[머신러닝 기반 데이터 분석] 01. 머신러닝 수행방법 계획하기

01. 머신러닝 수행방법 계획하기
02. 데이터 세트 분할하기
03. 지도학습 모델 적용하기
04. 자율학습 모델 적용하기
05. 모델성능 평가하기
06. 학습결과 적용하기

학습 목표
가. 머신러닝의 기본 개념에 대해 알아본다.

목차

1-1 머신러닝 정의

(가) 머신러닝이란?

- 머신러닝(Machine Learning)은 컴퓨터 과학의 영역에 속하는 인공지능의 한 분야
- 컴퓨터 프로그램이 어떤 것에 대한 학습을 통해 기존의 모델이나 결과물을 개선하거나 예측하게끔 구축하는 과정

(나) 머신러닝의 구성요소

- **작업 : 수행하고자 하는 작업**

- **모델(model) : 작업을 수행하기 위해 주어진 데이터로부터 학습한 '모델'**
- **특성(features) : 모델을 학습하는 과정에서 개체를 측정한 특성**

<표 1-1> 머신러닝에 대한 다양한 정의와 의미

머신러닝에 대한 다양한 정의	주요 제안자 및 출처
- 컴퓨터에게서 배울 수 있는 능력, 즉 코드로 정의하지 않은 동작을 실행하는 능력에 관한 연구 분야	Arthur Samuel (1959)
- 작업 T에 대한 성능을 P로 측정할 수 있고, 성능 P가 경험 E로 개선된다면 프로그램은 E,T,P 로부터 배웠다고 할 수 있다.	Tom Mitchell (1996, 1997)
- 정확한 '작업(Task)'을 성취할 수 있는 올바른 '모델(Model)'을 구축하기 위해 올바른 '특성(Features)'을 활용하는 것	Peter Flach (2012)
- 컴퓨터 프로그램이 어떤 것을 학습한 후에 최초 학습에 들인 시간과 노력보다 더 빠르고 수월하게 배운 것을 해낼 수 있게 하는 것	Jason Bell (2015)
- 인공지능의 한 분야로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야	Wekipedia

(ncs 모듈 교재 참조)

(다) 머신러닝의 활용 분야

스팸 메일 분류
주식매매 등의 알고리즘 트레이딩
추천시스템
컴퓨터 비전
무인 자동차
인공지능 알파고
...

*** 머신러닝은 명시적인 알고리즘을 설계하기 어렵거나 프로그래밍하기 어려운 작업을 해결하기 위해 주로 사용된다.**

(라) 통계학과 머신러닝

통계학은 전통적으로 데이터를 정보로 변환하기 위한 과학적이고 체계적인 방법을 제공하는 이론적 토대이다.

통계학

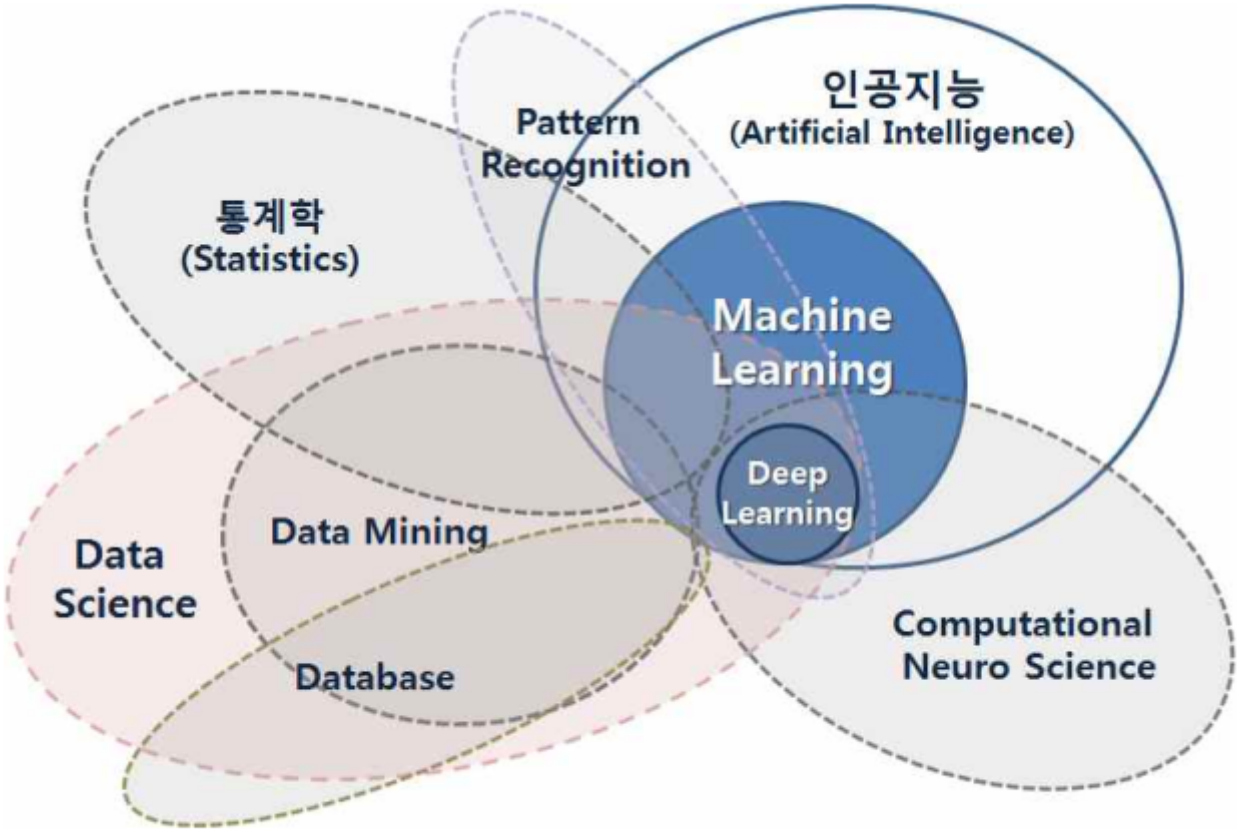
- **강조되는 영역은 추론과 검증이다.**
- **주어진 데이터가 가설과 이론에 얼마나 부합하는가 등을 설명하기 위해 다양한 방법론과 이론이 구축되어 있음.**

머신러닝

- 머신러닝은 명시적인 알고리즘을 설계하기 어렵거나 프로그래밍하기 어려운 작업을 해결하기 위해 주로 사용.
- A. 데이터가 생성한 잠재적인 메커니즘의 특징을 파악 후,
- B. 복잡한 관계를 정량화 한 후,
- C. 이 식별된 패턴을 사용하여 새로운 데이터의 예측

데이터 마이닝과 머신러닝

- **데이터 마이닝(data mining)**은 주로 대규모의 저장된 **데이터 안**에서 체계적이고 자동적으로 의미 있는 **규칙이나 패턴을 발견하고 이를 지식화**하는 과정
- **머신러닝(machine learning)**은 주어진 **입력 데이터**를 **컴퓨터 프로그램**이 학습하여 **예측을 수행**하고 스스로의 **예측 성능을 향상시키는 과정**과 **이를 위한 알고리즘을 연구**하고 **구축하는 기술**이다.



[그림 1-1] 머신러닝과 여러 학문 분야와의 연계성

1-2 머신러닝 데이터 분석 기법의 유형
지도 학습(Supervised Learning)
자율학습 혹은 비지도 학습(Unsupervised Learning)
강화학습(Reinforcement Learning)
준지도 학습(Semi-Supervised Learning)

(가) 지도 학습(Supervised Learning)
지도 학습은 설명변수(혹은 독립변수, 특성(Feature) 등으로 표현)와 목적변수(혹은 반응변수, 종속변수, 목표변수, 출력값 등으로 표현) 간의 관계성을 표현해내거나 미래 관측을 예측해 내는 것에 초점

적합한 분야 : 인식, 분류(classification), 진단, 예측(prediction) 등의 문제 해결

<표1-2> 지도 학습 기법의 여러 유형

분류	수치예측(혹은 회귀)
- K-최근접 이웃(K-Nearest Neighbors)	- 선형 회귀(Linear Regression)
- 로지스틱 회귀(Logistic Regression)	- 확장된 회귀분석(ex : 다항회귀, 비선형 회귀, 벌점화 회귀 등)
- 인공 신경망 분석(Artificial Neural Network)	- 인공 신경망 분석(Artificial Neural Network)
- 의사결정트리(Decision Tree)	- 의사결정트리(Decision Tree)
- 서포트 벡터 머신(Support Vector Machine)	- 서포트 벡터 머신(회귀) (Support Vector Machine (Regression))
- 나이브 베이즈(Naive Bayes)	- PLS(Partial Least Squares)
- 앙상블 기법(랜덤 포레스트 등)	- 앙상블 기법(랜덤 포레스트 등)

(나) 자율학습 혹은 비지도 학습(Unsupervised Learning)

자율학습 혹은 비지도 학습은
목적변수(혹은 반응변수, 종속변수, 목표변수, 출력값)에 대한 **정보 없이**
학습이 이루어지는 형태

대표적인 기법

- 군집화(Clustering)
- 차원축소기법(PCA)
- 연관관계분석(장바구니 분석)
- 자율학습 인공 신경망(SOM 등)

딥러닝기법에서 입력 특성들의 차원 축소 단계에서 자율학습 기법 적용

강화학습(Reinforcement Learning)
준지도 학습(Semi-Supervised Learning)

1-3 머신러닝 기반 데이터 분석 계획 및 절차
(가) 비즈니스 이해 및 문제 정의
A. 문제 정의를 해야 한다.
문제를 해결하기 위한 비즈니스 도메인 이해 및 **문제를 파악해 가는 과정**을 반복해 가면서
B. 문제를 재정의하고 **해결책을 모색**하는 단계를 반복적으로 거치게 된다.

문제 정의 과정과 필요한 데이터 형태에 따라 구상을 하는 과정에서 자연스럽게 어떤 머신러닝 기법을 적용하게 될지 1차적인 잠정적 의사결정을 거치게 된다.

(나) 데이터 수집
분석 이슈가 명확해 진다면, 분석을 위해 필요한 데이터를 수집하게 된다.

어디에서?
내부 데이터 저장소에서 SQL을 통해 데이터를 추출한다.
빅데이터 플랫폼에서 데이터를 추출하는 경우가 일반적이다.
어떤 경우는 **외부 데이터가 필요하면 웹 사이트에서 필요한 데이터를 스크래핑 형태로 수집**하거나
API등을 통해 데이터를 수집해야 할 경우가 생길 수 있다.

(다) 데이터 전처리와 탐색

- 머신러닝에 적용을 위한 적당한 형태로 데이터를 전처리하고 변환

- 머신러닝 기반 데이터 분석 결과의 질은 기법이나 알고리즘에 따라 좌우되지만, 데이터의 질에 따라 크게 좌우된다. 그래서 전처리와 변환 및 탐색 단계가 매우 중요하다.
- 대부분의 분석과 마찬가지로 전체 프로세스의 이 단계가 가장 많은 시간과 노력을 들이게 되는 단계이다.

(라) 모델 훈련

데이터 전처리와 탐색 과정 후, 머신러닝 기법을 적용하여 데이터를 학습을 수행한다.

지도학습의 경우, 모델 훈련을 위해 데이터
[데이터 세트 분할]
학습용 데이터와 검증용 및 평가용 데이터로 분할하거나, **교차 검증** 등에 대한 설계를 거친 후
모델 훈련을 수행한다.

자율학습의 경우,
목적값을 가지지 않기에
모델 훈련보다는 바로 분석을 통해 패턴 도출의 과정을 수행한다고 말할 수 있다.

(마) 모델 성능 평가

(가) 평가 데이터 세트를 이용하여 **모델의 정확도를 평가**한다.
(나) 자율학습의 경우, **평가 데이터 세트를 두지 않는** 경우가 일반적이다.
교차검증보다 분석과정에서의 통계치나 규칙들의 해석 가능성 등에 초점을 둔다. 이에 따라 평가 수행.

(바) 모델 성능 향상 및 현업 적용

일반적으로 단일 머신러닝 분석 프로세스로 해결하고자 하는 이슈가 단번에 해결되는 경우는 거의 없으며,
지속적으로 모델 파라미터나 추정방법 등을 변화시켜서 모델성능을 꾀하게 된다.

1-4 머신러닝 기반 데이터 분석을 위한 데이터 이해 및 전처리를 수행

(가) 프로그래밍 환경

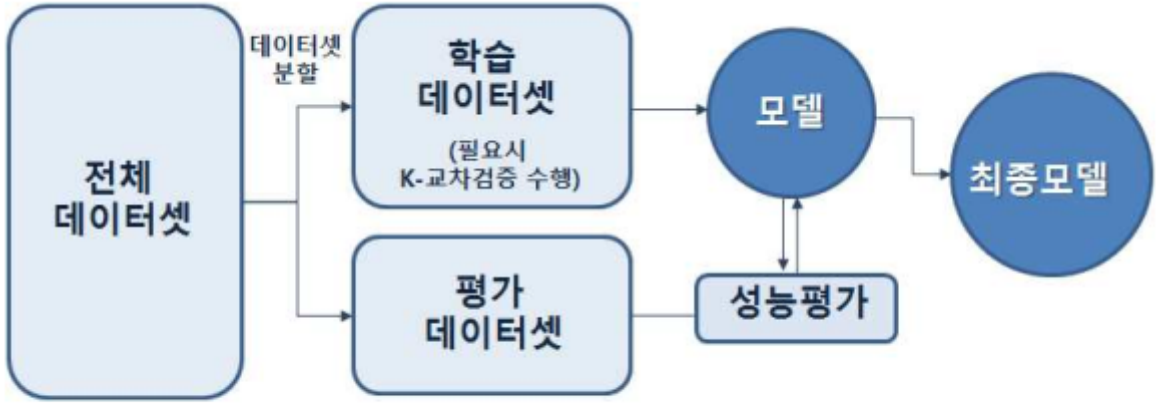
R, 파이썬(Python), SAS, JAVA, Scala 등이며 현업에서는 R과 파이썬이 많이 사용되고 있음.

(나) 시스템 환경

스파크, 머하웃 라이브러리 사용하여 머신러닝을 배치 혹은 실시간 형태로 수행.

(다) 머신러닝 적용을 위한 데이터 전처리와 기초 이해

A. 결측치 및 이상치 처리
B. 데이터 형태 변환, 수치 가공
C. 데이터의 정규화 및 표준화
D. 데이터 집계 및 요약
E. 탐색적 관점의 데이터 시각화



[그림 2-2] 훈련 데이터와 평가 데이터 분할 통한 머신러닝 모델링 수행절차

1-5 모델 성능 평가 지표를 선정하고 모델 성능을 개선

A. 모델 훈련 완료 후, 평가용 데이터를 통해 예측을 수행.

분류 모형의 경우, 예측결과와 실제 값의 대조를 위해
혼동행렬(Confusion Matrix)를 작성 후, 분류 정확도 등을 확인.
민감도(Sensitivity, Recall, Hit Ratio)나 정밀도(Precision)등을 이용.
=> 초반에 만족되지 않으므로 파라미터 튜닝과 여러 기법을 적용해 본다.

1-6 머신러닝 결과 도출된 산출물을 작성.

A. 머신러닝 기반 데이터 분석 계획서

B. 주요 사용 데이터 및 확보 방안
C. 머신러닝 모델 훈련 및 예측 결과
D. 비즈니스 성과 개선 및 기여 계량 자료