[머신러닝 기반 데이터 분석] 03. 지도학습 모델 적용하기

학습 내용

- 지도학습과 비지도학습에 대한 용어에 대해 이해한다.
- 지도학습의 종류에 대해 알아본다(분류, 회귀)
- 비지도학습에 대해 알아본다.

1-1 용어 이해하기

지도학습(Supervised Learning)

머신러닝을 통해 새로운 데이터의 목적변수(혹은 반응변수) 값을 추정하거나
분류하고자 하는 머신러닝 기법들을 지도학습(Supervised Learning) 기법이라 한다.

지도학습의 구분 - 분류, 회귀

- **목적변수의 형태**에 따라 분류와 회귀 문제로 구분이 가능.
- **목적변수가** '남성/여성', '스팸메일/햄 메일', '긍정/중립/부정' **이산형 혹은 명목형**- **분류기법**
- **목적변수가** 0~10, -500~500, -3.5~3.5 와 같은 **수치 형태** - **회귀문제**

<표 1-2> 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)

(NCS 교재 참조)

비지도학습(Unsupervised Learning)

- 자율학습 혹은 비지도 학습은 목적변수(혹은 반응변수, 종속변수, 목표변수, 출력값)에 대한 정보 없이 학습이 이루어지는 형태
- 예측의 문제보다 주로 현상의 기술(Description)이나 특징 도출, 패턴 도출 등의 문제에 많이 활용
- 데이터 마이닝의 성격이 강하다.

1-2 머신러닝 알고리즘 - 분류(Classification) 머신러닝 활용 영역

대표적인 예시

(1) 스팸 메일 분류
(2) 기업 부도 / 정상 예측
(3) 고객 이탈 / 유지 예측
(4) 고객 신용등급 판별
(5) 특정 질병(ex : 암, 심장병 등) 발생 여부 예측
(6) 특정 마케팅 이벤트에 대한 고객 반응 여부 예측
(7) 고객의 구매 여부 예측

<표 3-1> 분류 목적 주요 머신러닝 알고리즘 기법

종 류	개 념	비 고
K-최근접 이웃 (K-Nearest Neighbor)	특정 데이터 좌표점과 다른 나머지 데이터 좌표점들간의 거리에 기반을 두어 가장 가까운 K개점들의 목적변수(혹은 반응변수) 값들의 다수결로 분류하는 기법	게으른 학습(Lazy Learning)
나이브 베이즈 (Naive Bayes)	베이즈 정리에 근거하여, 목적변수(혹은 반응변수)가 발생할 조건부확률을 사전확률과 우도 함수의 곱으로 표현하여 어떤 분류항목에 속할지 확률이 높게 계산되는 쪽으로 분류하는 기법. 이때 모든 관측값은 서로 다른 관측값과 통계적으로 독립적으로 발생한다고 가정 (별다른 확신 없이 가정하므로 Naive 모형이라고 부름)	확률모형 (베이즈 정리 기반 조건부확률)
로지스틱 회귀 (Logistic Regression)	설명변수 값이 주어졌을 때, 목표변수 값이 특정 부류에 속할 확률이 로지스틱 함수 형태를 따른다고 가정하여 최대우도 추정방법으로 목표변수의 확률을 추정하는 기법	확률 모형 (최대우도 추정법)
의사결정트리 (Decision Tree)	목표변수와 가장 연관성이 높은 변수의 순서대로 불순도나 엔트로피 등이 낮아지는 방향으로 나무 형태로 가지를 분할하면서 분류 규칙을 만들어 내는 기법	분할 정복기법 (Divide & Conquer)

<표 3-1> 분류 목적 주요 머신러닝 알고리즘 기법 (계속)

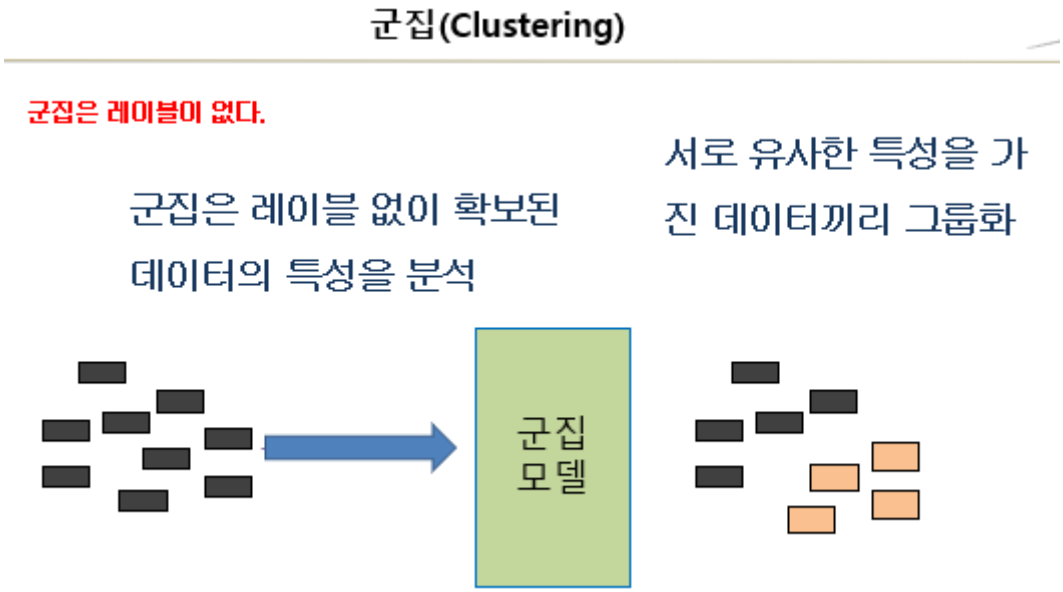
종 류	개 념	비 고
인공 신경망 분석 (Artificial Neural Network)	인간의 뇌의 뉴런 작용 형태에서 모티브를 얻은 기법으로서, 입력 노드와 은닉 노드, 출력 노드를 구성하여 복잡한 분류나 수치예측 문제를 해결할 수 있도록 하는 분석 기법	블랙박스기법
서포트 벡터 머신 (Support Vector Machine)	특정 데이터 들을 분류하는 데 있어, 서로 다른 분류에 속한 데이터 간의 간격(마진)이 최대화가 되는 평면을 찾아 이를 기준으로 분류하는 기법	선형 및 비선형(커널트릭)
랜덤 포레스트 (Random Forest)	주어진 데이터로부터 여러 개의 다양한 의사결정트리를 만들어 각 의사결정트리의 예측결과를 투표형식으로 집계하여 최종 분류 결과를 결정하는 앙상블 형태의 기법	앙상블 모형

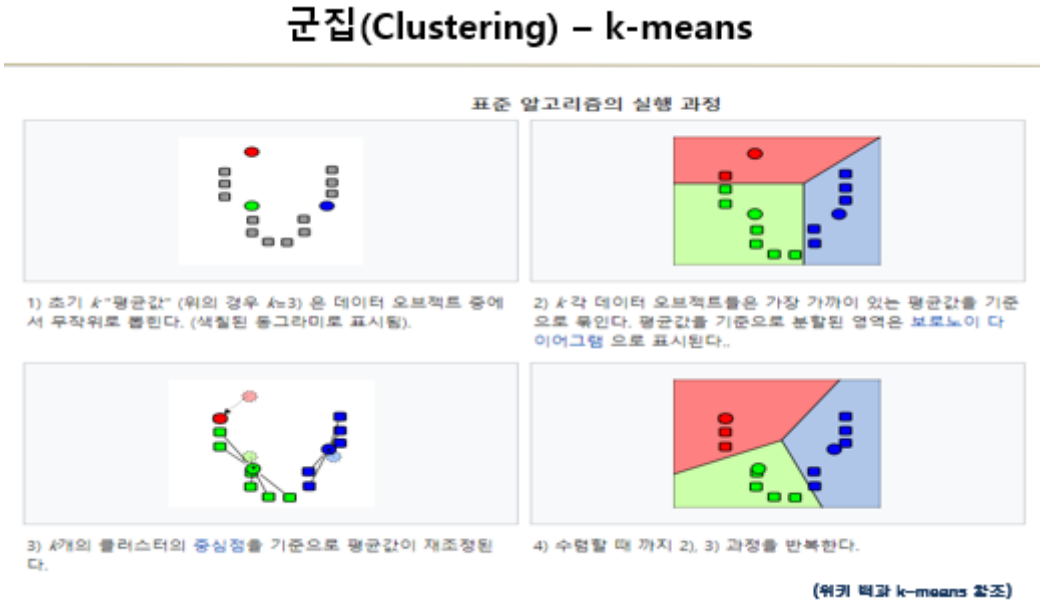
1-3 머신러닝 알고리즘 - 지도학습과 비지도학습

A. 지도학습

분류(Classification) – 이항분류와 다항분류

B. 비지도학습





1-4 머신러닝 알고리즘 - 지도학습과 비지도학습

가. 주어진 데이터가 예측하고자 하는 **목적 변수(혹은 반응변수) Y를 가질 경우**, 적용 알고리즘

--> 지도학습(Supervised) 기법 적용
--> KNN, 로지스틱 회귀분석, 나이브 베이즈, 의사결정트리등

나. 주어진 데이터가 분류 항목 표시나 **목적 변수(혹은 반응변수)가 없고,** 목적값 예측을 시도하는 경우가 아닐 경우

--> 비지도학습(UnSupervised, 자율학습) 기법 적용
--> 종류 : 군집(Clustering), 연관성 분석(Association), 차원축소(Dimension Reduction)