GNU

Lab Seminar: 2024. 02.08

Unsupervised Learning

(Clustering & K-means algorithm & Anomaly detection)

IDEALAB
Improving lives through learning

Su-Gyeong Ban
School of Computer Science
Gyeongsang National University (GNU)

GNU 경상국립대학교 Gyeongsang National University

Contents

- Feedback
- Unsupervised Learning
- K-means algorithm
- Gaussian distribution
- Anomaly detective algorithm

Feedback (1/3)

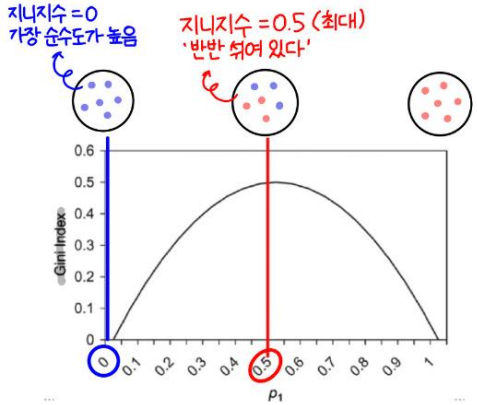
- **Bias and Variance**
  - high bias, high variance
    - 모델이 복잡하고 매우 유연할 수록 모두 높아질 가능성이 있음
    - 고차원의 복잡한 모델은 훈련에 과적합되기 쉬움
    - 훈련 데이터에서의 에러와 테스트 데이터에서 사이 에러 사이의 간격이 커짐
      - 즉, 훈련 데이터에서는 정확도가 높지만, 테스트 데이터에서는 성능이 떨어지는 경향이 있음

Feedback (2/3)

- **Generalization of Graphs**
  - 모델이 훈련 데이터에만 과도하게 맞추어져 있지 않고, 새로운 데이터에 대해서도 일반적인 패턴을 잘 학습했다는 것을 의미
    - 적절한 복잡성 : 모델이 너무 단순하거나 복잡하지 않음
    - 적은 과적합 : 모델은 훈련 데이터에만 과도하게 맞춰져 있지 않음
    - 훈련 데이터와 테스트 데이터 간의 에러 차이 감소 : 훈련 데이터 성능과 테스트 데이터 성능이 유사해짐
    - 일반화가 되었다는 것은 모델이 새로운 데이터에 대해 일반적인 패턴을 잘 학습하고, 신뢰할 만한 예측을 제공할 수 있다는 것을 의미함

Feedback (3/3)

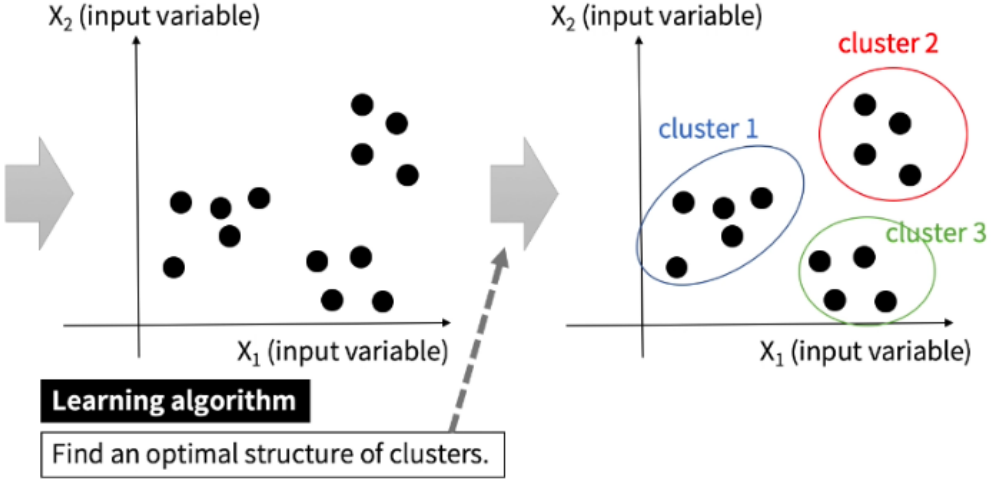
- **Gini Index**
  - 지니지수의 최댓값이 0.5인 이유
    - Decision tree에서 불순도를 측정할 때 이진 분류로 각 노드에서 두 개의 클래스로 데이터를 분할 함
    - 각 클래스가 반반씩 포함된 경우 최대 불순도를 가지고 이 때 지니지수는 0.5가 됨
    - 지니지수가 0.5일 때, 최대의 불확실성으로 클래스 간의 균형이 최악인 경우



Unsupervised Learning (1/2)

- **What is Unsupervised Learning?**
  - 레이블이 없는 데이터를 다루는 머신러닝의 한 분야
  - 데이터의 숨겨진 구조나 패턴 발견, 그룹화에 사용 됨

ID	X_1	X_2
1	128	142
2	...	...
3	...	...
4	...	...
...	...	...



Unsupervised Learning (2/2)

- **What is Clustering?**
  - 비지도 학습의 대표적인 알고리즘
  - 비슷한 특성을 가진 데이터들을 그룹으로 묶는 작업
    - 데이터 내에 숨겨진 구조나 패턴을 발견하고 이해하는데 사용 됨
    - 대표적인 알고리즘 : K-means

K-means algorithm (1/6)

- **What is K-means algorithm?**
  - 클러스터링 모델의 대표적인 종류
  - K개의 평균값을 중심으로 데이터를 군집화
    - 장점
      - 대용량 데이터에도 적용 가능하며, 계산 효율성 높음
    - 단점
      - 초기값 선택에 민감
      - 이상치에 민감하게 반응할 수 있음

K-means algorithm (2/6)

- **Algorithm process**
  - 임의의 K값을 초기 값으로 설정하고 K개의 군집 중심 선택
  - 각 데이터와 각 K개의 중심 거리가 얼마인지 계산하고 가장 가까운 중심 선택
  - 군집 중심을 다시 업데이트
  - 할당된 데이터의 군집이 군집 중심을 업데이트하기 전, 후로 변경되지 않을 때까지 2, 3 번째 단계 반복
    - 주의 할 점, 임의의 K값을 잘못 설정하면 매우 잘못된 결과로 이어질 수 있는 가능성

K-means algorithm (3/6)

- **Choosing the number of clusters**
  - 관점에 따라 초기값을 설정하는 방법이 달라짐
    - Hierarchical Clustering
      - bottom-up 방식으로 모두 계층화시킴
    - Elbow-method
      - Grid-search처럼 일일이 시행하면서 최적의 k의 초기값 설정
    - Information Criterion Approach
      - Gaussian Mixture model을 활용해 조건부확률 계산
    - Rule of thumb

K-means algorithm (4/6)

- **Hierarchical Clustering (1/2)**
  - 비슷한 군집끼리 묶어 가면서 최종적으로는 하나의 케이스가 될 때까지 군집을 묶는 클러스터링 알고리즘
  - 군집간의 거리를 기반으로 하는 알고리즘이며, K-means와 다르게 군집의 수를 미리 정해주지 않아도 됨
  - 병합형, 분할형 방법으로 나눠짐

K-means algorithm (5/6)

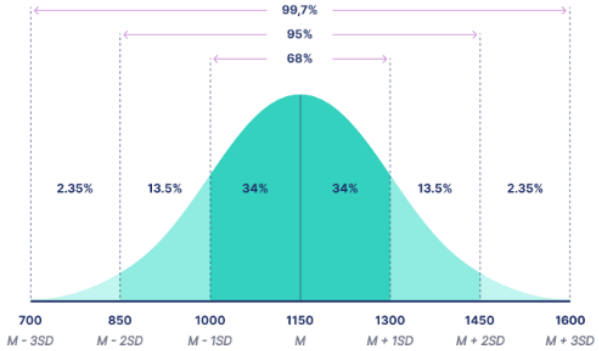
- **Hierarchical Clustering (2/2)**
  - **병합형 계층적 클러스터링**
    - 각 데이터 포인트를 하나의 클러스터로 시작하고, 가장 비슷한 두 클러스터를 병합하여 하나의 클러스터로만 만듦
    - 모든 데이터가 하나의 클러스터에 속할 때까지 이 과정을 반복
  - **분할형 계층적 클러스터링**
    - 모든 데이터 포인트를 하나의 클러스터로 시작하고 가장 다양한 데이터 포인트를 선택하여 둘로 분할 함
    - 각 클러스터가 단일 데이터 포인트가 될 때가지 분할 과정 반복

K-means algorithm (6/6)

- **Elbow-method**
  - K-means 알고리즘에서 최적의 클러스터 개수를 결정하기 위한 방법 중 하나
  - 클러스터 개수에 따른 왜곡의 변화를 그래프로 나타내고, 그래프에 꺾이는 지점을 찾아 최적의 k값 결정
    - K의 개수를 1부터 시작하여 점진적으로 증가시키면서 각각의 k에 대한 알고리즘 실행
    - 각 k값에 대한 클러스터링 결과의 왜곡 계산하고 그래프로 표현
    - 시각적으로 확인하여 값이 급격히 감소하는 지정을 찾고 이 지점에서 클러스터 개수 결정

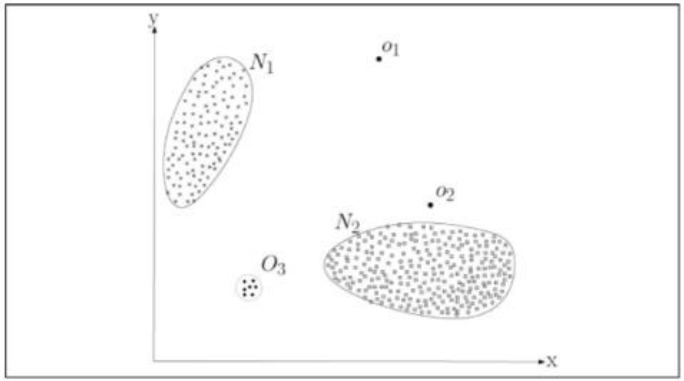
Gaussian distribution

- **What is Gaussian distribution?**
  - 통계학과 확률 이론에서 가장 잘 알려진 확률 분포 중 하나
  - 실제 데이터에서 발견되는 연속적인 확률분포로 널리 사용됨
    - 대칭성 : 평균을 중심으로 좌우 대칭적인 모양을 가짐
    - 평균 : 분포의 중심 경향성을 나타내며 이는 확률분포의 평균을 나타냄
    - 분산은 분포의 폭을 나타내고 표준편차는 분포의 넓이를 결정함



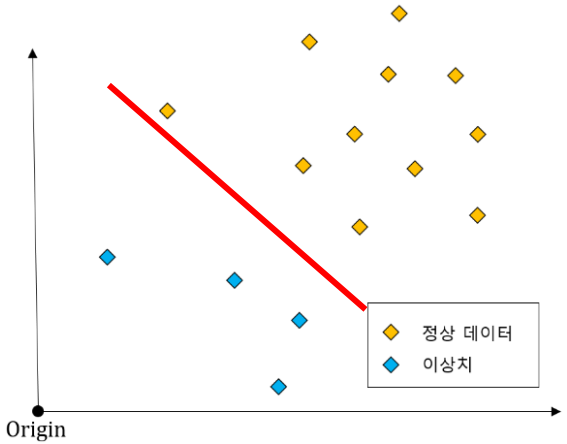
Anomaly detection algorithm (1/4)

- **What is Anomaly detection?**
  - 데이터에서 예상치 못한 이벤트나 패턴을 찾는 데 사용되는 기술
  - 학습 데이터를 기반으로 기존 데이터들과는 다른 특성을 갖는 데이터를 찾는 모형을 만드는 방법
  - 사이버 보안, 의학, 금융, 행동 패턴 분야 등 다양한 분야에 적용될 수 있음
    - One-Class SVM, Isolation Forest, Autoencoder 등



Anomaly detection algorithm (2/4)

- **One-Class SVM**
  - SVM 개념을 이용해 라벨링되어 있지 않은 데이터를 클러스터링하는 방법
    - 정상 데이터만을 사용하여 학습한 후, 데이터 공간에서 정상 데이터를 포함하는 최소 영역을 찾음
    - 새로운 데이터가 이 영역 밖에 위치하면 이를 이상 동작으로 간주




```
from sklearn.svm import OneClassSVM
X = [[0], [0.44], [0.45], [0.46], [1]]
clf = OneClassSVM(gamma='auto').fit(X)
clf.predict(X)
```

```
array([-1,  1,  1,  1, -1], dtype=int64)
```

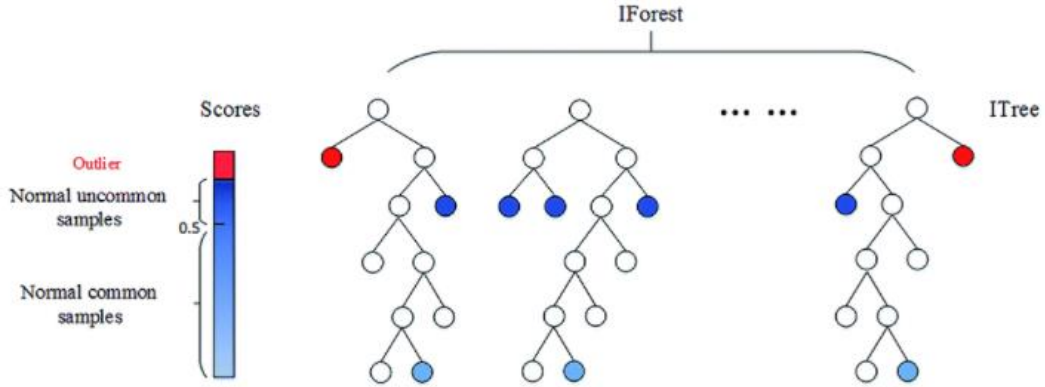
```
clf.score_samples(X)
```

```
array([1.77987316, 2.05479873, 2.05560497, 2.05615569, 1.73328509])
```

Scikit-learn 으로 One-Class SVM 실행

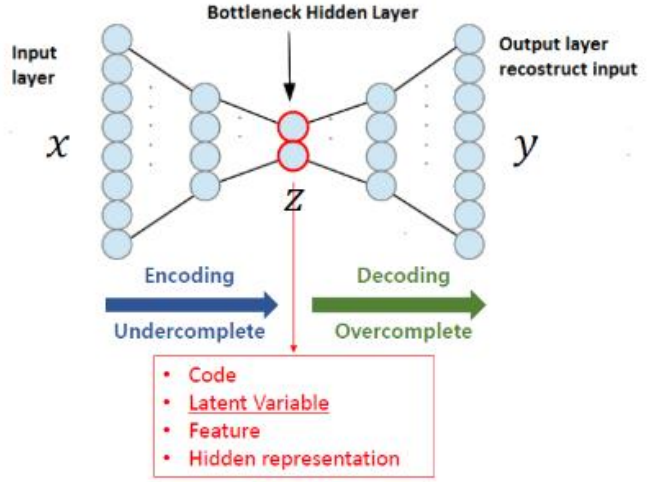
Anomaly detection algorithm (3/4)

- **Isolation Forest**
  - 여러 개의 의사결정나무를 종합한 앙상블 기반의 이상탐지 기법
    - 데이터를 트리 구조로 분할하고 이상치인 데이터를 더 적은 분할로 찾는 방식
    - 루트 노드까지의 거리가 짧을 수록 이상치가 높아지는 원리



Anomaly detection algorithm (4/4)

- **Autoencoder**
  - 신경망 기반의 비지도 학습 알고리즘
  - 입력 데이터를 재구성하는 방법으로 학습
    - 정상 데이터로 학습한 후, 새로운 데이터를 재구성하여 입력 데이터와의 차이를 측정하여 이상 동작 감지



경상국립대학교

Gyeongsang National University

Improving lives through learning

IDEALAB