BDA

Machine Learning Study

Assignment #4

7팀

박지수, 김인서, 박지영, 박혜원, 이주미

1. K-평균 군집화 모델의 평가지표인 Rand 지수와 실루엣 계수에 대해 설명해보시오.

두 계수 모두 군집화 기법의 성능을 평가할 때 사용된다는 공통점을 가지고 있다. Rand 지수는 데이터 분할에 대한 실제 레이블이 있는 경우에 사용되며, 실제 레이블 번호가 있다고 해서 군집 결과의 번호 자체를 비교하는 것이 아니라, 전체 데이터가 분할된 분포 또는 패턴을 분석한다는 특징을 가지고 있다. 수식 결과 값은 0부터 1 사이의 값을 가지며, 값이 클수록 유사도가 높다. 반면 실루엣 계수는, 데이터 분할에 대한 실제 레이블이 없을 경우에 사용된다. 같은 군집 내의 원소끼리는 가깝되 다른 군집의 원소와는 먼 것이 좋은 군집화라는 특징이 적용된다. 실루엣 계수는 -1에서 1사이의 값을 가지며, 좋은 군집화 결과일수록 높은 값을 가진다.

2. K-최근접 이웃 모델과 K-평균 군집화 모델의 차이를 서술하고 각 모델이 활용될 수 있는 사례를 적어보시오.

1) K-최근접 이웃 모델(K-NN) : 분류 및 회귀 문제에 사용되며, 주로 지도학습에 속한다. 특정 데이터 포인트의 분류를 결정할 때, 가장 가까운 'k'개의 이웃 데이터 포인트를 찾아 그들의 레이블을 바탕으로 예측을 수행한다. 예를 들어, 스팸 메일 필터링에서 K-NN 알고리즘이 사용될 수 있다. 스팸 메일과 비슷한 특성을 가진 이메일이 스팸인지 아닌지를 판단하는데 사용한다.

2) K-평균 군집화(K-Means Clusterin) : 비지도 학습에 속하며, 데이터를 k개의 서로 다른 군 집으로 분류하는 데 사용한다. 데이터의 특성만을 고려하여 군집을 형성하고 레이블을 필요로 하지 않는다. 소비자 행동 데이터를 분석하여 비슷한 구매 패턴을 가진 고객 그룹을 생성하는 고객 세분화 분야에 활용된다.

3. 계층적 군집화 모델에서 군집 사이의 거리를 정의하는 연결법들을 나열해보고 각 연결법들의 특징에 대해 서술해보시오.

1) Single linkage(단일 연결법)

: 두 군집 간 원소끼리의 거리를 모두 비교한 후, 최소거리를 군집 간 거리로 정의

- 큰 데이터셋에서 계층적 군집화를 수행할 때 적합
- 고립된 군집을 찾는 데에 중점

2) Complete linkage(최장 연결법)

: 두 군집 간 원소끼리의 거리를 모두 비교한 후, 최대거리를 군집 간 거리로 정의

- 군집들의 내부 응집성에 중점

3) Average linkage(평균 연결법)

: 두 군집 간 원소끼리의 거리를 모두 비교한 후, 그 평균 거리를 군집 간 거리로 정의

- 계산량이 불필요하게 많음

4) Centroid linkage(중심 연결법)

: 각 군집의 중심을 구한 후 중심 사이의 거리를 군집의 거리로 정의

- 군집 내 편차제곱합을 고려하므로 군집 간 정보 손실이 최소화됨

5) Ward linkage(와드 연결법)

: 두 군집을 병합했을 때 군집 내 분산의 증가분을 두 군집 사이의 거리로 정의

- 다른 연결법들과 달리 군집 간 거리가 아닌 군집 내 거리를 기반으로 한다는 특징을 지님
- 비슷한 크기의 여러 군집을 형성하고 싶을 때 적합한 연결법

4. 교재 346 페이지에 덴드로그램 시각화 예제를 실습해보면서 각 과정에 대해 설명해보시오.

```python
# 필요한 패키지를 불러오고 사용할 붓꽃 데이터셋 정의
import numpy as np
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram
from sklearn.datasets import load_iris
from sklearn.cluster import AgglomerativeClustering

X = load_iris().data # 피처의 군집화가 목표이므로 y는 정의하지 않음
```

✓ 7.2s

```python
# 병합적 군집화 모델을 생성하고 모델 학습
# distance_threshold=0 으로 설정하여 전체 계층 구조를 계산해야 함
model = AgglomerativeClustering(distance_threshold=0, n_clusters=None)
model = model.fit(X)
```

✓ 0.0s

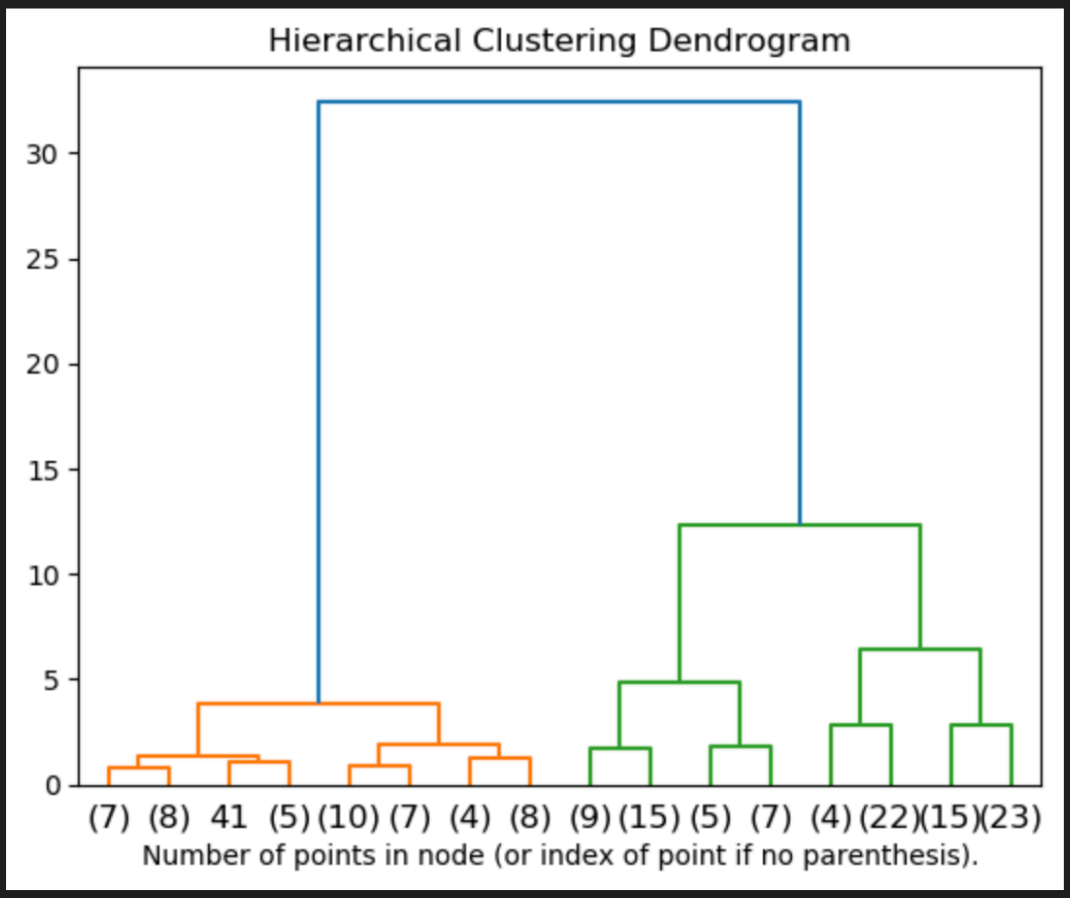
```python
# 학습한 모델로 linkage_matrix 계산
counts = np.zeros(model.children_.shape[0])
n_samples = len(model.labels_)
for i, merge in enumerate(model.children_):
    current_count = 0
    for child_idx in merge:
        if child_idx < n_samples:
            current_count += 1  # leaf node
        else:
            current_count += counts[child_idx - n_samples]
    counts[i] = current_count

linkage_matrix = np.column_stack([model.children_, model.distances_,
                                  counts]).astype(float)
```

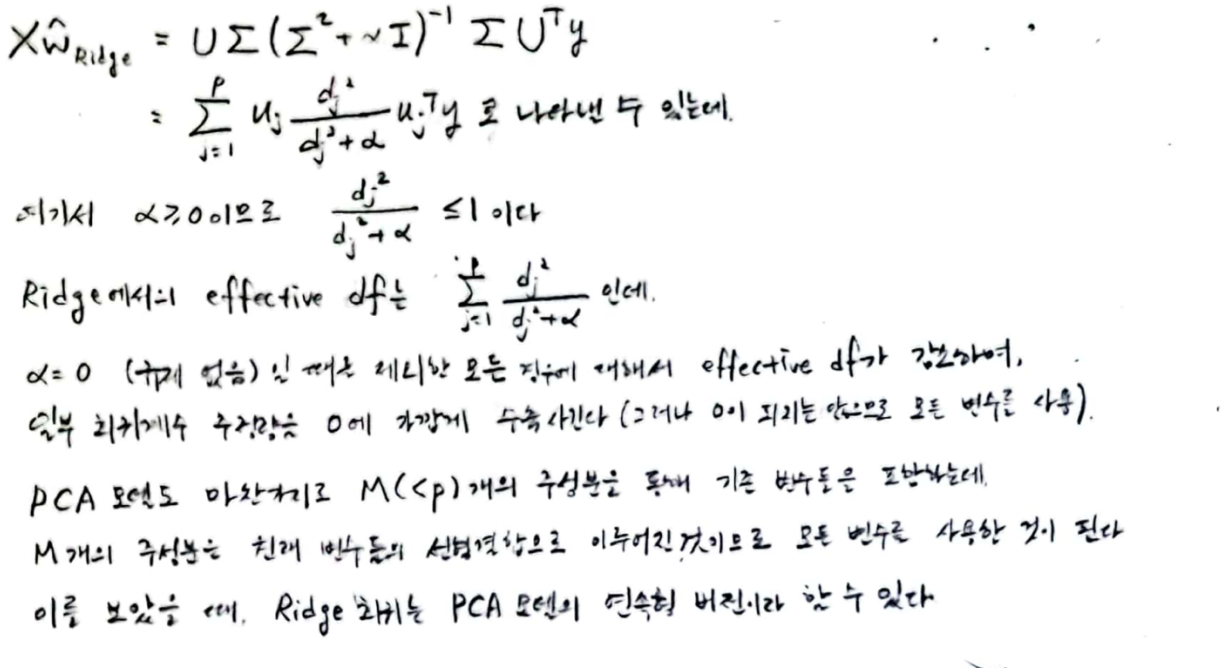
✓ 0.0s

```python
# linkage_matrix를 이용하여 덴드로그램 그리기
# truncate_mode='level', p=3 으로 설정하여 관찰 및 해석이 간단하게
dendrogram(linkage_matrix, truncate_mode="level", p=3)
plt.title("Hierarchical Clustering Dendrogram")
plt.xlabel("Number of points in node (or index of point if no parenthesis).")
plt.show()
```

✓ 0.1s



5. 이전 5장에서 릿지 회귀의 유일해는 $\hat{w}_{\text{Ridge}} = (X^T X + \alpha I)^{-1} X^T y$임을 알 수 있었다.이를 X의 특잇값 분해를 이용해 다시 나타내면, $X\hat{w}_{\text{Ridge}} = U\Sigma(\Sigma^2 + \alpha I)^{-1}\Sigma U^T y$이고, 규제가 없는 최소제곱법 모델의 해는 $X\hat{w} = UU^T y$이다. 이를 통해 PCA모델과 릿지회귀에는 어떠한 관련성이 있는지 설명해보이시오.



6. 차원 축소기법인 MDS, Isomap, LLE, -t-SNE 의 특징에 대해 서술해보시오.

- MDS (Multi-Dimensional Scaling) : MDS는 고차원 데이터의 유사성 또는 거리를 유지하 면서 데이터를 저차원 공간으로 변환하는 기법이다. 이 방법은 주로 시각화, 클러스터링, 또는 고차원 데이터의 패턴을 이해하는 데 사용된다.

- Isomap : Isomap은 주로 비선형 차원 축소에 사용되는 기법이다. 각 데이터 포인트 간의 최단 경로를 보존하려고 한다. 이 방법은 특히 데이터의 내재된 구조가 비선형일 때 유용하 다.

- LLE (Locally Linear Embedding): LLE는 비선형 차원 축소 기법 중 하나로, 각 데이터 포인트를 가장 가까운 이웃에 대한 선형 조합으로 표현한다. 이 방법은 데이터의 국소적인 구 조를 보존하는 데 초점을 맞추며, 데이터가 국소적으로 선형적으로 분포하는 경우에 효과적이 다.

- t-SNE (t-Distributed Stochastic Neighbor Embedding) : t-SNE는 고차원 공간의 데이 터 포인트 간의 확률적인 분포와 저차원 공간의 데이터 포인트 간의 분포를 비슷하게 만들려 는 차원 축소 기법이다. 이 방법은 특히 고차원 데이터의 시각화에 유용하며, 데이터의 구조 와 패턴을 이해하는 데 도움을 준다.