못하는 경우 솔루션을 찾는 방식으로 분석 과제를 수행함

- Insight : 분석 대상이 불분명하고, 분석방법을 알고 있는 경우 인사이트 도출
- Discovery : 분석 대상, 방법을 모른다면 발견을 통해 분석대상자체를 새롭게 도출

## 목표 시점 별 분석기획 방안

과제 중심적인 접근방식의 단기방안, 마스터플랜 단위의 중장기 방안으로 구분

	과제 단위 당면한 분석 주제의 해결	마스터플랜 단위 지속적 분석 문화 내재화
1차 목표	Speed & Test	Accuracy & Deploy
과제의 유형	Quick - Win	Long Term View
접근 방식	Problem Solving	Problem Definition

→ 두가지를 융합적으로 적용하는 것이 바람직함

\* Quick win

즉각적인 실행을 통한 성과 도출

프로세스 진행 과정에서 일반적인 상식과 경험으로 원인이 명백한 경우 바로 개선함으로써 과제를 단기로 달성하고, 추진하는 과정

## 분석 기획 시 고려사항

가용한 데이터, 적절한 유스케이스 탐색, 장애요소들에 대한 사전 계획 수립

### 1. 가용한 데이터(available data)

분석을 위한 데이터 확보

데이터 유형에 따라 적용가능한 solution 및 분석 방법이 다름

데이터의 유형분석이 선행적으로 이루어져야함 (정형, 비정형, 반정형)

### 2. 적절한 유스케이스 탐색(Proper Use-Case)

유사분석 시나리오 및 솔루션이 있다면 이것을 최대한 활용함

### 3. 장애요소들에 대한 사전 계획 수립(Low Barrier of Execution)

장애요소들에 대한 사전 계획 수립 필요

일회성 분석으로 그치지 않고 조직역량을 내재화 하기 위해서는 충분하고 지속적인 교육 및 활용방안 등의 변화관리가 고려되어야함

## 데이터 유형, 저장방식

### 1. 데이터 유형

1) 정형 데이터 : ERP, CRM Transaction data, Demand Forecast

2) 반정형 데이터 : Competitor Pricing, Sensor, machine data

3) 비정형 데이터 : e-mail, SNS, voice, IoT, 보고서, news

### 2. 데이터 저장 방식

1) RDB : 관계형 데이터를 저장, 수정, 관리 (ex. Oracle, MSSQL, MySQL)

2) NoSQL : 비관계형 데이터 저장소. (ex. MongoDB, Cassandra, HBase, Redis)

3) 분산파일시스템 : 분산된 서버의 디스크에 파일 저장 (ex. HDFS)

## 분석 방법론

### 1. 분석 방법론의 필요

- 데이터 분석을 효과적으로 기업에 정착하기 위해 데이터 분석을 체계화 하는 절차와 방법이 정리된 데이터 분석 방법론 수립이 필요

### 2. 분석방법론의 구성요소

- 상세한 절차, 방법, 도구와 기법, 템플릿과 산출물

### 3. 기업의 합리적 의사결정 장애요소

- 고정관념, 편향된 생각, 프레임링 효과(Framing Effect)

\* Framing Effect : 동일한 사건이나 상황임에도 불구하고 사람들의 선택이나 판단이 달라지는 현상으로, 특정 사안을 어떤 시각으로 바라보느냐에 따라 해석이 달라진다는 이론

### 4. 분석방법론의 모델 3 가지

1) 폭포수 모델 : 순차적, 하향식 진행, 문제점이 발견되면 전단계로 돌아가는 피드백 수행

2) 나선형 모델 : 반복을 통해 점진적으로 개발, 반복에 대한 체계가 효과적으로 갖춰지지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있음

3) 프로토타입 모델

사용자 요구사항이나 데이터를 규정하기 어렵고 데이터 소스도 명확히 파악하기 어려운 상황에서 사용

일단 분석을 시도해보고 그 결과를 확인해가면서 반복적으로 개선해나가는 방법

상향식 접근방법에 사용



# KDD(Knowledge Discovery in Database) 분석 방법론

데이터베이스에서 의미있는 지식을 탐색하는 데이터 마이닝 프로세스

분석 대상의 비즈니스 도메인에 대한 이해와 프로젝트 목표를 정확하게 설정

## 1. KDD 과정

1) 데이터 선택

2) 데이터 전처리 : 데이터셋에 포함되어 있는 잡음(Noise), 이상값(Outlier), 결측치(Missing Value)를 식별하고 필요시 제거

3) 데이터 변환 : 분석 목적에 맞는 변수 선택, 데이터의 차원 축소. 데이터 마이닝을 효율적으로 적용할 수 있도록 데이터셋 변경 작업

4) 데이터 마이닝 : 분석 목적에 맞는 데이터 마이닝 기법 및 알고리즘 선택, 데이터의 패턴을 찾거나 분류 또는 예측 등의 마이닝 작업 수행

5) 데이터 마이닝 결과 평가 : Interpretation / Evaluation, 분석 결과에 대한 해석과 평가, 활용

## CRISP-DM 분석 방법론

6 단계로 구성, 일방향으로 구성되어있지 않고 단계간 피드백을 통하여 단계별 완성도를 높이게 구성됨

### 1. CRISP-DM(Cross-Industry Standard Process for Data Mining)

- 6 단계 : 업무이해 - 데이터 이해 - 데이터 준비 - 모델링 - 평가 - 전개

1) 업무이해(Business Understanding)

비즈니스 관점 프로젝트의 목적과 요구사항 이해

도메인 지식을 데이터 분석을 위한 문제 정의로 변경하고 초기 프로젝트 계획을 수립하는 단계

업무목적 파악 -> 상황 파악 -> 데이터 마이닝 목표 설정 -> 프로젝트 계획 수립

2) 데이터 이해(Data Understanding)

분석을 위한 데이터 수집, 데이터 속성 이해를 위한 과정

데이터 품질에 대한 문제점 식별 및 숨겨져 있는 인사이트를 발견하는 단계

초기 데이터 수집, 데이터 기술 분석, 데이터 탐색, 데이터 품질 확인

3) 데이터 준비(Data Preparation)

KDD의 데이터 전처리 == CRISP-DM 분석 방법론의 데이터 준비

분석을 위해 수집된 데이터에서 분석기법에 적합한 데이터셋을 편성하는 단계

많은 시간소요

분석용 데이터셋 선택, 데이터 정제, 데이터 통합, 데이터 + 포매팅

4) 모델링(Modeling)

다양한 모델링 기법과 알고리즘을 선택

모델링 과정에서 사용되는 파라미터를 최적화해 나가는 단계

모델링 단계를 통해 찾아낸 모델은 테스트용 프로세스와 데이터셋으로 평가하여 모델 과적합(Overfitting) 등의 문제를 발견하고 대응 방안 마련

데이터 분석 방법론, 머신러닝을 이용한 수행 모델을 만들거나 데이터를 분할하는 부분

모델링 기법 선택, 모델링 작성, 모델 평가

\* 과적합(Overfitting) : 기계학습에서 학습 데이터를 과하게 학습하려는 것을 말함. 학습데이터는 실제 데이터의 부분집합이므로 학습데이터에 대해서는 오차가 감소하지만 실제데이터에 대해서는 오차가 증가함

5) 평가(Evaluation)

모델링 단계에서 얻은 모델이 프로젝트의 목적에 부합하는지 평가

데이터 마이닝 결과를 수용할 것인지 최종적으로 판단하는 과정

분석 결과 평가, 모델링 과정 평가, 모델 적용성 평가

\* 모델 평가는 '모델링' 단계에서, 모델링 과정 평가와 모델 적용성 평가는 '평가'단계에서

6) 전개(Deployment)

완성된 모델을 실제 업무에 적용하기 위한 계획 수립

전개 계획 수립, 모니터링과 유지보수 계획 수립, 프로젝트 종료 보고서 작성, 프로젝트 리뷰

## 빅데이터 분석 방법론

분석 기획	데이터 준비	데이터 분석	시스템 구현	평가 및 전개
비즈니스 이해 및 범위 설정	필요 데이터 정의	분석용 데이터 준비	설계 및 구현	모델 발전 계획
프로젝트 정의 및 계획 수립	데이터 스토어 설계	텍스트 분석	시스템 테스트 및 운영	프로젝트 평가 보고
프로젝트 위험 계획 수립	데이터 수집 및 적합성 점검	탐색적 분석		평가 및 전개
		모델링		
		모델 평가 및 검증		

### 분석 기획(Planning) 단계의 Task

#### 1. 비즈니스 이해 및 범위 설정

##### 1) 비즈니스 이해

분석 대상인 업무 도메인을 이해하기 위해 내부 업무 매뉴얼과 관련 자료, 외부의 관련 비즈니스 자료 조사 및 프로젝트 진행을 위한 방향 설정

##### 2) 프로젝트 범위 설정

프로젝트 목적에 부합하는 범위를 명확히 설정함

프로젝트에 참여하는 관계자들의 이해를 일치시키기 위하여 구조화된 프로젝트 범위 정의서 SOW(Statement of Work)를 작성

#### 2. 프로젝트 정의 및 계획 수립

##### 1) 데이터 분석 프로젝트 정의

상세 프로젝트 정의서 작성, 프로젝트의 목표를 명확화하기 위해 모델 이미지 및 평가기준 설정

##### 2) 프로젝트 수행 계획 수립

프로젝트 수행 계획서 작성, 프로젝트의 목적, 배경, 기대효과, 수행방법 일정 및 추진 조직 WBS 작성

\* WBS : Work breakdown structure, 작업 분할 구조도. 전체 업무를 분류하여 구성 요소로 만든 후 각 요소를 평가하고 일정별로 계획하며 그것을 완수할 수 있는 사람에게 할당해주는 역할

#### 3. 프로젝트 위험계획 수립

데이터 분석 위험 식별

계획 수립 단계에서 빅데이터 분석 프로젝트를 진행하면서 발생 가능한 모든 위험을 식별함

위험에 대한 대응 방법 : 회피(Avoid), 전이(Transfer), 완화(Mitigate), 수용(Accept)

### 데이터 준비(Preparing) 단계

#### 1. 필요 데이터 정의

##### 1) 데이터 정의

정형, 비정형, 반정형 등의 모든 내/외부 데이터를 포함하고 데이터의 속성, 데이터 오너, 데이터 관련 시스템 담당자 등을 포함하는 데이터 정의서 작성

예시 : 메타데이터 정의서, ERD(Entity Relationship Diagram) 포함

##### 2) 데이터 획득 방안 수립

내부 데이터 : 부서 간 업무 협조와 개인정보보호 및 정보보안과 관련한 문제점을 사전에 점검

외부 데이터 : 시스템 간 다양한 인터페이스 및 법적 문제점을 고려하여 상세한 계획 수립

#### 2. 데이터 스토어 설계

##### 1) 정형 데이터 스토어 설계

관계형 데이터베이스(RDBMS)를 사용하고, 데이터의 효율적 저장과 활용을 위해 데이터 스토어의 논리적 물리적 설계를 구분하여 설계함

##### 2) 비정형 데이터 스토어 설계

하둡, NoSQL 등을 이용하여 비정형 또는 반정형 데이터를 저장하기 위한 논리, 물리적 데이터 스토어 설계

#### 3. 데이터 수집 및 적합성 점검

##### 1) 데이터 수집 및 저장

크롤링 등의 데이터 수집을 위한 ETL 등의 다양한 도구와 API 스크립트 프로그램 등으로 데이터를 수집

수집된 데이터를 설계된 데이터 스토어에 저장

## 2) 데이터 정합성(무결성) 점검

데이터 스토어의 품질 점검을 통해 데이터의 정합성 확보

데이터 품질개선이 필요한 부분에 대해 보완 작업 진행

\* ETL(Extract Transformation Loading) : 다양한 데이터를 취합해 데이터 추출 후, 하나의 공통된 포맷으로 변환해 데이터 웨어하우스나 데이터 마트 등에 적재하는 과정을 지원하는 도구

\* API(Application Programming Interface) : 라이브러리에 접근하기 위한 규칙들을 정의한 것

## 데이터 분석 단계



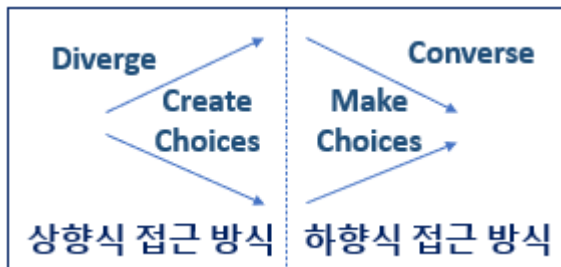
### 1. 데이터 분석

분석용 데이터를 이용한 가설 설정을 통해 통계 모델을 만들거나 기계학습을 이용한 데이터의 분류, 예측, 군집 등의 기능을 수행하는 과정

- 1) 분석용 데이터 준비
- 2) 텍스트 분석
- 3) 탐색적 분석(EDA)
- 4) 모델링
- 5) 모델 평가 및 검증

### 2. 분석 과제 도출 방법

- 1) 하향식 접근 방법 : 문제가 확실할 때 사용함, 문제가 주어지고 해법을 찾기 위해 사용함
- 2) 상향식 접근 방법 : 문제의 정의 자체가 어려운 경우 사용
- 3) 디자인 싱킹(Design Thinking)



중요한 의사결정시 상향식과 하향식을 반복적으로 사용

기존의 논리적인 단계적 접근법에 기반한 문제해결 방식은 최근 복잡하고 다양한 환경에서 발생하는 문제에 적합하지 않을 수 있음

"디자인 사고" 접근법을 통해 전통적인 분석적 사고를 극복하려함

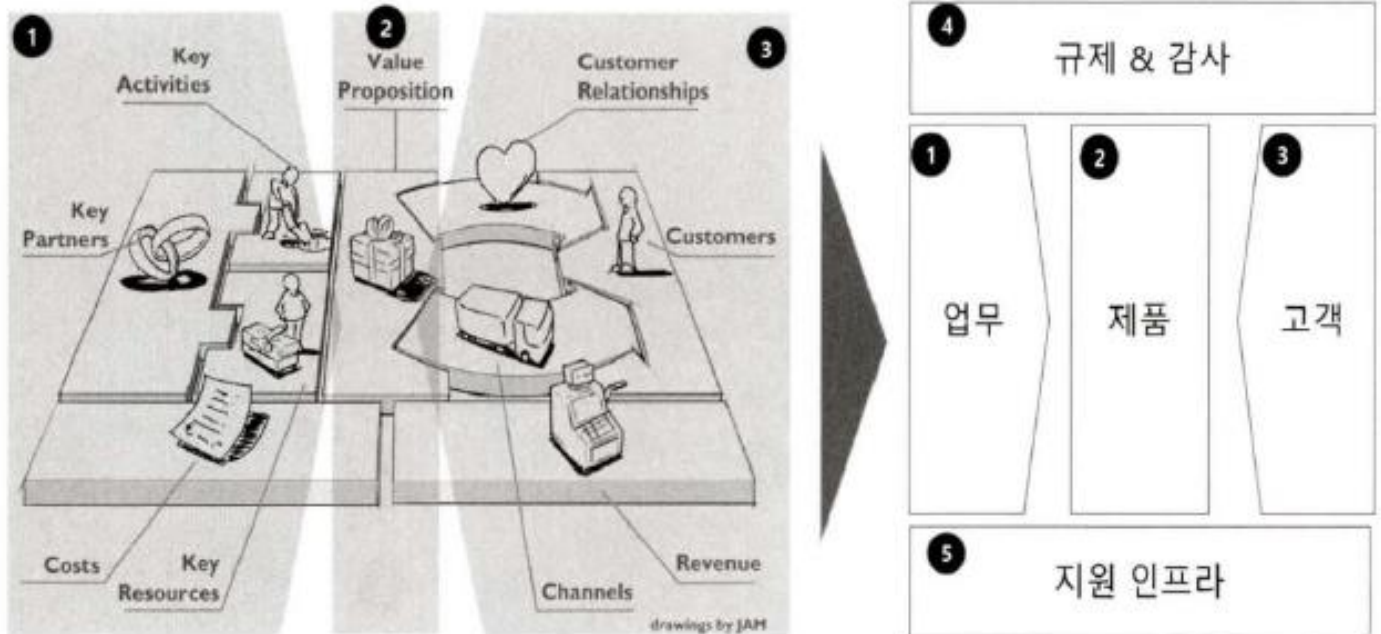
상향식 방식의 발산(Diverge) 단계와 도출된 옵션을 분석하고 검증하는 하향식 접근방식의 수렴(Converge) 단계를 반복하며 과제를 발굴함

## 하향식 접근 방식

하향식 접근 방식(Top-Down Approach)의 데이터 분석기획 단계



## 1. 문제탐색(Problem Discovery)



### 1) 비즈니스 모델 기반 문제 탐색

- 비즈니스 모델 캔버스를 활용하여 가치가 창출될 문제를 누락없이 도출할 수 있음
- 해당 기업의 사업 모델을 도식화한 비즈니스 모델 캔버스 블록을 단순화하여 업무, 제품, 고객단위로 문제를 발굴하고 이를 관리하는 지원 인프라, 규제와 감사 영역에 대한 기회를 추가로 도출하는 작업 수행 (5 가지 영역 : 업무, 제품, 고객, 지원 인프라, 규제와 감사)

### 2) 분석 기회 발굴의 범위 확장

- 거시적 관점의 요인 : STEEP - 사회, 기술, 경제, 환경, 정치 영역
- 경쟁자 확대 관점 : 대체제 영역, 경쟁자 영역, 신규진입자 영역
- 시장의 니즈 탐색 : 고객(소비자) 영역, 채널 영역, 영향자들 영역
- 역량의 재해석 관점 : 내부역량 영역, 파트너 네트워크 영역

### 3) 외부 참조 모델 기반 문제 탐색

- 유사 / 동종 사례 벤치마킹을 통한 분석 기회 발굴
- 제공되는 산업별, 업무 서비스별 분석 테마 후보 그룹을 통해 Quick & Easy 방식으로 필요한 분석 기회가 무엇인지에 대한 아이디어를 얻고 기업에 적용할 분석 테마 후보 목록을 빠르게 도출

### 4) 분석 유즈케이스

- 풀어야 할 문제에 대한 상세 설명 및 해당 문제를 해결했을 때 발생하는 효과를 명시
- 향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용하도록 함
- 예시 : (분석 유즈 케이스 : 구매최적화) = (설명 : 구매 유형과 구매자별로 과거실적과 구매조건을 비교/분석하여 구매 방안 도출) -> (효과 : 구매비용 절감)

## 2. 문제 정의(Problem Definition) 단계

식별된 비즈니스 문제를 데이터의 문제로 변환하여 정의하는 단계

- 1) 문제 탐색 단계 - 무엇(What)을 어떤 목적으로(Why) 수행해야하는지 관점
- 2) 문제 정의 단계 - 달성을 위해 필요한 데이터 및 기법(How)을 정의하기 위한 데이터 분석 문제로 변환을 수행



### 3. 해결 방안 탐색

어떤 데이터 또는 분석시스템을 사용할 것인지 검토하는 단계로 데이터 및 분석시스템에 따라 소요되는 예산 및 활용 가능 도구가 다름

분석 역량 (who)	확보	미확보
분석 기법 및 시스템		
기존 시스템	기존 시스템 개선 활용	교육 및 채용을 통한 역량 확보
신규 도입	시스템 고도화	전문업체(Sourcing)

### 4. 타당성 검토 단계

경제적 타당도 : 배용대비 편익 분석 관점의 접근

데이터 및 기술적 타당도 : 데이터 존재 여부, 분석 시스템 환경, 분석 역량

## 상향식 접근 방식

#### 1. 개념

- 문제의 정의 자체가 어려운 경우 상향식 접근 방식 사용
- 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식
- 상향식 접근방식의 데이터 분석은 비지도학습(Unsupervised Learning) 방법에 의해 수행됨
- 디자인 싱킹(Design Thinking)의 발산단계에 해당함
- 인사이트 도출 후 반복적인 시행착오를 통해 수정하며 문제를 도출하는 일련의 과정

#### 2. 지도학습 vs 비지도 학습

##### 1) 지도학습(Supervised Learning)

명확한 input, output 이 존재함

예측(Regression) : 데이터를 대표하는 선형모델 등을 만들고 그 모델을 통해 미래의 사건을 예측하는 것

분류(Classification) : 이전까지 학습된 데이터를 근거로 새로운 데이터가 기존에 학습된 데이터에 분류 여부

##### 2) 비지도학습(Unsupervised Learning)

컴퓨터가 알아서 분류를 하고, 의미 있는 값을 보여줌

데이터가 어떻게 구성되어 있는지 밝히는 용도로 사용함

군집화(Clustering)

## 분석프로젝트의 특징

#### 1. 분석프로젝트의 특징

- 분석 프로젝트는 다른 프로젝트 유형처럼 범위, 일정, 품질, 리스크, 의사소통 등 영역별 관리가 수행되어야한다
- 다양한 데이터에 기반한 분석기법을 적용하는 특성때문에 5 가지 주요 특성을 고려하여 추가적 관리가 필요하다 (분석과제 주요 특성 : Data Size, Data Complexity, Speed, Analytic Complexity, Accuracy & Precision)
- 분석 프로젝트는 도출된 결과의 재해석을 통한 지속적인 반복 및 정규화가 수행되기도 한다

#### 2. 분석과제의 주요 5 가지 특성 관리 영역

##### 1) Data Size

분석하고자하는 데이터의 양을 고려하는 관리방안 수립 필요

##### 2) Data Complexity

비정형데이터 및 다양한 시스템에 산재되어 있는 데이터들을 통합해서 분석 프로젝트를 진행할때는 해당 데이터에 잘 적용될 수 있는 분석모델선정에 대한 고려 필요

##### 3) Speed

분석결과 도출 후, 활용하는 시나리오 측면에서 일, 주 단위 실적은 배치 형태 작업, 사기 탐지, 서비스 추천은 실시간 수행되어야 함  
분석모델의 성능 및 속도를 고려한 개발 및 테스트가 수행되어야 함

##### 4) Analytic Complexity

정확도(Accuracy)와 복잡도(Complexity)는 트레이드 오프관계가 존재

분석 모델이 복잡할수록 정확도는 올라가지만 해석이 어려워짐

기준점을 사전에 정의해두어야함

##### 5) Accuracy & Precision

Accuracy : 분석의 활용적인 측면(모델과 실제값의 차이)

Precision : 분석의 안정성 측면(모델을 반복했을 때의 편차)

Accuracy, Precision 은 트레이드 오프인 경우가 많음

모델의 해석 및 적용 시 사전에 고려해야 함



# 10 개 주제별 프로젝트 관리 체계

1. 분석프로젝트의 경우 관리 영역에서 일반 프로젝트와 다르게 유의해야할 요소 존재
- 시간, 범위, 품질, 통합, 이해관계자, 자원, 원가, 리스크, 조달, 의사소통
- 1) 시간 : 프로젝트 활동의 일정을 수립, 일정 통제의 진척 상황 관찰
  - 2) 범위 : 작업과 인도물을 식별하고 정의하는데 요구되는 프로세스
  - 3) 품질 : 품질보증과 품질통제를 계획하고 확립하는데 요구되는 프로세스
  - 4) 통합 : 프로젝트와 관련된 다양한 활동과 프로세스를 도출, 정의, 결합, 단일화, 조정, 통제, 종료에 필요한 프로세스
  - 5) 이해관계자 : 프로젝트 스폰서, 고객사, 기타 이해관계자 식별, 관리에 필요한 프로세스
  - 6) 자원 : 인력, 시설, 장비, 자재, 기반시설, 도구와 같은 적절한 프로젝트 자원을 식별하고 확보하는데 필요한 프로세스
  - 7) 원가 : 개발 예산과 원가통제의 진척상황을 관찰하는 데 요구되는 프로세스
  - 8) 리스크 : 위험과 기회를 식별하고 관리하는 프로세스
  - 9) 조달 : 계획에 요구된 프로세스를 포함하며, 제품 및 서비스 또는 인도물을 인수하고 공급자와의 관계를 관리하는데 요구되는 프로세스
  - 10) 의사소통 : 프로젝트와 관련된 정보를 계획, 관리, 배포하는데 요구되는 프로세스

## 분석 마스터 플랜

### 1. 분석 마스터플랜 수립 프레임워크



- 중장기적 마스터 플랜 수립을 위해서는 분석 과제를 대상으로 다양한 기준을 고려해 적용할 우선순위를 설정할 필요가 있다
- 분석과제 수행의 선/후행 관계를 고려하여 우선순위를 조정해나간다
- 분석과제의 적용범위 및 방식에 대해서도 종합적으로 고려하여 결정한다

### 2. 수행 과제 도출 및 우선순위 평가

#### 1) 과정



\* ROI(Return On Investment) 관점에서의 빅데이터 4V

--- 투자비용 요소

Volume : 빅데이터의 크기/양

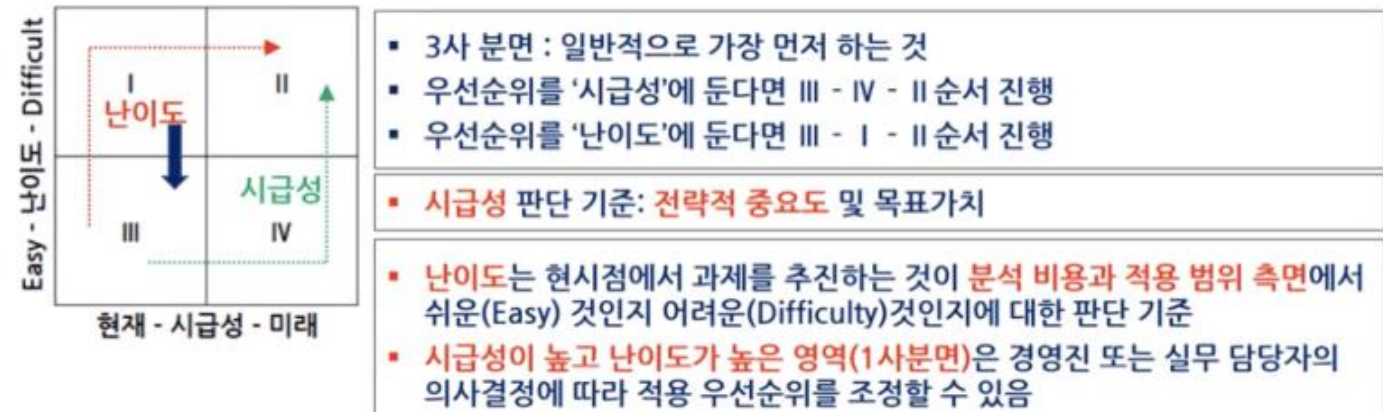
Variety : 데이터의 종류/유형

Velocity : 데이터의 생성/처리속도

--- 비즈니스 효과 요소

Value : 분석결과 활용 및 실행을 통한 비즈니스 가치

#### 2) 분석과제 우선순위 선정 기법



- 포트폴리오 사분면을 통한 과제 우선순위 선정

### 3. 이행계획 수립

#### 1) 로드맵 수립

결정된 과제의 우선순위를 토대로 분석 과제별 적용 범위 및 방식을 고려하여 최종적인 실행 우선순위를 결정 후 단계적 구현 로드맵 수립

#### 2) 세부 이행계획 수립

반복적인 정렬과정을 통해 프로젝트의 완성도를 높이는 방식을 주로 사용

모든 단계 반복보다 데이터 수집 및 확보와 분석데이터를 준비하는 단계를 순차적 진행하고 모델링 단계는 반복적으로 수행하는 혼합형을 많이 적용함

## 거버넌스 체계 개요

### 1. 거버넌스(Governance)

- Government와 같은 어원

- 더 폭넓은 의미로 진화하여 기업, 비영리 기관 등에서 규칙, 규범 및 행동이 구조화, 유지, 규제되고 책임을 지는 방식 및 프로세스를 지칭함

### 2. 분석 거버넌스

- 기업에서 데이터에 어떻게 관리, 유지, 규제되는지에 대한 내부적인 관리 방식이나 프로세스

#### 1) 분석 거버넌스 체계 구성요소 (분석비용 및 예산 없음에 주의)

- Process : 과제 기획 / 운영 프로세스

- Organization : 분석 기획 / 관리 및 추진 조직

- System : IT 기술 / 프로그램

- Human Resource : 분석 교육

- Data : 데이터 거버넌스

### 3. 데이터 거버넌스

- 데이터의 품질보장, 프라이버시 보호, 데이터 수명 관리, 전담조직과 규정정립, 데이터 소유권과 관리권 명확화 등을 통해 데이터가 적시에 필요한 사람에게 제공되도록 체계를 확립하는 것

- 데이터 거버넌스가 확립되지 못하면 빅브라더의 우려가 현실화될 가능성이 높음

- 빅브라더 : 정보의 독점으로 사회를 통제하는 관리 권력 혹은 그러한 사회체계

## 데이터 분석 수준 진단 & 진단 결과

### 1. 데이터 분석 수준 진단

- 데이터 분석 기법을 구현하기 위해 무엇을 준비하고 보완해야 하는지 등 분석의 유형 및 분석의 방향성 결정

- 분석 준비도와 분석 성숙도를 함께 평가함으로써 수행될 수 있음

#### 1) 분석 준비도 : 분석, 인력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라

분석 업무 파악	인력 및 조직	분석 기법	분석 데이터	분석 문화	분석 인프라
발생한 사실 분석 업무 예측 분석 업무 시뮬레이션 분석 업무 최적화 분석 업무 분석 업무 정기적 개선	분석 전문가 직무 존재 분석 전문가 교육 훈련 프로그램 관리자의 기본 분석 능력 전사 분석 업무 총괄 조직 존재 경영진 분석 업무 이해 능력	업무별 적합한 분석 기법 사용 분석 업무 도입 방법론 분석 기법 라이브러리 분석 기법 효과성 평가 분석 기법 정기적 개선	분석 업무를 위한 데이터 충분성 및 신뢰성 적시성 비구조적 데이터 관리 외부 데이터 활용 체계 기준 데이터 관리	사실에 근거한 의사 결정 관리자의 데이터 중시 회의 등에서 데이터 활용 경영진의 직관보다 데이터의 활용 데이터 공유 및 협업 문화	운영 시스템 데이터 통합 EAI, ETL 등 데이터 유통체계 분석 전용 서버 및 스토리지 빅데이터 분석 환경 비극

#### 2) 분석 성숙도 : 비즈니스 부문, 조직/역량 부문, IT 부문을 대상으로 도입단계, 활용단계, 확산단계, 최적화 단계로 구분해 살펴볼 수 있음

단계	도입 단계	활용 단계	확산 단계	최적화 단계
설명	분석을 시작하여 환경과 시스템 구축	분석 결과를 실제 업무에 적용	전사 차원에서 분석을 관리하고 공유	분석을 진화 시켜 혁신 및 성과 향상에 기여
비즈니스 부문	실적분석 및 통계 정기보고 수행 운영 데이터 기반	미래 결과 예측 시뮬레이션 운영 데이터 기반	전사 성과 실시간 분석 프로세스혁신 3.0 분석 규칙 관리, 이벤트 관리	외부환경 분석 활용 최적화 업무 적용, 실시간 분석 비즈니스 모델 진화
조직 역량 부문	일부 부서에서 수행 담당자 역량에 의존	전문 담당부서에서 수행 분석 기법 도입 관리자가 분석 수행	전사 모든 부서 수행 분석 CoE 조직 운영 데이터 사이언티스트 확보	데이터 사이언스 그룹 경영진 분석 활용 전략 연계
IT 부문	데이터 웨어하우스, 데이터 마트, ETL/EAI, OLAP	실시간 대시보드 통계분석 환경	빅데이터 관리 환경 시뮬레이션/최적화 비주얼 분석, 분석 전용 서버	분석 협업 환경, 분석 Sandbox 프로세스 내재화 빅데이터 분석

\* 온라인 분석 처리(Online Analytical Processing, OLAP) : 의사결정 지원 시스템 가운데 대표적인 예로, 사용자가 동일한 데이터를 여러 기준을 이용하는 다양한 방식으로 바라보면서 다차원 데이터를 분석할 수 있도록 도와준다

## 2. 분석 수준 진단 결과

### 1) 사분면 분석

- 분석 수준 진단 결과를 구분하여 향후 고려해야 하는 데이터 분석 수준에 대한 목표 방향을 정의하고 유형별 특성에 따라 개선방안을 수립할 수 있음



## 데이터 거버넌스 체계 수립

### 1. 데이터 거버넌스 체계요소

- 1) 데이터 표준화 : 데이터 표준용어 설정, 명명규칙 수립, 메타데이터 구축, 데이터 사전 구축
- 2) 데이터 관리체계 : 메타데이터와 데이터 사전(Data Dictionary)의 관리 원칙 수립
- 3) 데이터 저장소 관리 : 메타데이터 및 표준 데이터를 관리하기 위한 전사차원의 저장소를 구성
- 4) 표준화 활동 : 데이터 거버넌스 체계 구축 후, 표준 준수 여부를 주기적으로 점검, 모니터링

### 2. 데이터 거버넌스의 데이터 저장소 관리

- 메타데이터 및 표준데이터를 관리하기 위한 전사 차원의 저장소를 구성
- 저장소는 데이터 관리 체계 지원을 위한 워크플로우 및 관리용 응용 소프트웨어를 지원하고 관리 대상 시스템과의 인터페이스를 통한 통제가 이루어져야한다
- 데이터 구조 변경에 따른 사전영향평가도 수행되어야 효율적인 활용이 가능하다

## 데이터 분석을 위한 조직구조

### 1. 집중형 조직구조

- 조직내에 별도의 독립적인 분석 전담 조직 구성
- 분석 전담조직에서 회사의 모든 분석 업무를 담당함
- 일부 협업 부서와 분석업무가 중복 또는 이원화될 가능성이 있음

### 2. 기능중심 조직구조

- 별도로 분석 조직을 구성하지 않고, 각 해당 업무부서에서 직접 분석하는 형태
- 일반적인 분석 수행구조, 전사적 핵심 분석이 어려움

### 3. 분산 조직 구조

- 조직의 인력들이 협업부서에 배치되어 신속한 업무에 적합
- 전사 차원의 우선순위 수행, 부서 분석 업무와 역할 분담 명확히 해야함



# 분석 과제 관리 프로세스, 분석 교육 및 변화 관리

## 1. 분석과제 관리 프로세스

- 1) 과제 발굴 : 분석 아이디어 발굴, 분석 과제 후보 제안, 분석 과제 확장
- 2) 과제 수행 : 팀구성, 분석 과제 실행, 분석 과제 진행 관리, 결과 공유/개선

## 2. 분석 교육 및 변화 관리

- 예전에는 기업 내 데이터 분석가가 담당했던 일 -> 모든 구성원이 데이터를 분석하고 이를 바로 업무에 활용할 수 있도록 조직 전반에 분석 문화를 정착시키고 변화시키려는 시도
- 분석 조직 및 인력에 대한 지속적인 교육과 훈련이 필요함

# 빅데이터 거버넌스의 특징

## 1. 빅데이터 거버넌스 특징

- 기업이 가진 과거 및 현재의 모든 데이터를 분석하여 비즈니스 인사이트를 찾는 노력은 비용면에서 효율적이지 못함 -> 분석 대상 및 목적을 명확히 정의하고 필요한 데이터를 수집, 분석하여 점진적으로 확대해 나가는 것이 좋음
- 빅데이터 분석에서 품질관리도 중요하지만, 데이터 수명주기 관리방안을 수립하지 않으면 데이터 가용성 및 관리 비용 증대 문제에 직면할 수 있음
- ERD는 운영중인 데이터베이스와 일치하기 위해 계속해서 변경사항을 관리하여야 함
- 산업 분야별, 데이터 유형별, 정보 거버넌스 요소별로 구분하여 작성함
- 적합한 분석업무를 도출하고 가치를 높여줄 수 있도록 분석 조직 및 인력에 대해 지속적인 교육과 훈련을 실시함
- 개인정보보호 및 보안에 대한 방법을 마련해야함

# 관련 용어

다음의 용어는 단답형으로 기출되었음

## 1. Servitization

제조업과 서비스업의 융합을 나타내는 용어

ex. 웅진 코웨이의 코디

## 2. CoE(Center of Excellence)

구성원들이 비즈니스 역량, IT 역량, 분석역량을 고루 갖추어야 하며, 협업부서 및 IT 부서와의 지속적인 커뮤니케이션을 수행하는 조직 내 분석 전문조직을 말함

## 3. ISP(Information Strategy Planning, 정보 전략 계획)

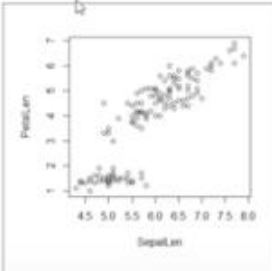
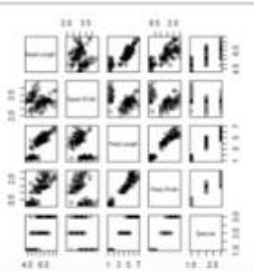
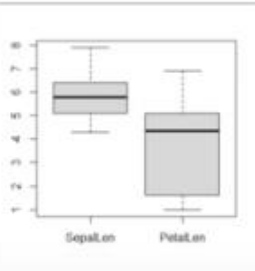
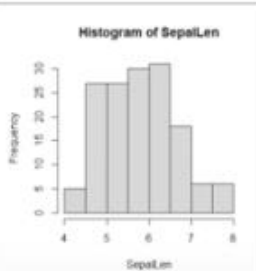
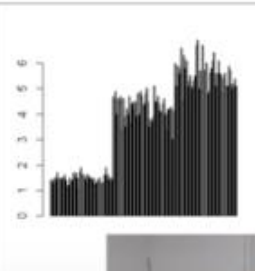
기업의 경영목표 달성에 필요한 전략적 주요 정보를 포착하고, 주요 정보를 지원하기 위해 전사적 관점의 정보구조를 도출하며, 이를 수행하기 위한 전략 및 실행계획을 수립하는 전사적인 종합추진 계획

## 4. Sandbox

보안모델, 외부 접근 및 영향을 차단하여 제한된 영역 내에서만 프로그램을 동작시키는 것

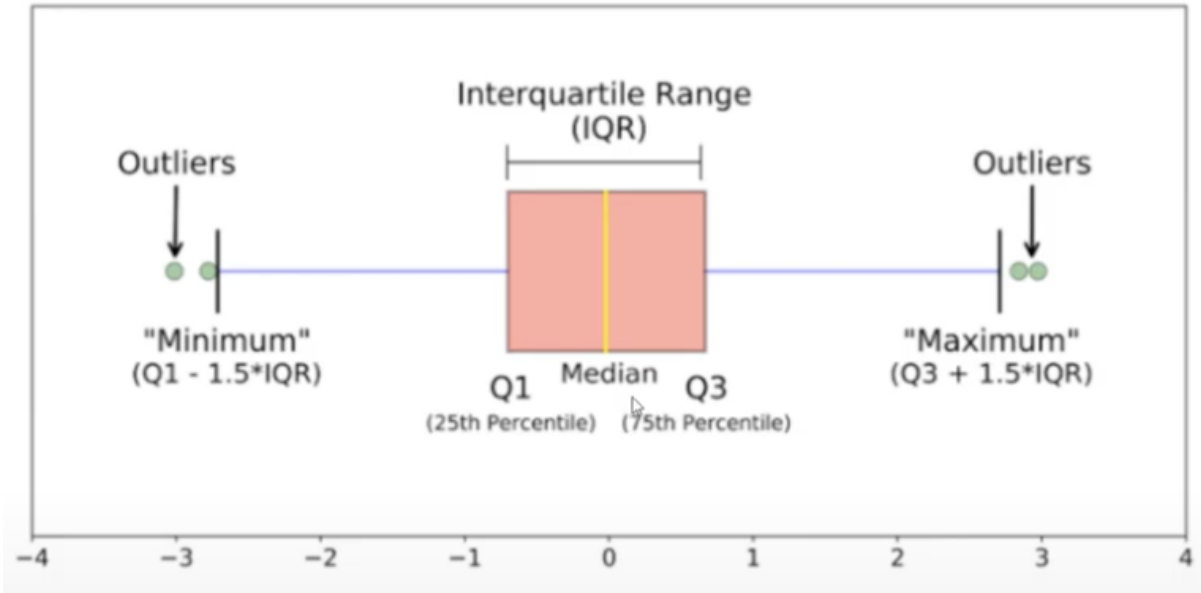
# 3. 데이터 분석

# 그래프 종류

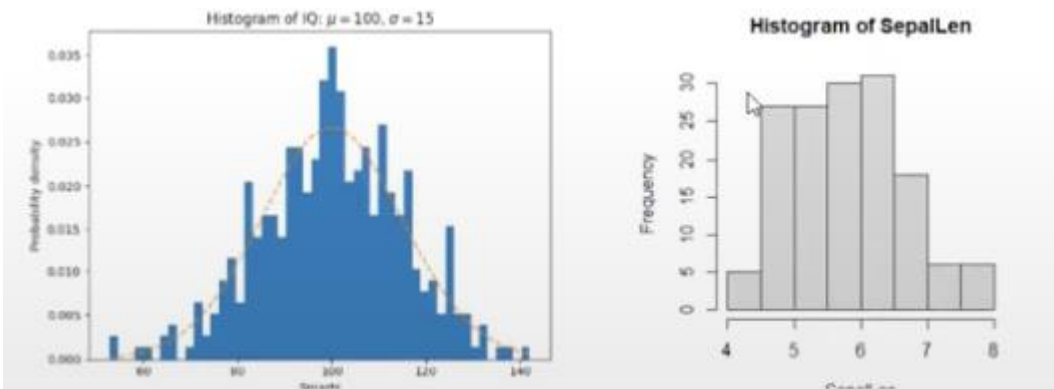
산점도	산점도 행렬	상자그림	히스토그램	막대그래프
plot(x, y)	pairs()	boxplot()	hist()	barplot()
2개 수치형 변수의 상관관계 알아보기	여러 개의 변수 관계 알아보기	이상치 존재 확인, IQR 길이, 최소, 최대, 1사분위, 3사분위, 중위값 확인 NA 제거하고 그려짐	연속형 수치에 적합 히스토그램의 사각형	명목형 변수의 빈도에 활용, 막대 사이가 끊겨 있는 모양
				

## 1. Boxplot

데이터의 분포를 파악하는 도구이다



## 2. 히스토그램



- 도수분포표의 각 계급을 가로축에 나타내고, 해당 계급에 속하는 측정값의 도수를 세로축에 표시하여 직사각형 모양으로 그림
- 왼쪽으로 치우친 모양이라면 데이터가 전체범위에서 수치가 낮은쪽에 몰려있음, 오른쪽이라면 높은쪽에 몰려있음을 의미
- 한쪽으로 치우치는 것 없이 비슷한 높이로 그려진다면 균일한 분포
- 막대 높이는 빈도를 나타내며, 폭은 의미가 없음
- 가로, 세로축 모두 연속적임, 범주형에는 막대그래프를 사용함
- 많은 데이터를 가지고 있는 경우 보다 정확한 관계 파악을 할 수 있음

\* apply 계열 함수

- 데이터 조작에 편리한 기능 제공한다
- for 문 없이 multi-core 사용으로 빠르게 연산 가능
- split -> apply -> combine 기능 제공 (데이터 분할 -> 함수 적용 -> 재결합)
- 배열 또는 행렬에 주어진 함수를 적용한 다음 그 결과를 벡터, 배열 또는 리스트로 반환함
- 함수

apply : (array) -> (vector, array)

lapply : (vector, list) -> (list)

sapply : (vector, list) -> (vector, matrix, array)

tapply : (vector, list, factor) -> (vector, array)

## 결측치 대처법

### 1. 결측치란?

대부분의 머신러닝 알고리즘은 Missing feature, 즉 누락된 데이터가 있을 때, 제대로 역할을 하지 못하기 때문에 먼저 Missing Feature(NA, Not available, 결측치)라고 하며 값이 표기되지 않은 값이다

1) 종류

- Random : 패턴이 없는 무작위 값
- No Random : 패턴을 가진 결측치

## 2) 결측치 처리 전략

- 제거(Deletion)
- 대치(Imputation)
- 예측 모델(Prediction model)

### 2. 단순 대치법(Simple Imputation)

#### 1) 완전히 응답한 개체 분석

- completes analysis, 불완전 자료는 모두 무시
- 부분적으로 관측된 자료를 무시하므로 생기는 효율성 상실, 통계적 추론의 타당성 문제 존재

#### 2) 평균 대치법

- 관측 또는 실험을 통해 얻어진 데이터의 평균으로 결측값 대치
- 비조건부 평균 대치법 : 관측 데이터의 평균값으로 대치
- 조건부 평균 대치법 : 회귀분석을 활용한 대치법

\* 회귀분석(regression analysis) : 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석방법

#### 3) 단순확률 대치법

- 평균대치법에서 추정된 표준오차의 과소추정문제를 보완하고자 고안됨
- Hot-deck, nearest neighbor 방법 등이 있음

\* Hot-deck : 전체 표본을 몇 개의 cell로 나눈 후 해당 cell 내의 응답자 가운데에서 랜덤하게 뽑아서 그 값을 대체하는 방법으로 1차년도 자료와 같이 과거자료가 없는 경우에 사용될 수 있다

### 3. 다중대치법

- 단순대치법을 한번이 아닌 m번을 수행하여 m개의 가상적 완전 자료를 만들
- 추정량 표준오차의 과소추정 또는 계산의 난해성 문제를 가지고 있음

## 이상값 검색

### 1. 이상값(Outlier)이란

- 의도하지 않게 잘못 입력된 경우(bad data)
- 분석 목적에 부합되지 않아 제거해야 하는 경우
- 의도되지 않은 현상이지만 분석에 포함해야 하는 경우

### 2. 이상값 판단

- ESD : 평균으로부터  $3 \times$  표준편차 밖의 값
- boxplot 사용 : IQR  $\times 1.5$  밖의 값
- summary() 사용 : 평균, 중앙값, IQR을 보고 판단함

### 3. 이상값 처리

- 이상값도 분석대상이 될 수 있어 무조건 삭제는 안됨

## 통계 분석 개요

Population, Parameter, Sample, Statistic



#### 1. 모집단(population)

잘 정의된 연구목적과 이와 연계된 명확한 연구대상(데이터 전체 집합)

ex. 대통령 후보의 지지율 - 유권자

#### 2. 모수(parameter)

모집단의 특성을 나타내는 수치들

모집단의 평균, 분산 같은 수치들을 모수(parameter)라고 함

#### 3. 표본(sampling)

모집단의 개체수가 많아 전부 조사하기 힘들 때 모집단에서 추출(sampling) 한 것

추출(sampling)한 표본으로 모집단의 특성을 추론(inference) 함 (오차 발생)

ex. 각종 여론조사에 참여한 유권자

#### 4. 통계량(statistic)

표본의 특성을 나타내는 수치들

표본의 평균, 분산 같은 수치를 통계량이라고 함

## 표본 추출

### 1. 확률적 표본추출법의 종류

#### 1) 단순 무작위 추출(Simple random sampling)

- 모집단의 각 개체가 표본으로 선택될 확률이 동일하게 추출되는 경우
- 모집단의 개체 수  $N$ , 표본 수  $n$  일때 개별 개체가 선택될 확률은  $n/N$

#### 2) 계통 추출(Systematic sampling)

- 모집단 개체에 1,2,...,N 이라는 일련번호를 부여한 후, 첫번째 표본을 임의로 선택하고 일정 간격으로 다음표본을 선택
- ex. 1~100 번호 부여후, 10 개 선택한다면, [1,11,21,31,...,91] 선택

#### 3) 층화 추출(Stratified sampling)

- 모집단을 서로 겹치지 않게 몇 개의 집단 또는 층(strata)로 나누고, 각 집단 내에서 원하는 크기의 표본을 단순 무작위추출법으로 추출함
- 층 : 성별, 나이대, 지역 등 차이가 존재하는 그룹

#### 4) 군집 추출(Cluster sampling)

- 모집단을 차이가 없는 여러개의 집단(cluster)로 나눔 (ex. 경상대학 내에 경영학과 경제학과)
- 이들 집단 중 몇 개를 선택한 후, 선택된 집단 내에서 필요한 만큼의 표본을 임의로 선택함

### 2. 비확률 표본추출법

특정 표본이 선정될 확률을 알 수 없어 통계학에서 사용 할 수 없음

## 척도의 종류

### 1. 명목척도(nominal scale)

- 단순히 측정 대상의 특성을 분류하거나 확인하기 위한 목적
- 숫자로 바꾸어도 그 값이 크고 작음을 나타내지 않고 범주를 표시함
- 성별, 혈액형, 출생지 등

### 2. 서열(순위) 척도(Ordinal scale)

- 대소 또는 높고 낮음을 순위만 제공할 뿐 양적인 비교를 할 수 없음
- 금,은,동메달, 선호도, 만족도 등

### 3. 등간척도(구간척도, Interval scale)

- 순위를 부여하되 순위 사이의 간격이 동일하여 양적인 비교가 가능함
- 절대 0 점이 존재하지 않음
- 온도계 수치, 물가지수

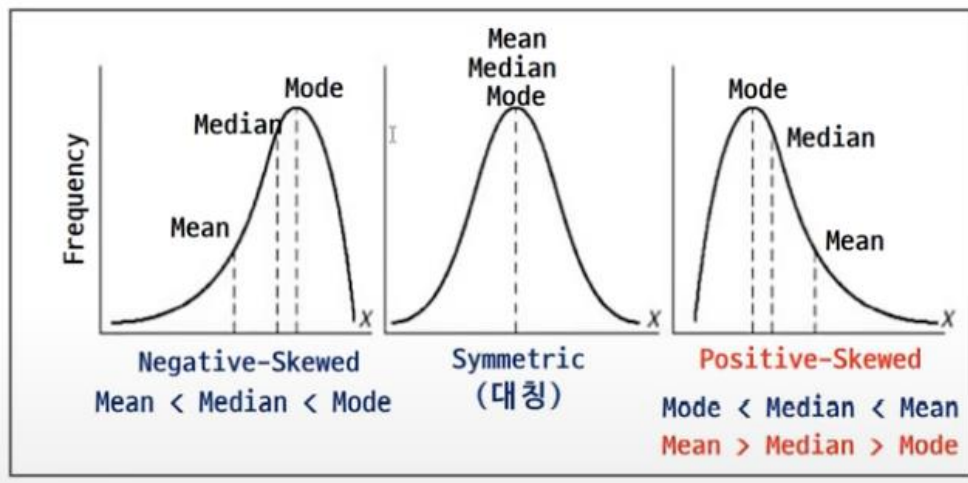
### 4. 비율척도 (Ratio scale)

- 절대 0 점이 존재하여 측정값 사이의 비율 계산이 가능한 척도
- 몸무게, 나이, 형제의 수, 직장까지의 거리

\* 절대 0 점 : 없음을 의미함(무) - 온도의 0 점은 상대 0 점으로 없음이 아니라 영상, 영하의 중간지점을 나타냄

## 집중화 경향 측정

### 1. 집중화 경향(Central Tendency) 측정에 사용되는 값들



1) 평균(Mean) : 값 들의 무게 중심이 어디인지를 나타내는 값, 산술 평균

2) 중앙값(Median) : 자료를 크기 순서대로 배열했을때, 중앙에 위치하게 되는 값

3) 최빈값(Mode) : 어떤 값이 가장 많이 관찰되는지 나타낸 값

# 데이터의 퍼짐 정도 측정

## 1. 산포도(dispersion)

- 자료의 변량들이 흩어져 있는 정도를 하나의 수로 나타낸 값
- 산포도가 크면 변량들이 평균으로부터 멀리 흩어져 있음, 변동성이 커짐
- 산포도가 작으면 변량들이 평균 주위에 밀집, 변동성이 작아짐
- 범위, 사분위수 범위, 분산, 표준 편차, 절대 편차, 변동 계수

## 2. 편차

- 어떤 자료의 변량에서 평균을 뺀 값을 편차라고 한다 (편차 = 변량 - 평균)
- 편차의 총합은 항상 0, 편차의 절댓값이 클수록 그 변량은 평균에서 멀리 떨어져 있고, 편차의 절댓값이 작을수록 평균에 가까이 있다

## 3. 분산( $s^2$ )

- 편차의 제곱의 합을  $n-1$ 로 나눈 것
- 데이터 집합이 얼마나 퍼져있는지 알아볼 수 있는 수치
- 평균이 같아도 분산은 다를 수 있음

## 4. 표준편차(s, Standard Deviation)

- 자료의 산포도를 나타내는 수치, 분산의 양의 제곱근
- 평균으로부터 각 데이터의 관찰 값까지의 평균 거리

# 데이터 표현 방법 - 기타

## 1. 표준오차(SE, Standards error)

표본 집단의 평균값이 실제 모집단의 평균값과 얼마나 차이가 있는지를 나타냄

오차 = 추정값 - 참값

모집단에서 샘플을 무한번 뽑아서 각 샘플마다 평균을 구했을 때, 그 평균들의 표준편차를 표준오차라 할 수 있음

표본평균이 모평균과 얼마나 떨어져있는가를 나타냄

모평균에 대해 추론할때 표본평균의 표준오차를 사용함

## 2. 표본오차(SE, Sampling Error)

표본을 샘플링할 때, 모집단을 대표할 수 있는 전형적인 구성요소를 선택하지 못함으로써 발생하는 오차

표본의 크기를 증가시키고, 표본 선택 방법을 엄격히 하여 줄일 수 있음

## 3. 오차한계

추정(estimation)을 할 때 모평균 추정구간의 중심으로부터 최대한 허용할 최대허용오차

추정문제에서 표본오차를 구하라는 것은 '오차한계'를 구하라는 것과 같음

오차한계 = 임계값(critical value) \* 표준오차(SE)

\* 임계값 : 표준정규분포에서는  $z$  값,  $t$  분포에서는  $t$  값, 카이제곱분포에서는 카이제곱값

## 4. 추정량(estimator)

- 추정이란 표본의 통계량(평균, 분산, 표준편차)를 가지고 모집단의 모수를 추측하여 결정하는 것
- 추정량 : 모수를 추정하기 위한 관찰 가능한 표본의 식 또는 표본의 함수
- 추정값 : 표본의 식 또는 함수에 실제 관찰치를 대입하여 계산한 값
- 좋은 추정량 판단 기준

일치성(consistency) : 표본의 크기가 커짐에 따라 표본오차가 작아져야한다

비편향성/불편성(unbiasedness) : 추정량의 기댓값이 모수의 값과 같아야 한다

효율성(efficiency) : 추정량의 분산이 될수있는데로 작아져야한다 (최소분산 추정량)

# 통계 기본 용어

## 1. 표본점

어떤 행위를 했을 때 나올 수 있는 값

주사위 굴리는 행위를 했다면 1,2,3,4,5,6 중 하나

## 2. 표본공간

- 모든 표본점의 집합
- 주사위를 굴리는 행위에 대한 표본공간  $S=\{1,2,3,4,5,6\}$

## 3. 사건

- 표본점의 특정한 집합
- 주사위를 한 번 굴렸을 때 홀수가 나오는 사건을 A 라고 한다면  $A=\{1,3,5\}$

## 4. 확률(probability)

- 사건이 일어날 수 있는 가능성을 수로 나타낸 것

- 어떤 사건을 A 라고 했을때, A 가 발생할 확률은  $P(A)$ 와 같이 표기함
- 확률 = 사건 / 표본공간
- 확률값  $= 0 \leq P(A) \leq 1$

## 5. 사건의 종류

독립사건	<ul style="list-style-type: none"> <li>A의 발생이 B가 발생할 확률을 바꾸지 않는 사건</li> <li>두 사건 A, B가 독립이면 <math>P(B A)=P(B)</math>, <math>P(A B) = P(A)</math>, <math>P(A \cap B) = P(A) \cdot P(B)</math> 성립</li> <li>예) 주사위 던져서 나오는 눈의 값과 동전을 던져 나오는 앞/뒤 사건</li> <li>예) 서로 다른 사람이 총을 쏘아 과녁에 명중할 사건</li> </ul>
배반사건	<ul style="list-style-type: none"> <li>교집합이 공집합인 사건, 한쪽이 일어나면 다른 쪽이 일어나지 않을 때의 두 사건</li> <li><math>P(A \cap B) = 0</math>, <math>P(A \cup B) = P(A) + P(B)</math></li> <li>예) 동전 하나를 던져 앞면 나오는 사건, 뒷면 나오는 사건</li> </ul>
종속사건	<ul style="list-style-type: none"> <li>두 사건 A와 B에서 한 사건의 결과가 다른 사건에 영향을 주는 사건</li> <li>예) 음주와 사고 사건, <math>P(A \cap B) = P(A B) \cdot P(B)</math></li> </ul>

- $P(A) = 3/6 = 1/2$
- $P(B) = 2/6 = 1/3$
- $P(A) \cdot P(B) = 1/2 \cdot 1/3 = 1/6$

독립사건

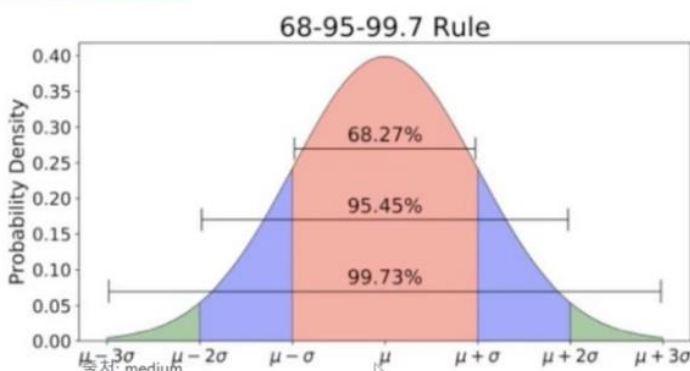
## 연속형 확률 분포 - 균등분포(uniform distribution)

### 연속형 확률분포의 선택



### 1. 정규분포

#### 3 시그마 규칙



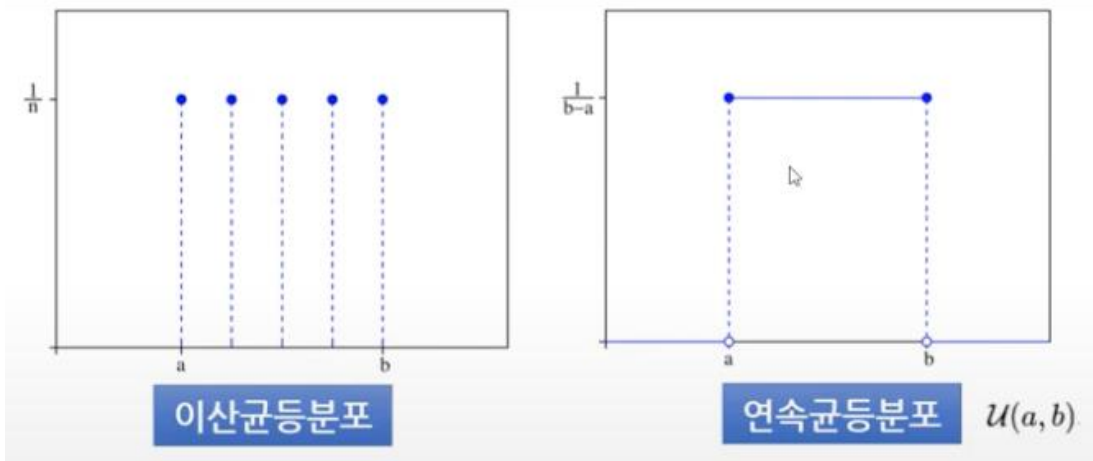
- 약 68%의 값들이 평균에서 양쪽으로 1 표준편차 범위( $\mu \pm \sigma$ )에 존재
- 약 95%의 값들이 평균에서 양쪽으로 2 표준편차 범위( $\mu \pm 2\sigma$ )에 존재
- 거의 모든 값들(실제로는 99.7%)이 평균에서 양쪽으로 3표준편차 범위( $\mu \pm 3\sigma$ )에 존재

- 가우스 분포라고도 하며, 수집된 자료의 분포를 근사하는데 자주 사용함
- 평균과 표준편차에 대해 모양이 결정됨
- 대부분의 측정값을 정규분포로 가정하는 이유 "정규분포의 당위성" -> 시행횟수 N 이 커질때 이항분포는 정규분포와 거의 같아짐





## 2. 균등분포



- 1) 이산균등분포 : 확률분포함수가 정의된 모든 곳에서 값이 일정한 분포
- 2) 연속균등분포 : 연속확률분포로 분포가 특정 범위 내에서 균등하게 나타나 있을 경우

## 3. 지수분포

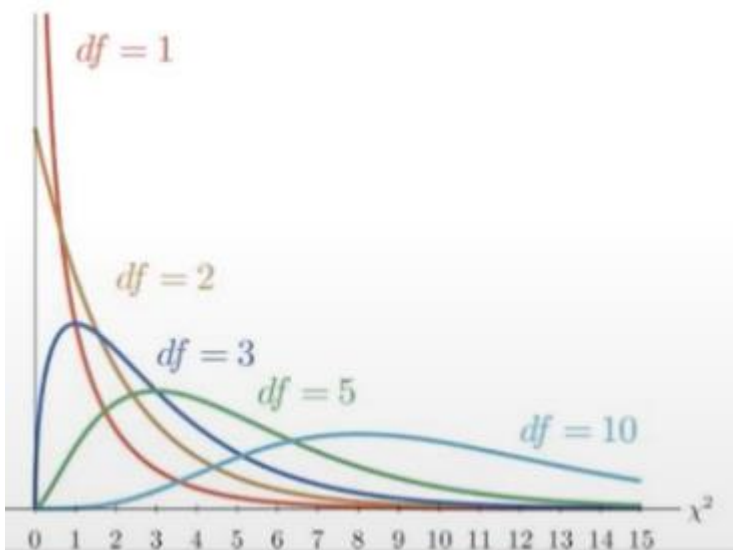
사건이 서로 독립적일 때 다음사건이 일어날 때 까지의 대기시간은 지수분포를 따름  
(일정시간동안 발생하는 사건의 횟수는 포아송 분포를 따름)

## 4. t-분포

정규분포는 표본의 수가 적으면 신뢰도가 낮아짐( $n$ 이 30 개 미만인 경우)  
표본을 많이 뽑지 못하는 경우에 대한 대응책으로 예측범위가 넓은 분포를 사용하며, 이것이 t-분포임  
t-분포는 표본의 개수에 따라 그래프의 모양이 변함 (표본의 갯수가 많아질수록 정규분포와 비슷해짐)  
t-분포는 표본의 개수가 30 개 미만일 때 사용하며, 신뢰구간 / 가설검정에 사용함

## 5. 카이제곱 분포( $\chi^2$ )

분산의 특징을 확률분포로 만든 것으로, 카이( $\chi$ )는 평균 0, 분산 1 인 표준 정규분포를 의미함  
카이제곱( $\chi^2$ )은 표준정규분포를 제곱한다는 의미가 내포되어 있음  
신뢰구간, 가설검정에 사용되며, 그래프의 x 축 좌표를 카이제곱값이라 부르며, 카이제곱분포표를 사용해 구하고 검정에 사용함  
카이제곱분포의 특징이 곧 분산(치우침정도)의 특징임



- 0 이상의 값만 가질 수 있으며, 오른쪽 꼬리가 긴 비대칭 모양  
0 의 오른쪽 부분에 분포가 많고, 0 에서 멀어질수록 분포 감소  
표본의 수가 많아지면 옆으로 넓적한 정규분포 형태가 됨

## 6. F 분포

카이제곱분포와 같이 분산을 다룰 때 사용하는 분포

카이제곱분포는 한 집단의 분산, F 분포는 두 집단의 분포를 다룸

두 집단의 분산 크기가 서로 같은지 또는 다른지 비교할때 사용

카이제곱과 비슷하게 비대칭 모양이며 양수만 존재함

두 분산의 나눗셈을 확률분포로 나타낸 것이 바로 F 분포임

표본의 수가 많아지면 1 을 중심으로 정규분포 모양이됨

분산분석에 F 분포를 사용하며, 그래프 x 축 좌표인 F 값을 활용하는데 F 분포표를 사용해 구함

## 통계적 추론의 분류

### 1. 모집단에 대한 가정 여부에 따른 통계적 추론의 분류



#### 1) 모수적 추론(Parametric Inference)

모집단에 특정 분포를 가정하고 모수에 대해 추론함

#### 2) 비모수적 추론(Non-Parametric Inference)

모집단에 대해 특정 분포를 가정하지 않음

### 2. 추론 목적에 따른 통계적 추론의 분류

#### 1) 추정(Estimation) : 통계량을 사용하여 모집단의 모수를 구체적으로 추측하는 과정

- 점추정(Point) : 하나의 값으로 모수의 값이 얼마인지 추측함

- 구간 추정(Interval) : 모수를 포함할 것으로 기대되는 구간을 확률적으로 구함

#### 2) 가설검정(Testing hypotheses)

모수에 대한 가설을 세우고 그 가설의 옳고 그름을 확률적으로 판정하는 방법론

## 모수적 추론

### 1. 개념

모집단에 특정 분포를 가정하고 모수에 대해 추론함

모수로는 평균, 분산 등을 사용

자료가 정규분포, 등간척도, 비율척도인 경우 (온도, 물가지수, 몸무게, 자녀수)

### 2. 모수적 검정

검정하고자 하는 모집단의 분포에 대해 가정을 하고, 그 가정하에서 검정 통계량과 검정통계량의 분포를 유도해 검정을 실시

### 3. 모수적 통계의 전제조건

표본의 모집단이 정규분포를 이루어야하며, 집단 내의 분산은 같아야함

변인(변수)는 등간척도나 비율척도로 측정되어야함 (아니면 비모수 통계 사용)

### 4. 모수 검정 방법

#### 1) T Test

평균값이 올바른지, 두 집단의 평균차이가 있는지를 검증하는 방법으로 t 값을 사용함

t 값이 커질수록 p-value 는 작아지며, 집단간 유의한 차이를 보일 가능성이 높아짐

- One sample t-test

단일 표본의 평균검정을 위한 방법

- Paired t-test (대응표본 t-검정)

동일 개체에 어떤 처리를 하기 전, 후의 자료를 얻을 때 차이값에 대한 평균 검정을 위한 방법

예시) 매일 1 시간 한달걸으면 2kg 이 빠진다 (걷기 수행 전 / 후)

가능한 동일한 특성을 갖는 두 개체에 서로 다른 처리를 하여 그 처리의 효과를 비교하는 방법(matching)

예시) X 질병 환자들을 두 집단으로 나누어 A,B 약을 투약해 약의 효과 비교

신뢰구간에 0 이 포함되지 않음



- Two Sample t-test (독립표본 t-검정)

서로 다른 두 그룹의 평균을 비교하여 두 표본의 차이가 있는지를 검정하는 방법

귀무가설 - 두 집단의 평균 차이 값이 0 이다

예시) 2 학년과 3 학년의 결석률은 같다

신뢰구간에 0 이 포함됨

2) ANOVA Test

## 비모수적 추론(Non-Parametric Inference)

### 1. 개념

모집단에 대해 특정 분포를 가정하지 않음

모수 자체보다 분포형태에 대한 검정을 실시함

가설을 "분포의 형태가 동일하다", "분포의 형태가 동일하지 않다"와 같이 분포형태에 대해 설정함

관측 값들의 순위나 두 관측 값 사이의 부호 등을 이용해 검정

모집단의 특성을 몇 개의 모수로 결정하기 어려우며 수 많은 모수가 필요할 수 있음

모수적 방법보다 훨씬 단순함, 민감성을 잃을 수 있음

표본수가 적고, 명목척도, 서열척도인 경우 (성별, 혈액형, 만족도, 메달)

### 2. 비수모적 검정의 종류

명목척도, 서열척도로 나뉜다

비교대상 집단수	관계	비모수-명목척도	비모수-서열척도	모수
1		카이스퀘어 검정	Kolmogorov-Smirnov test	One sample T test
2	독립	Crosstab	Mann-Whitney U test	Two sample T test
	대응 자료	McNemar test	Wilcoxon signed -rank test Sign test	Paired T test
k (다변량)	독립		Kruskal-Wallis H test	ANOVA test (분산분석)
	대응 자료	Cochran test	Friedman test	

#### \* 카이스퀘어 검정

적합도 검정 - 한개의 범주형 변수와 각 그룹 별 비율과 특정 상수비가 같은지 검정 (교배실험으로 얻은 완두콩 비율이 멘델의 법칙을 따르는지 검정)

동질성 검정 - 각 집단이 서로 유사한 성향을 갖는지 분석 (부모집단에 대해 열 변수의 분포가 동질한지 검정, 성별에 따라 음료 선호가 동일한지 검정)

독립성 검정 - 두 개 범주형 변수가 서로 독립인지 검정 (도로형태와 교통사고피해도의 관련성 검정)

#### \* 부호검정(Sign Test)

표본들이 서로 관련되어 있는 경우, 짝지어진 두 개의 관찰자들의 크고 작음을 +와 -로 표시하여 그 개수를 가지고 두 그룹의 분포 차이가 있는가에 대한 가설을 검증하는 방법

귀무가설 - 두 생산 라인의 일별 생산량 중 불량품 수의 분포는 동일하다

대립가설 - 두 생산 라인의 일별 생산량 중 불량품 수의 분포는 동일하지 않다

## 데이터의 정규성 검정

### 1. Q-Q plot

그래프를 그려서 정규성 가정이 만족되는지 시각적으로 확인하는 방법

대각선 참조선을 따라 값들이 분포하게 되면 정규성을 만족한다고 할 수 있음

### 2. Histogram

구간별 도수를 그래프로 표시하여 시각적으로 정규분포를 확인하는 방법

### 3. Shapiro-Wilk test

오차항이 정규분포를 따르는지 알아보는 검정

귀무가설은 정규분포를 따른다로 p-value 가 0.05 보다 크면 정규성을 가정하게 됨

회귀분석에서 모든 독립변수에 대해 종속변수가 정규분포를 따르는지 따르는지 알아보는 방법

### 4. kolmogorov-Smirnov test

K-S test, 두 모집단의 분포가 같은 지 검정하는 것

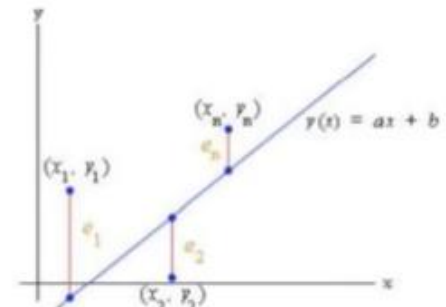
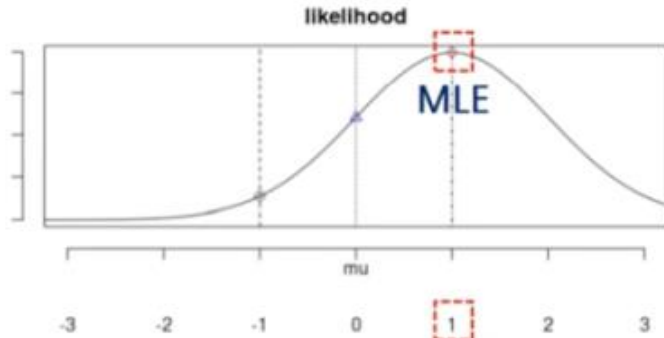
p-value 가 0.05 보다 크면 정규성을 가정하게됨

# 추론 목적에 따른 통계적 추론의 분류 (점추정 & 구간추정 & 가설검정)

## 1. 점추정

'모수가 특정한 값일것'이라고 추정하는 것

ex. A 과목 수강 전체 학생 중 50 명을 뽑아 조사한 결과 기말점수가 80 점이었다면, 50 명 뿐만 아니라 나머지 A 과목을 수강한 학생들의 점수도 90 점으로 추정하는 것



- 1) 적률법 : 표본의 기댓값을 통해 모수를 추정하는 방법
- 2) 최대가능도추정법(최대우도법) : 함수를 미분해서 기울기가 0 인 위치에 존재하는 MLE(maximum likelihood estimator)를 찾는 방법
- 3) 최소 제곱법 : 함수값과 측정값의 차이인 오차를 제곱한 합이 최소가 되는 함수를 구하는 방법

## 2. 구간추정(Interval estimation)

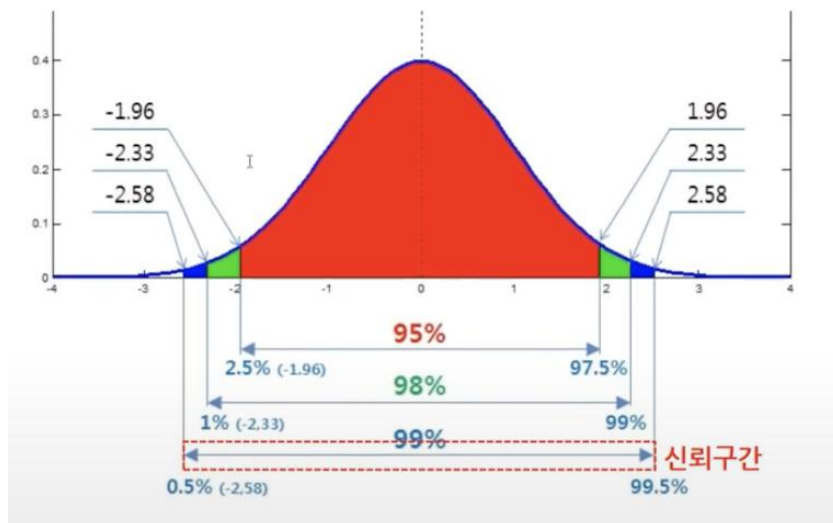
점추정의 정확성을 보완하는 방법

모수가 포함되리라고 기대되는 '범위'

모수값이 정해져 있을 때 다수 신뢰구간 중 모수값을 포함하는 신뢰구간이 존재할 확률

신뢰수준 95% 의미 : n 번 반복 추출하여 산정하는 신뢰구간들 중에서 평균적으로 95%는 모수값을 포함하고 있을것이라는 의미

\* 신뢰구간 : 99% 신뢰수준에 대한 신뢰구간이 95% 신뢰수준에 대한 신뢰구간보다 길다. 표본의 크기가 커지면 신뢰구간의 길이는 줄어든다



## 5. 가설검정(Statistical hypothesis testing)

모집단에 대한 어떤 가설을 설정한 뒤에 표본관찰을 통해 그 가설의 채택여부를 결정하는 통계적 추론 방법

### 1) 종류

- 귀무가설( $H_0$ , null hypothesis)

가설검정의 대상이 되는 가설, 연구자가 부정하고자 하는 가설

설정된 가설이 진실할 확률이 극히 적어 처음부터 버릴 것(기각)이 예상되는 가설

- 대립가설( $H_1$ , anti hypothesis)

귀무가설이 기각될 때 받아들여지는 가설

연구자가 연구를 통해 입증 또는 증명되기를 기대하는 예상이나 주장

ex. 범죄사건에서 용의자가 있을때 형사의 가설

귀무가설 : 용의자는 무죄이다

대립가설 : 용의자가 범죄를 저질렀다

### 2) 기각역(critical region)

검정통계량(t-value)의 분포에서 유의수준의 크기에 해당하는 영역

계산한 검정통계량의 유의성(귀무가설의 기각)을 판정하는 기준

### 3) 오류 종류

아래 오류가 작을수록 바람직함

두가지를 동시에 줄일 수 없기 때문에 1 종오류를 범할 확률의 최대 허용치를 미리 어떤 특정값(유의수준)으로 지정해놓고 제 2 종오류의 확률을 가장 작게해주는 검증방법을 사용함

- 제 1 종 오류 : 알파 에러, 귀무가설이 참인데 기각하게 되는 오류

- 제 2 종 오류 : 베타 에러, 귀무가설이 거짓인데 채택하는 오류

- 유의수준(significance level) : 제 1 종 오류의 최대 허용 단계 (ex. 유의수준 0.05 = 5% : 100 번 실험에서 1 종오류를 범하는 최대 허용한계가 5 번)

- 유의확률(p-value, probability value)

1 종오류를 범할 확률, 귀무가설을 지지하는 정도

귀무가설이 사실일때 기각하는 1 종오류시 우리가 내린 판정이 잘못되었을 확률

검정 통계량들은 거의 대부분이 귀무가설을 가정하고 얻게 되는 값

검정 통계량에 관한 확률로, 극단적인 표본 값이 나올 확률

p-value 가 작을수록 그 정도가 약하다고 보며,  $p\text{-value} < \alpha$  (유의확률이 유의수준보다 작으면) 귀무가설을 기각, 대립가설을 채택함

ex. p-value 가 0.05(5%) : 귀무가설을 기각했을때 기각결정이 잘못될 확률이 5%

### 4) 귀무가설을 이용한 가설검증프로세스



## 회귀분석(Regression Analysis)

### 1. 용어정리

- 독립변수 : 다른변수에 영향을 받지않고 독립적으로 변화하는 수, 설명변수라고도 함

- 종속변수 : 독립변수의 영향을 받아 값이 변화하는 수, 분석의 대상이 되는 변수

- 잔차(오차항) : 계산에 의해 얻어진 이론 값과 실제 관측이나 측정에 의해 얻어진 값의 차이 (오차(Error) - 모집단, 잔차(Residual) - 표본집합)

### 2. 회귀 분석

- 변수와 변수 사이의 관계를 알아보기 위한 통계적 분석 방법

- 독립변수의 값에 의해 종속변수의 값을 예측하기 위함

- 일반 선형회귀는 종속변수가 연속형 변수일때 가능

- 이산형 - 명목, 서열척도 / 연속형 - 구간, 비율척도

### 3. 회귀 모형

1) 선형회귀모형 : X와 Y가 1차식으로 나타날때의 모형

2) 단순회귀모형 : 독립변수 1개일때

- 최소자승법(Least Square Method) : (측정값-함수값)<sup>2</sup>의 합이 최소가 되는 직선의 그래프를 찾는것. 큰 폭의 잔차에 대해 더 큰 가중치를 둠

## 회귀 모형

### 1. 회귀 모형의 가정

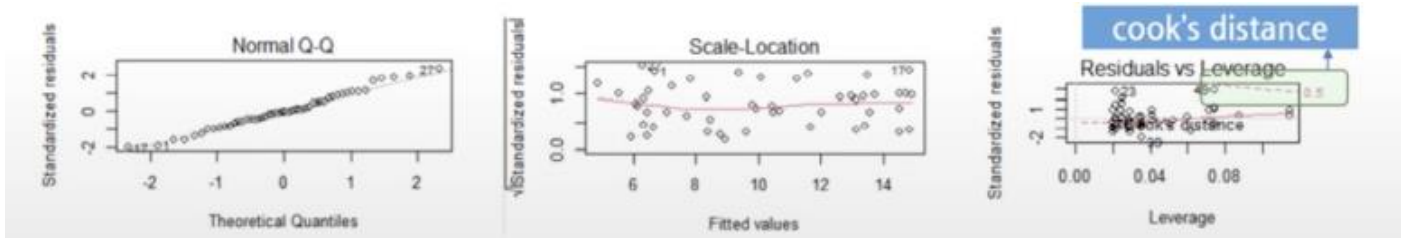
- 선형성 : 독립변수의 변화에 따라 종속변수도 변화하는 선형(linear) 모형

- 독립성 : 잔차와 독립변수의 값이 관련되어있지 않다

- 정규성 : 잔차항이 정규분포를 이뤄야 한다

- 등분산성 : 잔차항들의 분포는 동일한 분산을 갖는다

- 비상관성 : 잔차들끼리 상관이 없어야 한다 (Durbin-Watson 통계량 확인)



- 1) Normal Q-Q plot : 정규성(정상성), 잔차가 정규분포를 잘 따르고 있는지를 확인하는 그래프
- 2) Scale-Location : 등분산성, y 축이 표준화 잔차를 나타내며, 기울기 0 인 직선이 이상적임 (분산이 일정한 정규분포를 가정하므로)후
- 3) Cook's Distance : 일반적으로 1 이 넘어가면 관측치를 영향점(influence points)로 판별

## 2. 회귀모형 해석

### 1) 표본 회귀선의 유의성 검정

두 변수 사이에 선형관계가 성립하는지 검정하는 것으로 회귀식의 기울기 계수가 0 일때 귀무가설, 0 이 아닐때 대립가설로 설정한다

### 2) 회귀모형

모형이 통계적으로 유의미 한가? F 통계량, 유의확률(p-value)로 확인

회귀계수들이 유의미한가? 회귀계수의 t 값, 유의확률(p-value)로 확인

모형이 얼마나 설명력을 갖는가? 결정계수( $R^2$ ) 확인

모형이 데이터를 잘 적합하고 있는가? 잔차통계량 확인, 회귀진단 진행(선형성 ~ 정상성)

\* t 값 = Estimate(회귀계수) / Std.Error(표준오차)

t 값이 클수록 표준오차가 적다

t 통계량이 클수록 회귀계수가 유의하다

\* F 통계량

모델의 통계적 유의성을 검정하기 위한 검정 통계량 (분산 분석)

F 통계량 = 회귀제곱평균(MSR) / 잔차제곱평균(MSE)

F 통계량이 클수록 회귀모형은 통계적으로 유의하다

\* 결정계수( $R^2$ ) : 70~90%

회귀식의 적합도를 재는 척도

결정계수가 커질수록 회귀방정식의 설명력이 높아짐

## 다중 공선성(Multicollinearity)

### 1.개념

모형의 일부 설명변수(=예측변수)가 다른 설명변수와 상관되어 있을 때 발생하는 조건

중대한 다중공선성은 회귀계수의 분산을 증가시켜 불안정하고 해석하기 어렵게 만들기 때문에 문제가 됨

VIF(variance inflation factor)가 10 을 넘으면 다중공선성이 존재한다고 봄

### 2.해결방법

높은 상관관계가 있는 설명변수를 모형에서 제거하는 것으로 해결함

설명변수를 제거하면 대부분 R-square 가 감소함

단계적 회귀분석을 이용하여 제거함

### 3. 설명변수의 선택원칙

y 에 영향을 미칠 수 있는 모든 설명변수 x 들은 y 의 값을 예측하는 데 참여시킴

설명변수 x 들의 수가 많아지면 관리에 많은 노력이 요구되므로 가능한 범위 내에서 적은 수의 설명변수를 포함시켜야함

두 원칙이 이율배반적이므로 적절한 설명변수 선택이 필요함

### 4. 설명변수 선택방법

#### 1) 모든 가능한 조합

모든 가능한 독립변수들의 조합에 대한 회귀모형을 고려해 AIC, BIC 의 기준으로 선택

(AIC, BIC : 최소자승법의  $R^2$  와 비슷한 역할로 적합성을 측정해주는 지표로,  $R^2$  는 큰값이 좋지만 AIC, BIC 는 작은 값이 좋음)

#### 2) 후진제거법(Backward Elimination)

독립변수 후보 모두를 포함한 모형에서 출발해 영향이 적은 변수부터 하나씩 제거해나가는 방법

#### 3) 전진선택법(Forward Selection)

절편만 있는 모델에서 출발해 기준 통계치를 가장 많이 개선시키는 변수를 차례로 추가하는 방법

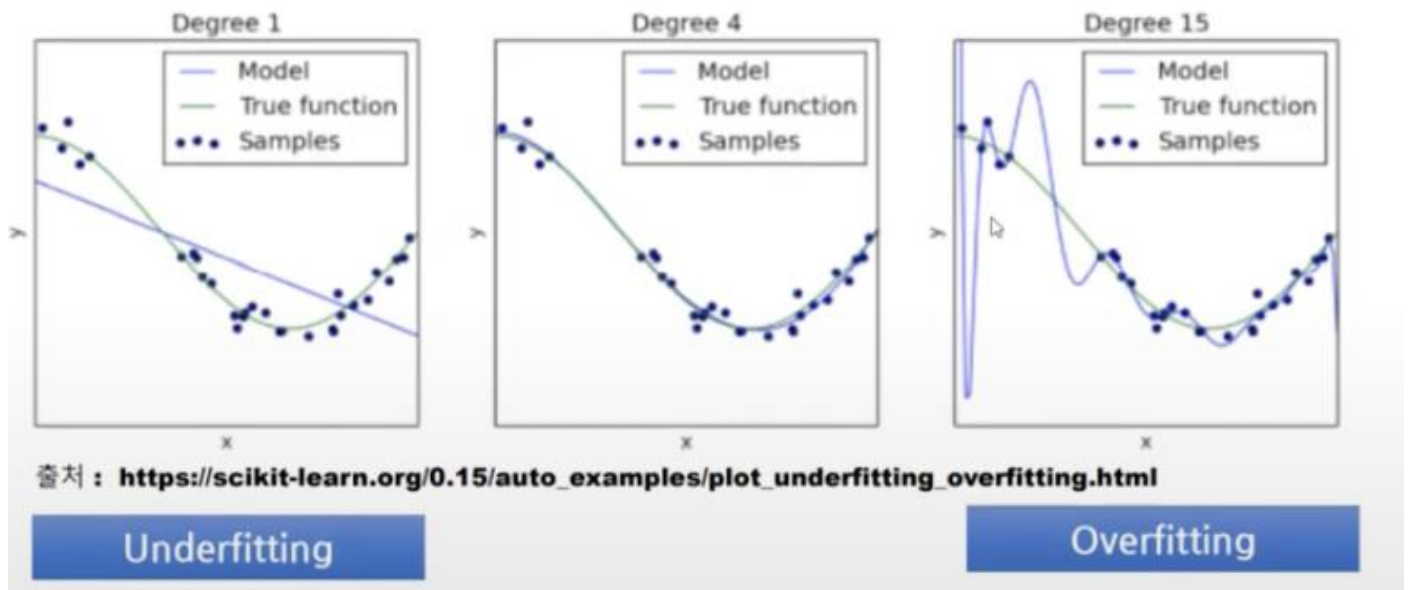
#### 4) 단계별 선택법(stepwise method)

모든 변수가 포함된 모델에서 출발해 기준 통계치에서 가장 도움이 되지 않는 변수를 삭제하거나, 모델에서 빠져있는 변수 중에서 기준

통계치를 가장 개선시키는 변수를 추가함

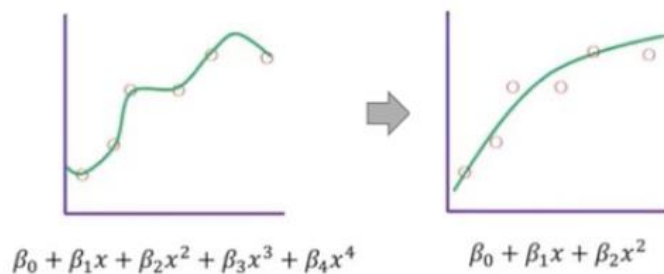
# 과적합(Overfitting)

## 1. 과적합의 문제와 해결방법



- 주어진 샘플들의 설명변수와 종속변수의 관계를 필요이상 너무 자세하고 복잡하게 분석
- 샘플에 심취한 모델로 새로운 데이터가 주어졌을 때 제대로 예측해내기 어려울 수 있음
- 해결 방법으로 Feature 개수를 줄이거나, Regularization 을 수행하는 방법이 있음

## 정규화(Regularization)



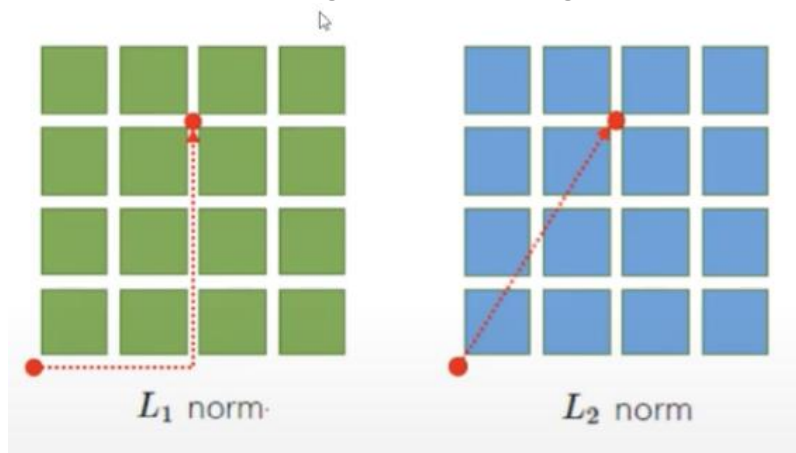
베타값에 제약을 주어 모델에 변화를 주는 것

제약값이 클수록 제약이 많아져 적은 변수가 사용되고, 해석이 쉬워지지만 underfitting 이 됨  
제약값이 작아질수록 제약이 적어 많은 변수가 사용되고, 해석이 어려워지며 overfitting 됨

## L1, L2 Norm

### 1. norm

선형대수학에서 벡터의 크기(magnitude) 또는 길이(length)를 측정하는 방법



- 1) L1 norm(Manhattan norm) : 벡터의 모든 성분의 절댓값을 더함
- 2) L2 norm(Euclidean norm) : 출발점에서 도착점까지의 거리를 직선거리로 측정함

# Regularized Linear Regression

## 1. 라쏘(Lasso) 회귀

### 1) 특징

변수선택이 가능하며, 변수간 상관관계가 높으면 성능이 떨어짐

L1 norm 을 패널티를 가진 선형회귀방법, 회귀계수의 절댓값이 클수록 패널티 부여

w의 모든 원소가 0 이되거나 0에 가깝게 되게 해야함 -> 불필요 특성 제거

어떤 특성은 모델을 만들때 사용하지 않게 됨

### 2) 장점

제약 조건을 통해 일반화된 모형을 찾는다

가중치들이 0 이 되게 함으로써 그에 해당하는 특성들을 제외해준다

모델에서 가장 중요한 특성이 무엇인지 알게 되는 등 모델 해석력이 좋아진다

## 2. Ridge 회귀 특성

L2 norm 을 사용해 패널티를 주는 방식

변수선택이 불가능, 변수 간 상관관계가 높아도 좋은 성능

라쏘는 가중치들이 0 이 되지만, Ridge 의 가중치들은 0에 가까워질뿐 0 이되진 않음

특성이 많은데 특성의 중요도가 전체적으로 비슷하다면 Ridge 가 좀 더 괜찮은 모델을 찾아줄 것이다

## 3. 엘라스틱넷 특성

L1, L2 norm regularization

변수선택 가능

변수간 상관관계를 반영한 정규화

# 데이터 스케일링(Scaling)

데이터 단위의 불일치 문제를 해결하는 방법

분석에 사용되는 변수들에 사용 단위가 다를 때 데이터를 같은 기준으로 만듦

원 데이터의 분포를 유지하는 정규화 방법

## 1. 정규화(normalization)

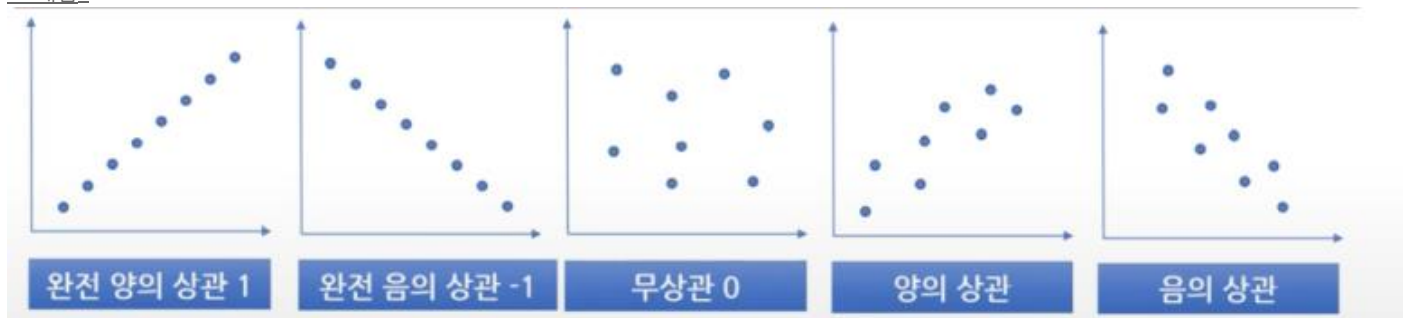
값의 범위를 [0,1]로 변환하는 것, min-max normalization

## 2. 표준화(standardization)

특성의 값이 정규분포를 갖도록 변환하는 것, 평균 0, 표준편차 1

# 상관분석

## 1. 개념



- 상관계수는 두 변수의 관련성 정도로 -1~1 의 값을 나타냄

- 두 변수의 상관관계가 존재하지 않을 경우 상관계수는 0 임

- 상관관계가 높다고 인과관계가 있다고 할수는 없음

- 귀무가설은 '상관계수가 0 이다', 대립가설은 '상관계수가 0 이아니다'

## 2. 종류

### 1) 피어슨 상관계수

대상자료는 등간척도, 비율척도 사용, 두 변수간의 선형적인 크기만 측정 가능

피어슨 상관계수 : x,y 의 공분산을 x,y 의 표준편차의 곱으로 나눈 값

\* 공분산(Covariance) : 2 개의 확률변수의 선형관계를 나타내는 값

### 2) 스피어만 상관계수

대상자료는 서열척도 사용, 비선형적인 관계도 가능

연속형 외에 이산형도 가능함

스피어만 상관계수는 원시데이터가 아니라 각 변수에 대해 순위를 매긴 값을 기반으로 함



# 자원축소 목표를 위해 개발된 분석방법

## 1. 주성분분석(PCA, Principal Component Analysis)

### 1) 개념

데이터를 분석할 때 변수의 개수가 많다고 모두 활용하는 것이 좋은거만은 아님

오히려 변수가 '다중공선성'이 있을 경우 분석결과에 영향을 줄 수도 있음

공분산행렬 또는 상관계수 행렬을 사용해 모든 변수들을 가장 잘 설명해주는 주성분을 찾는 방법

독립변수들과 주성분과의 거리인 '정보손실량'을 최소화

분산을 극대화

### 2) 공분산 행렬(default) vs 상관계수 행렬

공분산 행렬은 변수의 측정단위를 그대로 반영한 것이고, 상관계수 행렬은 모든 변수의 측정단위를 표준화한 것이다

공분산행렬을 이용한 경우 측정단위를 그대로 반영하였기 때문에, 변수들의 측정 단위에 민감하다

주성분 분석은 거리를 사용하기 때문에 척도에 영향을 받는다 (정규화 전후의 결과가 다르다)

설문조사처럼 모든 변수들이 같은 수준으로 점수화 된 경우 공분산행렬을 사용한다

변수들의 scale 이 서로 많이 다른 경우에는 상관계수행렬(correlation matrix)를 사용한다

### 3) 주성분 분석 해석

standard deviation(표준편차) : 자료의 산포도를 나타내는 수치로, 분산의 양의 제곱근, 표준편차가 작을수록 평균값에서 변량들의 거리가 가깝다

proportion of variance(분산비율) : 각 분산이 전체 분산에서 차지하는 비중

cumulative proportion(누적비율) : 분산의 누적 비율

## 2. 요인분석(Factor Analysis)

## 3. 판별분석(Discriminant Analysis)

## 4. 군집분석(Cluster Analysis)

## 5. 정준상관분석(Canonical Correlation Analysis)

## 6. 다차원척도법(Multi-dimensional scaling)

# 시계열

## 1. 시계열 모형

### 1) AR 모형 자기회귀모형

AR(p) : 현 시점의 자료가 p 시점 이전의 유한 개의 과거 자료로 설명될 수 있음

현시점의 시계열 자료에 과거 1 시점 이전의 자료만 영향을 주면 이를 1 차 자기회귀모형이라고 하고 AR(1)이라고 함

### 2) MA 모형 이동평균모형

최근 데이터의 평균을 예측치로 사용하는 방법, 각 과거치는 동일 가중치가 주어짐

현시점의 자료가 유한개의 과거 백색잡음(정상시계열)의 선형결합으로 표현되었기 때문에 항상 정상성을 만족함

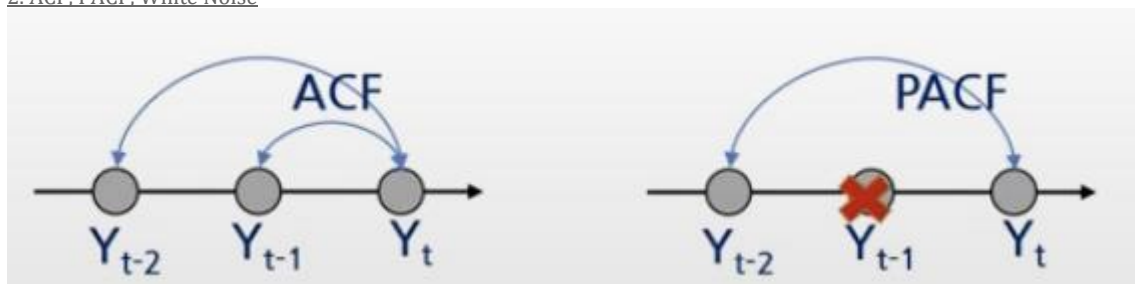
MA(p) : 과거 p 시점 이전 오차들에서 현재항의 상태를 추론한다

### 3) ARIMA 모형 자기회귀 누적 이동평균모형

현재와 추세간의 관계를 정의한 것, 많은 시계열 자료가 ARIMA 모형을 따름

ARIMA 모형은 비정상시계열 모형이며, 차분이나 변환을 통해 AR, MA, ARIMA 모형으로 정상화 할 수 있다

## 2. ACF, PACF, White Noise



### 1) 자기상관함수(ACF, Auto Correlation Function)

시계열 데이터의 자기상관성을 파악하기 위한 함수

시계열의 관측치 간의 상관계수를 k 의 함수 형태로 표시한 것 (k:시간단위)

k 가 커질수록 ACF 는 0 으로 수렴

$-1 \leq ACF \leq 1$

### 2) 부분자기상관함수(PACF, Partial ACF)

시계열의 관측치 중간에 있는 값들의 영향을 제외시킨, 관측값 사이의 직접적 상관관계를 파악하기 위한 함수

### 3) 백색잡음(White Noise)

시계열 자료 중 자기상관이 전혀 없는 특별한 경우  
시계열의 평균이 0, 분산이 일정한 값, 자기공분산이 0 인 경우  
현재값이 미래예측에 전혀 도움이 되지 못함, 회귀분석의 오차항과 비슷한 개념

### 3. 분해시계열

#### 1) 개념

시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법

#### 2) 분해시계열 분해 요인

- 추세요인 (Trend Factor) : 자료의 그림을 그렸을 때 그 형태가 오르거나 내리는 등 자료가 어떤 특정한 형태를 취할 때
- 계절요인 (Seasonal Factor) : 계절에 따라, 고정된 주기에 따라 자료가 변화하는 경우
- 순환요인 (Cyclical Factor) : 물가상승률, 급격한 인구 증가 등의 이유로 알려지지 않은 주기를 가지고 자료가 변화하는 경우
- 불규칙요인 (Irregular Factor) : 위 3 가지 요인으로 설명할 수 없는 회귀분석에서 오차에 해당하는 요인에 의해 발생하는 경우

## 대표적 데이터마이닝 기법

### 1. 분류(Classification)

새롭게 나타난 현상을 검토하여 기존의 분류, 정의된 집합에 배정하는 것  
의사결정나무, memory-based reasoning 등

### 2. 추정(Estimation)

주어진 입력 데이터를 사용하여 알려지지 않은 결과의 값을 추정하는 것  
연속된 변수의 값을 추정, 신경망 모형

### 3. 연관분석(Association Analysis)

같이 팔리는 물건과 같이 아이템의 연관성을 파악하는 분석  
카탈로그 배열 및 교차판매, 공격적 판촉행사 등의 마케팅 계획

### 4. 예측(Prediction)

미래에 대한 것을 예측, 추정하는 것을 제외하면 분류나 추정과 동일한 의미  
장바구니 분석, 의사결정나무, 신경망 모형

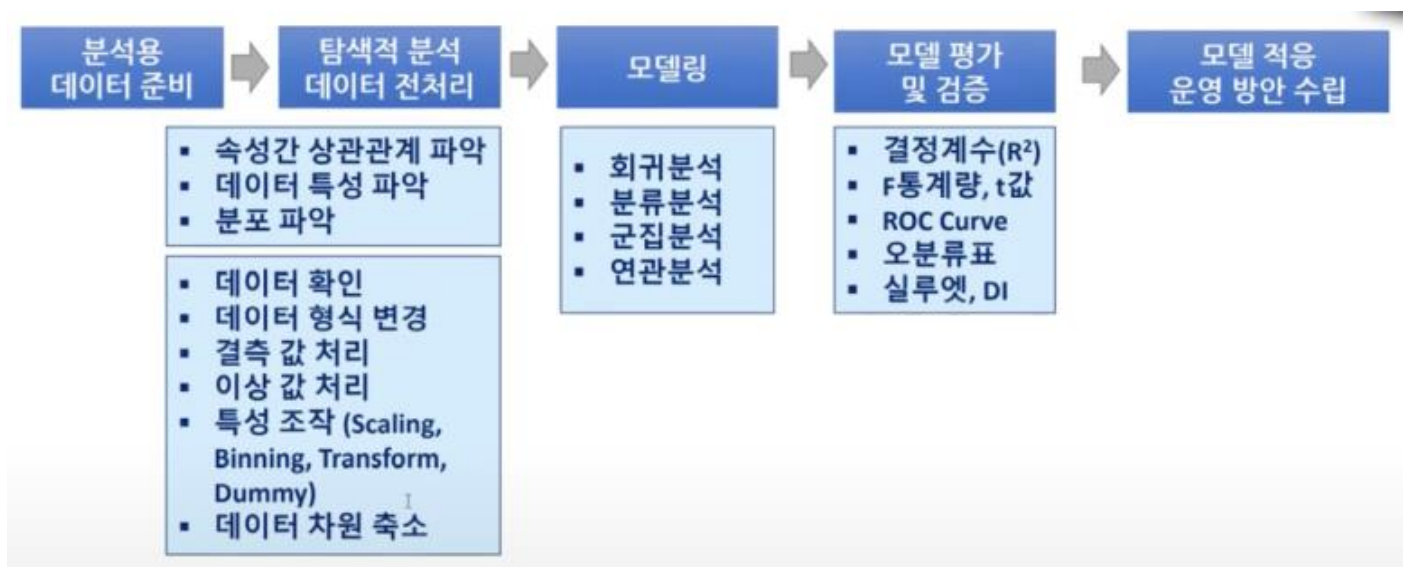
### 5. 군집(Clustering)

미리 정의된 기준이나 예시에 의해서가 아닌 레코드 자체가 가진 다른 레코드와의 유사성에 의해 그룹화되고 이질성에 의해 세분화 됨  
데이터 마이닝이나 모델링의 준비단계로서 사용됨

### 6. 기술(Description)

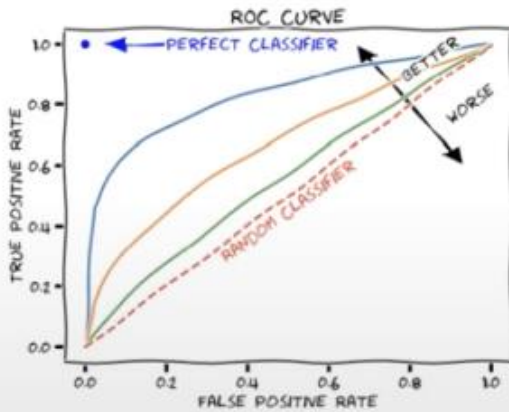
데이터가 가진 특징 및 의미를 단순하게 설명하는 것  
데이터가 암시하는 바에 대해 설명 및 그에 대한 답을 찾아낼 수 있어야함

## 데이터 분석 순서



ROC(Receiver Operation Characteristic) curve





- X축 : False positive rate (1 - Specificity)
- Y축 : True positive rate (Sensitivity)

Perfect classifier : 긍정, 부정 모두 다 맞추는 위치로 classification 성능이 우수하다고 봄,  $x=0, y=1$ 인 경우

- 분류 모형 성능 평가
- 다양한 threshold에 대한 이진분류기의 성능을 한번에 표시한 것이다
- X 축은 FP Rate(1-Specificity), Y 축은 민감도(Sensitivity)를 나타내 이 두 평가 값의 관계로 모형을 평가함
- ROC 그래프의 밑 부분 면적(AUC, Area Under the Curve)이 넓을 수록 좋은 모형으로 평가함

## Machine Learning Algorithms

	Supervised	Unsupervised
Continuous	<ul style="list-style-type: none"> <li>▪ <b>Regression</b> <ul style="list-style-type: none"> <li>▪ Linear</li> <li>▪ Polynomial</li> <li>▪ Ridge, Lasso</li> <li>▪ <u>kNN</u>, SVM</li> <li>▪ Decision Trees</li> <li>▪ Random Forests</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>▪ <b>Association Analysis</b> <ul style="list-style-type: none"> <li>▪ Apriori</li> <li>▪ FP-Growth</li> </ul> </li> <li>▪ <b>Dimensionality Reduction</b> <ul style="list-style-type: none"> <li>▪ PCA</li> <li>▪ SVD</li> </ul> </li> </ul>
Categorical	<ul style="list-style-type: none"> <li>▪ <b>Classification</b> <ul style="list-style-type: none"> <li>▪ Logistic Regression</li> <li>▪ Naïve-Bayes</li> <li>▪ <u>kNN</u>, SVM</li> <li>▪ Decision Trees</li> <li>▪ Random Forests</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>▪ <b>Clustering</b> <ul style="list-style-type: none"> <li>▪ K-means</li> <li>▪ DBSCAN</li> </ul> </li> </ul>

인공신경망 = 회귀, 분류 모두 가능

\* supervised(지도학습)

사전데이터를 기반으로 충분히 학습시키는 방법

Label(분류결과)과 Feature(결과에 영향을 주는 요소)를 파악

1. SVD(Singular value decomposition, 특이값 분해)

행렬을 특정한 구조로 분해하는 방식

2. PCA(Principal Component Analysis, 주성분 분석)

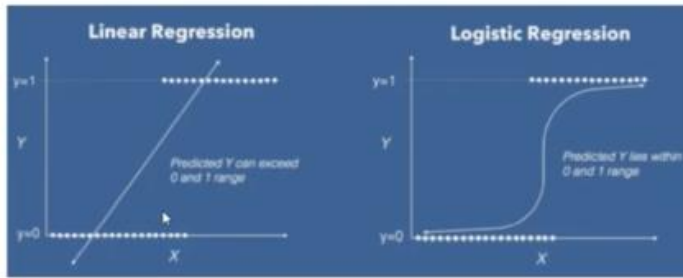
분산을 최대한 보존하면서 서로 직교하는 축을 찾아 고차원 공간의 표본들을 선형연관성이 없는 저차원공간으로 변환하는 기법

3. K-means (K 평균 군집화)

대표적인 분리형 군집화 알고리즘 가운데 하나

각 군집은 하나의 중심을 가지고, 각 개체는 가장 가까운 중심에 할당되며 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성  
각 군집중심의 초기값을 랜덤하게 정하는 알고리즘인데, 초기값 위치에 따라 원하는 결과가 나오지 않을수도 있음

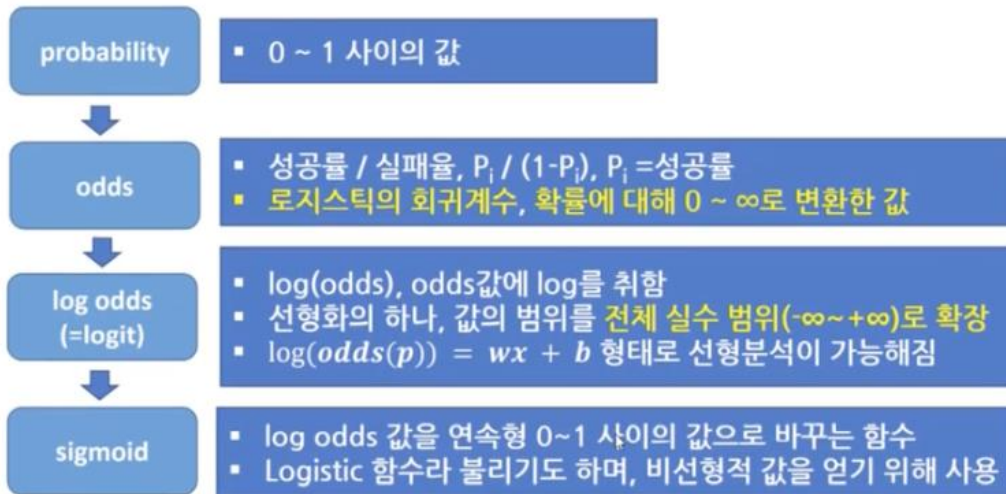
3. Regression



- 선형회귀
  - x값의 변화에 따른 y값의 변화를 알아냄
  - x가 1증가할 때, y는 회귀계수 만큼 증가함
- 로지스틱회귀
  - x값에 따른 y값의 변화량의 문제가 아님!
  - 회귀계수를 해석할 때 문제가 생김!

	일반 선형 회귀분석	로지스틱 회귀분석
종속변수	연속형 변수	이산형(범주형) 변수
모형 탐색 방법	최소자승법(LSM, 최소제곱법)	최대우도법(MLE), 가중최소자승법
모형 검정	F-test, T-test	$\chi^2$ test

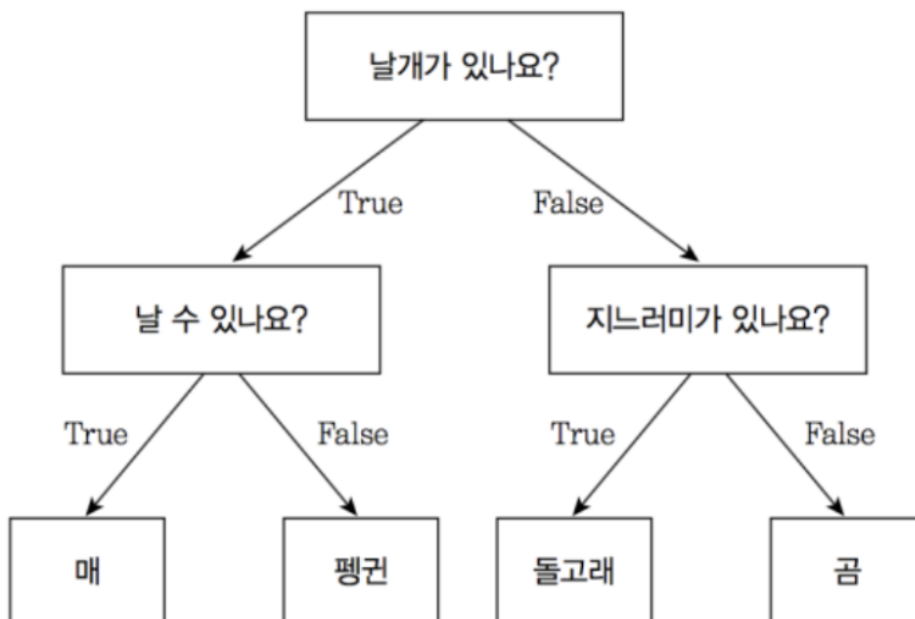
- 1) 일반선형 회귀분석
- 2) 로지스틱 회귀분석



회귀 식에 대한 해석 방법이 선형회귀와 다름

독립변수의 선형결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계기법  
 종속변수가 성공/실패, 사망/생존과 같이 이항변수(0,1)로 되어있을때 종속변수와 독립변수 간의 관계식을 이용하여 두 집단 또는 그 이상의 집단을 분류하고자 할때 사용되는 분석기법

#### 4. Decision Trees (의사결정트리)



- 1) 개념

분류(Classification)과 회귀(Regression) 모두 가능한 지도 학습 모델 중 하나  
 의사결정 규칙을 나무구조로 나타내 전체 자료를 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석 방법  
 분석과정이 직관적이고 이해하기 쉬움  
 목적은 새로운 데이터를 분류하거나 값을 예측하는 것이다  
 분리변수 P 차원공간에 대한 현재 분할은 이전 분할에 영향을 받는다  
 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해나간다(불순도감소)

## 2) 종류

독립변수	종속변수
<ul style="list-style-type: none"> <li>설명변수</li> <li>예측변수</li> <li>Feature</li> </ul>	<ul style="list-style-type: none"> <li>목표변수</li> <li>반응변수</li> <li>Label</li> </ul>

분류나무(classification) : 목표변수(=종속변수)가 이산형(범주형)인 경우  
 회귀나무(regression tree) : 목표변수가 연속형인 경우

## 3) 장점

비모수적 모형으로 선형성, 정규성, 등분산성 등의 수학적 가정이 불필요함  
 범주형(이산형)과 수치형(연속형) 변수를 모두 사용할 수 있음

## 4) 단점

분류기준값의 경계선 부근의 자료 값에 대해서는 오차가 큼(비연속성)  
 로지스틱회귀와 같이 각 예측변수의 효과를 파악하기 어려움  
 새로운 자료에 대한 예측이 불안정할 수 있음

## 5) 분리기준 (split criterion)

순수도가 높아지는 방향으로 분리  
 불확실성이 낮아지는 방향

## 6) 정지규칙(stopping rule)

더이상의 분리가 일어나지 않고 현재의 마디가 최종마디가 되도록 하는 규칙  
 '불순도 감소량'이 아주 작을때 정지함

## 7) 가지치기 규칙(pruning rule)

어느가지를 쳐내야 예측력이 좋은 나무가 될까?  
 최종 노드가 너무 많으면 Overfitting 가능성이 커짐, 이를 해결하기 위해 사용  
 가지치기의 비용함수(Cost Function)을 최소로 하는 분기를 찾아내도록 학습  
 Information Gain 이란 어떤 속성을 선택함으로써 인해 데이터를 더 잘 구분하게 되는 것을 의미함(불확실성 감소)

## 8) 불순도 측정 지표

목표변수가 범주형일때 사용하는 지표 (분류나무에서 사용)  
 지니지수 : 불순도 측정 지표, 값이 작을수록 순수도가 높음(분류 잘됨)  
 엔트로피 지수 : 불순도 측정 지표, 가장 작은 값을 갖는 방법 선택  
 카이제곱 통계량의 유의 확률(p-value) : 가장 작은 값을 갖는 방법 선택

## 9) 알고리즘

의사결정 나무를 위한 알고리즘은 CHAID, CART, C5.0 가 있으면, 하향식 접근방법을 이용한다  
 정지기준변수 선택법

알고리즘	이산형 목표변수 (분류나무)	연속형 목표변수 (회귀나무)
<b>CART (Classification And Regression Tree)</b>	<b>지니지수</b>	<b>분산 감소량</b>
<b>C5.0</b>	<b>엔트로피지수</b>	
<b>CHAID (Chi-squared Automatic Interaction Detection)</b>	<b>카이제곱 통계량의 p-value</b>	<b>ANOVA F-통계량 - p-value</b>

## 5. Random Forests

분류(Classification)과 회귀(Regression) 모두 가능한 지도 학습 모델 중 하나  
 다수의 결정트리들을 학습하는 앙상블 방법이다

### \* 앙상블

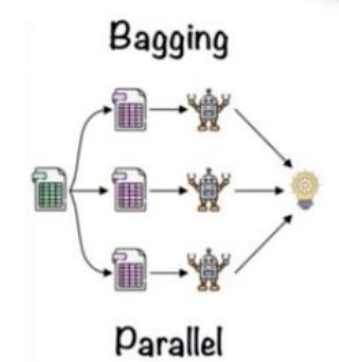
여러개의 분류 모형에 의한 결과를 종합하여 분류의 정확도를 높이는 방법

성능을 분산시키기 때문에 과적합(overfitting) 감소 효과가 있음

서로 다른 여러개 알고리즘 분류기 사용

각 모델의 결과를 취합하여 많은 결과 또는 높은 확률로 나온 것을 최종 결과로 채택하는 것

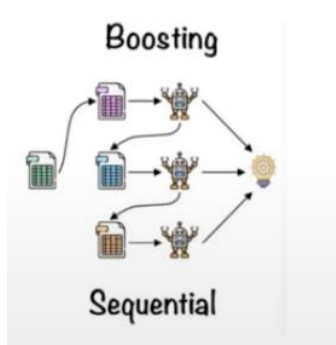
1) 배깅(Bagging, Bootstrap AGGregatING)



서로 다른 훈련 데이터 샘플로 훈련, 서로 같은 알고리즘 분류기 결합

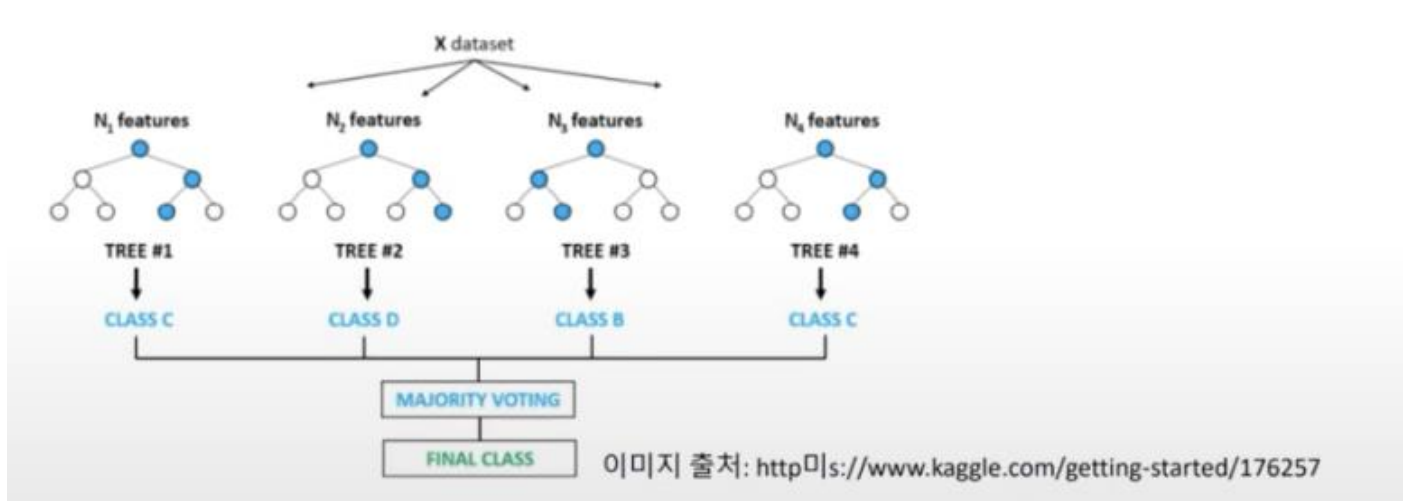
여러모델이 병렬로 학습, 그 결과를 집계하는 방식

2) 부스팅(Boosting)



여러 모델이 순차적으로 학습

3) 랜덤 포레스트(Random forest)



배깅에 랜덤과정을 추가한 방법

설명변수의 일부분만을 고려

여러개 의사결정나무를 사용해, 과적합 문제를 피할 수 있음

## 6. KNN (k 최근접 이웃 알고리즘)

1) 개념

분류(Classification)과 회귀(Regression) 모두 가능한 지도 학습 모델 중 하나

두 경우 모두 입력이 특징공간 내 k 개의 가장 가까운 훈련데이터로 구성되어있다

k 값에 따라 소속되는 그룹이 달라질 수 있음 (k 값은 hyper parameter)

거리를 측정해 이웃들을 뽑기 때문에 스케일링이 중요함

모형을 미리 만들지 않고, 새로운 데이터가 들어오면 그때부터 계산을 시작하는 lazy learning(게으른 학습)이 사용되는 지도학습 알고리즘

2) 종류 (출력 : knn 이 분류로 사용되었는지 회귀로 사용되었는지에 따라 다르다)

분류 : 반응변수가 범주형

회귀 : 반응변수가 연속형

## 7. SVM (support vector machine)

분류(Classification)과 회귀(Regression) 모두 가능한 지도 학습 모델 중 하나

주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진선형분류모델을 만들고, 만들어진 분류모델은 데이터가 사상된 공간에서 경계로 표현되는데, SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘

## 8. Apriori(연관규칙분석)

두 아이템 집합이 빈번히 발생하는가를 알려주는 일련의 규칙들을 생성하는 알고리즘

장바구니 분석으로 널리 알려져있는 방법론

## 9. FP-Growth

빈번한 패턴 마이닝에 널리 사용되는 Apriori Algorithm의 개선된 버전

데이터 세트에서 빈번한 패턴이나 연관성을 찾는 분석 프로젝트로 사용됨

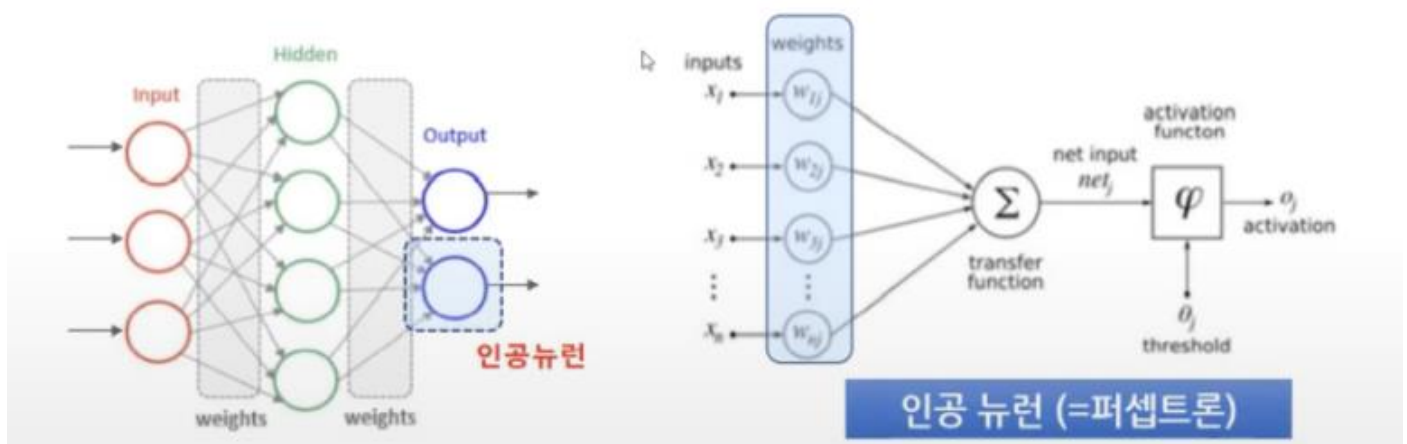
## 10. Naive-Bayes (나이브 베이즈)

스팸 메일 필터, 텍스트 분류, 감정분석, 추천 시스템 등에 광범위하게 활용되는 분류기법

베이즈 정리에 기반한 통계적 분류 기법으로 가장 단순한 지도학습(supervised learning) 중 하나이다

빠르고 정확하며 믿을만한 알고리즘이며 대용량 데이터에 대해 속도도 빠름

# 인공신경망(ANN) 모형



## 1. 개념

- 인공신경망을 이용하면 분류 및 군집을 할 수 있음
- 인공신경망은 입력층, 은닉층, 출력층 3 개의 층으로 구성되어 있음
- 각 층에 뉴런(노드)이 여러 개 포함되어 있음
- 학습 : 입력에 대한 올바른 출력이 나오도록 가중치(weight)를 조절하는 것
- 가중치 초기화는 -1.0~1.0 사이의 임의 값으로 설정하며, 가중치를 지나치게 큰 값으로 초기화하면 활성화 함수를 편향 시키게 되며, 활성화 함수가 과적합 되는 상태를 포화상태라고 함

## 2. 장점

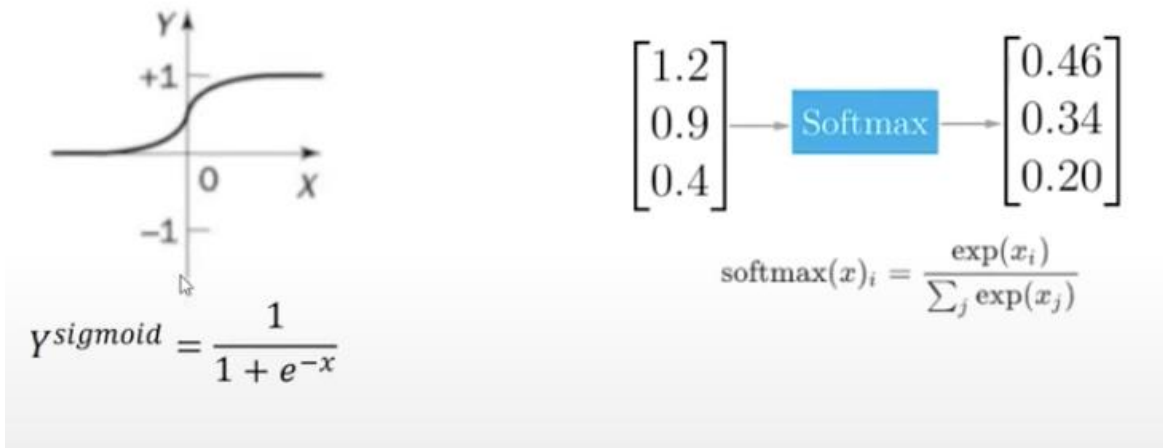
- 변수의 수가 많거나 입,출력변수 간에 복잡한 비선형 관계에 유용
- 이상치 잡음에 대해서도 민감하게 반응하지 않음
- 입력변수와 결과변수가 연속형이나 이산형인 경우 모두 처리 가능

## 3. 단점

- 결과에 대한 해석이 쉽지 않음
- 최적의 모형을 도출하는 것이 상대적으로 어려움
- 모형이 복잡하면 훈련 과정에 시간이 많이 소요됨
- 데이터를 정규화하지 않으면 지역해(local minimum)에 빠질 위험이 있음

## 4. 신경망 활성화 함수(activation function)





- 결과값을 내보낼때 사용하는 함수로, 가중치 값을 학습할 때 에러가 적게 나도록 도움
- 풀고자 하는 문제 종류에 따라 활성화 함수 선택이 달라짐
- 목표 정확도와 학습시간을 고려하여 선택하고 혼합 사용도 함
- 문제 결과가 직선을 따르는 경향이 있으면 '선형함수'를 사용
- sigmoid 함수(= Logistic 함수) : 연속형 0~1
- softmax 함수 : 모든 logits의 합이 1이 되도록 output을 정규화, sigmoid 함수의 일반화된 형태로 결과가 다 범주인 경우 각 범주에 속할 사후 확률(posterior probability) 제공하는 활성화 함수

### 5. 신경망 은닉 층, 은닉 노드

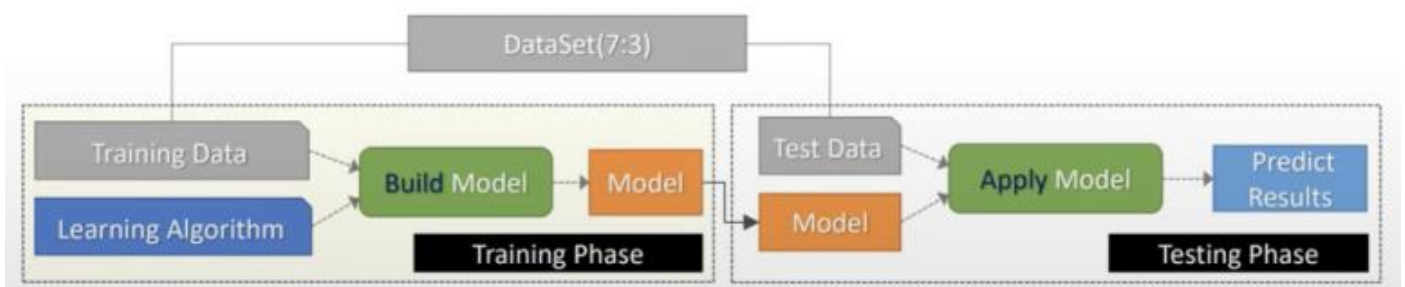
- 다층신경망은 단층신경망에 비해 훈련이 어려움
- 은닉 층 수와 은닉 노드 수의 결정은 '분석가가 분석경험에 의해 설정'함
- 은닉 층 노드가 너무 적으면 : 네트워크가 복잡한 의사결정 경계를 만들 수 없음, Underfitting 문제 발생
- 은닉 층 노드가 너무 많으면 : 복잡성을 잡아낼 수 있지만, 일반화가 어렵다 / 레이어가 많아지면 기울기 소실문제가 발생할 수 있다 / 과적합(Overfitting) 문제 발생
- 역전파 알고리즘(Backpropagation Algorithm) : 출력층에서 제시한 값에 대해, 실제 원하는 값으로 학습하는 방법으로 사용 / 동일 입력층에 대해 원하는 값이 출력되도록 개개의 weight를 조정하는 방법으로 사용됨
- 기울기 소실 문제(Vanishing Gradient Problem) : 다층신경망에서 은닉층이 많아 인공신경망 기울기 값을 베이스로하는 역전파 알고리즘으로 학습시키려고 할 때 발생하는 문제

## 경사하강법(Gradient descent)

- 함수 기울기를 낮은 쪽으로 계속 이동시켜 극값에 이를때까지 반복시키는 것
- 제시된 함수의 기울기를 최소값을 찾아내는 머신러닝 알고리즘
- 비용함수(cost function)을 최소화 하기 위해 parameter를 반복적으로 조정하는 과정
- 과정
  - 1) 임의의 parameter 값으로 시작
  - 2) Cost Function 계산, cost function - 모델을 구성하는 가중치 w의 함수, 시작점에서 곡선의 기울기 계산
  - 3) parameter 값 갱신 :  $W = W - \text{learning rate} * \text{기울기 미분값}$
  - 4) n 번의 iteration, 최소값을 향해 수렴함. learning rate가 적절해야함

## 모형 평가

### 1. 홀드아웃(Hold Out)



- 원천 데이터를 랜덤하게 두 분류로 분리하여 교차검정을 실시하는 방법으로 하나는 모형 학습 및 구축을 위한 훈련용 자료로 다른 하나는 성과평가를 위한 검증용 자료로 사용하는 방법
- 과적합 발생여부를 확인하기 위해서 주어진 데이터의 일정 부분을 모델을 만드는 훈련 데이터로 사용하고, 나머지 데이터를 사용해 모델을 평가

- 잘못된 가정을 하게 되는 2종 오류의 발생을 방지
- iris 데이터를 7:3 비율로 나누어 Training에 70%, Testing에 30% 사용하도록 하는 것

## 2. 교차검증(Cross Validation)

- 데이터가 충분하지 않을 경우 Hold out으로 나누면 많은 양의 분산 발생
- 이에 대한 해결책으로 교차검증을 사용할 수 있음. 그러나 클래스 불균형 데이터에는 적합하지 않음
- 주어진 데이터를 가지고 반복적으로 성과를 측정하여 그 결과를 평균한 것으로 분류 분석 모형의 평가 방법
- k-fold cross validation

전체 데이터를 shuffle

k 개로 데이터를 분할

k 번째의 하부집합을 검증용 자료, K-1 개는 훈련용자료로 사용하여 k 번 반복 측정

결과를 평균낸 값을 최종 평가로 사용함

## 3. 부트스트랩(Bootstrap)

- 평가를 반복하는 측면에서 교차검증과 유사하지만, 훈련용 자료를 반복 재선정한다는 점에서 차이가 있는 평가 방법
- 부트스트랩은 관측치를 한 번 이상 훈련용 자료로 사용하는 복원추출법에 기반함
- 전체 데이터 양이 크지 않을 경우와 모형 평가에 가장 적합
- 훈련데이터를 63.2% 사용하는 0.632 부트스트랩이 있음

## 4. 데이터 분할 시 고려사항

- class 비율이 한 쪽에 치우쳐 있는 클래스 불균형 상태라면 다음 기법 사용을 고려한다
- under sampling : 적은 class의 수에 맞추는 것
- over sampling : 많은 class의 수에 맞추는 것

## 오분류표를 활용한 평가지표

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

T/F P/N	
실제 == 예측 : True	True 예측 : Positive
실제 != 예측 : False	False 예측 : Negative

T	P	T	N	F	P	F	N
---	---	---	---	---	---	---	---

confusion matrix		실제값	
		Y	N
예측값	Y	True Positive	False Positive
	N	False Negative	True Negative

## 1. 오분류표를 활용한 평가지표

### 1) 정밀도(Precision)

- 예측값이 True 인 것에 대해 실제값이 True 인 지표

### 2) 재현율(Recall), 민감도(Sensitivity)

- 실제값이 True 인 것에 대해 예측값이 True 인 지표

### 3) F1

- 데이터가 불균형할때 사용한다

- 오분류표 중 정밀도와 재현율의 조화평균을 나타내며 정밀도와 재현율에 같은 가중치를 부여하여 평균한 지표 :  $2 * (정밀도 * 재현율) / (정밀도+재현율)$

### 4) Accuracy

- 전체 예측에서 옳은 예측의 비율 :  $(TP+TN) / (TP+FP+FN+TN)$

### 5) Error Rate

- 전체예측에서 틀린 예측의 비율 :  $(FP+FN) / (TP+FP+FN+TN)$

### 6) 특이도(Specificity)

- 실제로 N 인 것들 중 예측이 N 으로 된 경우의 비율 :  $(TN) / (TN+FP)$

7) FP Rate

- 실체가 N 인데 예측이 P 로 된 비율 (Y 가 아닌데 Y 로 예측된 비율, 1 종 오류) :  $(FP) / (FP+TN)$

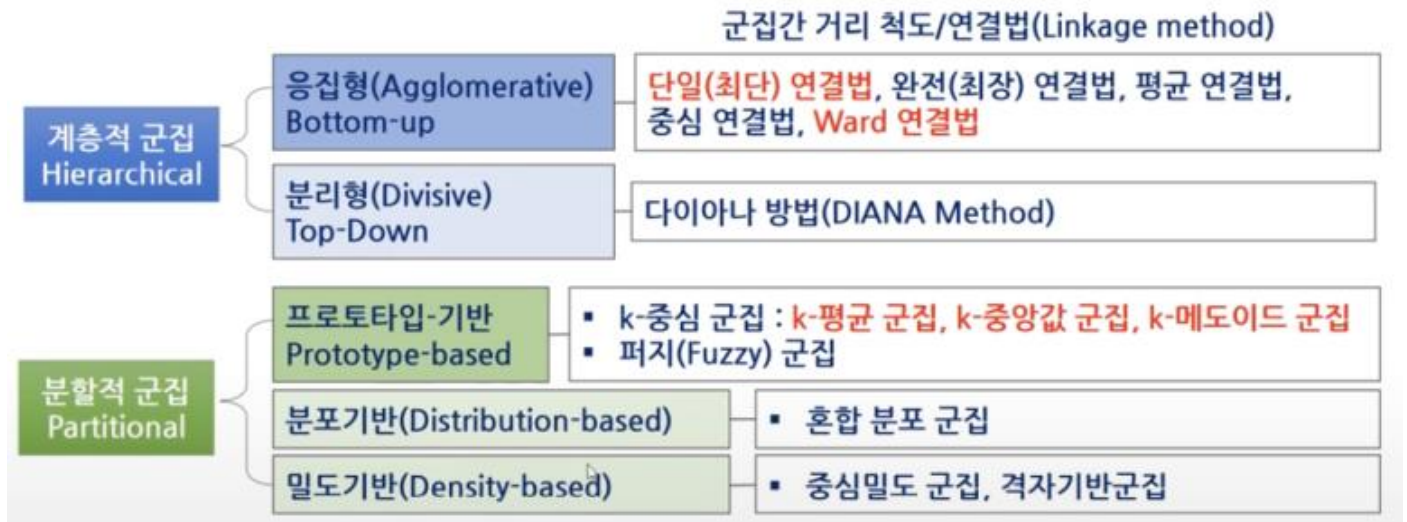
8) kappa

- 두 평가자의 평가가 얼마나 일치하는지 평가하는 값으로 0~1 사이의 값을 가짐 :  $(Accuracy - P(e)) / (1 - P(e)) * P(e)$  : 우연히 일치할 확률

9) F2

- 재현율에 정밀도의 2 배 만큼 가중치를 부여하는 것

## 군집분석(Clustering Analysis)의 종류

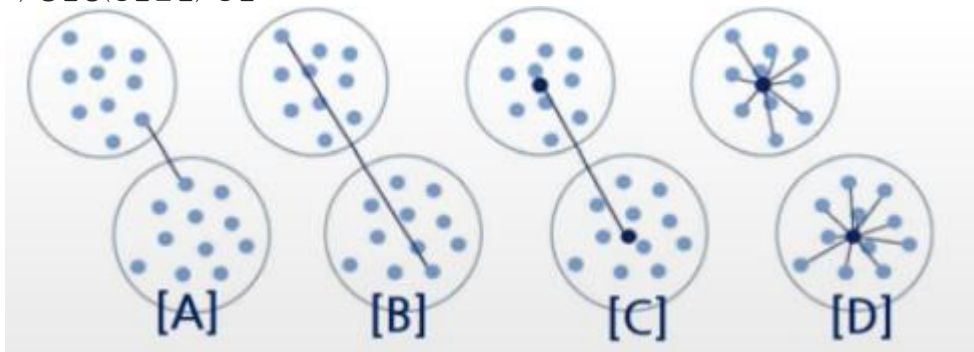


- 여러 변수 값들로부터 n 개의 개체를 유사한 성격을 가지는 몇 개의 군집으로 집단화하고 형성된 군집들의 특성을 파악해 군집들 사이의 관계를 분석하는 다변량분석 기법

### 1. 계층적 군집(Hierarchical Clustering)

- 가장 유사한 개체를 묶어 나가는 과정을 반복하여 원하는 개수의 군집을 형성하는 방법
- 유사도 판단은 두 개체 간의 거리에 기반하므로 거리 측정에 대한 정의가 필요함 (유클리드, 맨해튼, 민코프스키, 마할라노비스)
- 이상치에 민감함
- 사전에 군집수 k 를 설정할 필요가 없는 탐색적 모형
- 군집을 형성하는 데 매 단계에서 지역적 최적화를 수행해 나가는 방법을 사용하므로 그 결과가 전역적인 최적해라고 볼 수 없음
- 병합적 방법에서 한 번 군집이 형성되면 군집에 속한 개체는 다른 군집으로 이동할 수 없음

#### 1) 응집형(병합군집) 방법

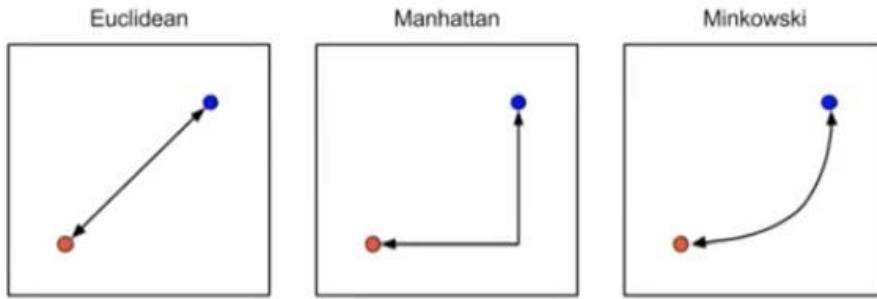


- A 최단 연결법 : 단일연결법이라고도 하며, 거리의 최솟값 측정
- B 최장 연결법 : 완전연결법이라고도 하며, 거리의 최댓값을 측정
- C 중심 연결법 : 두 군집 중심간의 거리를 측정함
- D 와드 연결법 : 오차제곱합에 기초하여 군집을 수행하는 군집방법
- E 평균 연결법 : 모든 항목에 대한 거리 평균을 구하면서 군집화, 계산량이 많아질 수 있음

#### 2) 계층형 군집의 거리

- 수학적 거리 개념





[https://subscription.packtpub.com/book/big\\_data\\_and\\_business\\_intelligence/9781785882104/6/ch06lvl1sec40/measuring-distance-or-similarity](https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785882104/6/ch06lvl1sec40/measuring-distance-or-similarity)

유클리드(Euclidean) : 두 점 사이의 직관적이고 일반적인 거리 개념, 방향성이 고려되지 않음

맨해튼(Manhattan) : 두 점 각 성분별 차의 절대 합

민코프스키(Minkowski) : 거리 차수와 함께 사용 ( $q=2$  이면 Euclidean,  $q=1$  이면 Manhattan Distance)

- 통계적 거리 개념

표준화 : 각 변수를 해당 변수의 표준편차로 척도 변환한 후에 유클리드 거리를 계산한 것으로 통계적 거리(Statistical distance)라고도 함.  
척도의 차이, 분산의 차이로 인한 왜곡을 피할 수 있음

마할라노비스 : 변수의 표준화와 함께 변수 간의 상관성을 동시에 고려한 통계적 거리

- 함수

dist 함수 : 거리측정에 사용하는 함수로 사용가능한 거리 개념으로 유클리드, 맨해튼, 민코프스키, maximum, canberra, binary 등이 있음  
코사인(cosine) 거리 : 두 벡터 사이의 사잇각을 계산해서 유사한 정도를 구하는 것, 1 인 경우 유사도가 크며, -1 인 경우 유사도가 매우 작음을 의미

## 2. 비계층형 군집(Non-Hierarchical Clustering)

### 1) k-means

- 사전에  $k$  를 정해줘야함 ( $k$  : hyper parameter)
- 군집수  $k$  가 원데이터 구조에 적합하지 않으면 좋은 결과를 얻을 수 없음
- 알고리즘이 단순하며 빠르게 수행되어 계층형 군집보다 많은 양의 자료를 처리
- k-means 군집은 잡음이나 이상값에 영향을 받기 쉬움
- k-means 분석 전에 이상값을 제거하는 것도 좋은 방법
- 평균대신 중앙값을 사용하는 k-medoids 군집을 사용할 수 있음
- 절차

초기군집의 중심으로  $k$  개의 객체를 임의로 선택

각 자료를 가장 가까운 군집의 중심에 할당

각 군집내의 자료들의 평균을 계산하여 군집의 중심을 갱신

군집중심의 변화가 거의 없을때까지 반복한다

- DBSCAN

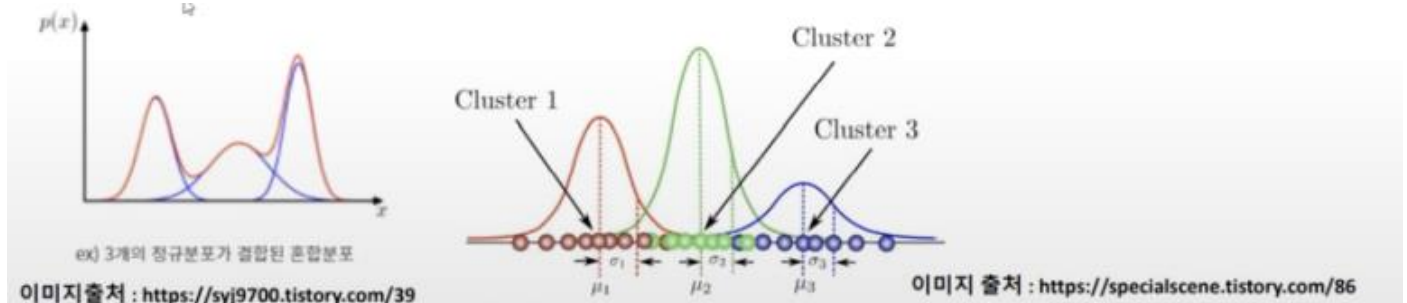
밀도기반 클러스터링으로 점이 세밀하게 몰려있어 밀도가 높은 부분을 클러스터링함

어느 점을 기준으로 환경내에 점이  $n$  개 이상 있으면 하나의 군집으로 인식하는 방식

Gaussian 분포가 아닌 임의적 모양의 군집분석에 적합함

$k$  값을 정할 필요가 없음, outlier에 의한 성능 하락을 완화할수 있음

### 2) 혼합분포군집



- 데이터가 봉우리가 2 개인 분포, 도넛형태의 분포 등 복잡한 형태를 가진 분포의 경우 여러분포를 확률적으로 선형 결합한 혼합분포로 설명될 수 있음

- 데이터가  $k$  개의 모수적 모형의 가중합으로 표현되는 모집단 모형에서 나왔다는 가정하에, 추정된  $k$  개의 모형 중 어느 모형으로부터 나왔을 확률이 높은지에 따라 군집분류를 수행

- 모수와 가중치 추정에 EM 알고리즘이 사용됨 (Expected Maximization)

### \* EM 알고리즘

모수(평균, 분산, 혼합계수)에 대해 임의의 초기값을 정함

잠재변수(latent variable) : 어느 집단에 속하는지에 대한 정보를 갖는 변수

E-step : k 개의 모형 군집에 대해 모수를 사용해 각 군집에 속할 사후확률을 구함

M-step : 사후확률을 이용해 최대 우도 추정으로 모수를 다시 추정하고, 이를 반복함

### 3. 군집화 평가지수

#### 1) 실루엣 계수(Silhouette Coefficient)

1에 가까울수록 군집화가 잘 되었다고 함

0.5 보다 크면 결과가 타당한 것으로 평가

#### 2) Dunn Index(DI)

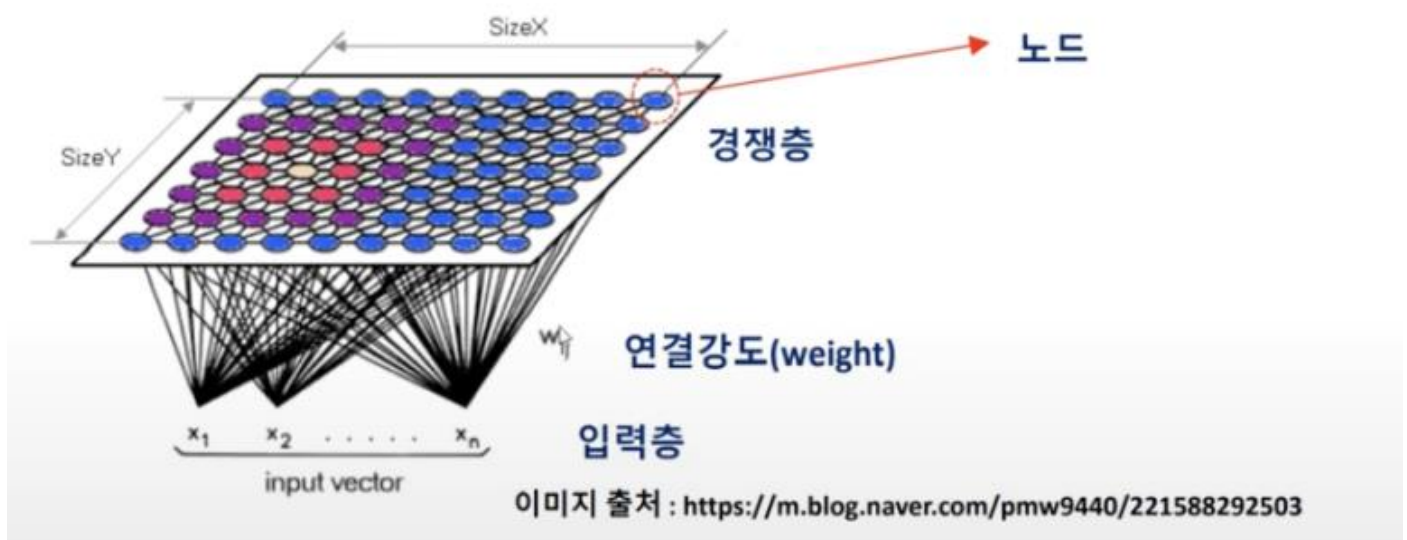
군집과 군집 사이 거리 중 최솟값 / 군집 내 데이터들 거리 중 최댓값

분자가 클 수록 군집간 거리가 멀고, 분모가 클수록 군집 내 데이터가 모여 있음

Dunn Index 가 클수록 군집화가 잘 되었다고 평가

## SOM(Self-Organizing Maps)

### 1. 개념



- 자기조직화지도
- 인공신경망의 한 종류로, 차원축소와 군집화를 동시에 수행하는 기법
- 비지도학습(Unsupervised Learning)의 한 기법
- 고차원으로 표현된 데이터를 저차원으로 변환해서 보는데 유용함
- 입력층과 2 차원의 격자 형태의 경쟁층(출력층)으로 이루어져 있음(2 개의 층으로 구성)

### 2) SOM Process

단계 1 : SOM 의 노드에 대한 연결강도(weight) 초기화

단계 2 : 입력 벡터와 경쟁 층 노드 간의 거리 계산 및 입력벡터와 가까운 노드 선택 -> 경쟁

단계 3 : 경쟁에서 선택된 노드와 이웃 노드의 가중치(연결강도) 갱신 -> 협력 및 적응

단계 4 : 단계 2 로 가서 반복

-> 승자만이 출력을 내고, 승자와 그의 이웃만이 연결강도를 수정하는 승자 독점구조로 인해 경쟁층에서는 승자누런만 나타남

### 3) SOM vs 신경망 모형

신경망 모형은 연속적인 layer 로 구성된 반면, SOM 은 2 차원의 그리드(격자)로 구성

신경망 모형은 에러수정을 학습하는 반면 SOM 은 경쟁학습 실시

신경망은 역전파 알고리즘이지만, SOM 은 전방패스를 사용해 속도가 매우 빠름

## 연관분석(Association Analysis)

### 1. 연관분석

- 항목들간의 '조건-결과' 식으로 표현되는 유용한 패턴
- 이러한 패턴 규칙을 발견해 내는 것을 연관분석이라함
- 장바구니 분석이라고함(미국 마트에서 기저귀를 사는 고객은 맥주를 동시에 구매한다는 연관규칙을 알아낸 것에 기인함)

### 2. Apriori 알고리즘

연관규칙의 대표적 알고리즘으로 현재도 많이 사용됨

데이터들에 대한 발생빈도를 기반으로 각 데이터 간의 연관관계를 밝히는 방법  
데이터셋이 큰 경우 모든 후보 itemset에 대해 하나하나 검사하는 것이 비효율적임

### 3. FP Growth

Apriori 단점을 보완하기 위해 FP-tree와 node, link라는 특별한 자료구조를 사용

### 4. 장점

조건반응(if-then)으로 표현되는 연관분석의 결과를 이해하기 쉬움

강력한 비목적성 분석 기법이며, 분석 계산이 간편함

### 5. 단점

분석 품목수가 증가하면 분석 계산이 기하급수적으로 증가함

너무 세분화된 품목을 가지고 연관규칙을 찾으려면 의미없는 분석 결과가 도출됨

상대적 거래량이 적으면 규칙 발견 시 제외되기 쉬움

### 6. 연관규칙 측정 지표

#### 1) 지지도(Support)

전체 거래항목 중 상품 A와 B를 동시에 포함하여 거래하는 비율

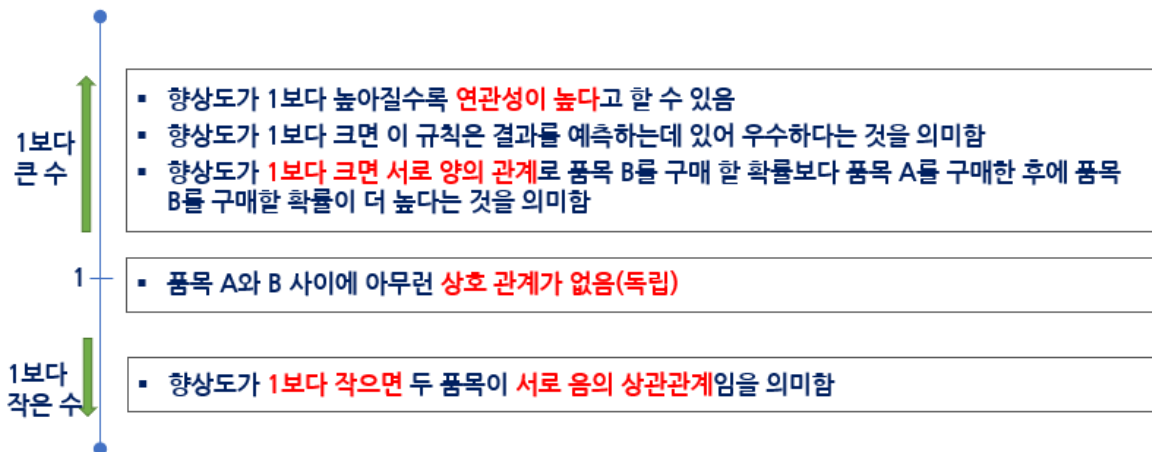
지지도 =  $P(A \text{ AND } B)$  : A와 B가 동시에 포함된 거래수 / 전체 거래수

#### 2) 신뢰도(Confidence)

상품 A를 포함하는 거래 중 A와 B가 동시에 거래되는 비율

신뢰도 =  $P(B|A) = P(A \text{ AND } B) / P(A)$  : A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수

#### 3) 향상도(Lift)



A가 주어지지 않았을 때 B의 확률 대비 A가 주어졌을 때 B 확률 증가 비율

품목 B를 구매한 고객 대비 품목 A를 구매한 후 품목 B를

품목 B를 구매한 고객 대비 품목 A를 구매한 후 품목 B를 구매하는 고객에 대한 확률

향상도 =  $P(B|A)/P(B) = P(A \text{ AND } B) / (P(A) * P(B))$