1강. 데이터분석 개요

1. 데이터마이닝 개요와 프로세스

1.1 데이터마이닝의 개요: 개념 및 배경

❖ 데이터마이닝(Datamining)이란?

- 대용량의 데이터 창고[광석 더미]로부터 유용한 정보[금 조각]를 캐내는(mining) 작업을 의미
- 대용량 데이터에 존재하는 데이터 간의 관계, 패턴, 규칙 등을 찾아내고 모형화해서 기업의 경쟁력 확보를 위한 의사결정을 돕는 일련의 과정

❖ 데이터마이닝 도입 배경

- 치열한 경쟁상황 하에서의 정보/지식의 필요성 증대 (What happen? → What will happen?)
- 데이터웨어하우징(Data Warehousing) 기술의 발달로 체계적으로 데이터를 집적/정리할 수 있게 됨.
- 데이터 분석 및 컴퓨팅 기술의 발전 (예: SAS e-Miner, SPSS Modeler)

1. 데이터마이닝 개요와 프로세스

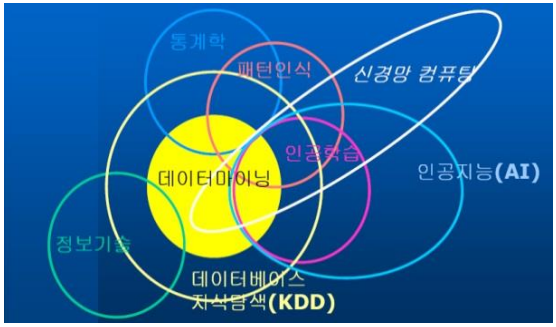
1.1 데이터마이닝의 개요: 개념 및 배경

❖ 통계분석 Vs. 데이터마이닝

- **전통적 통계분석**
  - 대상집단이 있으며, 모집단의 분포 혹은 모형 등 여러 가지 가정을 전제로 하게 되며 이 전제 조건하에서 분석을 실시 → 표본(Sample)의 관찰을 통해 모수(Population) 전체를 추론(Inference)하는 과정
- **데이터마이닝**
  - 표본조사/실험에서 필연적으로 수반되는 분포라든가 모형에 대한 전제조건이 필요하지 않음 → 모집단의 전체자료를 이용하여 필요한 정보/지식을 추출하는 과정
  - 대용량 자료여야 한다는 전제조건 있음

❖ 데이터마이닝은 어느 영역인가?

- 전산학? / 인공지능? / 통계학? /경영학? → 통합영역
- 출발이 어디건 활용되는 곳이 중요하다!



1. 데이터마이닝 개요와 프로세스

1.1 데이터마이닝의 개요: 데이터마이닝 기법

Data Mining 문제

Supervised

예측 (Prediction)

주어진 Data에 근거하여 Model을 만들고 이 Model을 이용하여 새로운 Case들에 대한 예측을 하는 작업

예) 조업에 투여되는 성분, 제조 환경(압력, 온도,습도)에 의해 품질spec을 예측

분류 (Classification)

일련의 범주들이 사전에 분류되어 있을 때, 특정한 case가 어디에 속하는 가를 결정

(예) 품질등급이 양호/보통/불량일 때 새로운 제품의 품질등급 결정

Unsupervised

군집 (Clustering)

데이터의 여러 속성들을 비교하여 유사한 특성을 갖는 항목들을 함께 군집시키는 것

예) 여러 공정 중 유사 특성을 갖는 공정간의 군집화

연관규칙 (Association Rule)

한 패턴의 출현이 다른 패턴 출현을 암시하는 속성이나 항목간의 관계를 파악

예) 한 공정에서 이상패턴이 전체공정에서 나타날 현상을 예측 (조업편차 세분화)

1. 데이터마이닝 개요와 프로세스

1.1 데이터마이닝의 개요: 대표적 분석방법

- **지도학습모형(Supervised Learning Model) – Top down / Goal Driven / 연역적 방법**
  - **회귀(Regression)**
    - 통계적/수량적 모형
    - 예: 개별고객 수치화/점수화, 고객 기여도 평가 및 예측, 고객생애가치(Lifetime Value; LTV) 예측 등
  - **분류(Classification)**
    - 사전에 미리 정의된 집단 중 어디에 해당할 것인지를 예측, 분류기준이 미리 정해짐
    - 예: 가치 있는 고객과 가치 없는 고객, 반응하는 고객과 반응 없는 고객, 충실한 고객과 곧 이탈할 고객

로지스틱 회귀분석

의사결정나무, 인공신경망 분석

1. 데이터마이닝 개요와 프로세스

지도학습모형의 비교 및 평가

1.1 데이터마이닝의 개요: 대표적 분석방법

❖ 참고: 지도학습모형(Supervised Learning Model)이란?

- 훈련 데이터(Training Data)로부터 하나의 함수(모형)를 유추해내기 위한 학습(Learning)의 한 방법
- 훈련 데이터로부터 도출(적합)된 모형을 활용하여 테스트 데이터(Test Data)로 평가진행
  - 미래에만 가능한 평가를 현재시점에서 진행하고 과대적합(overfitting)현상을 방지하기 위함
- 지도학습모형에서 변수선정 및 역할
  - 목표변수 : 데이터마이닝의 맥락에서 지도자에 해당하는 변수(종속변수, 결과변수, 내생변수)
  - 설명변수 : 목표변수 값의 예측에 쓰이는 변수(독립변수, 입력변수, 외생변수)
  - 예측모델 : 목표변수와 설명변수간의 관계를 나타내는 모형-시행착오 통한 오차최소화

어린아이가 말을 배우는 과정 (엄마가 Supervisor 역할)

과일 중에서 처음으로 '사과'를 알게 된 어린아이는 토마토를 보고도 '사과'라고 할 것이다. 그러면 엄마는 그것은 사과가 아니고 토마토라고 말해준다. 어린아이는 그것이 자기가 이제까지 알고 있던 사과와 어떻게 다른가를 생각해 보고 엄마의 지도 하에 여러 차례의 시행착오를 통해 토마토와 사과를 구별할 수 있게 된다. 이렇게 어린아이는 엄마로부터 온갖 사물의 개념과 분류를 경험적으로 배우는 것이다. 이러한 학습방법을 **지도학습(Supervised Learning)**이라고 하는데 지도자(교사)를 필요로 하는 점이 본질적 특성이다. -> 아이는 시행착오를 통한 오차 최소화 과정을 거쳐 제대로 된 학습이 이루어짐.

1. 데이터마이닝 개요와 프로세스

1.1 데이터마이닝의 개요: 대표적 분석방법

- **비지도학습모형(Unsupervised Learning Model) – Bottom up / Data Driven / 귀납적 방법**
  - **군집화(Clustering)**
    - 분류기준이 없이 비슷한 것끼리 묶은 후 의미부여
    - 예: 유사한 고객들끼리 한 그룹이 되도록... 시장 세분화 (Market Segmentation)
  - **연관(Association)**
    - 비슷한 상품들을 찾아내는 유사행태 집단화(affinity grouping)의 한 방법
    - 장바구니 리스트 중 동시에 구매가능성이 높은 품목에 대한 규칙을 찾는 방법
    - 예: 맥주와 기저귀, 피자 산 사람의 85%는 콜라도 같이 산다.
  - **순차적 패턴(Sequential Pattern)**
    - 비슷한 상품들을 찾아내는 유사행태 집단화(affinity grouping)의 한 방법
    - 시간의 흐름을 고려
    - 예: 집 산 사람의 65%는 2주 안에 냉장고를 산다, 꽃 등 선물용품 -> 웨딩용품 -> 아기용품 -> 어린이 교육 용품

군집분석 (계층적 / K-평균 / 코호넨)

장바구니 분석

순차 연관성 분석

1. 데이터마이닝 개요와 프로세스

1.1 데이터마이닝의 개요: 활용분야

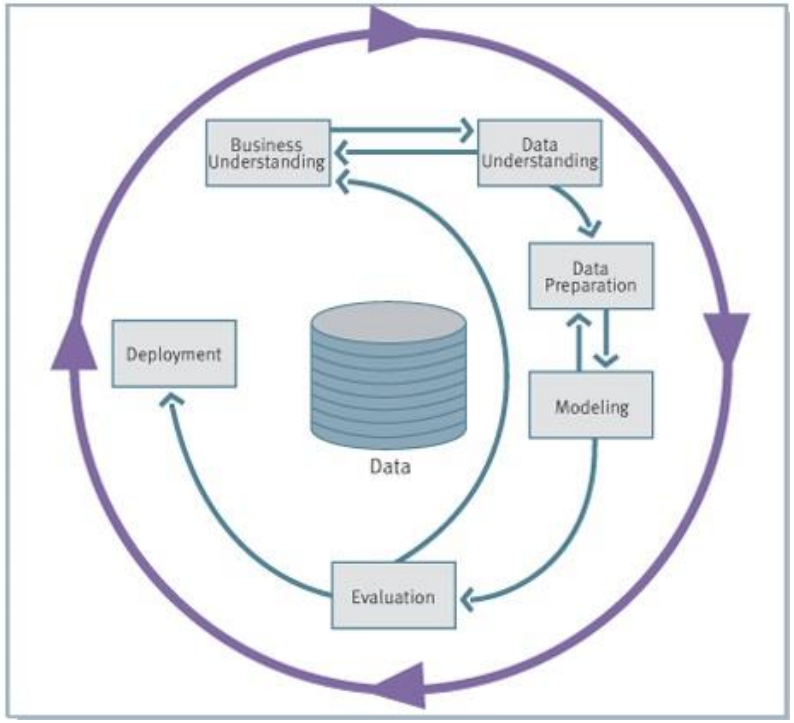
- **데이터마이닝의 활용분야는 대용량 데이터베이스가 구축된 거의 전 분야!!**
  - 금융업(은행/보험/신용카드), 이동통신업, 유통업, 공공기관, 대학 등
- **대표적 활용 예**
  - 카드 도용사고 방지(fraud detection)
  - 위험 관리(risk management)
  - 고객 불만 관리(claim prevention)
  - 고객 유지
    (customer retention, churn management)
  - 고객 유치(customer acquisition)
  - 고객 세분화 및 프로파일링
    (customer segmentation and profiling)
  - 수요 및 판매 예측(forecasting)
  - 가격 산출(pricing)
  - 마케팅 효과 관리(campaign effect analysis)
  - 타겟 마케팅(target marketing)
  - 텔레 마케팅(tele marketing)
  - 다이렉트 메일링(direct mailing)
  - 교차 판매(cross-selling/up-selling)

1. 데이터마이닝 개요와 프로세스

1.2 데이터마이닝 프로세스

❖ 데이터마이닝의 업계 표준 실행: CRISP-DM (IBM SPSS MODELER)

- 데이터마이닝은 대용량의 자료를 이용한 정보화 과정이기 때문에 여러 단계의 절차에 의해 수행됨. 이를 위해 제정된 업계 표준 프로세스인 CRISP-DM(cross-industry standard process for data mining)은 다음 여섯 단계로 구성됨.



데이터 마이닝 표준 실행 과정

- **단계 1: 비즈니스 이해(Business Understanding)**
  각종 참고 자료와 현업 책임자와의 커뮤니케이션을 통해 해당 비즈니스를 이해하는 단계
  반드시 Field Knowledge를 가진 그 분야의 전문가가 함께 참여해야 함.

- **단계 2: 데이터 이해(Data Understanding)**
  레코드 수, 변수(필드) 종류, 자료의 질 등, 현업이 보유 관리하고 있는 데이터를 이해하는 단계

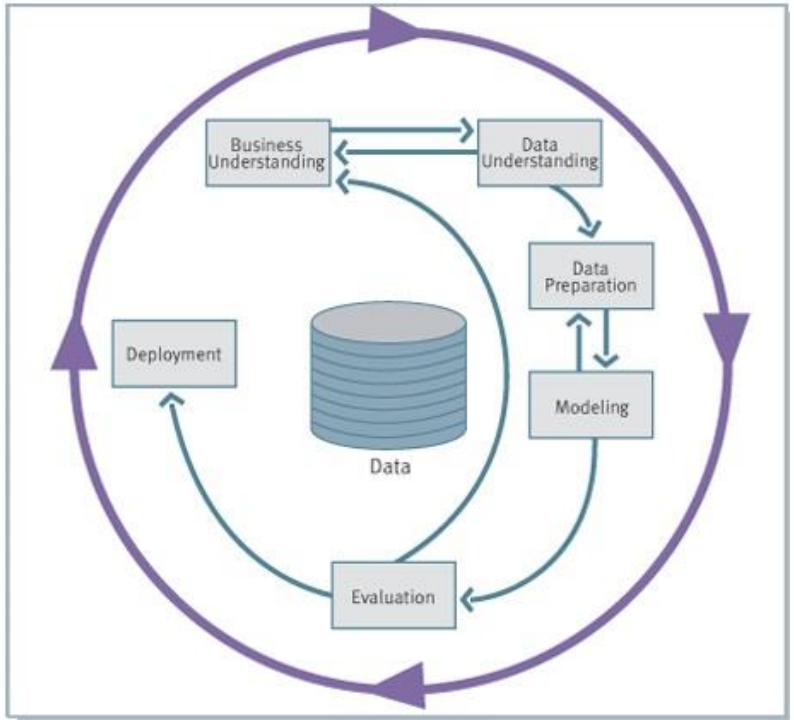
- **단계 3: 데이터 준비(Data Preparation)**
  데이터의 정제, 새로운 데이터 생성, 데이터 업데이트 등, 자료를 분석 가능한 상태로 만드는 단계

1. 데이터마이닝 개요와 프로세스

1.2 데이터마이닝 프로세스

❖ 데이터마이닝의 업계 표준 실행: CRISP-DM

- 데이터마이닝은 대용량의 자료를 이용한 정보화 과정이기 때문에 여러 단계의 절차에 의해 수행됨. 이를 위해 제정된 업계 표준 프로세스인 CRISP-DM(cross-industry standard process for data mining)은 다음 여섯 단계로 구성됨.



데이터 마이닝 표준 실행 과정(계속)

- **단계 4: 모델링(Modeling)**
  자료 기술 및 탐색을 포함하여 필요한 각종 모델링을 하는 단계
  로지스틱 회귀, 의사결정나무, 인공 신경망, K-평균 군집화, 장바구니 분석 모델 등 사용
- **단계 5: 평가(Evaluation)**
  모형의 해석 가능 여부, 독립적인 새 자료에 적용되는 경우에도 재현 가능한지를 검토하는 단계
  예측이 얼마나 잘되었느냐? 현실적으로 모델링이 얼마나 잘 맞는가? What-if 혹은 Sensitivity Analysis 수행 단계
  필요하다면 모형 재구축
- **단계 6: 전개(Deployment)**
  각 관리자에게 전달하여 필요한 조치를 취하는 등 검토가 끝난 모형을 실제 현업에 적용하는 단계

1. 데이터마이닝 개요와 프로세스

1.2 데이터마이닝 프로세스

❖ 데이터마이닝의 업계 표준 실행: SEMMA (SAS Enterprise Miner)

- SEMMA 방법론이란 SAS 기업에서 개발한 데이터마이닝 표준 가이드로써 Sample, Explore, Modify, Model, Assess의 단계로 되어있으며 각 5단계의 약자를 따서 만들었음
- SEMMA는 데이터마이닝을 구현하는데 있어서 하나의 가이드 역할을 할 수 있으며 그림1-3와 같이 5단계를 순차적으로 이루어져 있음 (Sample, Explore, Modify, Model, Assess)
- SEMMA는 데이터 마이닝의 방법론이 아니라 SAS Enterprise Miner Tool의 작업을 수행하기 위한 기능적 논리 구성이다. 또한, SEMMA는 데이터 마이닝 모델 개발 측면에서 초점을 맞추고 있음.

1. 데이터마이닝 개요와 프로세스

1.3 데이터마이닝의 활용사례: 고객 획득

[출처: 허명회, 이용구, "데이터 마이닝 모델링과 사례, 제2판," 한나래, 2008.]

- A은행은 매년 25회 DM(Direct Mailing) 캠페인을 실시한다. 1회 캠페인에서 100만 명을 대상으로 신용카드 가입신청서를 발송한다. 한 건당 1달러의 비용이 발생한다. 100만 명 중 6만 명이 카드를 신청하지만(6% 응답률), 신용조건에 의해 1만 명만 발급된다(1% 성공률). 수익은 카드발급 1인당 125달러 발생한다. ➔ **카드를 받게 될 1만 명에게 효율적으로 접근하는 방법이 필요!**
- **지출-수익 구조개선을 위한 효율적인 데이터마이닝**
  - **1단계:** 5만 명을 대상으로 테스트 DM발송하여 얻은 데이터를 통해 성공(응답-카드신청)에 관한 예측 모형 개발
  - **2단계:** 95만 명에 대해 예측 모형을 적용하여 성공 가능성을 개인별로 산출(70만 명을 선발하여 DM발송)
  - **3단계:** 9,000명이 카드 발급 하지만, 포기한 25만 명에 대한 1,000명 손실
  - **4단계:** 투자비용 및 수익 계산 모형 개발을 통해 85,000달러의 수익 창출

	새 방식의 캠페인	옛 방식의 캠페인	차이
우편 수	750,000건	1,000,000건	-250,000건
우편 비용 (1)	750,000달러	1,000,000달러	-250,000달러
카드 발급자 수 (2)	9,000명	10,000명	-1,000명
카드발급자 1인당 수익 (3)	125달러	125달러	0달러
총수익 (4) : (2)*(3)	1,125,000달러	1,250,000달러	-125,000달러
순수익 (5) : (4)-(1)	375,000달러	250,000달러	125,000달러
마이닝 비용 (6)	40,000달러	0달러	40,000달러
최종 순수익 (5)-(6) :	335,000달러	250,000달러	**85,000달러**

➔ **데이터마이닝 모형화에 따른 비용(소프트웨어 및 인력) 40,000달러를 투자하여 얻은 순수익은 85,000달러이므로 ROI는 200%가 넘는다.**

1. 데이터마이닝 개요와 프로세스

1.3 데이터마이닝의 활용사례: 고객 유지

[출처: 허명회, 이용구, "데이터 마이닝 모델링과 사례, 제2판," 한나래, 2008.]

- B통신은 매월 총 가입자의 8%에 해당하는 8만 명의 고객을 잃는데, 1명의 신규 고객을 확보하기 위해선 200달러의 비용이 들기 때문에 잃는 고객만큼 신규고객을 매월 보충하려면 16,000,000달러가 투입되어야 한다. ➔ **고객의 이탈을 막아 회사의 수익성이 하락하는 것을 막을 방법이 필요! (Churn Management)**

- **우수고객 유지활동을 위한 데이터마이닝**
  - **수익성이 있는 고객들을 식별해내기 위해 데이터마이닝 활용**
    - 매월 인터넷 사용량, 사용 요금 등을 활용하여 현재 수익성 높은 고객과 생애가치(LTV)가 높은 고객을 추출
  - **우수고객에 대한 이탈여부 기록을 분석하여 이탈고객을 예측하는 통계적 모형 구축**
    - 매월 인터넷 사용량과 수리서비스 기록 등을 활용한 결과 3개월에 걸쳐 사용량 추세가 감소하거나 수리서비스 요청건수가 일정 수준 이상, 혹은 20대 여성이면 이탈 가능성이 높다는 사실 발견
      ➔ 개별 고객에 대하여 이탈가능성 점수를 산출
  - **고객이 채택한 요금제가 적절한가를 분석하여 고객에게 유리한 요금제 추천, 품질 개선, 마일리지 보상 등 추진**
  - 일부 고객을 특별 관리하여 이탈율을 8%에서 7%로 1% 하락시키는데 성공 (1만명에 해당)

➔ **비용: 이탈예상고객 특별관리비용[10달러 * 80,000 = 800,000달러] + 마이닝 비용 [200,000달러]**

➔ 효과: 200달러(1인당 신규고객확보비용)*10,000 = 2,000,000달러 절감

➔ **순수효과: 1,000,000달러**

1. 데이터마이닝 개요와 프로세스

1.3 데이터마이닝의 활용사례: 고객 확장

[출처: 허명회, 이용구, "데이터 마이닝 모델링과 사례, 제2판," 한나래, 2008.]

- C쇼핑은 가전제품을 전문으로 파는 업체이다. 잠재고객들에게 정기적으로 카탈로그를 보내며, 한번에 대략 1천 2백 만 가구에 발송한다. 카탈로그를 보고 고객이 전화주문을 하면, 교차판매를 시도한다. 교차판매로 매출이 10% 늘었지만, 그에 못지않게 불평이 많이 늘었다. ➔ **교차판매 마케팅 전략에서 고객의 불평 없이 매출상승효과를 내는 방법이 필요!**

- **교차판매를 위한 데이터마이닝**
  - **교차판매 모형1**
    - 교차판매 시도를 달갑지 않게 생각하는 고객들이 누구인가 하는 것을 알고자, 소규모 면접조사를 실시
    - 조사자료를 활용하여, 교차판매 시도를 기피하는 사람들과 선호하는 사람들을 판별하는 통계적 모형 개발
      ➔ 교차판매 추천을 싫어하는 사람들의 특성을 파악하여, 이 분류에 속한 사람들에게는 교차판매를 추천하지 않음
  - **교차판매 모형2**
    - 현재 주문한 상품을 조건화하여 어떤 상품을 추천할 것인가에 관한 통계적 규칙(연관성규칙)을 도출
      ➔ 일반적으로 해당 고객이 주문한 상품과 같이 주문이 많이 되는 상품을 추천하여 동시구매를 유도

➔ **C쇼핑은 이러한 교차판매 모형을 고객의 전화 주문 시에 적용함으로써 매출을 20% 늘릴 수 있었으며, 고객의 불만도 대폭 줄일 수 있었다.**