2025.02.08

15기 신입교육세션

5주차 - 비지도학습

B.a.f

CONTENTS

1. 모델링2 과제 피드백

[1] 주요 피드백 정리

B.a.f

주요 피드백 내용

(1)타이타닉 : 분류 문제 / 따릉이 : 회귀 문제
-> 분류 모델과 회귀 모델을 잘 구분해야함 (모델 선정, 평가지표)

분류

최종모델 - test 예측

```
#### 모델 선언
final_model = RandomForestClassifier(random_state=42, max_depth= 9,
                                     min_samples_split= 2,n_estimators= 200)

#### 모델 학습
```

회귀

```
# 학습 진행
model = RandomForestRegressor(random_state=0, n_estimators=250, max_depth = 7, mi

model.fit(x_train, y_train)
print(model.score(x_train, y_train))
print(model.score(x_valid, y_valid))
```

(2) Test 셋에 Encoder를 fit 시키는 것은 X

```
ohe = OneHotEncoder(sparse_output=False)
result_ohe_bike = ohe.fit_transform(bike[['Seasons', 'Holiday', 'Functioning Da
                                         'month', 'day', 'Hour', 'Rainfall_Bin
df_ohe_bike = pd.DataFrame(result_ohe_bike, columns=ohe.get_feature_names_out([

ohe = OneHotEncoder(sparse_output=False)
result_ohe_test = ohe.fit_transform(test[['Seasons', 'Holiday', 'Functioning Da
                                         'month', 'day', 'Hour', 'Rainfall_Bin
df_ohe_test = pd.DataFrame(result_ohe_test, columns=ohe.get_feature_names_out([
```

scaling할 때와 마찬가지로
train - fit_transform()
test - transform() 사용

주요 피드백 내용

(1) GridSearch 진행 시, 시간이 너무 오래 걸려 진행이 되지 않은 경우

!! 튜닝할 하이퍼파라미터의 개수와, 각 파라미터에 대해 여러 값을 넣어 진행할 경우
컴퓨터 자원에 따라 오랜 시간이 걸릴 수 있음.
>> 시간이 오래 걸려 되지 않으면, 튜닝할 하이퍼파라미터의 수를 줄이거나 경우의 수를 줄여볼 것

주요 피드백 내용

(1) 전처리 / EDA 시에 모든 변수 특성을 하나하나 확인하기

(2) 결측치/이상치 처리는 필수.

무작정 제거하기보다는 특성을 확인하면서 채우는 것이 중요

(3) 분류 모델링 – 트리 계열 모델은 스케일링 필요 X

(4) 회귀 모델링 – 이상치에 민감하기 때문에 스케일링 필수

(5) 변수 중요도 확인

(6) 마크다운, 주석 활용

2. 클러스터링

[1] 클러스터링

[2] k-means

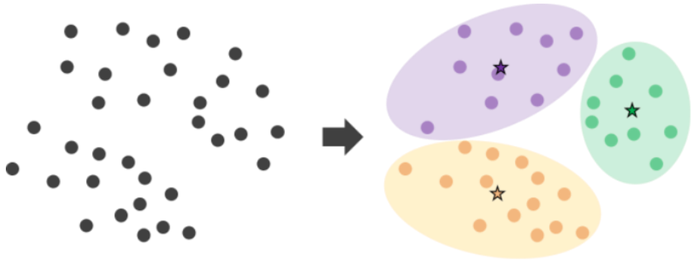
[3] DBSCAN

[4] 클러스터링 활용

B.a.f

클러스터링 - 분석 소개

군집화 (clustering)



- 데이터셋을 여러 군집(cluster)로 나누는 작업
- 각 데이터의 유사성을 측정하여 집단으로 분류하는 비지도 학습

클러스터링 종류

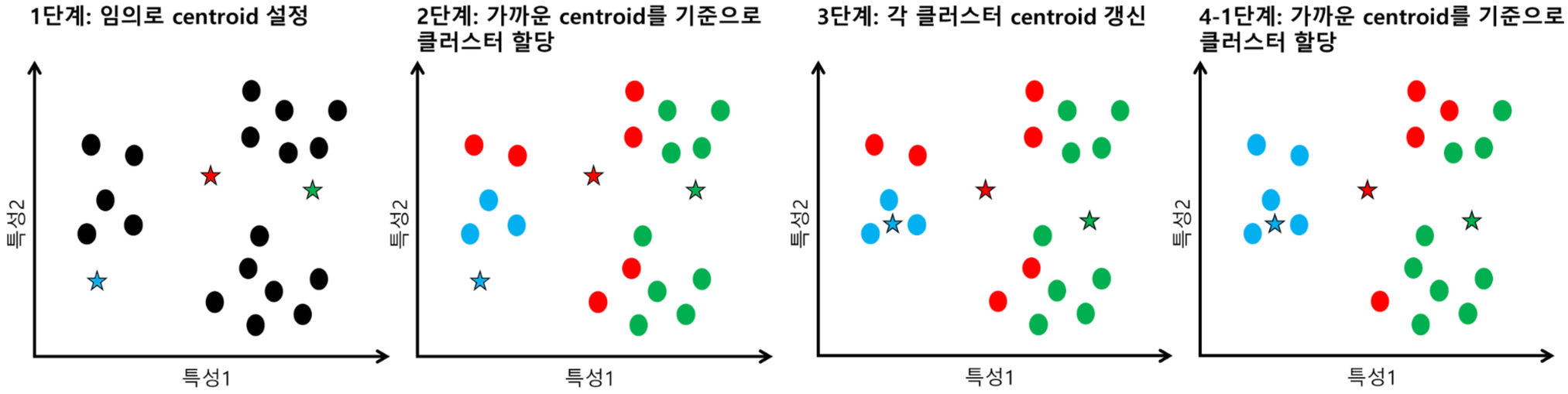
- k-means clustering
- DBSCAN
- GMM

k-means 클러스터링

K개의 중심점을 정하여 각 샘플로부터 그 샘플이 속한 군집(클러스터)의 중심까지의 평균거리를 최소화 시키는 알고리즘

☑ iterative method

1. K개의 중심점 정하기
2. 각 샘플들을 가장 가까운 중심점과 같은 그룹에 할당
3. 그룹의 중심점을 다시 업데이트
4. 업데이트 된 중심점을 기준으로 다시 클러스터 할당



k-means 클러스터링

장점

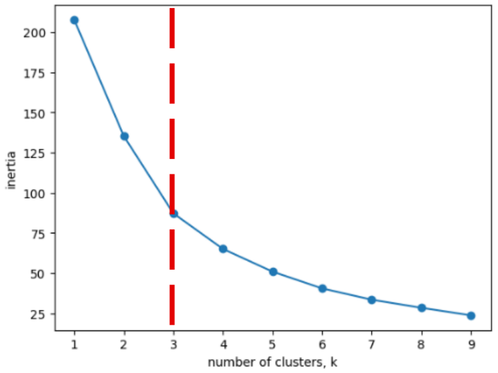
계산이 빠르고 큰 데이터셋에도 사용 가능
시각화에도 적합하고 해석도 쉬움

단점

초기 중심점 선택에 따라 결과에 영향을 미침
이상치에 민감
클러스터 개수인 K를 직접 선택해야함
거리를 기반으로 하기 때문에 변수들의 단위를 통일 시켜야함

군집 수 k 정하기

1. Elbow Method



군집 수에 따라 SSE의 변동을 비교하여 SSE가 급격하게 감소할 때(팔꿈치 모양)의 k를 선정

** SSE : 각 샘플과 해당 클러스터 중심점 사이의 거리를 제곱한 값의 합

2. 실루엣 계수

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

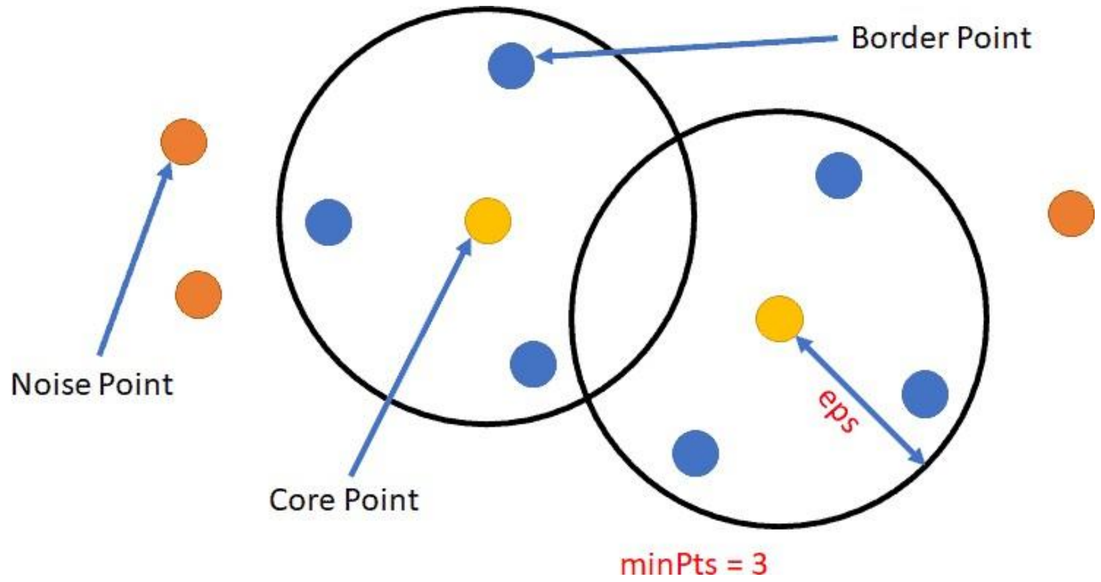
a(i) : 같은 클러스터에 있는 다른 샘플과의 거리
b(i) : 다른 클러스터에 있는 샘플과의 거리

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

실루엣 계수가 1에 가까울수록 군집화가 뚜렷하게 잘 된 것
따라서 실루엣 계수가 높을 때의 k를 선정

DBSCAN

밀도 기반으로 서로 가까운 데이터 포인트를 함께 그룹화하는 알고리즘



Hyperparameter

1. Epsilon : 클러스터를 구성하는 최소의 거리
2. Min points : 클러스터에 필요한 최소 데이터 샘플 수

DBSCAN에서 각 샘플들은 3가지 유형으로 분류

1. 정해진 반지름 Epsilon 내에 이웃하는 점이 Min points개 이상이면, core point
2. core point의 반지름 Epsilon 안에 있으나, 이웃하는 점의 개수가 Min points보다는 작다면 border point
3. core/neighboring point가 아닌 점들은 noise point

DBSCAN

알고리즘

1. Core point 별로 독립된 cluster를 형성, Epsilon 안에 여러개의 core point가 존재하면 이를 연결하여 클러스터 형성
2. 각 Border point를 core point에 맞는 cluster로 할당
3. 1~2를 반복

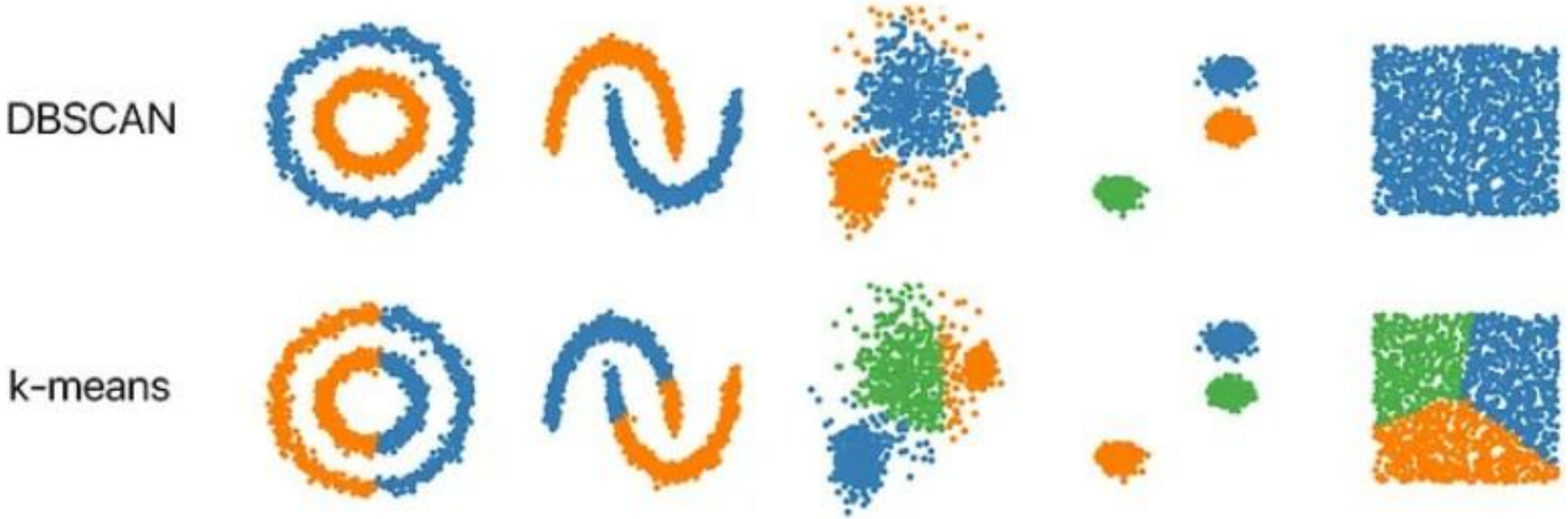
장점

- 군집 수를 미리 설정해주지 않아도 됨
- 이상치를 제외하고 클러스터링을 진행
- 밀도에 따라 클러스터링을 하여 기하학적 모양을 갖는 군집도 잘 찾아낼 수 있음

단점

- 고차원 데이터에서 적절한 Epsilon을 찾기 어려움
- 밀도가 높은 곳에 집중하기 때문에 상대적으로 밀도가 낮은 부분에 대해 모두 noise point라고 판단하는 경우도 있음

k-means VS DBSCAN



클러스터링 활용

- 고객분류 » 하나에 시장에 대해 고객의 특성, 구매패턴에 따른 군집화 가능
- 입지선정 » 본인이 세운 기준에 맞는 지역을 군집화를 통해 뽑아낼 수 있음

  ex) 집 값이 높고 교통수단이 잘 되어있는 자치구를 선별
- 이상치 판별
- 군집화한 특성으로 파생변수 생성

3. 잠재변수모델

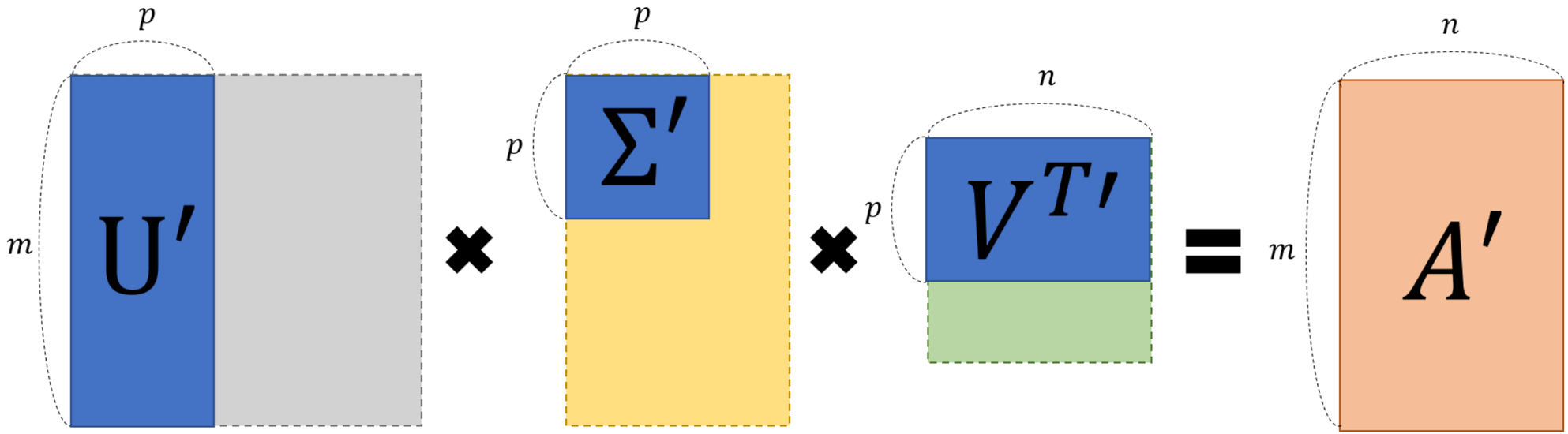
[1] SVD

[2] PCA

[3] LDA

B.a.f

SVD(특이값 분해)

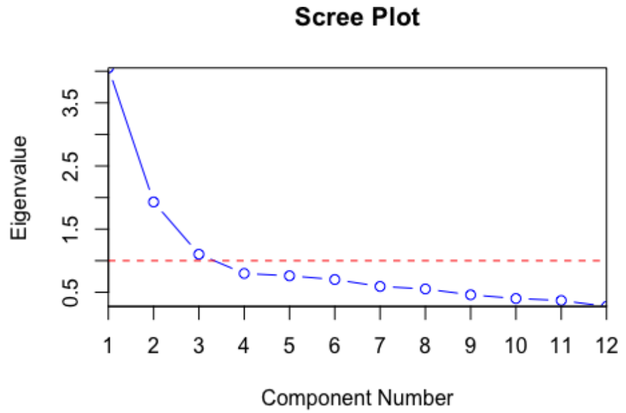


PCA(주성분 분석)

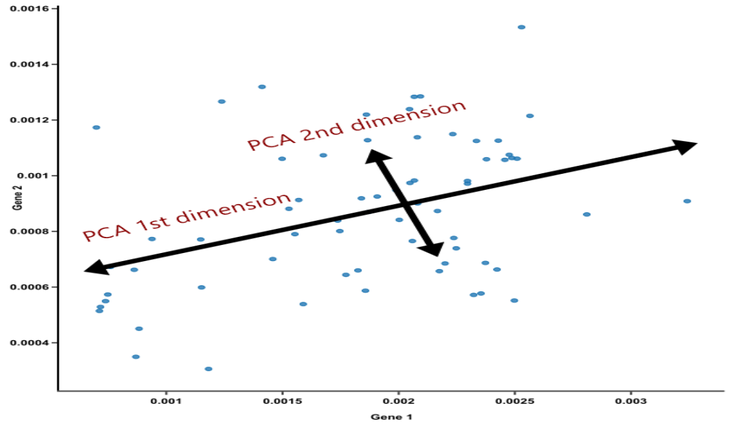
- 차원축소 기법 중 하나
- 다중공선성 문제를 해결할 수 있음

주의할 점

1. 원래 변수들 사이에 있던 상관관계를 없애고 독립으로 만들어줌 : 주성분끼리는 독립
2. 선형결합으로 차원 축소 -> 해석에 용이 **원 변수의 단위가 같아야함 : 표준화 진행
3. 변수를 몇개로 줄일 지는 공분산을 기준으로 정함



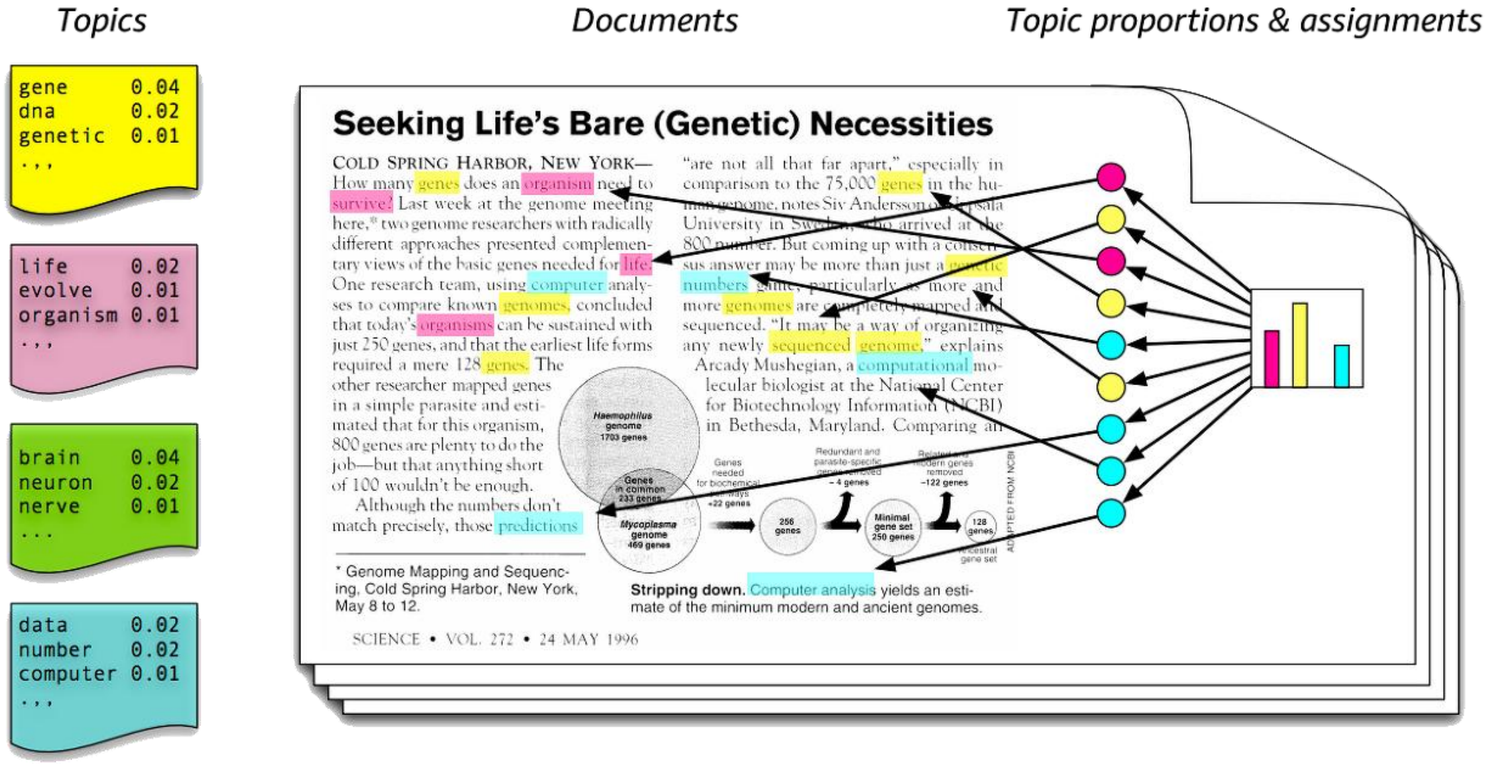
- 주성분의 개수 정하기



- 원 변수의 분산을 최대한 반영

LDA(잠재 디리클레 할당)

- 주제 찾기에 활용
- PCA를 사용한 기법
- 문서 안의 단어에 대한 확률분포를 활용



4. 이상치 탐지

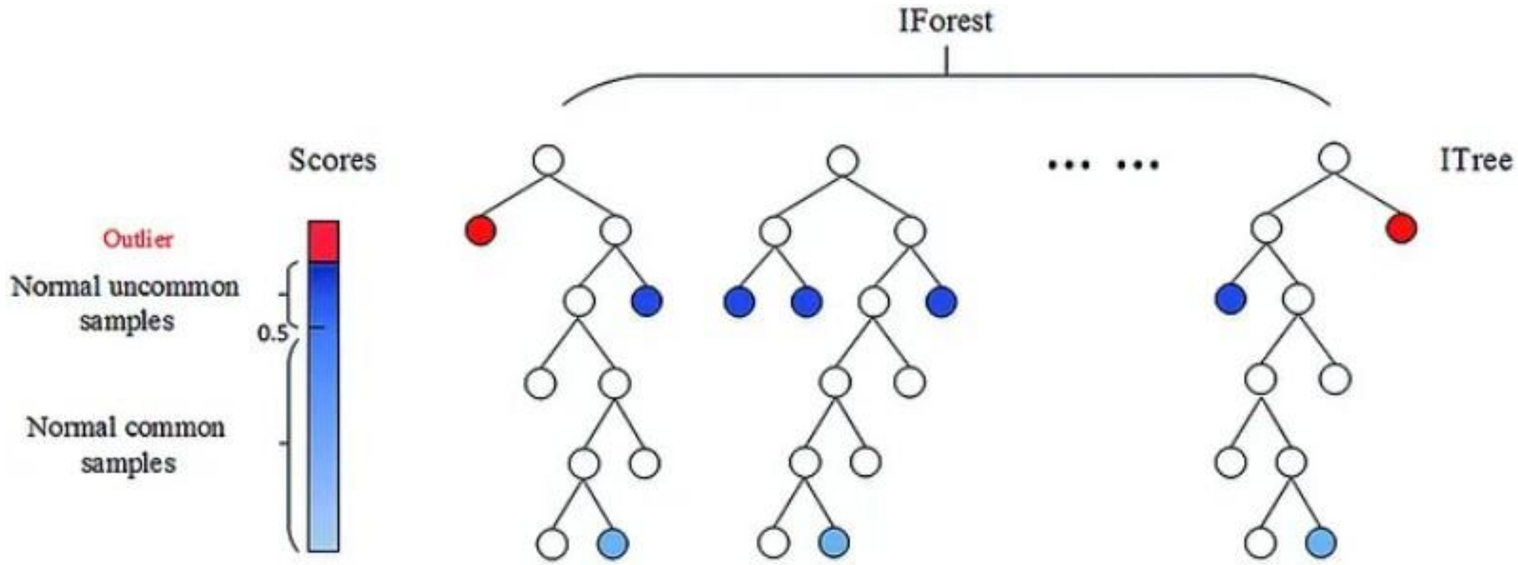
[1] Isolation Forest

[2] LOF

B.a.f

Isolation Forest

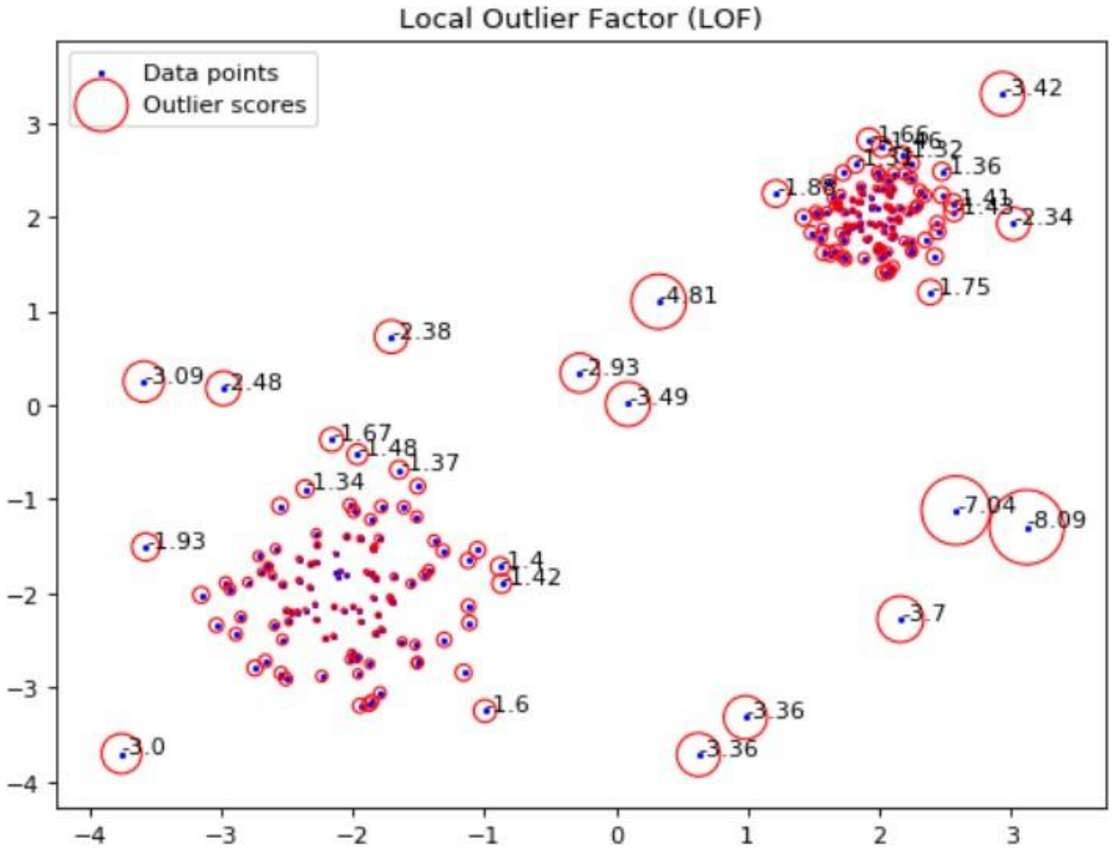
이상치는 정상 데이터에 비하여 이진 탐색 나무로 고립이 잘 될 것이라는 아이디어에서 나온 비지도 학습 알고리즘



- 다차원 데이터셋에서 효율적으로 작동하는 아웃라이어 제거 방법
- 의사결정 트리기반 이상탐지 기법
- 랜덤하게 칼럼을 선택하고, 선택된 칼럼의 최대값과 최소값을 분리하는 값을 랜덤으로 선택하는 방법

LOF

주어진 데이터가 이상치라면 해당 데이터의 밀도가 주변 이웃의 밀도보다 작을 것이라는 아이디어에 착안하여 만들어진 밀도 기반 이상치 탐지 기법



- 데이터 분포에 대한 가정 필요 없음
- 정답 라벨이 없는 데이터셋에 사용 가능한 비지도 학습 방법
- 데이터 포인트의 국소적 밀도를 기반으로 이상치를 탐지하기 때문에 다양한 밀도를 갖는 클러스터에서도 작동 가능

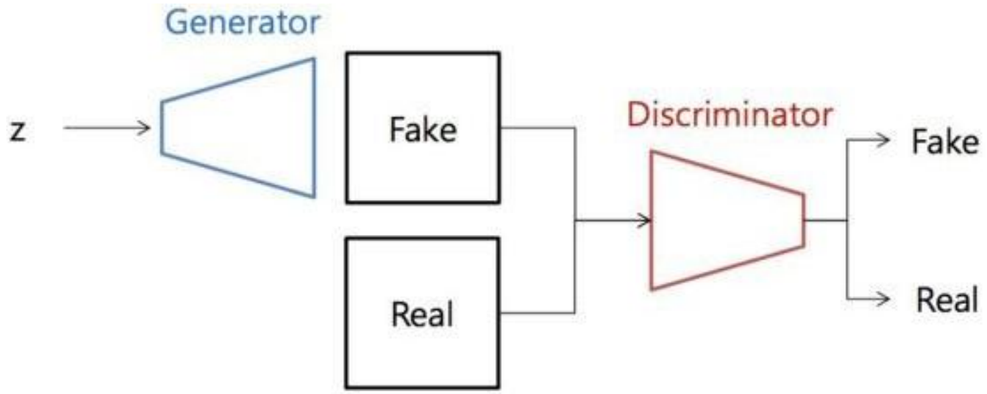
5. 인공신경망

[1] GAN

B.a.f

GAN

'생성적 적대 신경망'의 약자로, 생성자와 식별자가 서로 경쟁하며 데이터를 생성하는 모델



- Generator(생성자) : 생성된 z를 받아 실제 데이터와 비슷한 데이터를 만들어내도록 학습
- Discriminator(구분자) : 실제 데이터와 생성자가 생성한 가짜 데이터를 구별하도록 학습

활용사례

- 이미지 복원
- 얼굴 변환
- 음성 변조

감사합니다

B.a.f