3장 머신러닝 핵심 알고리즘

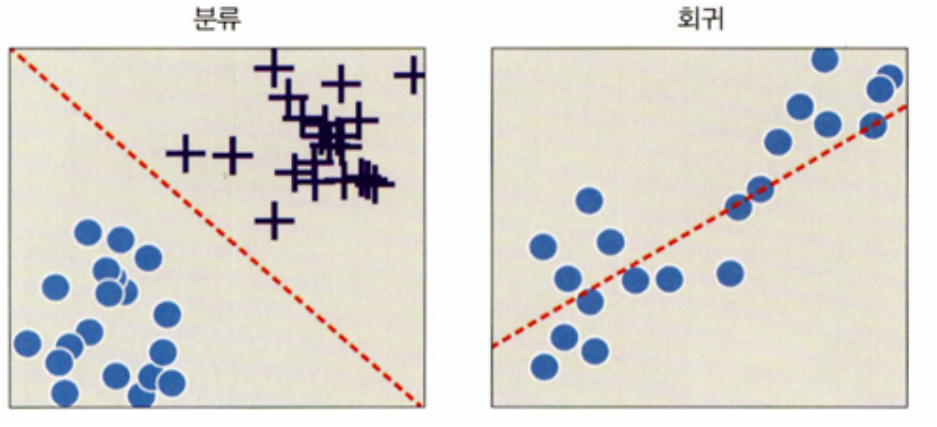
생성일	@2024년 9월 22일 오후 2:11
주차	Week 2
완료여부	☐

3.1 지도학습

정답 레이블을 컴퓨터이 미리 알려주고 데이터를 학습시키는 방법으로 지도 학습에는 회귀 & 분류가 있음

▼ 표 3-1 분류와 회귀 차이

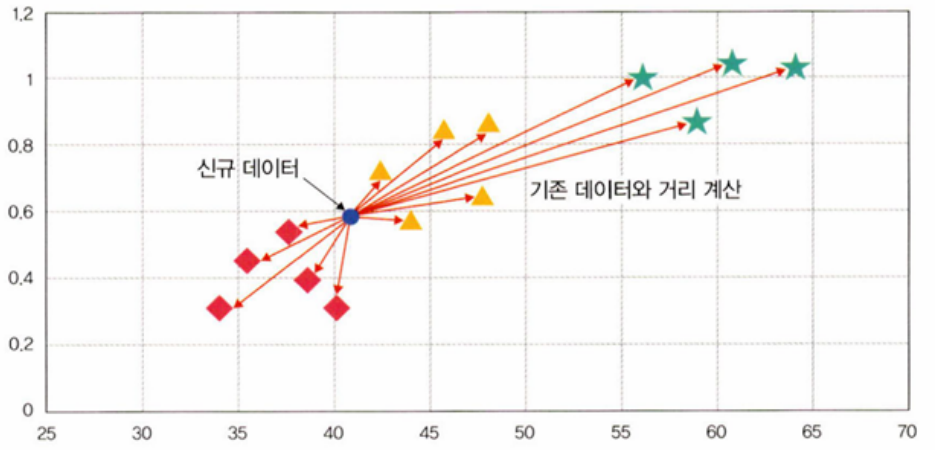
구분	분류	회귀
데이터 유형	이산형 데이터	연속형 데이터
결과	훈련 데이터의 레이블 중 하나를 예측	연속된 값을 예측
예시	학습 데이터를 A · B · C 그룹 중 하나로 매핑 예 스팸 메일 필터링	결괏값이 어떤 값이든 나올 수 있음 예 주가 분석 예측



3.1.1 K-최근접 이웃

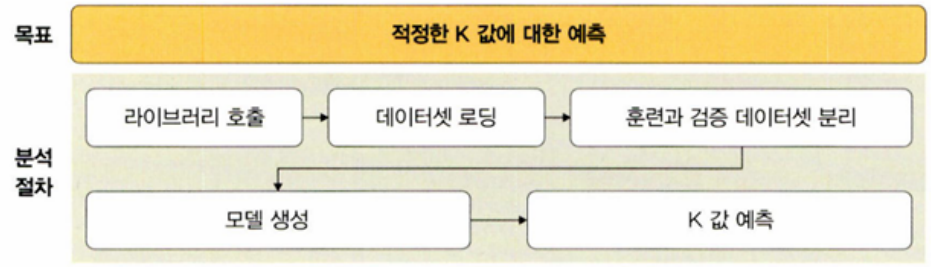
새로운 입력(test data)를 받았을 때 기존 클러스터에서 모든 데이터와 instance 기반 거리를 측정한 후 가장 많은 속성을 가진 클러스터에 할당하는 분류 알고리즘

❯ 그림 3-2 K-최근접 이웃



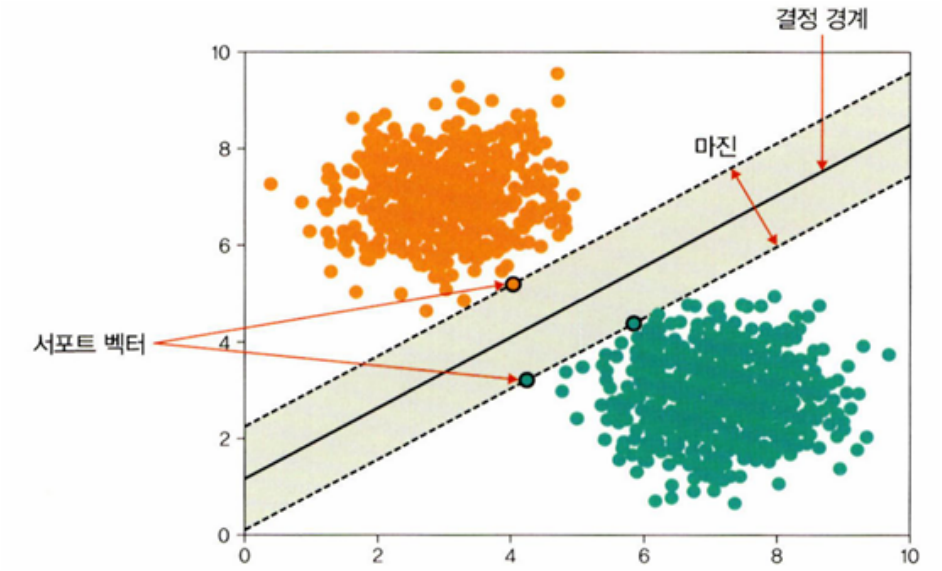
k=5로 지정했을 때 거리상 가장 가까운 데이터 5개를 선택하여 해당 클러스터에 할당

❯ 그림 3-4 K-최근접 이웃 예제



3.1.2 서포트 벡터 머신 Supporter Vector Machine

- 분류를 위한 기준선을 정의하는 모델로 분류되지 않은 새로운 데이터가 나타나면 결정 경계를 기준으로 경계의 어느쪽에 속하는지 분류하는 모델
- 이때 결정경계는 마진(결정경계와 서포트 벡터(결정경계와 가까이 있는 데이터)사이의 거리)을 최대로 해야 함



- 이때 이상치를 허용하지 않는 것이 hard margin, 어느 정도 이상치들이 마진 안에 포함되는 것을 허용하는게 soft margin이라고 함
- 이때 SVM은 선형 분류와 비선형 분류를 지원함
- 비선형에 대한 커널은 선형으로 분류될 수 없는 문제 때문에 발생함
- 비선형문제를 해결하는 가장 기본적인 방법은 저차원 데이터를 고차원으로 보내는 것인데 이때 많은 수학적 계산이 필요하기 때문에 성능 문제를 줄 수 있음
- 이때 도입된게 kernel trick으로 선형을 위한 커널 → linear, 비선형을 위한 커널 → 가우시안 RBF, polynomial
    - 다항식 커널) 다항식 특성을 추가해서 엄청난 수의 특정 조합이 생기는 효과로 고차원 데이터로 매핑이 가능해짐

$$K(a,b) = (\gamma a^T \cdot b)^d$$

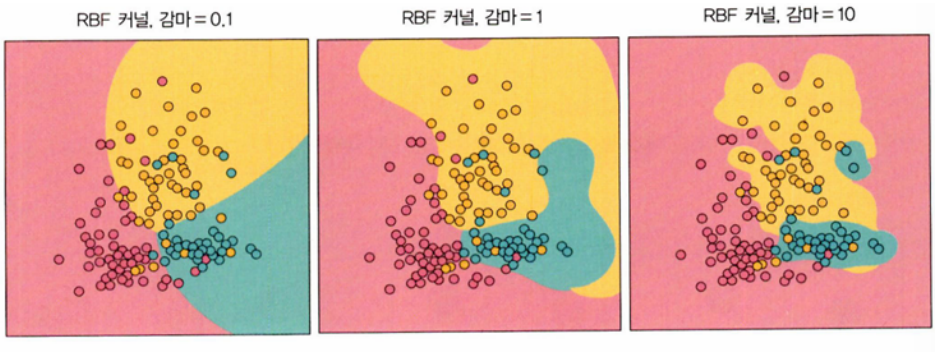
(a, b: 입력 벡터, γ: 감마, d: 차원, 이때 γ, d는 하이퍼파라미터)

    - 가우시안 RBF 커널) 다항식 커널의 확장으로 입력 벡터를 차원이 무한한 고차원으로 매핑하는 것으로 모든 차수의 다항식을 고려함

$$K(a,b) = \exp(-\gamma \| a-b \|^2)$$

(이때 γ는 하이퍼파라미터)

- 이때 C값은 오류를 어느정도 허용할지 지정하는 파라미터로 C값이 클수록 하드마진, 작을수록 소프트 마진임
- 이때 gamma는 결정경계를 얼마나 유연할지 지정하는 파라미터로 감마값이 높으면 과적합이 발생할 수 있음



감마값이 높을 수록 결정경계과 굉장히 유연해짐

3.1.3 결정트리

- 트리구조를 통해 데이터를 분류하거나 결과를 예측하는 분석기법
- 데이터를 1차로 분류한 후 각 영역의 순도 homogeneity는 증가하고 불순도 impurity, 불확실성 uncertainty는 감소하는 방향으로 학습을 진행함 → information gain
- Entropy: 확률변수의 불확실성을 수치화한 것으로 엔트로피가 높을 수록 불확실성이 높다는 의미, 즉 순도가 높다는 것을 의미함

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2(p_k)$$

(P_k=A 영역에 속하는 데이터 가운데 k 범주에 속하는 데이터 비율)

레코드 m개가 A 영역에 포함되어 있을 때

- Gini index: 불순도를 측정하는 지표로 통계적 분산 정도를 정량화 한 것. 원소 n개 중에 임의로 2개를 추출했을 때 이 두개가 서로 다른그룹에 속해 있을 확률을 의미함

$$G(S) = 1 - \sum_{i=1}^{c} p_i^2$$

(S: 이미 발생한 사건의 모음, c: 사건 개수)

3.1.4 로지스틱 회귀와 선형 회귀

- 로지스틱 회귀는 분석하고자 하는 대상들이 두 집단 혹은 그 이상의 집단으로 나누어진 경우, 개별 관측치들이 어느 집단으로 분류될 수 있는지 분석하고 이를 예측하는 모형을

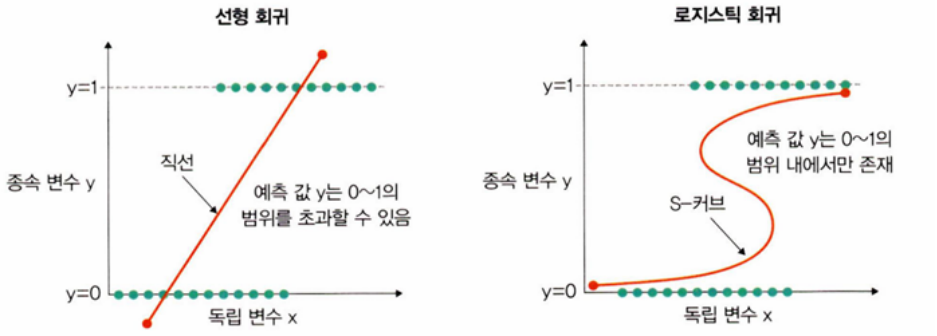
개발

표 3-7 일반 회귀 분석과 로지스틱 회귀 분석 차이

구분	일반적인 회귀 분석	로지스틱 회귀 분석
종속 변수	연속형 변수	이산형 변수
모형 탐색 방법	최소제곱법	최대우도법
모형 검정	F-테스트, t-테스트	X^2 테스트

- 로지스틱 회귀의 분석 절차 → 1) 각 집단에 속하는 확률의 추정치를 예측, 2) 분류값 기준을 설정한 후 특정 범주로 분류
- 선형 회귀는 종속변수와 독립변수 사이의 관계를 설정하는데 사용됨 (로지스틱 회귀는 사건의 확률을 확인하는데 사용)

그림 3-22 선형 회귀와 로지스틱 회귀



3.2 비지도학습

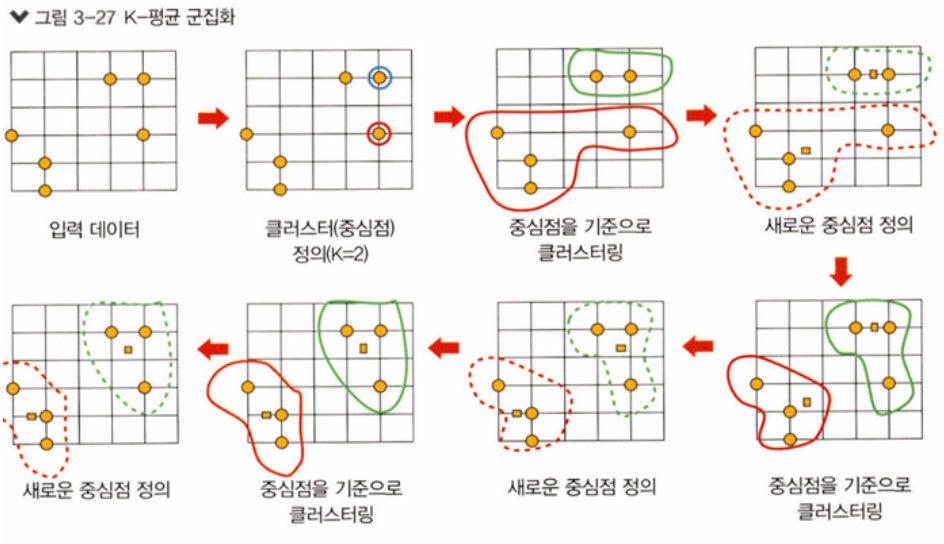
- 레이블이 필요하지 않고 정답이 없는 상태에서 훈련시키는 방식으로 군집화와 차원축소가 있음

표 3-9 비지도 학습 군집과 차원 축소 비교

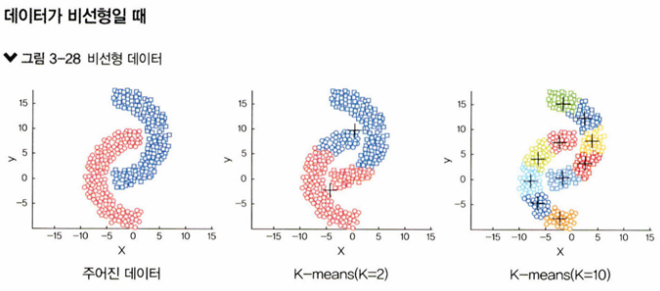
구분	군집	차원 축소
목표	데이터 그룹화	데이터 간소화
주요 알고리즘	K-평균 군집화(K-Means)	주성분 분석(PCA)
예시	사용자의 관심사에 따라 그룹화하여 마케팅에 활용	• 데이터 압축 • 중요한 속성 도출

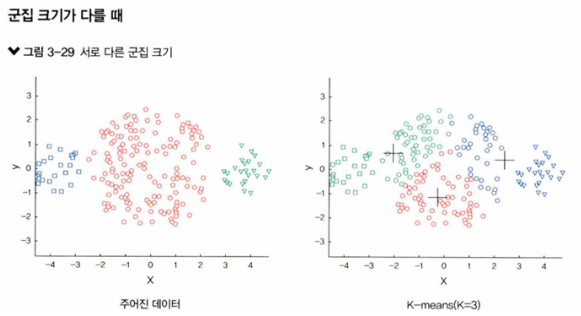
3.2.1 K-Means Clustering

- 데이터를 입력받아 소수의 그룹으로 묶는 알고리즘
- 랜덤하게 초기 중심점 (디폴트는 k=2) 선택→ k개의 중심점과 각 개별 데이터간의 거리 측정 후 가장 가까운 중심점을 기준으로 데이터를 할당 → 클러스터 형성 → 새로운 중심점 계산 x 반복



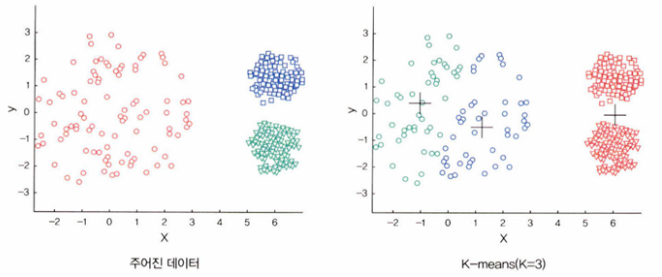
이때 k-means clustering은 데이터가 비선형일 때, 군집 크기가 다를 때, 군집마다 밀집도가 다를 때 성능이 좋지 않을 수 있음





군집마다 밀집도(density)와 거리가 다를 때

❯ 그림 3-30 밀집도와 거리가 다른 군집

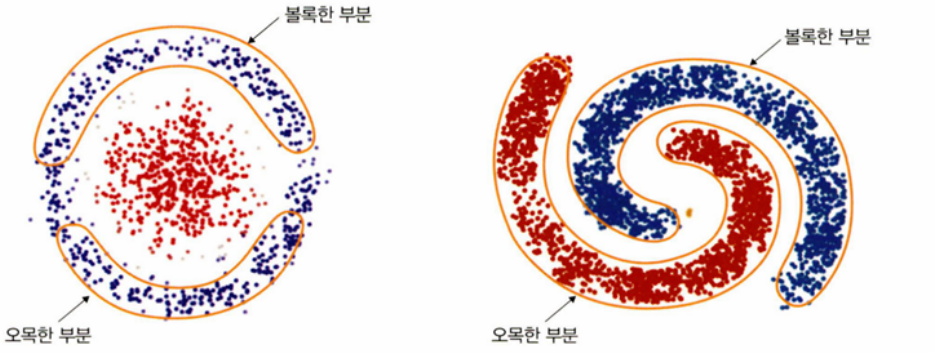


- 거리제곱합 SSD의 경우 x,y 데이터의 차이를 구해서 제곱한 값을 모두 더한 것으로 k가 증가할 수록 감도하는 경향이 있음

3.3.2 밀도 기반 군집분석

- Density-Based Spatial Clustering of Application with Noise → DBSCAN이란 일정 밀도 이상을 가진 데이터를 기반으로 군집을 형성하는 것
- 노이즈에 영향을 받지 않고 k-means clustering에 비해 연산량이 많지만 비선형 데이터 등을 처리하는데 유용함

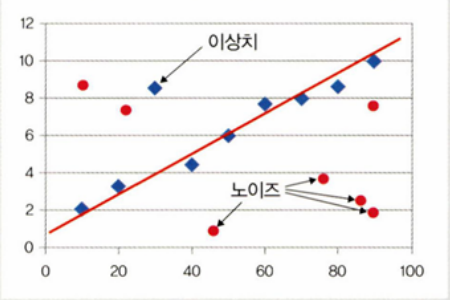
❯ 그림 3-36 밀도 기반 군집 분석의 데이터 표현



Note ≡ **노이즈와 이상치 차이**

노이즈는 주어진 데이터셋과 무관하거나 무작위성 데이터로 전처리 과정에서 제거해야 할 부분입니다. 이상치는 관측된 데이터 범위에서 많이 벗어난 아주 작은 값이나 아주 큰 값을 의미합니다.

❯ 그림 3-37 노이즈와 이상치



1. 어떤 점에서 거리 epsilon 내에 점이 m개 이상 있으면 하나의 군집으로 인식한다고 할 때, 이 점을 core point라고 함
2. 새로운 군집 생성 후 주위의 값들로 다시 core point를 만들어 군집확장

→ 이를 반복하여 군집 정의, 노이즈 정의함 (어떤 군집에서 포함되지 않은 데이터 의미)

3.2.3 주성분 분석 (PCA) - Principal Component Analysis

- 변수가 많은 고차원 데이터를 저차원으로 축소시켜 데이터가 가진 대표 특성만 추출하는 기법
- 데이터의 분포 특성을 잘 설명하는 벡터 2개를 선택 후 이를 위한 적정한 가중치를 찾을 때까지 학습을 진행함
- 데이터 하나하나 성분을 분석하는게 아니라 여러 데이터가 모여 하나의 분포를 이룰 때 이 분포의 주성분을 분석하는 방법