






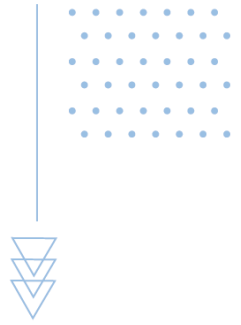
IT와 비즈니스혁신

W2. 데이터 마이닝 개요 및 프로세스



Contents

- I. 데이터 마이닝 개요
 - II. 데이터 마이닝 적용 분야
 - III. 데이터 마이닝 유형
 - IV. 데이터 마이닝 프로세스
- 
- 
- 

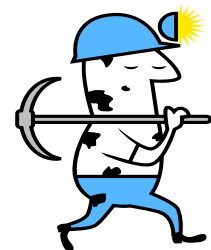


I . 데이터 마이닝 개요

■ 데이터 마이닝 (Data Mining)

- 대용량의 데이터 창고(광석 더미)로부터 유용한 정보(금 조각)를 캐내는(mining, 마이닝) 작업을 의미
- 대용량 데이터에 존재하는 데이터 간의 **관계, 패턴, 규칙** 등을 찾아내고 모형화해서 기업의 경쟁력 확보를 위한 의사결정을 돕는 일련의 과정
- (Berry & Linoff, 2004)
 - 기업이 보유하고 있는 대용량 데이터에서 쉽게 드러나지 않으나 존재하고 있는 **유용한 정보**를 찾아내고 분석하여 기업의 **경쟁력** 확보를 위한 의사 결정에 직접적인 도움이 되는 지식으로 변환하는 일련의 과정
 - 좁은 관점에서는 도구(tools)와 기술(techniques)을 의미
 - 넓은 관점에서는 도구와 기술을 적용시키는 하나의 과정(process) 또는 방법론(methodology)

데이터 속의 **유용한 패턴(규칙, 관계)**을 찾고 이를 이용하여 직접적으로 도움이 되는 **행동(action)**을 취할 수 있어야 함



데이터의 양



컴퓨팅 성능과 데이터 저장기술 향상 및 비용 감소 (Moore's Law)

- Computing power doubles every 18 months.
- The price of computing falls by half every 18 months

데이터의 급격한 증가로 기존 분석 방법으로는 한계가 있음

데이터의 다양성

다양한 형식의 데이터 (오디오, 비디오, 센서, 문자, 수치 데이터 등)를 분석에 이용함으로써 더 많은 수의 차원 (분석에 사용되는 변수) 데이터들을 사용하게 됨
차원의 증가는 더 복잡한 분석기법을 필요로 함

복잡한 문제 해결을 위한 알고리즘

데이터 분석에 대한 관심이 높아지고, 데이터로부터 추출해야 하는 정보 또한 복잡해짐
다양한 고급 분석 알고리즘이 개발되고 이를 활용할 수 있는 상업용 소프트웨어가 일반화됨

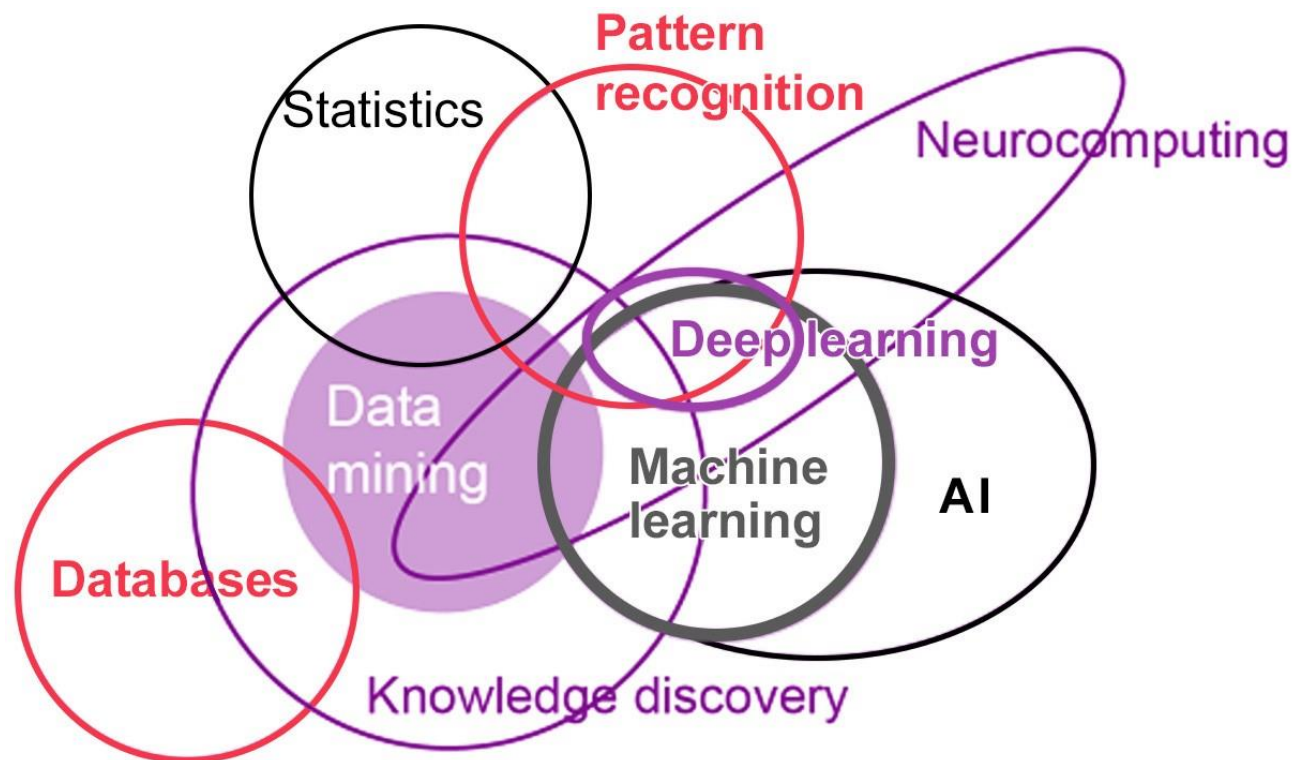
■ 데이터 마이닝 (Data Mining)

- 통계, 인공지능, 기계 학습, 데이터베이스, 패턴인식 분야에서 시작되어 발전
- 데이터로부터 **새로운 지식 발견/추출**이 주요 목적 (알지 못했던 정보를 발견하는 것에 집중)

■ 기계 학습 (Machine Learning)

- 데이터를 통해 학습된 알려진 속성을 기반으로 모델, 패턴, 규칙성 등을 발견하고 결과를 **예측**하는 기법 (**알고리즘, 시스템**에 초점)
- 기계는 기존 데이터를 이용하여 스스로 학습하고 알고리즘 성능 향상

통계, 인공지능, 기계학습, 데이터베이스 등 기존 기술을 기반으로 발전하여 서로 유사한 부분이 있음



How AI intersects with other branches of computer science. Adapted from pwc

■ 기술 통계 (descriptive statistics)

- 평균, 표준편차와 같이 데이터를 요약하여 정량화
- 데이터를 이해하는데 필수적이므로 데이터 전처리, 후처리 단계에서 필수적

■ 탐색적 시각화

- 시각적으로 데이터 표현
- 데이터 전처리, 후처리 단계에서 필수적

■ 차원 슬라이싱

- 차원(상품, 지역, 날짜 등)별로 정량적 데이터(수익, 수량 등)를 보여주나 일종의 정보 검색으로 볼 수 있음

■ 가설 검정

- 통계 검정으로 실험에 사용된 데이터가 가설을 지원할 만한 증거가 충분한지 평가 (유의성)

■ 질의 (Query)


- 데이터베이스에 서 데이터베이스 언어를 이용하여 보고자 하는 정보를 요청
 - 예) 매출액이 높은 상위 5개 제품은 무엇인가


데이터 마이닝은 데이터의 기본적인 요약 정보가 아닌, 새로운 지식 추출이 목표임


지하철 자리앉기 알고리즘 (이번 역에서 내릴 확률 예측)

Logistic Regression, 크리스탈 크리스탈


종속 변수 Y = 이번 역에 내릴 사건

독립 변수 X_1 =  짐보유 여부

X_2 =  소지품

X_3 =  신문

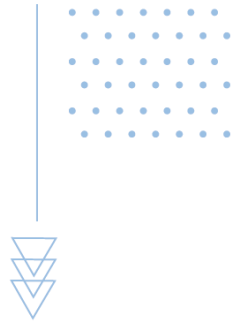
\vdots

X_n =  복장

$P(Y=1 \mid X_1, X_2, X_3, \dots, X_n)$
= 이번 역에 내릴 **확률**

이번 역에 내릴

이 변수들



II. 데이터 마이닝 적용 분야

데이터 마이닝은 이미 마케팅, 금융분야에서 널리 사용되고 있으며, 헬스케어 등 다한 분야로 확장 적용되는 추세

■ 고객 유지를 위한 데이터 마이닝 적용 사례

- S통신은 매월 총 가입자의 8%에 해당하는 8만 명의 고객을 잃는데, 1명의 신규 고객을 확보하기 위해선 \$200의 비용이 들기 때문에 잃는 고객만큼 신규고객을 매월 보충하려면 \$1600만 가 투입되어야 한다.

고객의 이탈을 막아 회사의 수익성이 하락하는 것을 막을 방법이 필요! (Churn Management)

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 요청건 수가 일정 수준 이상, 혹은 20대 여성이면 이탈 가능성이 높다는 사실 발견 -> 개별 고객에 대하여 이탈가능성 점수 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[\$10 * 8만 = \$80만] + 마이닝 비용 [\$20만] = \$100만

효과: \$200(1인당 신규고객확보비용)*1만 = \$200만 절감 -> 순수효과: \$100만

■ 고객 확장을 위한 데이터 마이닝 적용 사례

- M쇼핑은 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!

교차판매 모형1

- 교차판매 시도를 달갑지 않게 생각하는 고객들이 누구인가 하는 것을 알고자, 소규모 면접조사를 실시
- 조사자료를 활용하여, **교차판매 시도를 기피하는 사람들과 선호하는 사람들을 판별**하는 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 **연관성규칙**을 도출
- > 일반적으로 해당 고객이 주문한 상품과 같이 주문이 많이 되는 상품을 추천하여 동시구매를 유도

위의 교차판매 모형을 고객의 전화 주문 시에 적용함으로써 매출 20% 늘릴 수 있었으며 고객의 불만도 대폭 줄일 수 있었다.