강원지역혁신플랫폼

기계학습

Machine Learning

K-최근접 이웃 알고리즘

학습목표

- K-최근접 이웃 알고리즘을 이해하고 구현할 수 있습니다.

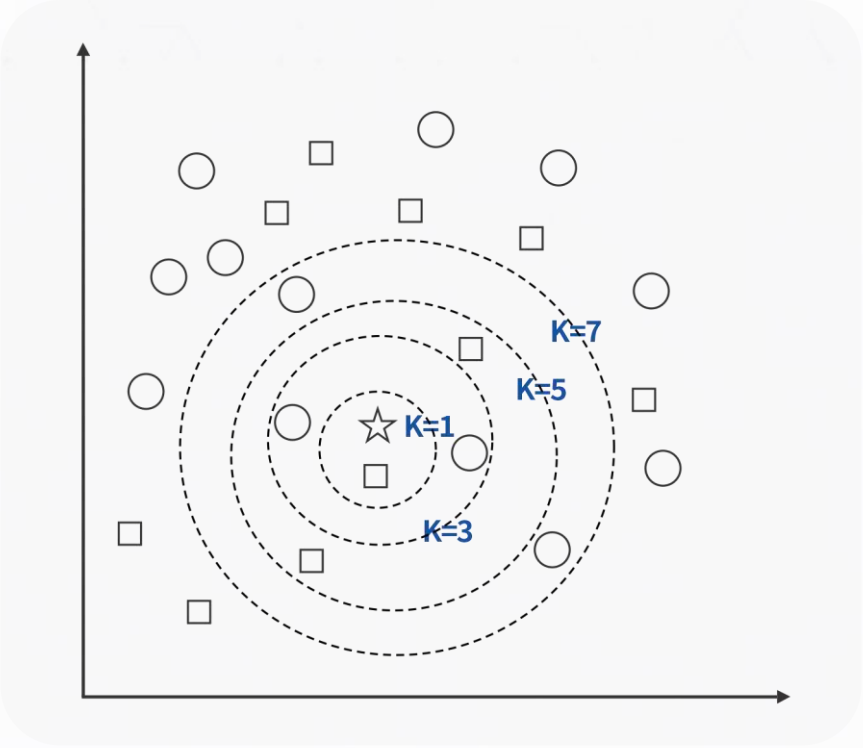
01 | K-최근접 이웃(K-NN) 알고리즘

K-최근접 이웃(K-Nearest Neighbor, K-NN)

- K-최근접 이웃 알고리즘은 비슷한 특성을 가진 데이터는 비슷한 범주에 속하는 경향이 있다는 가정하에 사용함
  - 이 알고리즘은 가장 간단한 기계학습 알고리즘임
    - 그리고, 분류(classification) 알고리즘임

01 | K-최근접 이웃(K-NN) 알고리즘

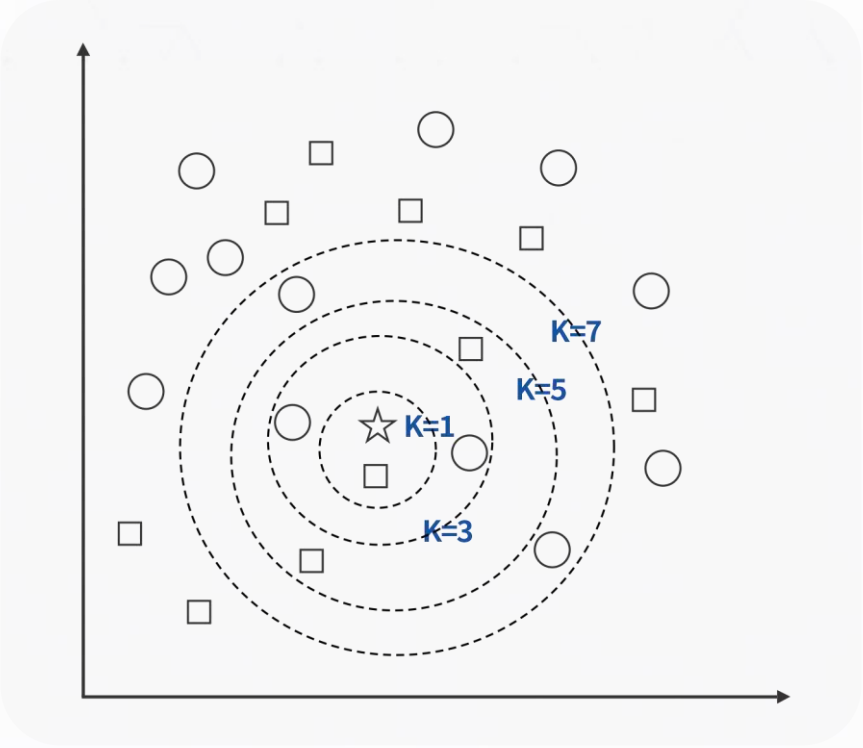
아래 그림과 같이 주변의 가장 가까운 K개의 데이터를 보고
데이터가 속할 그룹을 판단하는 알고리즘이 K-NN 알고리즘임



01 | K-최근접 이웃(K-NN) 알고리즘

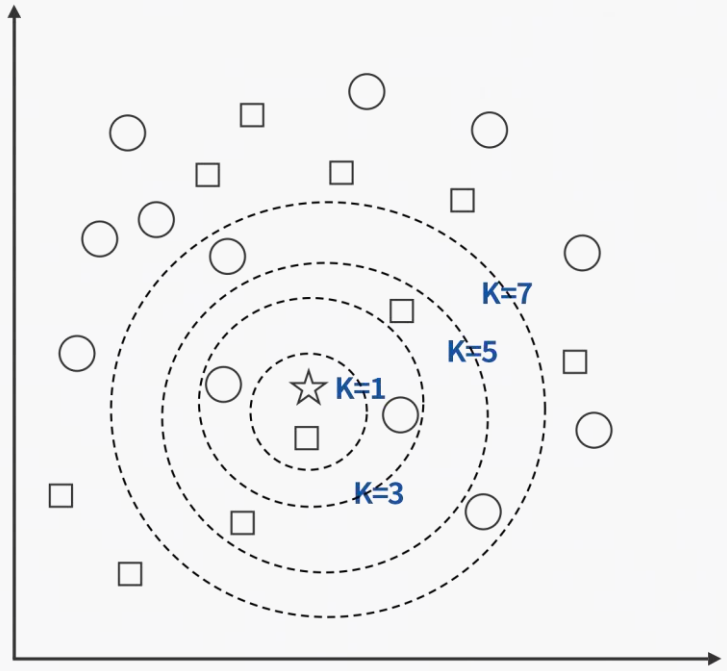
아래의 그림처럼 새로운 데이터가 들어왔을 때를 가정해보자.

- 기존 데이터와 비교하여 가장 가까운 K개의 이웃의 정보를 기반으로 새로운 데이터를 예측함
  - 아래 그림에서 새로운 데이터는 ☆ 임



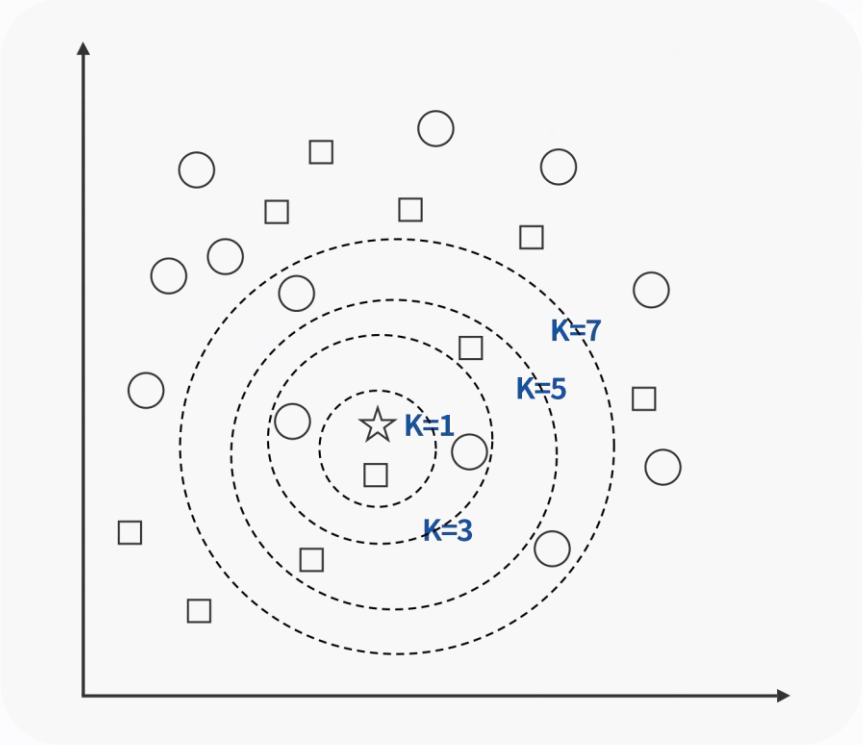
01 | K-최근접 이웃(K-NN) 알고리즘

- K-최근접 이웃 알고리즘은 새로운 데이터 점(아래 그림에서 ☆)과 주변 데이터 세트 간의 유사성(similarity)을 측정하여 최종적으로 목표변수의 범주를 분류할 때, 몇 개를 기준(K=?)으로 주변 데이터 세트를 판단할 것인가에 대한 기준이 필요함
  - 아래 그림은 K=1, K=3, K=5, K=7인 경우의 K-NN 기법의 예시임



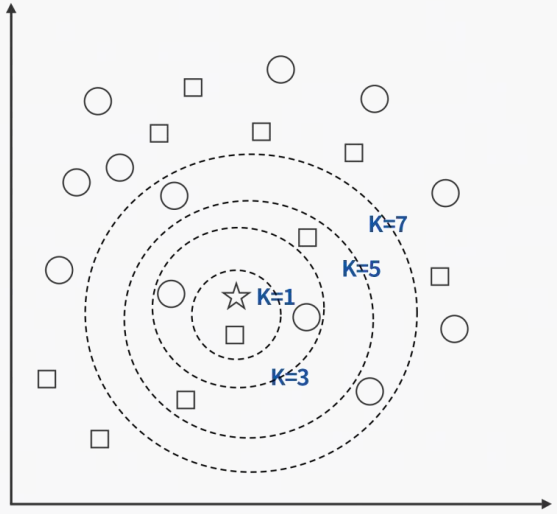
01 | K-최근접 이웃(K-NN) 알고리즘

- K-최근접 이웃 알고리즘은 새로운 데이터 점(아래 그림에서 ☆)과 유사한 K개의 주변 데이터 점에서 다수결의 원칙에 따라 새로운 범주를 결정하는 방식임
  - 아래 그림에서 새로운 데이터 '☆'이 다수결의 원칙에 따라 '□' 또는 '○'로 분류됨



01 | K-최근접 이웃(K-NN) 알고리즘

- 아래 그림은 K-NN에서 K(=1,3,5,7)값 설정에 따른 목표변수 값 변화를 개념적으로 표현하였음
  - ‘☆’은 새롭게 분류해야 할 데이터 값이고, ‘□’와 ‘○’는 주변에 존재하는 데이터 값임
    - 여기서 K=1로 설정하면, ‘☆’과 가장 가까운 데이터 값은 ‘□’이므로 ‘☆’의 목표변수는 ‘□’로 분류됨
    - 이것을 반복하여 K=3, K=5, K=7로 설정하면, 다수결의 원칙에 의하여 ‘☆’의 목표변수를 예측 분류할 수 있음
    - 여기서 주목할 점은 K값을 어떻게 정하느냐에 따라 목표변수의 범주 예측 결과가 크게 달라질 수 있다는 것임



01 | K-최근접 이웃(K-NN) 알고리즘

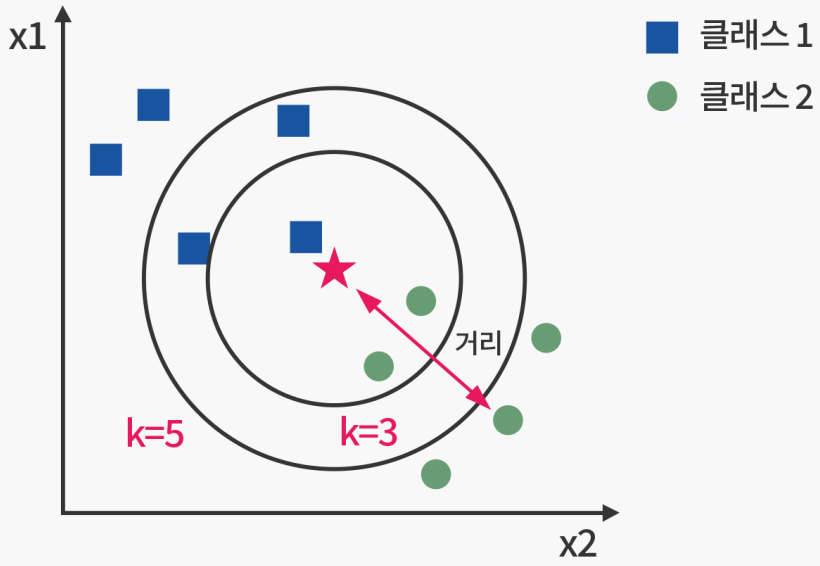
- K-NN 기법에서 적절한 K값을 정하는 것이 매우 중요함
  - K가 작을 경우 데이터의 지역적 특성을 지나치게 반영하여 과적합이 발생함
    - 반대로, K가 너무 클 경우 모델이 지나치게 정규화되어 과소적합이 발생할 수 있음
      - 여기서 적절한 K를 선택하기 위한 방법을 생각해 보자.

01 | K-최근접 이웃(K-NN) 알고리즘

- 적절한 K를 선택하기 위해서는 훈련 데이터 셋과 테스트 데이터 셋을 나눔
  - K를 바꿔가면서 실험을 거쳐 최적의 K를 찾아야 함
    - 다만 K값은 $\sqrt{\text{관측치의 개수}}$ 보다는 작은 것이 좋다고 알려져 있음
      - 일반적으로 K=3에서 K=9사이의 범위 내에서 분류 성능을 테스트해보면서 최적의 K값을 정하는 방법을 사용하기도 함

01 | K-최근접 이웃(K-NN) 알고리즘

- K-NN에서 **거리 측정 방법**은 또 하나의 하이퍼파라미터(Hyper-parameter)이며 **거리**를 **측정하는 방법**에 따라 **결과**가 **크게 달라짐**
  - 거리 측정 방법에는 유클리디안(Euclidean), 맨해튼(Manhatan), 마할라노비스(Mahalanobis)거리 등이 있음
    - **일반적**으로 **거리**를 **측정**할 땐 **유클리드 거리**(Euclidean distance)를 **주로 사용**함

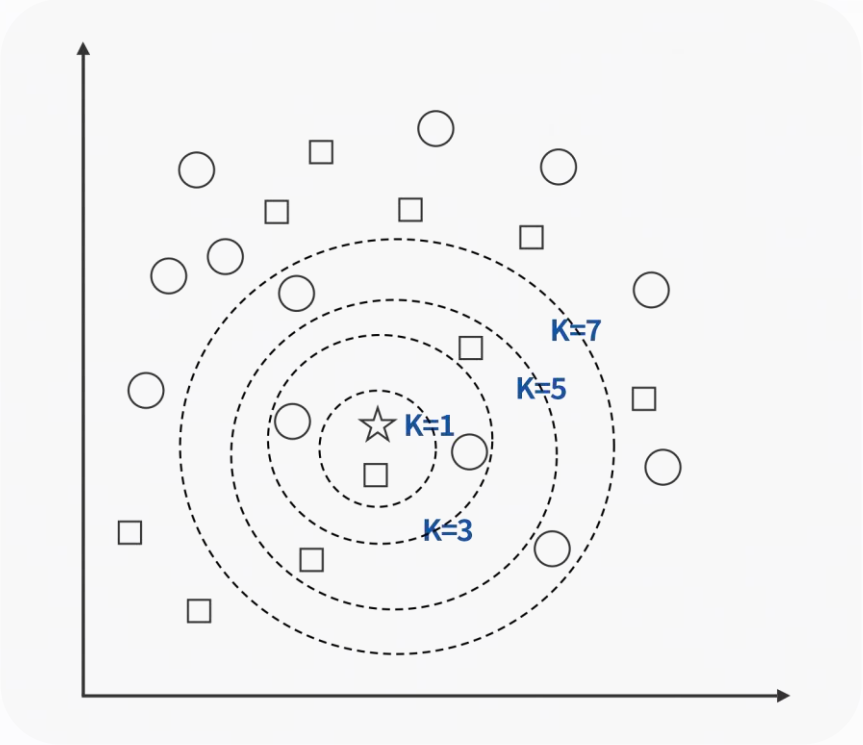


출처: https://june7song.blogspot.com/2020/11/knnk-meanssklearn-knn-k-means.html

01 | K-최근접 이웃(K-NN) 알고리즘

K-최근접 이웃(K-NN) 모델의 특징

- K-NN의 특징은 **학습**이 따로 **필요 없다는 것**임
  - **새로운 데이터**가 주어지면 그때 **기존 데이터**에서 **가까운 이웃**을 뽑고 **예측**할 뿐임



01 | K-최근접 이웃(K-NN) 알고리즘

K-NN 알고리즘의 장 · 단점

구 분	설 명
장 점	❖ 알고리즘이 **이해하기 쉽고 직관적**이다. ❖ 데이터 세트의 **확률분포** 등에 대한 **가정이 필요하지 않다.** ❖ 사전 모형 설정 및 **모수 추정이 필요 없다**(비모수 방식). ❖ **훈련**(학습) **시간**이 **빠르다.** ❖ **노이즈 데이터**의 **영향**을 **크게 받지 않는다**(Robust하다). ❖ 분류 뿐만 아니라 **회귀 문제**에도 **K-NN 방법론**을 **적용할 수 있다.**
단 점	❖ K값에 대한 **명확한 기준이 없어 시행착오적 접근**이 **필요하다.** ❖ 특정한 가설이나 모형 없이 **주어진 데이터**를 통해 **범주의 분류결과만 판단함**으로 분석을 통한 통찰력을 얻기 어렵다. ❖ **새로운 데이터**가 주어질 때마다 **모든 데이터**와의 **유사도**를 **계산해야 함**으로 그만큼 **시간 소요가 많다** (이런 특성 때문에 **게으른 학습**(Lazy learning)으로 불림). ❖ 데이터 세트의 모든 데이터들과 **거리 계산**을 위해 **메인 메모리**에 **가져와야 함**으로 **많은 메모리**가 **필요**하다.

01 | K-최근접 이웃(K-NN) 알고리즘

- K-NN 알고리즘으로 학습하기 전에는 변수의 범위(scale)를 축소해야 함
  - 변수마다 측정 범위가 다를 경우에 범위가 큰 변수가 모델에 과도하게 큰 영향을 미침
    - 따라서, 범위가 작은 변수는 무시될 수 있기 때문임

01 | K-최근접 이웃(K-NN) 알고리즘

예를 들어 다음과 같은 가상 데이터가 있다고 가정하자.

- 아래 표를 기준으로는 거리를 측정할 때 직원 수 정보는 전혀 반영되지 않을 것임
  - 직원 수 단위가 매출의 단위보다 훨씬 작기 때문임

매출 (원)	직원 수 (명)	분류
50,000,000	20	A
55,000,000	40	A
60,000,000	50	A
70,000,000	30	B
30,000,000	60	B
75,000,000	70	B

01 | K-최근접 이웃(K-NN) 알고리즘

- 아래의 데이터로 K-NN모델을 만들면 매출에 강한 영향을 받게 됨
  - 따라서, 각 변수의 값의 범위를 동등하게 조정해야 함
    - 변수 별로 평균과 분산을 일치시키는 등의 정규화 작업이 필요함

매출 (원)	직원 수 (명)	분류
50,000,000	20	A
55,000,000	40	A
60,000,000	50	A
70,000,000	30	B
30,000,000	60	B
75,000,000	70	B

01 | K-최근접 이웃(K-NN) 알고리즘

데이터의 값 범위를 변환하는 것을 스케일링(scaling) 또는 정규화라고 하며 자주 사용되는 정규화 방법은 다음과 같음

1 최대최소 스케일링(MinMax Scaling)

› 모든 변수의 값을 0~1 사이에 위치하도록 조정함

- 새로운 데이터 x'는 원래 값 x_i에서 변수 x의 최솟값을 뺀 값을 변수 x의 최대값과 최솟값의 차이로 나눈 값으로 변환한 것임

$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

01 | K-최근접 이웃(K-NN) 알고리즘

2 표준정규분포로 표준화(Standardization)

› 곱셈과 덧셈만으로 변환하는 선형변환을 통해 각 특성의 평균을 0, 분산을 1로 변경하여 모든 특성이 같은 크기를 가지게 함

- 그러나, 이 방법은 특성은 최소값과 최대값 크기를 제한하지는 않음
- 새로운 데이터 x'는 원래 값에서 변수 x의 평균을 뺀 값을 변수 x의 표준편차(STDEV, Standard Deviation)로 나눈 값으로 변환한 것임

→ 이 값을 표준점수 혹은 Z-점수(Z-Score)라고 부름

$$x' = \frac{x_i - mean(x)}{STDEV(x)}$$