1과목

1장 데이터의 이해

1. 데이터와 정보

• 데이터의 유형

- 정성적 데이터 : 저장, 검색, 분석에 많은 비용이 소모되는 언어, **문자** 형태의 데이터
  (예. 회사 매출이 증가함)
- 정량적 데이터 : 정형화된 데이터로 수치, 도형, 기호 등의 형태를 가진 데이터 **도형, 기호, 수치**
  (예. 나이, 몸무게, 주가 등)

• 지식영영의 핵심 이슈

구분	의미	특징	상호작용
암묵지	학습과 경험을 통해 개인에게 체화되어 있지만 겉으로 드러나지 않는 지식 (김장김치 담그기, 자전거 타기)	사회적으로 중요하지만 공유되기 어려움	공통화 내면화
형식지	문서나 매뉴얼처럼 형상화된 지식 (교과서, 비디오, DB)	전달과 공유가 용이함	표출화 연결화

• DIKW 피라미드

2. 데이터베이스 정의와 특징

• 용어의 연혁

1950 년대 미국에서 군대의 군비상황을 집중 관리하기 위하여 컴퓨터 도서관 설립
1975 년 미국의 CAC가 KORSTIC을 통해 서비스되면서 우리나라에서 데이터베이스 이용이 이루어짐
1980 년대 중반 국내의 데이터베이스 관련 기술의 연기, 개발

• 데이터베이스 정의

EU	체계적이거나 조직적으로 정리되고 전자식 또는 기타 수단으로 개별적으로 접근할 수 있는 독립된 저작물, 데이터 또는 기타 소재의 수집물
국내 저작권법	소재를 체계적으로 배열 또는 구성한 편집물로 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것
국내 컴퓨터용어사전	동시의 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합

• 데이터베이스 특징

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

• OLTP 데이터 갱신 위주

• OLAP 조회 위주, 다차원의 데이터를 대화 식으로 분석하기 위한 기술

2장 데이터의 가치와 미래

1. 빅데이터의 이해

• 빅데이터의 정의

(1) 관점에 따른 정의

- 데이터 규모에 중점을 둔 정의 (Mckinsey,2011)
  일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터
- 분석 비용 및 기술에 초점을 둔 정의 (IDC, 2011)
  다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고,
  데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍쳐
- 가트너 그룹 더니 래그의 3V
  Volune 규모 / Variety 유형과 소스 / Velocity 수집과 처리

(2) 빅데이터 정의의 범주 및 효과

데이터 변화 (3V)

→ 기술 변화 (데이터 처리, 저장, 분석기술 및 아키텍처 / 클라우드 컴퓨팅 활용)

→ 인재, 조직 변화 (data scientist, 데이터 중심 조직)

• 출현 배경과 변화

- 산업계 : 고객 데이터 축척, 보유를 통해 데이터에 숨어있는 가치 발굴
- 학계 : 거대 데이터를 다루는 학문 분야가 늘어나며 필요한 기술 및 통계 도구의 발전
  - 기술발전 : 관련기술 (저장, 인터넷 보급, 클라우드 컴퓨터, 모바일 혁명)의 발달

• 빅데이터에 거는 기대의 비유적 표현

- 산업혁명의 석탄과 철(서비스) / 21세기의 원유(생신상) / 렌즈(생물학) / 플랫폼(공동활용)

• 빅데이터가 만들어 내는 본질적인 변화

사전처리 → 사후처리 / 표본조사 → 전수조사 / 질 → 양 / 인과관계 → 상관관계

2. 빅데이터의 가치와 영향
• 빅데이터의 가치 산정이 어려운 이유
데이터 활용방식, 새로운 가치 창출, 분석 기술 발전
• 빅데이터의 영향
- 기업 : 혁신, 경쟁력제고, 생산성 향상
– 빅데이터를 활용해 소비자의 행동 분석, 시장 변동 예측, 비즈니스 모델 혁신하거나 신사업 발굴
- 정부 : 환경 탐색, 상황분석, 미래 대응
– 기상,인구이동, 각종 통계, 법제 데이터 등을 수집해 사회 변화를 추정, 정보 추출
- 개인 : 목적에 따른 활용
– 비용이 지속적으로 하락하여 정치인, 가수 등이 인지도 향상에 활용

3. 비즈니스 모델
• 빅데이터 활용사례
- 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 위해 실시간 모니터링 실시
- 개인 : 정치인, 가수
• 빅데이터 활용 기본 테크닉
연관 규칙 학습, 군집분석, **유전 알고리즘**, 기계학습, 회귀분석, 감정분석, 소셜네트워크분석
*유전자 알고리즘 : 생명의 진화를 모방하여 최적해를 구하는 알고리즘, 존 홀랜드 발명

4. 위기요인과 통제 방안
사생활 침해 → 동의에서 책임으로
책임 원칙 훼손 → 결과 기반 책임 원칙 고수
데이터 오용 → 알고리즘 접근 허용 (전문가 알고리즈미스트 필요)

5. 미래의 빅데이터
• 빅데이터 활용의 3요소
데이터 : 모든 것의 데이터화
기술 : 진화하는 알고리즘, 인공지능
인력 : 데이터 사이언티스트, 알고리즈미스트

3장 가치 창조를 위한 데이터 사이언스와 전략 인사이트
1. 빅데이터 분석과 전략 인사이트
• 빅데이터 회의론의 원인과 진단
- 투자효과를 거두지 못했던 부정적 학습효과 – 과거의 고객관계관리CRM
- 빅데이터 성공사례가 기존 분석 프로젝트를 포함해 놓은 것이 많다 (굳이 빅데이터가 필요 없는 경우 등)
: 분석을 통해 가치를 만드는 것에 집중해야

• 빅데이터 분석
- 빅데이터에 대한 관심 증대
- 빅데이터 프로젝트에 거는 기대
- 빅데이터 분석의 가치 (데이터는 크기의 이슈가 아닌, 그 데이터로부터 얻을 수 있는 시각과 통찰이 관건)
• 전략적 통찰이 있는 분석을 해야 한다
• 일차원적인 분석과 전략도출을 위한 가치기반 분석
- 산업별 일차원적 분석 애플리케이션
금융서비스: 신용점수 산정, 사기 탐지, 고객 수익성 분석
소매업: 재고 보충, 수요 예측
제조업: 맞춤형 상품 개발, 신상품 개발
에너지: 트레이딩, 공급/수요 예측
온라인: 웹 매트릭스, 사이트 설계, 고객 추천

- 전략도출 가치기반 분석
중요한 기회 발굴 및 주요경영진의 지원을 얻기에 강력한 모멘텀 생성
일차원적인 분석으로 경험을 쌓고, 성공을 거두면 범위를 넓혀 전략적으로 변화 모색

2. 전략 인사이트 도출을 위한 필요 역량
• 데이터 사이언스의 의미와 역할
- 의미
데이터 공학, 수학, 통계학, 컴퓨터공학, 시각화, 해커의 사고방식, 해당 분야의 전문지식을 종합한 학문
다양한 유형의 데이터를 대상으로 분석할 뿐만 아니라 효과적으로 구현하고 전달화는 과정까지를 포함

- 역할
비즈니스 성과를 좌우화는 핵심이슈에 답을 하고, 사업의 성과를 견인해 나갈 수 있어야 한다.

- 구성요소
전략 컨설턴트 : 비즈니스 분석(스토리텔링, 시각화 등) + analytics (수학, 확률모델, 머신러닝)
IT 컨설팅 : 비즈니스 분석 + IT (시그널 프로세싱, 프로그래밍 등)

- 데이터 사이언티스트의 역할
데이터 소스 찾기, 복잡한 대용량 데이터 구조화, 불완전한 데이터 연결

- 데이터 사이언티스트의 요구 역량
Hard skil 빅데이터 이론 지식, 분석 기술에 대한 숙련 I
Soft skill 통찰력, 설득력 있는 전달, 다분야간 협력
가트너 : 비즈니스 분석, 데이터관리, 분석모델링, soft skill

3. 빅데이터 그리고 데이터 사이언스의 미래
- 데이터 사이언스의 한계와 인문학
분석 과정에서는 인간의 해석이 개입될 수밖에 없음, 아무리 정략적인 분석이라도 모든 분석은 가정에 근거한다

외부 환경적 측면에서 본 인문학 열풍의 이유
: 컨버전스 → 디버전스 / 생산 → 서비스 / 생산 → 시장창조

관계형 DBMS	데이터를 컬럼과 로우를 이루는 하나 이상의 테이블로 정리
객체지향 DBMS	멀티미디어 등 복잡한 데이터 구조 관리, 정보를 객체로 표현
네트워크 DBMS	레코드는 노드, 레코드 관계는 간선으로
계층형 DBMS	트리 구조를 기반으로

- SQL 데이터 베이스의 하부 언어. 데이터 베이스와 통신을 하기 위해 고안
  COUNT 수치/문자형 모든 데이터 타입에서 사용 가능

- 개인정보 비식별 기술 (개인을 알아볼 수 없도록)

* 난수화 : 개인정보를 무작위 처리 (사생활 침해 방지, 본래 목적 외에 가공되고 처리되는 것 방지)

데이터 마스킹 (익명)	홍길동 → 홍**
가명처리 (다른 값)	홍길동,35세 → 임꺽정, 30대
총계처리 (총합 값)	홍길동 180, 김수빈 175 → 물리학과 학생 키 합 255
데이터값 삭제	홍길동, 35세 → 35세
데이터 범주화	홍길동, 35세 → 홍씨, 30~40세

- 데이터 레이크

의미있는 내용을 찾기 위해 방식에 상관없이 데이터를 저장하는 시스템. 정형/비정형, 저장/접근

- 데이터의 유형

	형태	연산	저장	
정형 데이터	O	O	관계형 DBMS	ERP,CRM,SCM
반정형 데이터	O	X	파일	로그/모마일/센싱 데이터
비정형 데이터	X	X	NoSQL	소셜데이터(트위터, 페북), 영상, 이미지, 텍스트

- 메타 데이터 : 데이터에 관한 구조화된 데이터, 다른 데이터를 설명해주는 데이터
- 인덱스 : 데이터베이스 내의 데이터를 신속하게 정렬하고 탐색하게 해주는 구조

2과목 데이터 분석 기획

1장 데이터 분석 기획의 이해

1. 분석기획의 방향성 도출

• 분석기획

실제 분석을 수행하기에 앞서 분석을 수행할 과제 정의, 의도했던 결과를 도출할 수 있도록 사전에 계획

어떠한 목표를 달성하기 위하여 어떠한 데이터를 가지고 어떤 방식으로 수행할 지

• 분석의 대상 WHAT

known	unknown		
Optimization 최적화	Insight 통찰	Known	분석의 방법 HOW
Solution 솔루션	Discovery 발견	Unknown	

• 목표 시점 별 분석 기획 방안

- **과제 중심적인 접근 방식**

과제 단위 / speed&test / quick&win / problem solving

- **장기적인 마스터 플랜 방식**

지속적 분석 문화 내재화 / accuracy&deploy / long term view / problem definition

- 의미있는 분석을 위해서는 분석 기술, IT 및 프로그래밍, 분석 주제에 대한 도메인 전문성, 의사소통이 중요

• 분석 기획시 고려사항

(1) 가용 데이터 available data

분석을 위한 데이터의 확보, 데이터의 유형에 따라 적용 가능한 솔루션 및 분석 방법이 다르다

(2) 적절한 활용방안과 유즈케이스 proper business use case

(3) 장애요소들에 대한 사전계획 수립 low barrier of execution

충분하고 계속적인 교육 및 활용 방안 등 의 변화 관리 (일회성에 그치지 않도록)

* 장애요소

비용대비 효과의 적정한 비용, 분석 모형의 안정적 성능 확보, 조직 역량으로 내재화를 위한 변화 관리

2. 분석 방법론

• 분석 방법론 개요

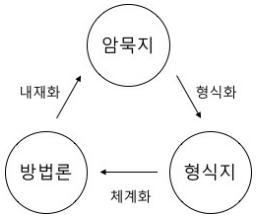
개인의 역량, 조직의 우연한 성공 < 일정한 수준의 품질을 갖춘 산출물, 성공가능성

경험과 감 < 데이터 기반의 의사 결정

• 기업의 합리적 의사결정을 가로막는 장애요소

- 고정관념stereo type, 편협된 생각bias, 프레이밍 효과framing effect

(3) 방법론의 생성과정



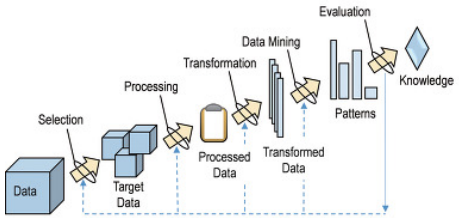
(4) 방법론의 적용 업무의 특성에 따른 모델

- **폭포수 모델** 단계를 순차적으로 진행, 문제가 발견될 시 피드백 과정이 수행

- **프로토타입 모델** 폭포수 모델의 단점 보완, 점진적으로 시스템 개발. 고객의 요구를 완전하게 이해하고 있지 못하거나 완벽한 요구 분석의 어려움을 해결하기 위해 일부분을 우선 개발하여 사용자에게 제공, 시험 사용 후 사용자의 요구를 분석하거나 요구 정당성의 점검, 성능을 평가하여 그 결과를 통한 개선 작업 시행

- ***나선형 모델** 반복을 통해 점증적으로 개발, 처음 시도하는 프로젝트에 적용 용이, 관리 체계를 효과적으로 갖추지 못하면 복잡도 상승하여 프로젝트 진행이 어려울 수 있다.

• **KDD 분석 방법론** 탐색을 통해서 패턴을 찾아냄



- 데이터 선택 selection 분석 대상의 비즈니스 도메인에 대한 이해. 프로젝트 목표 설정.
  데이터베이스 또는 원시 데이터에서 분석에 필요한 데이터 선택. 데이터 마이닝에 필요한 목표 데이트를 구성
- 데이터 전처리 preprocessing 추출된 분석 대상용 데이터 셋을 정제 (잡음, 이상치, 결측치 식별).
- 데이터 변환 transformation
  데이터 전처리 과정을 통해 정제된 데이터에 분석 목적에 맞게 변수 생성,선택, 데이터의 차원 축소
  (효율적인 데이터마이닝을 위한 준비) 학습용 데이터와 검증용 데이터로 데이터를 분리
- 데이터 마이닝 data mining
- 데이터 마이닝 결과 평가 interpretation/evaluation

• **CRISP-DM 분석 방법론**

- 업무이해 business understanding
  비즈니스 관점에서 프로젝트의 목적과 요구사항을 이해하기 위한 단계
  도메인 지식을 데이터 분석을 위한 문제 정의로 변경, 초기 프로젝트 계획 수립
  : 업무 목적 파악, 상황 파악, 데이터마이닝 목표 설정, 프로젝트 계획 수립

- 데이터 이해 data understanding
  데이터를 수집, 데이터 속성을 이해. 데이터 품질에 대한 문제점 식별, 숨겨져 있는 인사이트 발견
  : 초기 데이터 수집, 데이터 기술 분석, 데이터 탐색, 데이터 품질 확인

- **데이터 준비** data preparation : **분석용 데이터 셋 선택, 데이터 정제, 분석용 데이터 셋 편성, 데이터 통합**
  수집된 데이터에서 분석기법에 적합한 데이터를 편성

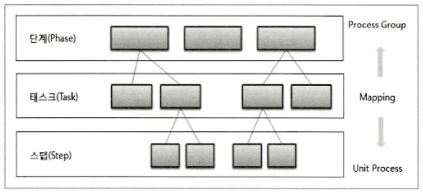
- **모델링 modeling**: **모델링 기법 선택, 모델 테스트 계획 설계, 모델 작성, 모델 평가**
  다양한 모델링 기법과 알고리즘을 서낵, 모델링 과정에서 사용되는 파라미터를 최적화
  모델링 과정에서 데이터 셋이 추가로 필요한 경우 데이터 준비단계 반복 수행,
  모델링 결과를 테스트용 데이터 셋으로 평가하여 모델의 과적합 문제 확인

- 평가 evaluation
  모델링 결과가 프로젝트 목적에 부합하는지 평가
  : 분석결과 평가, 모델링 과정 평가, 모델 적용성 평가

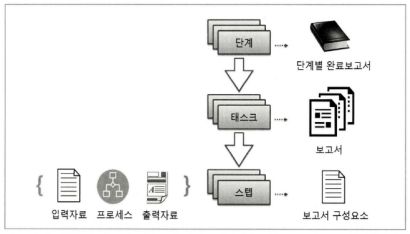
- 전개 deployment
  완성된 모델을 실 업무에 적용하기 위한 계획 수립, 모델의 유지보수 계획 마련
  : 전개 계획 수립, 모니터링과 유지보수 계획 수립, 프로젝트 종료보고서 작성, 프로젝트 리뷰

• 빅데이터 분석 방법론 : 지금까지의 방법론을 정리해서 일반화

(1) 빅데이터 분석의 계층적 프로세스



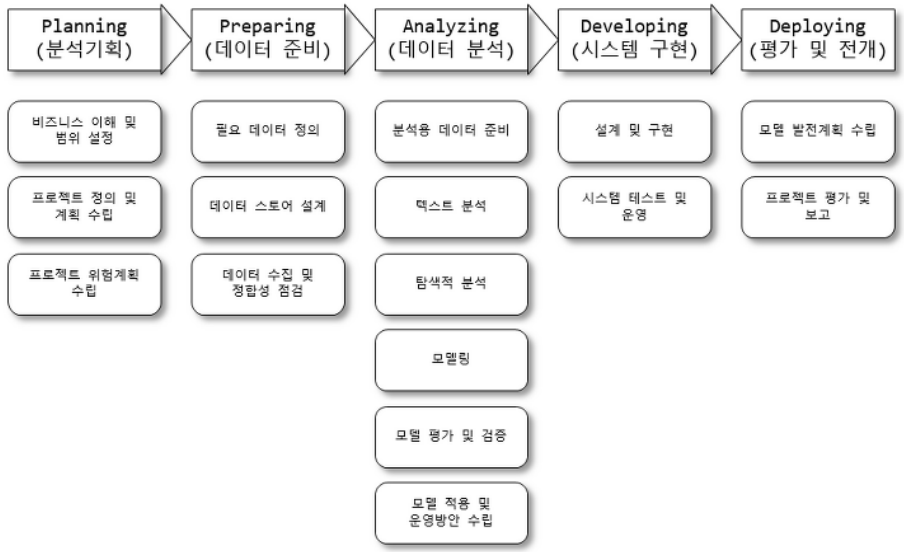
[그림 Ⅲ-1-12] 빅데이터 분석 방법론 3개층 구조



[그림 Ⅲ-1-8] 방법론의 구성

- 단계 phase : 각 단계별 산출물 생성/각 단계는 기준선으로 설정되어 관리/ 버전관리 등을 통하여 통제
- 태스크 task : 각 단계를 구성하는 단위 활동, 물리적 또는 논리적 단위로 품질검토의 항목이 될 수 있다
- 스텝 step : WBS의 워크패키지에 해당, 입력자료, 처리 및 도구, 출력자료로 구성된 단위 프로세스

(2) 빅데이터 분석 방법론



3. 분석 과제 발굴
- 디자인 사고 : 상향식 접근 방식의 발산과 하향식 접근 방식의 수렴 단계를 반복적으로 수행

• **하향식 접근법 problem solving**
현황 분석을 통해 기회나 문제를 탐색 – 해당 문제를 정의 – 해결방안을 탐색 – 타당성 평가
분석 대상 명확

① **문제탐색** 문제를 해결함으로써 발생하는 가치에 중점을 두는 것이 중요
- **비즈니스 모델 기반 문제 탐색 비즈니스 모델 캔버스 *(제품, 업무, 고객, 지원인프라, 규제와 감사)**
- **분석 기회 발굴의 범위 확장**
거시적 관점의 메가 트렌드 Social Technology Economic Environmental Political
경쟁자 확대 관점 (대체재, 경쟁자, 신규진입자)
시장의 니즈 탐색 과정 (고객, 채널, 영향자)
역량의 재해석 관점 (내부역량, 파트너와 네트워크)
- **외부참조 모델기반 문제탐색** 유사동종 사례 벤치마킹, 필요한 분석기회가 무엇인지
- **분석 유즈 케이스** 현재의 비즈니스 모델 및 유사 동종사례 탐색을 통해서 빠짐없이 도출한 분석 기회들

② **문제정의** 식별된 비즈니스 문제 →데이터의 문제로 변환하여 정의
달성하기 위해서 필요한 데이터 및 기법을 정의하기 위한 데이터 분석의 문제로의 변환을 수행
예. 고객 이탈의 증대 (비즈니스 문제) – 고객 이탈에 영향을 미치는 요인 식별, 이탈가능성 예측 (데이터 분석 문제)

③ **해결방안 탐색**
④ 타당성 검토
- 경제적 타당성 (비용대비 편익 분석 관점의 접근)
- 데이터 및 기술적 타당성 : 데이터 존재 여부, 분석 시스템 환경 그리고 분석 역량
문제발생 포인트에 대한 데이터가 아닌 모든 데이터 확보

• **상향식 접근법 problem creating / 비지도학습 / 프로토타입 방식**
다양한 원천 데이터를 대상으로 분석을 수행하여 가치 있는 모든 문제를 도출하는 과정

• 비지도 분석 – 상향식 접근방식의 데이터 분석
데이터 자체의 결합, 연관성, 유사성 등을 중심으로 데이터의 상태를 표현하는 것
인자들 간의 유사성을 바탕으로 수행하는 군집화
목표값을 사전에 정의하지 않고 데이터 자체만을 가지고 그룹들을 도출
빅데이터 환경에서 인과관계 뿐만 아니라 상관관계 분석 또는 연관 분석을 통해 다양한 문제 해결

• 지도 분석
데이터 분석의 목적이 명확히 정의된 형태의 특정 필드 값을 구하는 것
명확한 목적 하에 데이터 분석 실시
O 와 X 를 구분 짓게 하는 분류
도출되는 값에 대하여 사전에 인지하고 어떠한 데이터를 넣었을 때 어떠한 결과가 나올지를 예측
통계적 분석에서 인과관계 분석을 위해 가설을 설정하고 이를 검정하기 위해 모집단으로부터 추출한 표본으로 가설 검정 실시

• 시행착오를 통한 문제 해결
- 정의
프로토타이핑 접근법 사용자가 요구사항이나 데이터를 정확히 규정하기 어렵고 데이터 소스도 명확히 파악하기 어려운 상황에서 일단 분석을 시도해 보고 그 결과를 확인하면서 반복적으로 개선해 나가는 방법
완전하지는 못해도 신속하게 해결책이나 모형을 제시, 문제를 좀 더 명확하게 인식하게 필요한 데이터를 식별하여 구체화할 수 있게.
가설의 생성 – 디자인에 대한 실험 – 실제 환경에서의 테스트 – 테스트 결과에서의 통찰 도출 및 가설 확인
(cf. 하향식 접근방식 – 문제가 정형화되어 있고 문제해결을 위한 데이터가 완벽하게 조직에 존재할 경우)

- 빅데이터 분석 환경에서 프로토타이핑의 필요성
문제에 대한 인식 수준 (문제 정의가 불명확할 경우 프로토 타입을 이용하여 문제 이해)
필요 데이터 존재 여부의 불확실성 (데이터의 집합이 모두 존재하지 않을 경우)
데이터 사용 목적의 가변성 (데이터의 가치는 변할 수 있음, 기존의 데이터 정의 재검토 후 사용목적과 범위 확대)

(4) 분석과제 정의
분석과제 정의서를 통해 분석별로 필요한 소스 데이터, 분석방법, 데이터 입수 및 분석의 난이도, 분석 수행주기, 분석결과에 대한 검증 오너십, 상세 분석 과정 등을 정의.

• 분석과제 관리를 위한 5가지 주요 영역
분석 과제를 수행할 때 고려해야 할 주요 속성
Speed / data complexity / data size / analytic complexity / accuracy&precision

• 분석 프로젝트의 특성
• 분석 프로젝트의 관리방안*
범위 / 시간 / 원가 / 품질 / 통합 / 조달 / 자원 / 리스크 / 의사소통 / 이해관계자

2장 분석 마스터 플랜

1. 마스터 플랜 수립 프레임 워크

분석 과제를 대상으로 다양한 기준을 고려해 적용 우선순위 설정

: **전략적 중요도 / 비즈니스 성과,ROI / 실행 용이성**

적용 범위 및 방식에 대해서 고려하여 데이터 분석 구현을 위한 로드맵 수립

: 업무 내재화 적용 수준 / 분석 데이터 적용 수준 / 기술 적용 수준

* **정보전략계획 ISP** 조직 내외부 환경을 분석하여 기회나 문제점 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터 플랜을 수립하는 절차

• 수행 과제 도출 및 우선순위 평가

(1) 우선순위 평가 방법 및 절차

분석 과제 도출 – 우선순위 평가 – 우선순위 정련(순위 조정)

(2) 전략적 중요도 (전략적 필요성, 시급성), 실행 용이성 (투자, 기술 용이성)

(3) ROI관점에서 빅데이터의 핵심 특징 **4V**

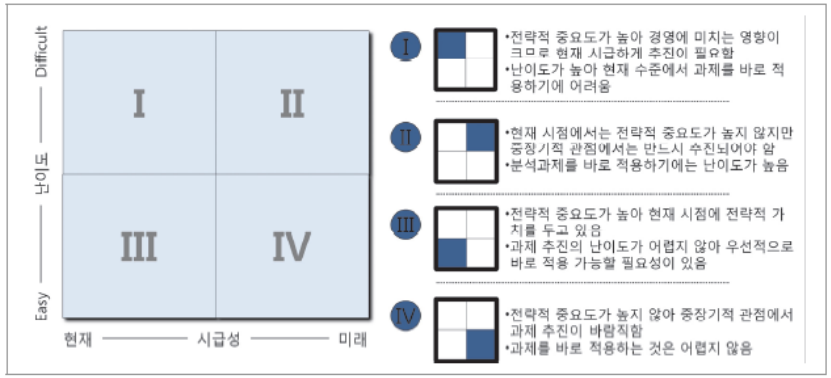
3V (Volume, Variety, Velocity) : 투자비용 요소

+ Value : 비즈니스 효과 요소 (분석 결과를 활용하거나 실질적인 실행을 통해 얻게 되는 비즈니스 효과)

(4) 데이터 분석과제 추진 시 고려해야 하는 우선순위 평가 기준 **시급성 / 난이도**

- 시급성 : 전략적 중요도, 목표가치 (Value)
- 난이도 : 데이터 획득, 저장, 가공비용 / 분석 적용 비용 / 분석 수준 (3V)

(5) 포트폴리오 사분면 분석을 통한 과제 우선순위 선정



가장 우선적인 분석 과제 적용이 필요한 영역 3

시급성에 우선순위 기준 : 3 – 4 – 2

난이도에 우선순위 기준 : 3 – 1 – 2

• 이행계획 수립

(1) 로드맵 수립

(2) 세부 이행계획 수립

데이터 분석체계는 고전적인 폭포수 방식도 있으나 프로토 타입을 통해 프로젝트의 완성도를 높음

데이터 수집 및 확보, 분석데이터 준비 단계는 순차적 진행, 모델링 단계를 반복적으로 수행

2. 분석 거버넌스 체계 수립

• 거버넌스 체계의 구성요소 Organization / Process / System / Data / Human Resource

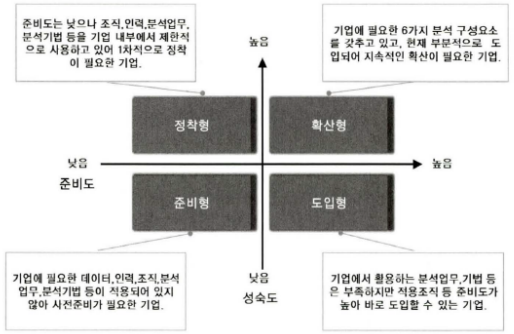
• **분석 준비도**

- 분석 업무 : 발생한 사실, 예측, 시뮬레이션, 최적화, 분석업무 정기적 개선
- 분석 인력
- 분석기법 : 업무별 적합한 기법, 분석업무 도입 방법론, 분석기법 라이브러리, 분석기법 효과 평가·정기적 개선
- 분석 데이터 : 데이터 충분성·신뢰성·적시성, 비구조적 데이터 관리, 외부 데이터 활용 체계, 기준 데이터 관리
- 분석 문화
- 분석 인프라

• **분석 성숙도** by. CMMI 모델 / 도입 – 활용 – 확산 – 최적화

단계	도입단계	활용단계	확산단계	최적화단계
설명	분석을 시작하여 환경과 시스템을 구축	분석 결과를 실제 업무에 적용	전사 차원에서 분석을 관리하고 공유	분석을 진화시켜서 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 • 프로세스 혁신 3.0 • 분석규칙 관리 • 이벤트 관리	• 외부 환경분석 활용 • 최적화 업무 적용 • 실시간 분석 • 비즈니스 모델 진화
조직·역량 부문	• 일부 부서에서 수행 • 담당자 역량에 의존	• 전문 담당부서에서 수행 • 분석기법 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 • 통계 분석 환경	• 빅데이터 관리 환경 • 시뮬레이션·최적화 • 비주얼 분석 • 분석 전용 서버	• 분석 협업환경 • 분석 Sandbox • 프로세스 내재화 • 빅데이터 분석

(3) **분석 수준 진단 결과**



• 분석지원 인프라 방안 수립

장기적, 안정적으로 활용할 수 있는 확장성을 고려한 플랫폼 구조를 도입

- 플랫폼

분석 서비스를 위한 응용프로그램이 실행될 수 있는 기초를 이루는 컴퓨터 시스템을 의미.

일반적으로 하드웨어에 탑재되어 데이터 분석에 필요한 프로그래밍 환경과 실행 및 서비스 환경을 제공하는 역할 수행. 새로운 분석 니즈가 존재할 경우 서비스를 추가적으로 제공하는 방식으로 확장성을 높일 수 있음.

• **데이터 거버넌스 체계 수립**

표준화된 관리체계 수립, 운영을 위한 프레임워크 및 저장소 구축

(1) 구성요소 Principle Process Organization

(2) 체계

- **데이터 표준화**

데이터 표준 용어 설정, 명명 규칙 수립, 메타데이터 및 데이터 사전 구축

메타 데이터 : 데이터에 관한 구조화된 데이터, 다른 데이터를 설명해주는 데이터

- **데이터 관리 체계**

데이터 정합성 및 활용의 효율성을 위하여 표준 데이터를 포함한 메타 데이터, 데이터 사전의 **관리 원칙 수립**

- **데이터 저장소 관리**

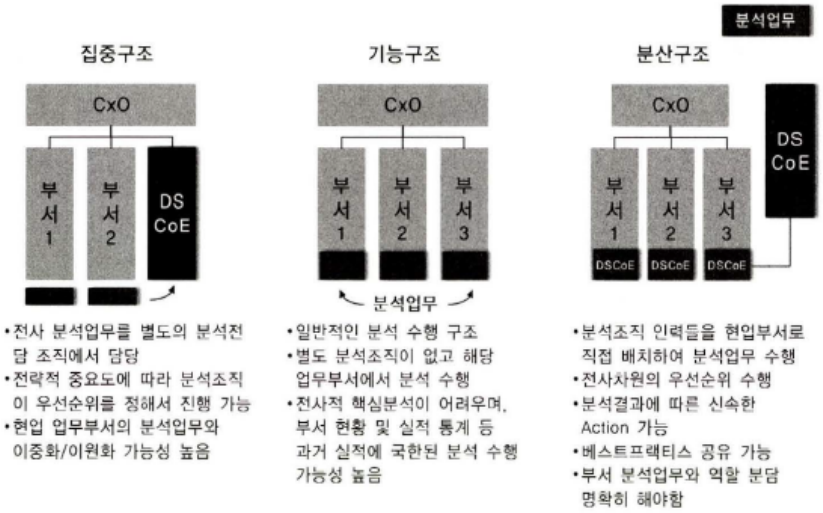
워크 플로우 및 관리용 소프트웨어 지원, 데이터 구조 변경에 따른 **사전 영향 평가**도 수행

- **표준화 활동**

데이터 거버넌스 체계를 구축한 후 표준 준수 여부를 주기적으로 점검, 모니터링 실시

• 데이터 조직 및 인력방안 수립

데이터를 효과적으로 분석, 활용하기 위해 기획, 운영 및 관리를 전달할 수 있는 전문 분석 조직의 필요성



※ DSCoE : Data Science Center of Excellence

• 분석과제 관리 프로세스 수립

과제 발굴 : 분석 아이디어 발굴 – 과제화 – 분석 과제 풀로 관리 – 분석 프로젝트 선정

과제 수행 및 모니터링 : 팀구성 – 분석과제 실행 – 분석과제 진행관리 – 결과 공유 및 개선

해당 과제를 진행하면서 만들어진 시사점을 포함한 결과물을 풀pool에 잘 축적, 관리.

향후 유사한 분석과제 수행 시 사용 가능

• 분석 교육 및 변화 관리

데이터 활용을 통한 비즈니스 가치를 전사적으로 확대하기 위해

단순 툴 교육이 아닌 분석 역량을 확보하고 강화하는 것에 초점

3과목 데이터 분석

- 데이터분석 기법의 이해

데이터 처리 과정	- 데이터 분석을 위해서는 데이터 웨어하우스(DW)나 데이터마트(DM)을 통해 분석데이터 구성 - 신규데이터나 DW에 없는 데이터는 기존 운영시스템(Legacy)에서 직접 가져오거나 운영 데이터 저장소(ODS)에서 정제된 데이터를 가져와서 DW의 데이터와 결합하여 사용
시각화 기법	가장 낮은 수준의 분석이지만 효율적이며, 대용량 데이터 다룰 때는 필수
공간분석	공간적 차원과 관련된 속성들을 시각화하는 분석으로 지도위에 관련된 속성들을 생성하고 크기모양, 선 굵기 등을 구분하여 인사이트를 얻음
탐색적 자료분석 (EDA)	**- 다양한 차원과 값을 조합해가며 특이점이나 의미있는 사실 도출** **- EDA 4가지 주제 : 저항성의 강조, 잔차 계산, 자료변수의 재표현, 그래프를 통한 현시성**
데이터마이닝	대용량의 자료로부터 정보요약, 미래예측 목표로 자료에 존재하는 관계, 패턴, 규칙 등을 탐색 이를 모형화 함으로써 이전에 알지 못한 유용한 지식을 추출하는 분석 기법 - 방법론: 기계학습(인공신경망, 의사결정나무, 클러스터링, SVM), 패턴인식(연관규칙, 장바구니분석)

- 배치 실행

매일 실행해야 하는 프로그램을 시스템에서 자동으로 구동하는 작업 R CMD BATCH

- 기본적인 통계량 계산

기능	함수	기능	함수
평균	Mean()	중간값	Median()
표준편차	Sd()	분산	Var()
공분산	Cov()	상관계수	Cor()

- 날짜

R표현	표시 형태	R표현	표시 형태
%b	축약된 월 이름 (“Jan”)	%B	전체 월 이름 (“January”)
%d	두 자리 숫자로 된 일 (“31”)	%m	두 자리 숫자로 된 월 (“12”)
%y	두 자리 숫자로 된 년 (“14”)	%Y	네 자리 숫자로 된 년 (“2014”)

3장 데이터 마트

1절 데이터 변경 및 요약

1.R reshape를 이용한 데이터 마트 개발

데이터 웨어하우스 : 기업의 의사결정 과정을 지원하기 위해 주제 중심적으로 통합적이며 시간성을 가지는 비휘발성 데이터의 집합

데이터 마트 : 데이터 웨어하우스와 사용자 사이의 중간층에 위치한 것으로, 하나의 주제 또는 하나의 부서 중심의 데이터 웨어하우스라고 할 수 있다

요약변수 : 수집된 정보를 분석에 맞게 종합한 변수, 공통 사용 가능, 재활용성이 높다

파생변수 : 특정조건을 만족하거나 특정함수에 의해 값을 만들어 의미 부여, 주관적 - 논리적 타당성 갖춰야.

Reshape의 활용: 녹이고 melt ( ) – 원하는 형태로 변형 cast ( )

2. sqldf : R에서 sql의 명령어를 사용가능하게 해주는 패키지

3. plyr : apply 함수에 기반해 데이터와 출력변수를 동시에 배열로 치환하여 처리하는 패키지
   데이터를 분리하고 처리한 다음, 다시 결합하는 등 필수적인 데이터 처리 기능 제공

4. 데이터 테이블 : 가장 많이 사용됨. 탐색, 연산, 병합하는데 유용
   빠른 그루핑과 ordering, 짧은 문장 지원 측면에서 데이터프레임보다 유용

2절 데이터 가공

1. data exploration : head, tail, summary 함수
2. 변수 중요도 klaR패키지
3. 변수의 구간화 : 연속형 변수를 구간화
   구간화 방법 : binning, 의사결정나무

3절 기초 분석 및 데이터 관리

1. 데이터 EDA (탐색적 자료 분석)
2. 결측값 인식
3. 결측값 처리
   (1) **단순대치법**
   - complete analysis 결측값 존재하는 레코드 삭제
   - 평균대치법: 조건부(회귀분석), 비조건(관측데이터의 평균)
   - 단순확률대치법: 평균대치법에서 추정량 표준 오차의 과소 추정문제 보안

   (2) **다중대치법 (대치-분석-결합)**
4. 결측값처리

complete.cases ( ) 결측값 있으면 F

is.na ( ) 결측값 있으면 T

centralImputation ( ) 가운데 값으로 대치

knnImputation ( ) 이웃 분류 알고리즘, 가중평균

amelia ( ) 랜덤포레스트

5. 이상값인식과 처리 (부정사용방지 시스템) – 데이터의 손실율이 낮아 설명력이 높아지는 장점
   이상값의 인식 방법: ESD (3표준편차), IQR
   극단값 절단trimming
   극단값 조정winsorizing

4장 통계 분석

1절 통계분석의 이해

1. 통계
2. 통계자료의 획득방법 : 전수조사, 표본조사, **표본추출방법 (단순랜덤/계통/집락/층화추출법)**
   - **명목척도(집단)**
     순서척도(서열) 만족도, 학년, 신용등급
     구간척도(간격) 온도, 지수
     비율척도 (비율) 절대적 기준 0 존재, 무게, 나이, 시간, 거리
3. 통계분석 : 기술통계, 통계적 추론(모수추정, 가설검정, 예측)
4. 확률 및 확률분포

- 확률: 표본공간, 사건, 원소, 확률변수
- 확률분포 : **이산형 확률변수 (베르누이, 이항분포, 기하분포, 다항분포, 포아송분포)**

연속형 확률변수 (균일분포, 정규분포, 지수분포, t, 카이제곱, f)

5. 추정과 가설검정

- 추정의 개요 : 확률표본, 추정(점추정, 구간추정)
- 가설검정

귀무가설

대립가설 alternative hypothesis

검정통계량 관찰된 표본으로부터 구하는 통계량, 검정 시 가설의 진위를 판단하는 기준

유의수준a 귀무가설을 기각하게 되는 확률의 크기 (귀무가설이 옳은데도 이를 기각하는 확률의 크기)

기각역 귀무가설이 옳다는 전제 하에서 구한 검정통계량의 분포에서 확률이 유의수준a인 부분
대립가설이 맞을 때 그것을 받아들이는 확률

P값 귀무가설이 옳다는 가정하에 얻은 통계량이 귀무가설을 얼마나 지지하는지를 나타낸 확률
귀무가설이 사실인데도 불구하고 사실이 아니라고 판정할 때 실제 확률
(제 1종 오류에서 우리가 내린 판정이 잘못되었을 실제 확률)

- **비모수 검정**

- 비모수정 방법: 모수적방법, 모집단의 분포에 대한 아무런 제약 가하지 않음
특정분포를 따른다고 가정할 수 없는 경우,
관측된 자료 수가 적거나 서열관계를 나타내는 경우
- 가설의 검정: 분포의 형태에 대해 가설 설정, 관측값의 순위나 두 관측값 차이의 부호로 검정
- 비모수 검정의 예
부호검정, 윌콕슨의 순위합검정, 부호순위합검정, 만-위트니의 U검정, 런검정, 스피어만의 순위상관계수

2절 기초 통계분석

1. 기술통계

- 통계량에 의한 자료정리

중심위치의 측도 (자료, 표본평균, 중앙값)
산포의 측도 (분산, 표준편차, 사분위수, 백분위수, 변동계수, 평균 표준오차)
분포의 형태에 관한 측도 (왜도, 첨도)
그래프를 이용한 자료 정리: 막대그래프(범주형), 히스토그램(연속형), 줄기잎그림, 상자그림

2. 인과관계의 이해

종속변수, 독립변수, 산점도 /
공분산cov (+ 두 변수는 양의 방향성, 0이면 독립)

3. **상관분석(상관계수)**
   - **피어슨 : 등간척도** 이상으로 측정된 두 변수들의 상관관계, 연속형 변수, 정규성 가정
   - **스피어만 : 서열척도인** 두 변수, 순서형 변수, 비모수적 방법, 순위상관관계

상관분석을 위한 R코드 var, cov, cor, rcorr
Cov(x,y, method=c("pearson","kendall","spearman"))

- 상관분석의 가설 검정

상관계수가 0이면 입력변수와 출력변수 사이에는 아무런 관계 없음
T 검정통계량을 통해 얻은 p값이 0.05 이하인 경우, 대립가설을 채택

3절 회귀분석

1. 회귀분석의 개요

회귀분석의 변수: 반응변수, 설명변수

선형회귀분석의 가정: 선형성, 등분산성, 독립성, 비상관성, 정상성(정규성)

2. 단순선형회귀분석

- 검토사항

회귀계수들의 유의미한가: t통계량의 p값

설명력을 갖는가: 결정계수

데이터를 잘 적합하고 있는가: 잔차를 그래프로 그리고 회귀진단을 한다

- 회귀계수의 추정 (**최소제곱법**): 측정값을 기초로 하여 제곱합을 만들고 그것을 최소하는 값을 구함
- 회귀분석의 검정

회귀계수 이 0이면 x – y 아무런 관계 없다 / 적합된 추정식 의미 없다.

귀무가설

결정계수

회귀직선의 적합도 검토: 다변량 회귀분석에서는 독립변수가 많아지면 결정계수가 높아진다.

* 이 독립변수가 유의하든, 유의하지 않든 수가 많아지면.

3. 다중선형회귀분석

- 다중회귀식
- 모형의 통계적 유의성

- 유의수준 5% 하에서 p값이 0.05보다 작으면 추정된 회귀식은 통계적으로 유의
- F통계량이 크면 p값이 작아지고 귀무가설을 기각. 즉, 모형이 유의하다.

- 회귀계수의 유의성: t통계량
- 모형의 설명력: 결정계수R
- 모형의 적합성: 잔차와 종속변수의 산점도

데이터가 전제하는 가정을 만족시키는가? (선형성, 독립성, 등분산성, 비상관성, 정상성)

다중공선성 : 설명변수들 사이에 선형관계가 존재하면 정확한 추정 곤란

4. 회귀분석의 종류 (단순, 다중, 로지스틱, 다항, 곡선, 비선형 회귀)

- 최적회귀방정식

설명변수 선택(가능한 적은 수의 설명변수 포함)

모형선택

단계적 변수선택 변수 선택법 함수 step ( ): direction="forward"/"backward"/"stepwise"

: 전진선택법(상수모형에서 시작), 후진제거법(독립변수 후보 모두 포함한 모형에서 시작), 단계선택법(넣었다 뺐다)

벌점화된 선택기준 AIC, BIC 값이 최소가 되는 모형

4절 시계열 분석 시간의 흐름에 따라 관찰된 값, 미래의 값을 예측.

1. 시계열 자료: 비정상성 시계열 자료, **정상성 시계열 자료**
2. **정상성**: 평균, 분산이 일정. 공분산도 시점이 아닌 시차에만 의존 (**정상 시계열**)

3. 시계열자료 분석방법

- 자료 형태에 따른 분석방법 : 일변량 시계열분석, 다중 시계열분석
- 이동평균법 : 일정기간의 이동평균 계산, 추세를 파악하여 예측

간단하고 쉽게 미래예측, 자료의 수가 많고 안정된 패턴을 보이는 경우 예측 품질 높음

특정기간 안에 속하는 시계열에 대해서는 동일한 가중치

- 지수평활법 : 모든 시계열 자료를 사용하여 평균, 최근 시계열에 가중치

4. 시계열모형

자기회귀모형 : p 시점전의 자료가 현재 자료에 영향을 주는 존재

이동평균모형 : 유한한 개수의 백색잡음의 결합, 언제나 정상성 만족.

같은 시점의 백색잡음과 바로 전 시점의 백색잡음의 결합으로 이뤄진 모형

자기회귀누적이동평균 모형

분해시계열: 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법

추세요인	형태가 오르거나 또는 내리는 추세, 선형, 이차식, 지수형태
계절요인	요일, 월, 사분기 별로 변화하여 고정된 주기에 따라 자료가 변화
순환요인	명백한 경제적, 자연적 이유 없이 알려지지 않은 주기로 자료가 변화
불규칙요인	위 세가지 요인으로 설명할 수 없는 회귀분석에서 오차에 해당하는 요인

5절 다차원척도법

개체간 근접성을 시각화, 2차원 공간상에 점으로 표현.

유클리드 거리 행렬을 활용한다.

- 다차원척도법 종류 (계량적MDS: 구간척도, 비율척도/ 비계량적MDS: 순서척도)

6절 주성분 분석

1. 정의

여러 변수들의 변량을 주성분이라는 서로 상관성이 높은 변수들의 선형결합으로 만들어

기존의 상관성이 높은 변수들을 요약, 축소하는 기법

2. 주성분분석의 목적

연관성이 높은 변수 – 주성분 분석(차원축소) – 군집분석 수행 – 군집화 결과와 연산속도 개선

3. 주성분분석과 요인분석
4. **누적기여율cumulative proportion**이 85% 이상이면 주성분의 수로 결정할 수 있다.

Scree plot 을 활용하여 고유값이 수평을 유지하기 전단계로 주성분의 수를 선택

5장 정형 데이터마이닝

1절 데이터마이닝의 개요

데이터마이닝: 대용량 데이터에서 의미있는 패턴 파악, 예측하여 의사결정에 활용
분석 목적에 따른 작업 유형과 기법 : 예측(분류 규칙), 설명(연관, 연속규칙, 데이터군집화)
데이터마이닝 추진단계 (목적설정, 데이터준비, 가공, 기법 적용, 검증)
데이터마이닝을 위한 데이터 분할: 구축용/검정용/시험용 (홀드아웃, 교차확인 방법)

- **성과분석**

정분류율
특이도
민감도
재현율
정확도
F1 score

- **ROC커브** (FPR:1-특이도, TPR:민감도), AUROC

2절 분류분석

1. 분류분석(범주형 속성), 예측분석(연속형 속성)
2. 로지스틱 회귀분석

새로운 설명변수가 주어질 때 반응변수의 각 범주에 속할 확률이 얼마인지를 추정 (사후확률)
의 부호에 따라 양수일 때 S자, 음수일 때 역S자.
* 오즈비: 성공할 확률이 실패할 확률의 몇 배인지 나타낸다.

3. 의사결정나무

예측력과 해석력
의사결정나무의 활용: 세분화, 분류, 예측, 차원축소 및 변수선택, 교호작용효과의 파악
의사결정나무의 특징
- 비정상 잡음 데이터에 대해서도 민감함이 없이 분류할 수 있다
  한 변수와 상관성이 높은 다른 불필요한 변수가 있어도 크게 영향을 받지 않는다.
- 새로운 자료에 대한 과대 적합이 발생할 가능성, 설명변수 간의 중요도 파악 어렵.

의사결정나무의 분석과정 (성장 – 가지치기 – 타당성평가 – 해석 및 예측)
나무의 성장: 분리규칙, 분리기준, 정지규칙
나무의 가지치기 : 너무 크면 과대적합, 너무 작으면 과소 적합

4. 불순도의 여러가지 측도

- 카이제곱 통계량
- 지니지수
- 엔트로피 지수

5. 의사결정나무 알고리즘

- CART
- C4.5와 C5.0
- CHAID

3절 앙상블분석

앙상블 기법의 종류

- **배깅** 각 붓스트랩 자료에 예측모형을 만든 후 결합, 동일한 크기의 표본을 랜덤 복원추출
- **부스팅** 약한 모형들을 결합
- **랜덤포레스트** 더 많은 무작위성을 주어 약한 학습기들을 생성, 선형 결합하여 최종 학습기

*붓스트랩은 주어진 자료에서 단순랜덤 복원추출방법을 활용하여 동일한 크기의 표본을 여러 개 생성하는 샘플링 방법. 붓스트랩을 통해 100개의 샘플을 추출하더라도 샘플에 한 번도 선택되지 않는 원데이터가 발생할 수 있는데 전체 샘플의 약 36.8%가 이에 해당한다.

4절 인공신경망 분석

인공신경망
인공신경망의 특징: 구조, 뉴런의 계산
뉴런의 활성화 함수 (계단, 부호, **시그모이드**, relu, **softmax 함수**)
*softmax 함수: 표준화지수 함수로도 불리며, 출력값이 여러개로 주어지고
목표치가 다범주인 경우 각 범주에 속할 사후확률을 제공하는 함수
단일 뉴런의 학습 (단층퍼셉트론)
신경망 모형 구축시 고려사항: 입력변수, 가중치의 초기값과 다중 최소값 문제, 학습모드,
은닉층과 은닉노드의 수 많으면 과대적합문제, 적으면 과소적합

5절 군집분석

1. 군집분석
2. 거리

(1) 연속형 변수
유클리디안 거리 데이터간의 유사성 측정, 산포정도 감안 X
표준화 거리 표준편차로 척도 변환 후 유클리디안 거리
마할라노비스 거리 통계적개념이 포함된 거리, 변수들의 산포를 고려하여 표준화. 벡터사이의 거리/표본공분산
체비셰프 max
캔버라
민코우스키

(2) 범주형 변수
자카드 거리
자카드 계수
코사인 거리
코사인 유사도

3. 계층적 군집분석 (합병형, 분리형)

최단연결법/최장연결법/평균연결법/와드연결법/군집화

4. 비계층적 군집분석 – k평균 균집분석

과정 (개수 정하기 – seed 선정)
특징
- 초기 중심값의 선정에 따라 결과가 달라짐
- 연속형 변수에 활용이 가능

5. 혼합 분포 군집 - EM알고리즘 사용

6**. SOM** (self organizing map)

비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬

구성 (입력층/경쟁층)

특징

- 지도 형태로 형상화, 시각적으로 이해가 쉽다.
- 입력 변수의 위치 관계를 그대로 보존, 실제 데이터가 유사하면 지도상에서 가깝게 표현
- 단 하나의 전방 패스를 사용

6절 연관분석

1. 연관규칙 – 장바구니 분석, 서열분석(A 산 다음에 B 산다)

연관규칙의 측도

- 지지도 (교집합)
- 신뢰도 (조건부확률)
- 향상도

연관규칙의 절차

순차패턴 (+시간)

2. 기존 연관분석의 이슈 – 대용량 데이터에 대한 연관성분석 불가능
3. 최근 연관성분석의 동향 Apriori 알고리즘