3-1 데이터 마이닝 개요

데이터 마이닝

- 데이터 마이닝(Data Mining)이란 '광산에서 광물을 캐내는 일'을 의미하는 'mining'에서 유래했으며, 대규모의 데이터에서 유용한 인사이트(Insight)를 발견하고 분석하여 가치(Value)를 창출하며, 그 결과를 의사결정에 반영하는 과정

(1) 데이터 마이닝의 기능

기능	설명	예시 또는 유형
분류 (Classification)	데이터를 미리 정의된 여러 그룹 또는 클래스에 할당하는 것 (**주로 범주형**)	개와 고양이의 분류
예측 (Prediction) 또는 회귀 (Regression)	미래의 양상을 예측하거나 미래의 값을 추정하는 것 (**주로 연속형**)	주식 가격의 예측
연관분석 (Association Analysis)	같이 팔리는 물건과 같이 아이템의 연관성을 파악하는 분석	시장바구니분석
군집분석 (Clustering Analysis)	이질적인 모집단을 동질성을 지닌 그룹별로 세분화하는 것을 의미	온라인샵고객군집분석
기술 (Description)	데이터의 특징 및 의미를 표현하거나 설명하는 것	

(2) 데이터 마이닝의 단계

목준가적검

- **목적 정의 → 데이터 준비 → 데이터 가공 → 데이터마이닝 기법 적용 → 검증**

단계	설명
목적 정의	- 데이터마이닝 도입의 목적을 분명히 정의하는 단계 - 가능하면 전문가가 참여하여 사용할 데이터마이닝 모델과 데이터를 정의하는 것이 바람직함

단계	설명
데이터 준비	- 고객정보와 거래정보, 상품 마스터 정보 등 데이터마이닝 수행에 필요한 데이터를 수집하는 단계 - 데이터는 대부분 용량이 크므로 IT부서와 협의하고 도움을 요청함 - 데이터 정제를 통해 품질을 보장하고, 필요하면 보강 작업을 거쳐 데이터 양을 충분히 확보
데이터 가공	- 데이터마이닝 기법 적용이 가능하도록 수집된 데이터를 가공 - 모델링 목적에 따라 목적 변수를 정의하고, 필요한 마이닝 소프트웨어를 적용할 수 있도록 적합한 형식으로 가공
데이터마이닝 기법 적용	- 데이터마이닝 기법을 적용해 목적하는 정보를 추출 - 데이터마이닝 적용 목적, 데이터, 산출물의 조건에 따라 적절한 소프트웨어와 기법 선정
검증	- 데이터마이닝으로 추출한 정보를 검증하는 단계 - 최적의 모델을 선정하고 결과를 업무에 적용

지도 학습과 비지도 학습

1. 지도 학습 (Supervised Learning)

- 레이블 또는 정답이 주어진 데이터를 사용하여 모델을 학습시키는 방법이다.
- 주어진 입력 데이터(X)에 대한 적절한 출력 데이터(Y)를 찾는 모델을 찾는 것이 목표이다.
- **분류(Classification)**, **회귀(Regression)** 등을 수행한다.

2. 비지도 학습 (Unsupervised Learning)

- 레이블 또는 정답이 주어지지 않은 데이터를 사용하여 모델을 학습시키는 방법이다.
- 데이터의 구조나 패턴을 찾아내는 것이 목표이다.
- **연관분석(Association Analysis)**, **군집분석(Clustering Analysis)** 등을 수행한다.

분류

- 분류는 새롭게 나타난 현상을 검토하여 기존의 분류 정의된 집합의 배정하는 것을 의미
- 분류 작업은 잘 정의된 분류 기준과 선분류 되어진 검증 집합에 의해 완성
- 완성된 분류 작업을 통해 모형이 구축되면 그 모형을 이용하여 분류되지 않은 다른 현상들을 분류할 수 있음

군집

- 군집은 이질적인 모집단을 동질성을 지닌 그룹별로 세분화하는 것을 의미
- 군집과 분류와의 차이점을 살펴보면 군집은 선분류 되어 있는 기준에 의존하지 않는다
- 다시 말하면 미리 정의된 기준이나 예시에 의해서가 아닌 레코드 자체가 지니고 있는 다른 레코드와의 유사성에 의해 그룹화되고 이질성에 의해 세분화된다

지도 학습 알고리즘

분류	알고리즘
분류	- 로지스틱 회귀분석 (Logistic Regression) - 나이브 베이즈 (Naive Bayes)
회귀	- 선형 회귀분석 (Linear Regression) - 릿지 회귀분석 (Ridge Regression) - 라쏘 회귀분석 (Lasso Regression) - 엘라스틱넷 (ElasticNet)
분류 + 회귀	- 의사결정나무 (Decision Tree) - 서포트 벡터 머신 (SVM) - K-최근접 이웃 (KNN: K-Nearest Neighbors) - 앙상블 기법: 배깅, 부스팅, 랜덤 포레스트 - 인공 신경망 (Neural Networks)

비지도 학습 알고리즘

유형	알고리즘
군집	- 계층적 군집분석 (Hierarchical Clustering) - K-평균 군집분석 (K-Means Clustering) - K-중앙값 군집분석 (K-Medoids Clustering) - 혼합 분포 군집분석 (Gaussian Mixture Models) - 밀도 기반 군집분석 (Density-Based Clustering) - 자기 조직화 맵 (SOM: Self-Organizing Maps)
연관	- Apriori - FP(Frequent Pattern)-Growth

종료