[머신러닝 기반 데이터 분석] 02. 데이터 세트 분할하기

학습 목표

가. 지도학습과 비지도학습의 차이를 이해해 본다.
나. KNN에 대해 이해해 본다.

목차

▶ 대표적인 지도학습(supervised learning)

(가) 회귀(예측) - Regression

(나) 분류(Classification)

공통점 : 입력 및 **목표 변수의 값**을 이용하여
주어진 **입력변수에 대한 목표변수의 값을 예측**하는
모형을 개발한다.

▶ 대표적인 지도학습(supervised learning)

(가) 회귀(예측) - Regression

(나) 분류(Classification)

차이점 :

A. 목표 변수의 형태가 회귀의 경우 연속형이다.

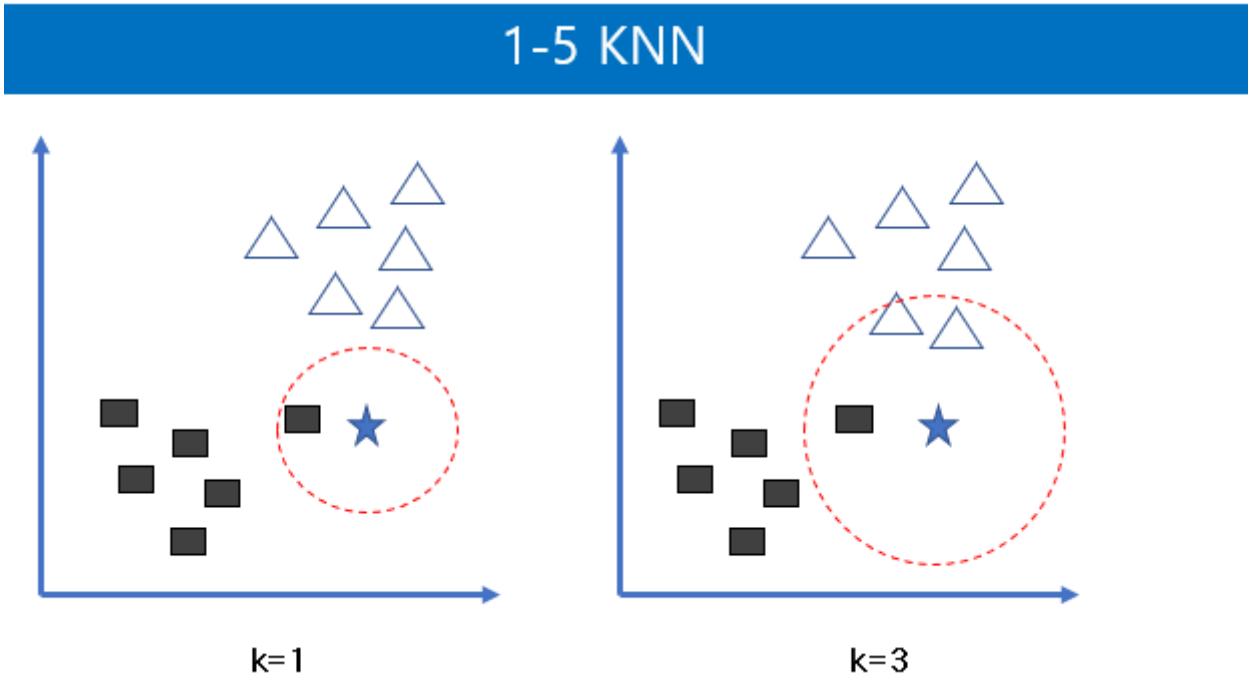
B. 분류의 경우는 범주형이다.

> 비지도학습

주어진 데이터에서 분류항목 표시나 **목적변수(혹은 반응변수)가 없고** 목적 값 예측을 시도하는 것이 아닐 경우, 자율학습 혹은 **비지도 학습 (Unsupervised Learning)** 기법 적용

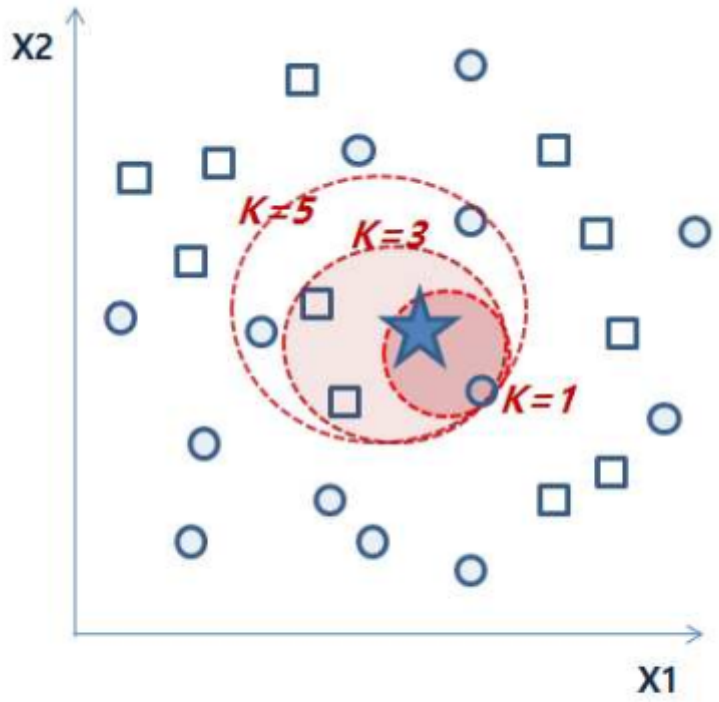
(나) 분류 목적 머신러닝 주요 기법 핵심 개념 및 특징

1. K-최근접 이웃(K-Nearest Neighbor) 기법



K-최근접 이웃 기법은 **목표변수의 범주를 알지 못하는** 데이터 세트의 분류를 위해
해당 데이터 세트와 가장 유사한 주변 데이터 세트의 범주로 지정하는 방식으로 분류 예측을 한다.

주변 데이터 세트를 몇 개로 기준으로 판단할 것인가에 대한 기준 필요(K개)



[그림 3-1] K-최근접 이웃 기법에서의 K 값 변화에 따른 목표변수 범주 변화

(NCS 모듈 교재)

K=1로 설정

'☆'와 가장 가까운 데이터 점은'◯'이므로'☆'의 목표변수는 **'◯'로 분류**

K=3으로 설정

'☆'와 가장 가까운 3개의 점을 고려하게 되므로
'□'가 2개,'◯'가 1개이므로 이때는**'☆'점은'□'로 분류**

K=5로 설정

'☆'와 가장 가까운 3개의 점을 고려하게 되므로'□'가 2개,'◯'가 1개이므로
이때는 **'☆'점은'□'로 분류**

K-최근접 이웃(K-Nearest Neighbor) 기법 장단점

장 점	단 점
• 알고리즘 이해하기 쉽고 직관적 • **빠른 훈련(학습)시간** • **데이터 세트의 확률분포 등에 대한 가정이 필요없다.**	• 많은 메모리 소요(대용량 데이터 불리) • **느린 분류(예측) 소요시간** **새로운 데이터가 주어질 때마다 모든 데이터와의 유사성 계산을 해야 한다. 게으른 학습(Lazy Learning)으로 불린다.** • 모델(모형)이 없으므로 변수들간의 관계 등 가설검증이나 구조 등에 대한 분석을 통해 통찰력을 얻기 어렵다. **(주어진 데이터 통해 범주의 분류 결과 판단)**

K-최근접 이웃(K-Nearest Neighbor) 활용 분야
온라인 및 모바일 서비스 추천 시스템에서 K-최근접 이웃 기법에 근거한 상품 및 서비스 추천 등이 대표적인 분야라고 할 수 있다.

2. 로지스틱 함수(Logistic function) 기법

- 예측하고자 하는 것이 목표변수(Y)가 아닌, 목표변수 Y가 특정 범주(i)가 될 확률-P(Y=i)이다.

만약 **목표변수 Y의 범주가 0,1 두 가지만 있다고 가정**한다.
목표변수 P(Y=1) = P(Y)로 표기하면, 이를 회귀식으로 다음과 같이 표현한다.

$$\frac{P(Y)}{1-P(Y)} = \exp(\beta_0 + \beta_1 X)$$

여기서 **좌편은 오즈(Odds)**라고 한다.

좌편은 확률들의 비율이고, 우변은 지수함수의 형태이다.
따라는 이는 **값의 범위는 (0, ∞)**의 범위를 갖는다.

양쪽 좌변, 우변의 같은 값의 범위를 가지게 하기 위해 **양변에 Log 함수**를 취한다.

$$\log\left(\frac{P(Y)}{1-P(Y)}\right)=\beta_0+\beta_1 X$$

log로 인해 우변은 선형모델이 되어 범위가 (-∞,∞)가 되고,
좌변도 동일하다.
좌변의 log(P(Y)/(1-P(Y)) 를 우리는 **로짓(logit)함수**라 한다.

위의 식을 exp를 취하고 다시 P(Y=1)로 정리하면

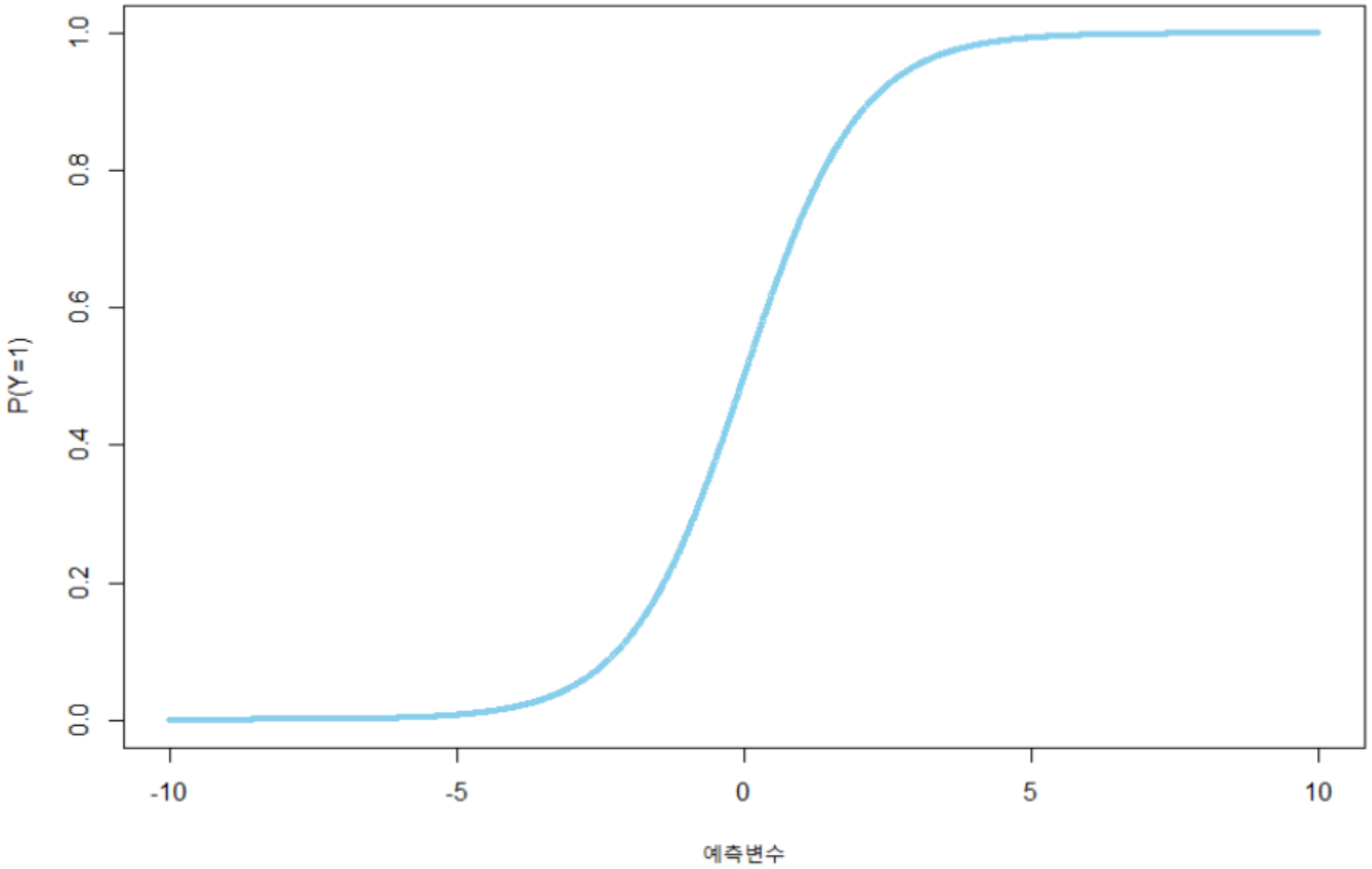
$$P(Y=1)=\frac{e^{(\beta_0+\beta_1 X)}}{1+e^{(\beta_0+\beta_1 X)}}=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

다음과 같이 식이 만들어진다.

로지스틱 회귀분석은 주어진 훈련 데이터에서 목표변수가 Y가 범주값 1을 가질 확률 P(Y=1)를

위의 식의 로지스틱 함수를 이용하여 모델을 만들고 모수 β_0, β_1 들을 추정하는 알고리즘이다.

β_0, β_1 추정에는 일반적으로 **최대우도추정법(Maximum Likelihood Estimation)을 사용**한다.



[그림 3-2] 로지스틱 회귀 곡선 그래프

(NCS 교재 참조)

<표 3-4> 로지스틱 회귀분석의 장단점

장 점	단 점
• 선형통계모형의 이론에 기반한 정교하고 체계적인 모수 추정이 가능하다. • 확률모형이므로, 목표변수의 범주 확률값을 추정할 수 있다. • 추정된 모형의 계수에 대한 해석이 가능하며, 독립변수들의 유의성 및 영향력 등 결과 분석 시 유용한 해석이 가능하다.	• 데이터 세트의 차원이 매우 많을 때 모형의 추정 정확도가 다른 머신러닝 기법에 비해 좋지 않다. • 추정 방법상 x값이 매우 커지거나 작아지면 확률값이 1(혹은 0)에 매우 가까워져서 수치계산 정확도가 떨어지게 되며, 반복 계산 시 오버 피팅이 빈번하게 발생한다. • 복잡한 비선형적 분류가 필요한 경우에는 분류 정확도가 좋지 않다.

(history) 2019.01.01 logistic 추가