week5. 비지도 학습

비지도 학습

: 훈련 데이터에 타깃이 없는 머신러닝의 종류. 타깃이 없기에, 외부의 도움 없이 스스로 학습해야한다.

대표적으로 군집, 차원 축소 등이 있다.

- 픽셀의 평균을 계산하는 방법 : axis=0으로 지정한 뒤, bar() 함수를 이용한다.

히스토그램

: 구간별로 값이 발생한 빈도를 그래프로 표현한 것.

보통 x축이 값의 구간(계급) 이고 y 축은 발생 빈도 (도수) 이다.

군집

: 비슷한 샘플끼리 하나의 그룹으로 모으는 대표적인 비지도 학습.

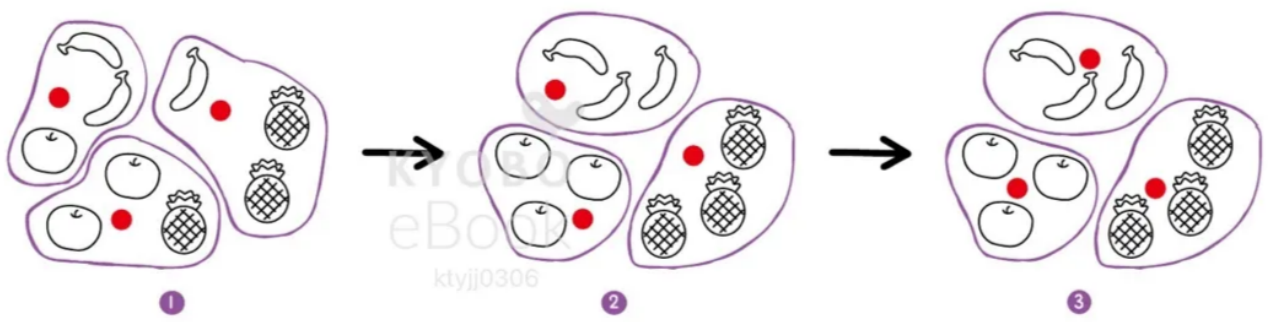
군집 알고리즘으로 모은 샘플 그룹을 클러스터 라고 부름.

k-평균 알고리즘

KMeans : k-평균 알고리즘 클래스

: 처음에 랜덤하게 클러스터 중심을 선택하고 점차 가장 가까운 샘플의 중심으로 이동하는 알고리즘.

1. 무작위로 k개의 클러스터 중심을 정한다.
2. 각 샘플에서 가장 가까운 클러스터 중심을 찾아 해당 클러스터의 샘플로 지정한다.
3. 클러스터에 속한 샘플의 평균값으로 클러스터 중심을 변경한다.
4. 클러스터 중심에 변화가 없을 때까지 2번으로 돌아가 반복한다.



n_cluster : 클러스터 개수 지정, 기본값은 8

처음에 랜덤하게 센트로이드를 초기화하기 때문에 여러 번 반복하여 이너셔를 기준으로 가장 좋은 결과를 선택. n_init는 반복 횟수를 지정.

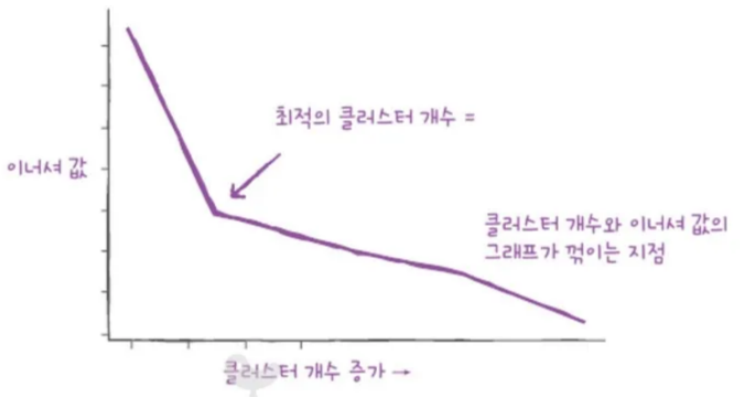
max_iter는 k-평균 알고리즘의 한 번 실행에서 최적의 센트로이드를 찾기 위해 반복할 수 있는 최대 횟수.

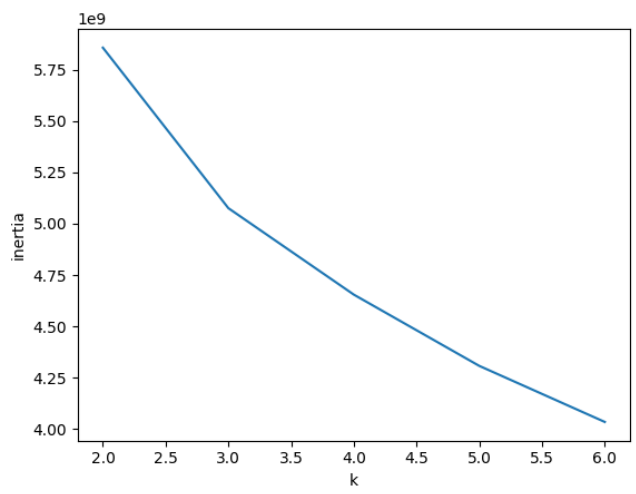
기본 값은 200.

```
import matplotlib.pyplot as plt

def draw_fruits(arr, ratio=1):
```

```
    n = len(arr)    # n은 샘플 개수.
    # 한 줄에 10개씩 이미지 그림.
    # 샘플 개수를 10으로 나누어 전체 행 개수를 계산.
    rows = int(np.ceil(n/10))
    # 행이 1개 이면 열 개수는 샘플 개수. 그렇지 않으면 10개.
    cols = n if rows < 2 else 10
    fig, axs = plt.subplots(rows, cols,
                            figsize=(cols*ratio, rows*ratio), squeeze=False)
    for i in range(rows):
        for j in range(cols):
            if i*10 + j < n:
                axs[i, j].imshow(arr[i*10 + j], cmap='gray_r')
            axs[i, j].axis('off')
    plt.show()
```





주성분 분석 (PCA) : 데이터에 있는 분산이 큰 방향을 찾는 것.

차원 축소 알고리즘의 한 종류.

데이터에서 가장 분산이 큰 방향을 찾는 방법이다. 이런 방향을 '주성분' 이라고 부르며,

원본 데이터를 주성분에 투영하여 새로운 특성을 만들 수 있다.

일반적으로 주성분은 원본 데이터에 있는 특성 개수보다 적다.

비지도 학습이기에, fit() 메서드에 타깃값을 제공하지 않음.

n_components 는 주성분의 개수를 지정한다.

기본값은 None으로, 샘플 개수와 특성 개수 중 작은 것의 값을 사용한다.

random_state : 넘파이 난수 시드 값을 지정

components_ 속성 : 훈련 세트에서 찾은 주성분 저장

explained_variace_ 속성 : 설명된 분산 저장, explained_variance_ratio_ 에는 설명된 분산의 비율 저장.

inverse_transform() 메서드 : transform() 메서드로 차원을 축소시킨 데이터를 다시 원본으로 복원.

```
import matplotlib.pyplot as plt

def draw_fruits(arr, ratio=1):
    n = len(arr)
    rows = int(np.ceil(n/10))
    cols = n if rows < 2 else 10
    fig, axs = plt.subplots(rows, cols,
                            figsize=(cols*ratio, rows*ratio), squeeze=False)
    for i in range(rows):
        for j in range(cols):
            if i*10 + j < n:
```

```
                axs[i, j].imshow(arr[i*10 + j], cmap='gray_r')
            axs[i, j].axis('off')
    plt.show()
```

차원 축소 :

원본 데이터의 특성을 적은 수의 새로운 특성으로 변환하는 비지도 학습의 한 종류.

차원 축소는 저장 공간을 줄이고 시각화를 쉽게 돕는다.

또한, 다른 알고리즘의 성능을 높인다.

설명된 분산 :

주성분 분석에서 주성분이 얼마나 원본 데이터의 분산을 잘 나타내는지 기록한 것.

사이킷런의 PCA 클래스는 주성분 개수나 설명된 분산의 비율을 지정하여 주성분 분석을 수행한다.