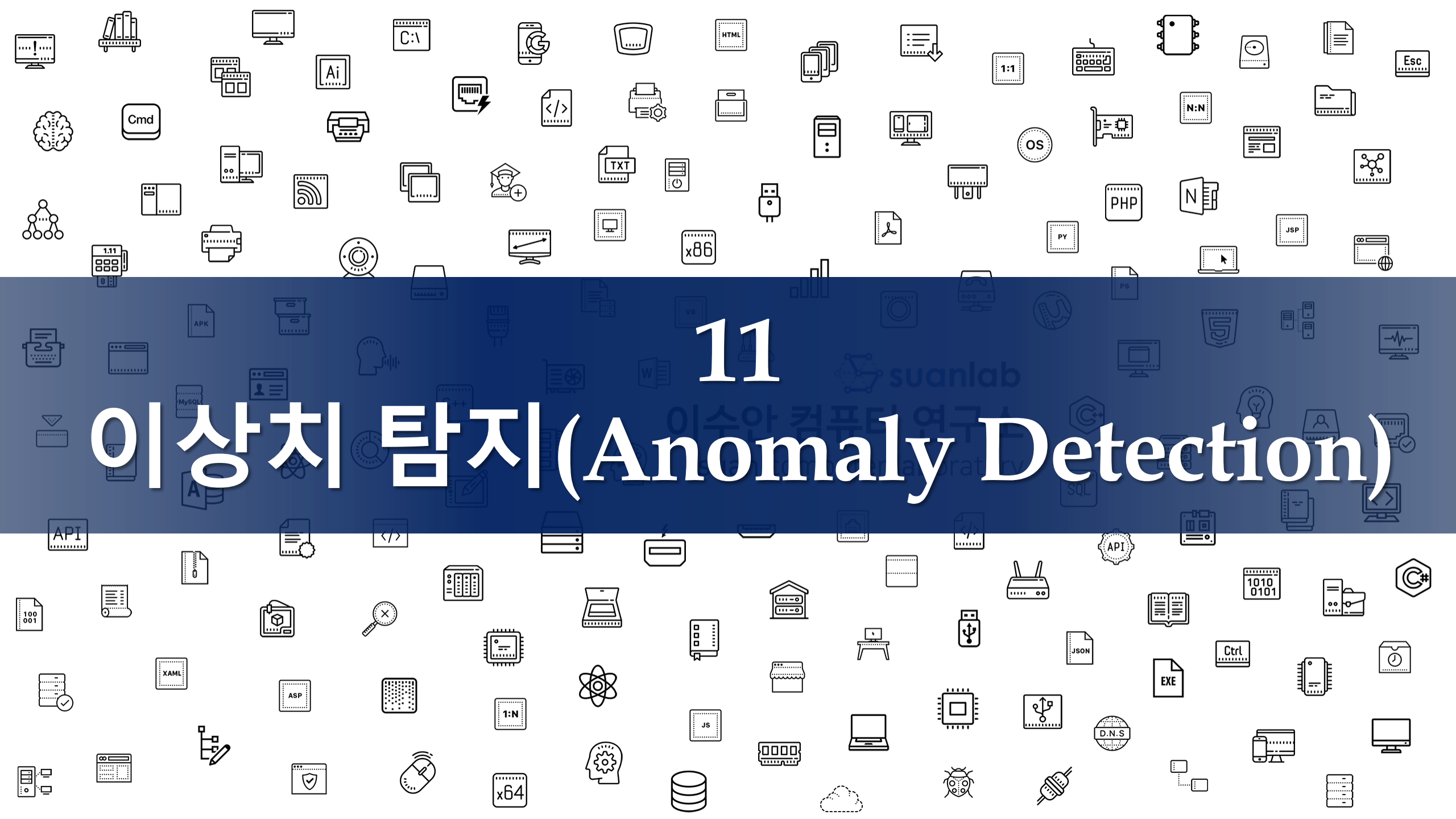
데이터 전처리
Data Preprocessing



11
이상치 탐지(Anomaly Detection)

목차

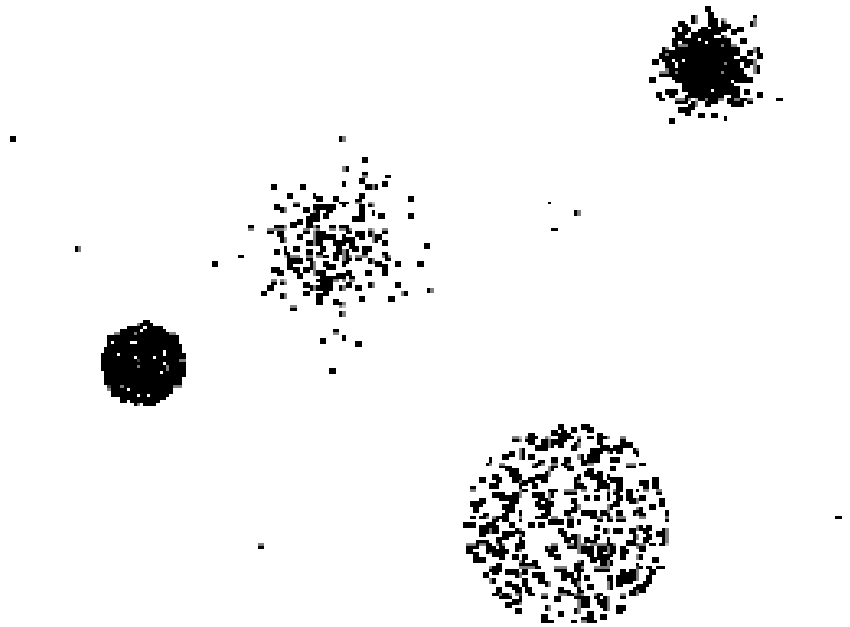
1. 이상치 탐지

이상치 탐지 Anomaly/Outlier Detection

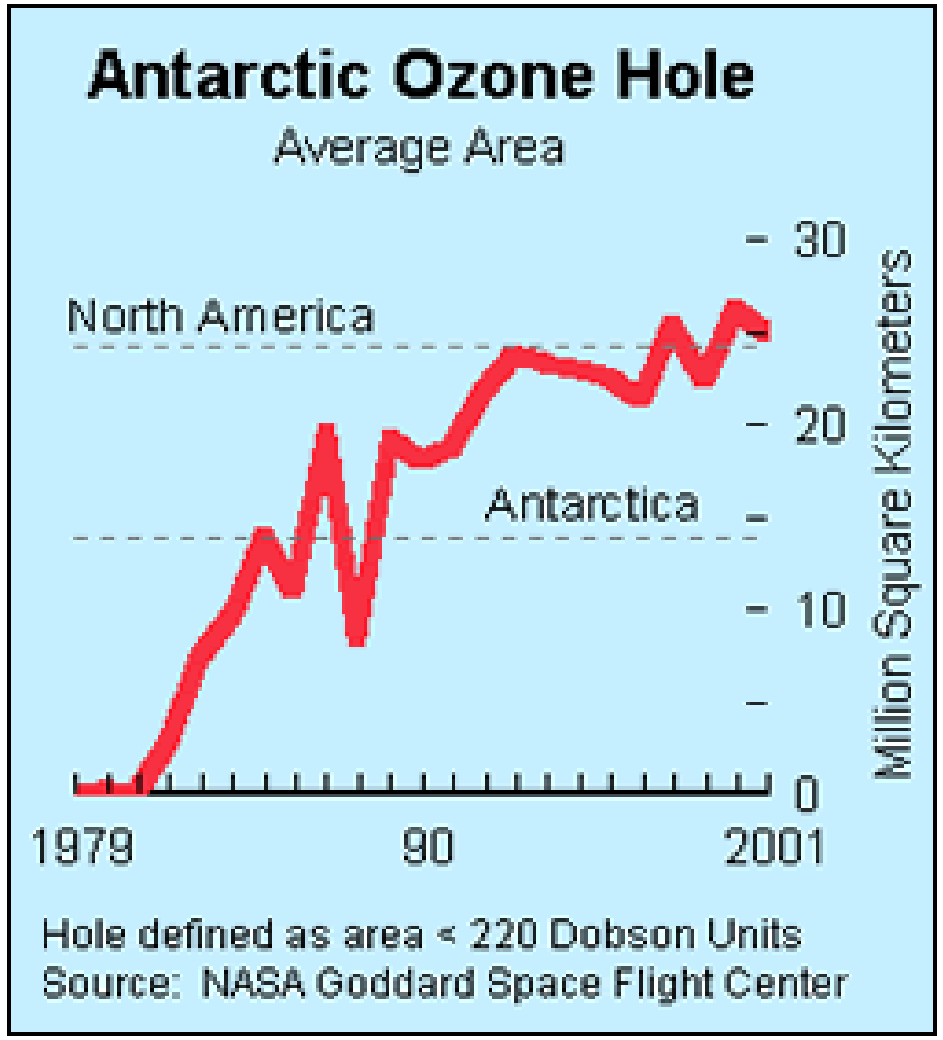
- 이상치anomalies/outliers 란 무엇인가?
  - 데이터의 나머지 부분과 상당히 다른 데이터 요소 집합

- 자연적 함의Natural implication가 이상한 것은 상대적으로 드문현상
  - 수 많은 데이터가 있는 경우, 수천 개 중에 하나가 자주 발생
  - 상황이 중요, 예: 7월에 기온이 몸시 추움

- 중요하거나 방해가 될 수 있음
  - 10 피트(3.048 미터) 키, 2살
  - 비정상적으로 높은 혈압



이상치 탐지의 중요성



오존층 파괴의 역사

- 1985 년에 3 명의 연구자 (Farman, Gardinar 및 Shanklin)가 영국 남극 조사 (British Antarctic Survey)에 의해 수집된 자료에 따라 남극 대륙의 오존 농도가 정상 수준보다 10 % 떨어 졌음을 보여줌
- 왜 Nimbus 7 인공위성은 오존 수준을 기록하기 위한 장비를 가지고 있었지만 비슷한 낮은 오존 농도를 기록하지 않았는가?
- 위성에 기록된 오존 농도는 너무 낮아서 컴퓨터 프로그램에 의해 이상치로 취급되어 폐기됨!

Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

이상치의 원인

- 다른 클래스의 데이터
  - 오렌지의 무게를 측정하지만 자몽이 몇 개 섞여 있음

- 자연 변형Natural variation
  - 비정상적으로 키가 큰 사람들

- 데이터 오류Data errors
  - 200 파운드 (약 90kg), 2살

노이즈Noise와 이상치Anomalies의 구분

- 노이즈는 잘못되었거나, 임의적이거나, 값이 있거나 오염된 객체
  - 무게가 잘못 기록됨
  - 오렌지와 섞인 자몽
- 노이즈가 반드시 비정상적인 값이나 객체를 생성하지는 않음
- 노이즈는 흥미롭지 않음
- 이상치가 노이즈의 결과가 아닌 경우는 흥미로울 수 있음
- 노이즈와 이상치는 관련이 있지만 별개의 개념

일반적인 이슈: 속성의 수

- 많은 이상치가 하나의 속성으로 정의
  - 신장Height
  - 모양Shape
  - 색깔Color
- 모든 속성을 사용하여 이상치를 찾기가 어려울 수 있음
  - 노이즈 또는 관련 없는 속성
  - 객체는 일부 속성과 관련해서만 이상치를 가짐
- 그러나 어떤 속성에서는 객체가 이상치가 아닐 수도 있음

일반적인 이슈: 이상치 점수

- 많은 이상치 탐지 기술은 단지 이진 분류만을 제공
  - 객체가 이상치이거나 그렇지 않음
  - 특히 분류 기반 접근법에 해당

- 다른 접근법은 모든 포인트에 점수를 할당
  - 점수는 객체가 비정상인 정도를 측정
  - 객체의 순위를 매길 수 있음

- 결국 이진 결정이 필요할 수 있음
  - 이 신용 카드 거래가 신고되어야 하나?
  - 여전히 점수를 얻는 데 유용

- 얼마나 많은 이상치가 있나?

이상치 탐지를 위한 기타 이슈

- 한번에 한 가지 또는 한번에 하나씩 모든 이상치를 찾기
  - 쇄도Swamping
  - 차폐Masking

- 평가Evaluation
  - 성능을 어떻게 측정하나?
  - 지도Supervised vs. 비지도unsupervised 상황

- 효율Efficiency

- 상황Context
  - 프로 농구팀

이상치 탐지 문제의 변형

- 데이터 집합 D가 주어지면, 어떤 임계값 t보다 큰 이상치 점수를 갖는 모든 데이터 포인트 $x \in D$를 찾음

- 데이터 집합 D가 주어지면, top-n개의 가장 큰 이상치 점수를 갖는 모든 데이터 포인트 $x \in D$를 찾음

- 데이터 집합 D가 주어지면, 대부분 일반적인 (그러나 레이블이 없는) 데이터 포인트와 테스트 포인트 x를 포함하고, D에 대한 x의 이상치 점수를 계산

모델 기반 이상치 탐지

- 데이터에 대한 모델을 생성하고 확인
- 비지도Unsupervised
  - 이상치는 잘 맞지 않는 포인트
  - 이상치는 모델을 왜곡시키는 포인트
  - 예제:
    - 통계 분포Statistical distribution
    - 클러스터Clusters
    - 회귀 분석Regression
    - 기하학Geometric
    - 그래프Graph
- 지도Supervised
  - 이상치는 희귀한 등급으로 간주
  - 학습 데이터가 필요

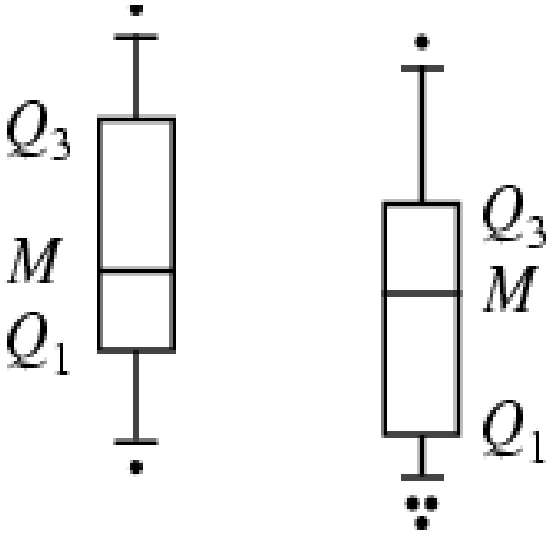
추가적인 이상치 탐지 기술

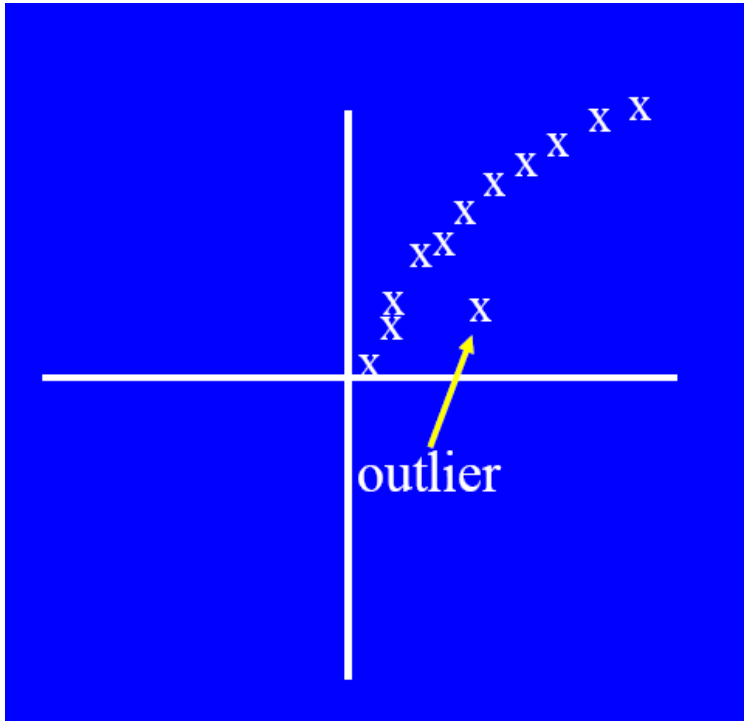
- 근접 기반Proximity-based
  - 이상치는 다른 포인트와 멀리 떨어진 지점
  - 일부 경우 그래픽로도 감지 가능
- 밀도 기반Density-based
  - 저밀도 포인트는 이상치
- 패턴 매칭Pattern matching
  - 이례적이지만 중요한 이벤트 또는 객체의 프로필이나 템플릿 생성
  - 이러한 패턴을 탐지하는 알고리즘은 일반적으로 간단하고 효율적

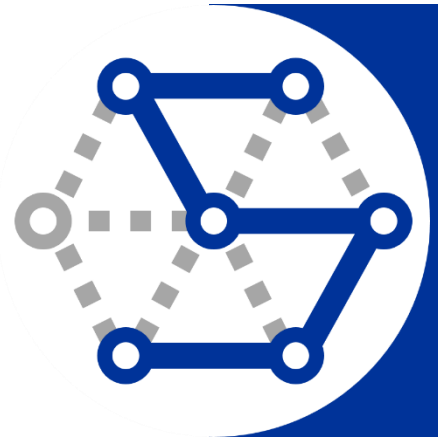
시각적 접근방법

- 박스 플롯Boxplots 또는 분산형 플롯(산포도)scatter plots

- 한계
  - 자동이 아님
  - 주관적Subjective





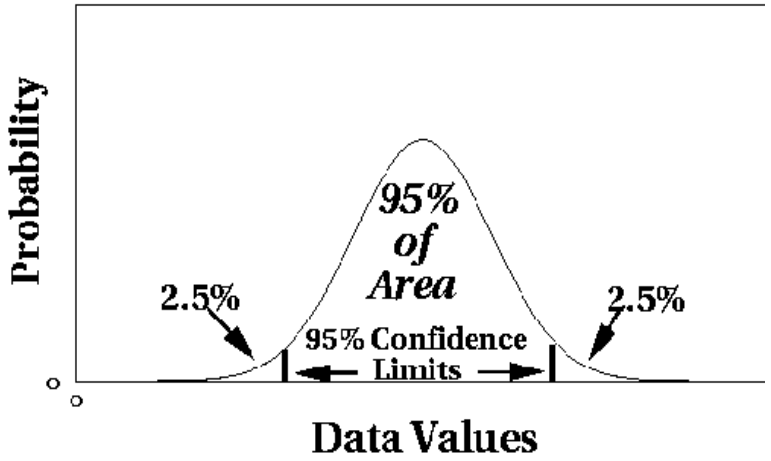


2. 통계 접근방식

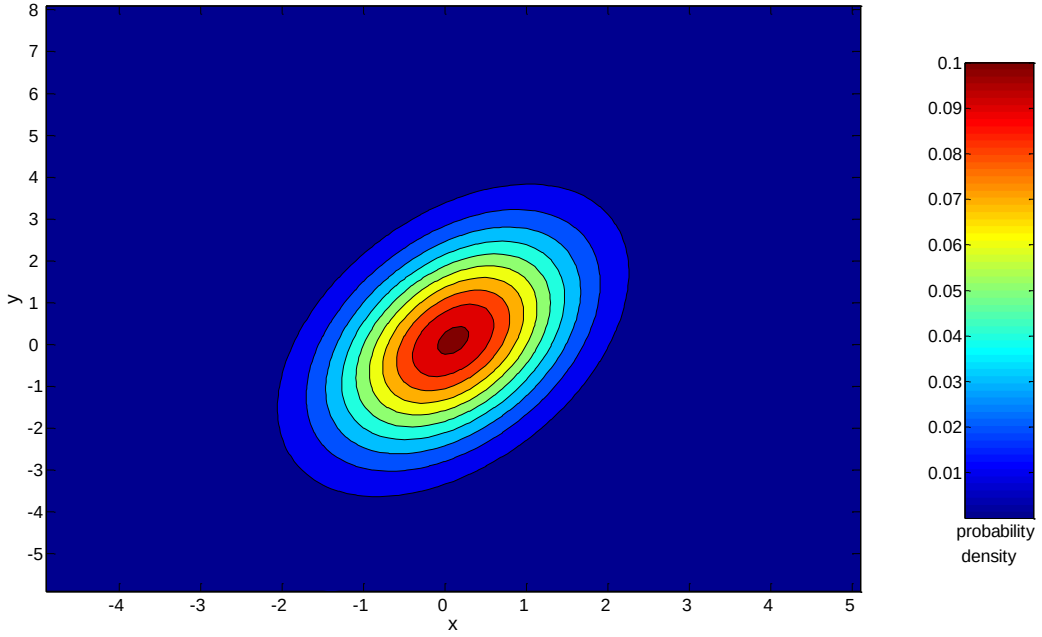
통계적 접근법

- 이상치의 확률적 정의:
- Probabilistic definition of an outlier: An outlier is an object that has a low probability with r espect to a probability distribution model of the data.
- Usually assume a parametric model describing the distribution of the data (e.g., normal di stribution)
- Apply a statistical test that depends on
  - Data distribution
  - Parameters of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)
- Issues
  - Identifying the distribution of a data set
    - Heavy tailed distribution
  - Number of attributes
  - Is the data a mixture of distributions?

정규 분포Normal Distributions



1차원 가우시안Gaussian



2차원 가우시안Gaussian

Grubbs' Test

- 단변량 데이터univariate data의 이상치 검출
- 데이터가 정규 분포에서 온다고 가정
- 한번에 하나의 이상치 탐지, 이상치 제거, 반복
  - H_0: 데이터에 이상치가 없음
  - H_A: 데이터에 이상치가 적어도 하나 이상 있음
- Grubbs' test 통계:

$$G = \frac{\max\left|X - \overline{X}\right|}{s}$$

- 다음 경우엔 H_0 거부:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

통계 기반Statistical-based – 가능성 접근Likelihood Approach

- 데이터 집합 D가 2개의 확률 분포의 혼합된 샘플을 포함한다고 가정:
  - M (다수 분포majority distribution)
  - A (비정상 분포anomalous distribution)
- 일반적인 접근General Approach:
  - 처음에는 모든 데이터 포인트가 M에 속한다고 가정
  - $L_t(D)$를 시간 t에서 D의 log likelihood
  - M에 속한 각 포인트 x_t 에 대해 A로 이동
    - $L_{t+1}(D)$를 새로운 log likelihood
    - 차이를 계산, $\Delta = L_t(D) - L_{t+1}(D)$
    - 만약 $\Delta > c$ (임계값)인 경우, x_t 는 이상치로 선언하고 M에서 A로 영구적으로 이동

통계 기반Statistical-based – 가능성 접근Likelihood Approach

- 데이터 분포Data distribution, $D = (1 - \lambda) M + \lambda A$
- M은 데이터로부터 추정된 확률 분포probability distribution
  - 어떤 모델링 방법(naïve Bayes, maximum entropy, etc)을 기반으로 할 수 있음
- A는 초기에 균일 분포로 가정
- 시간 t에서의 Likelihood:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

통계적 접근의 강점/약점

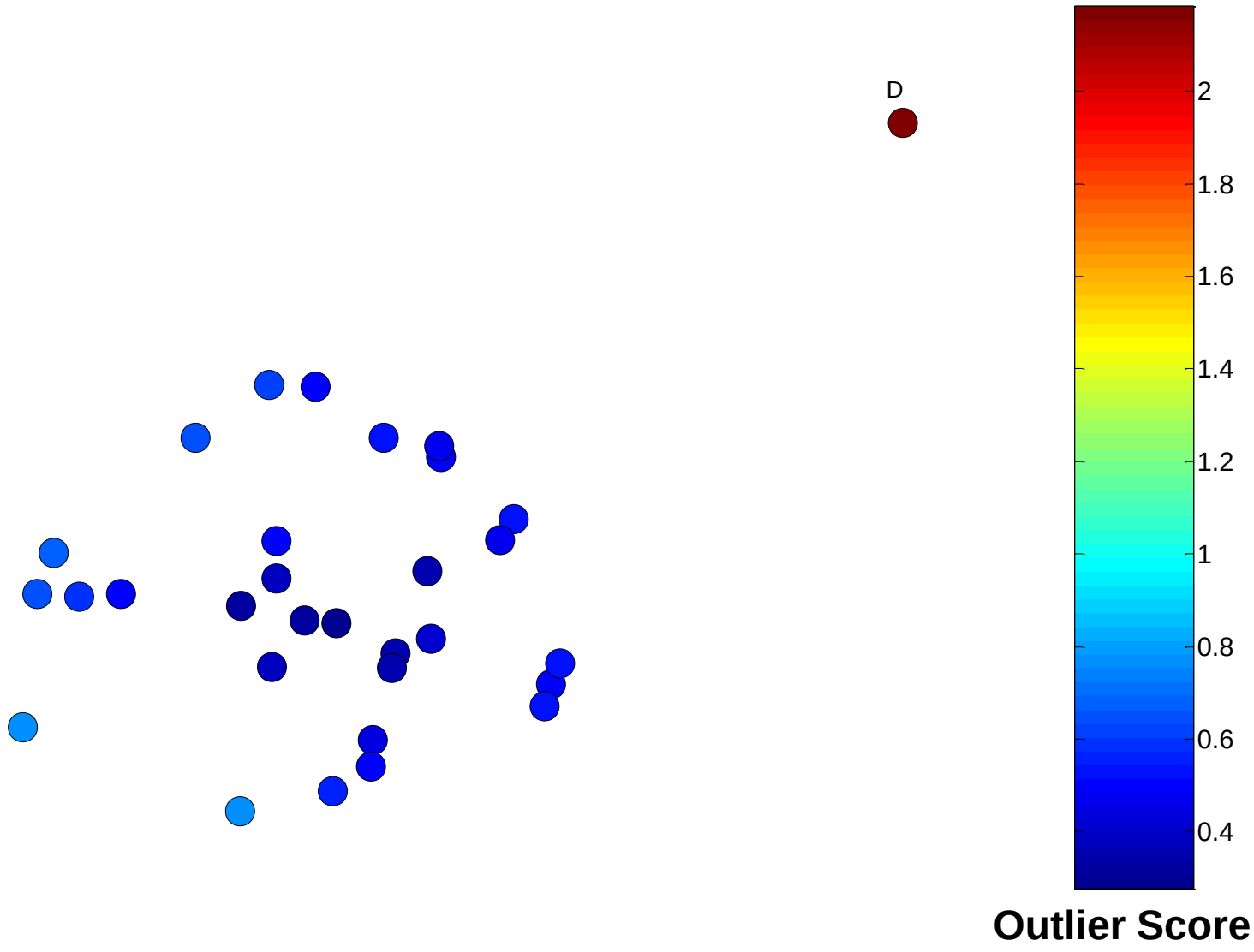
- 변동 없는 수학 기초
- 매우 효율적일 수 있음
- 분포가 알려진 경우 좋은 결과
- 많은 경우, 데이터 분포가 알려지지 않을 수 있음
- 고차원 데이터의 경우 실제 분포를 예측하기 어려울 수 있음
- 이상치는 분포의 매개 변수를 왜곡시킬 수 있음

3. 근접성 기반 이상치 탐지

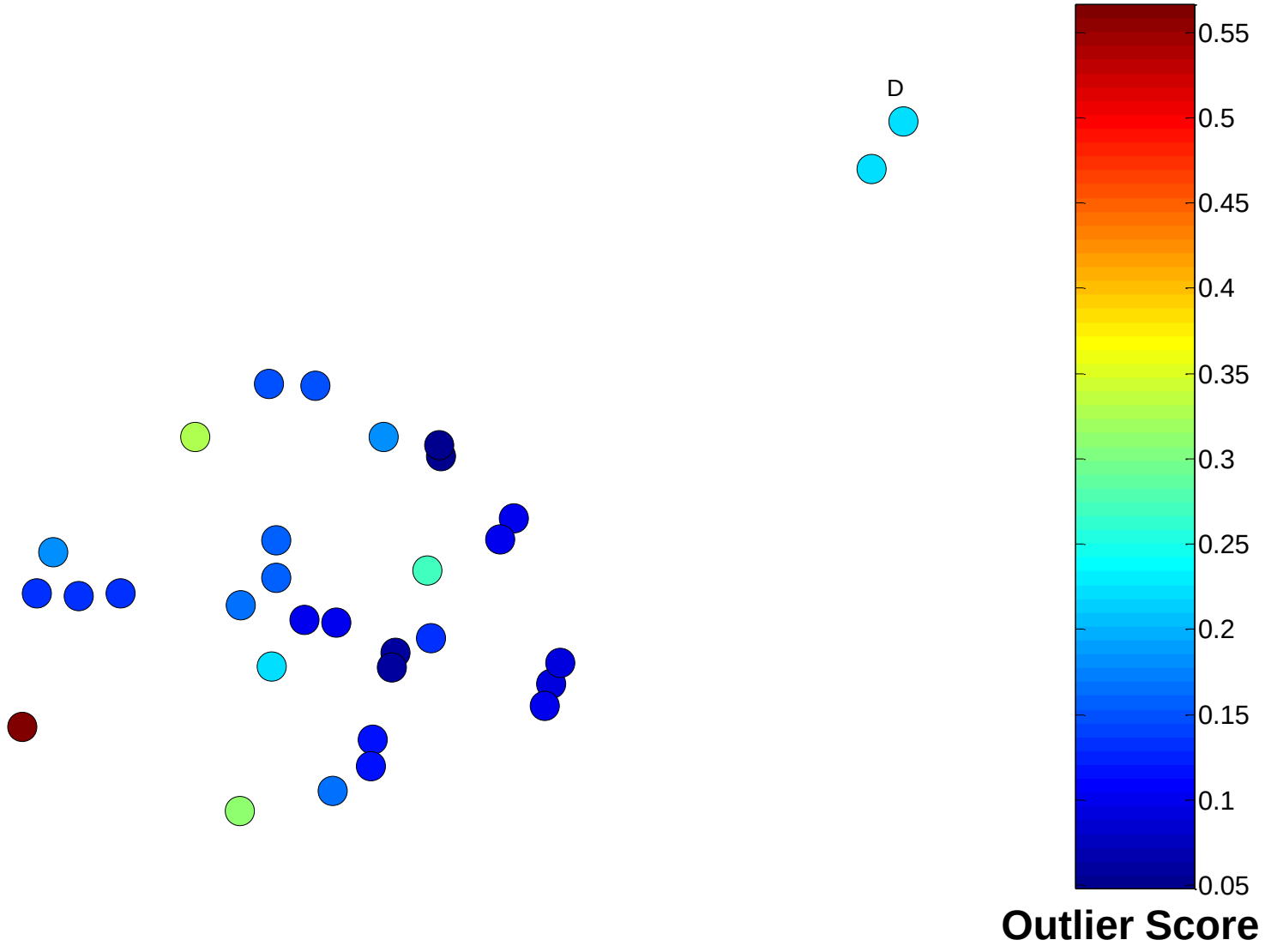
거리 기반 접근Distance-Based Approaches

- 여러 가지 기술
- 객체의 특정 부분이 지정된 거리보다 멀리 떨어져 있으면 객체가 이상치 (Knorr, Ng 1998)
  - 일부 통계적 정의는 이것의 특별한 경우
- 객체의 이상치 점수는 k번째 가장 가까운 이웃까지의 거리

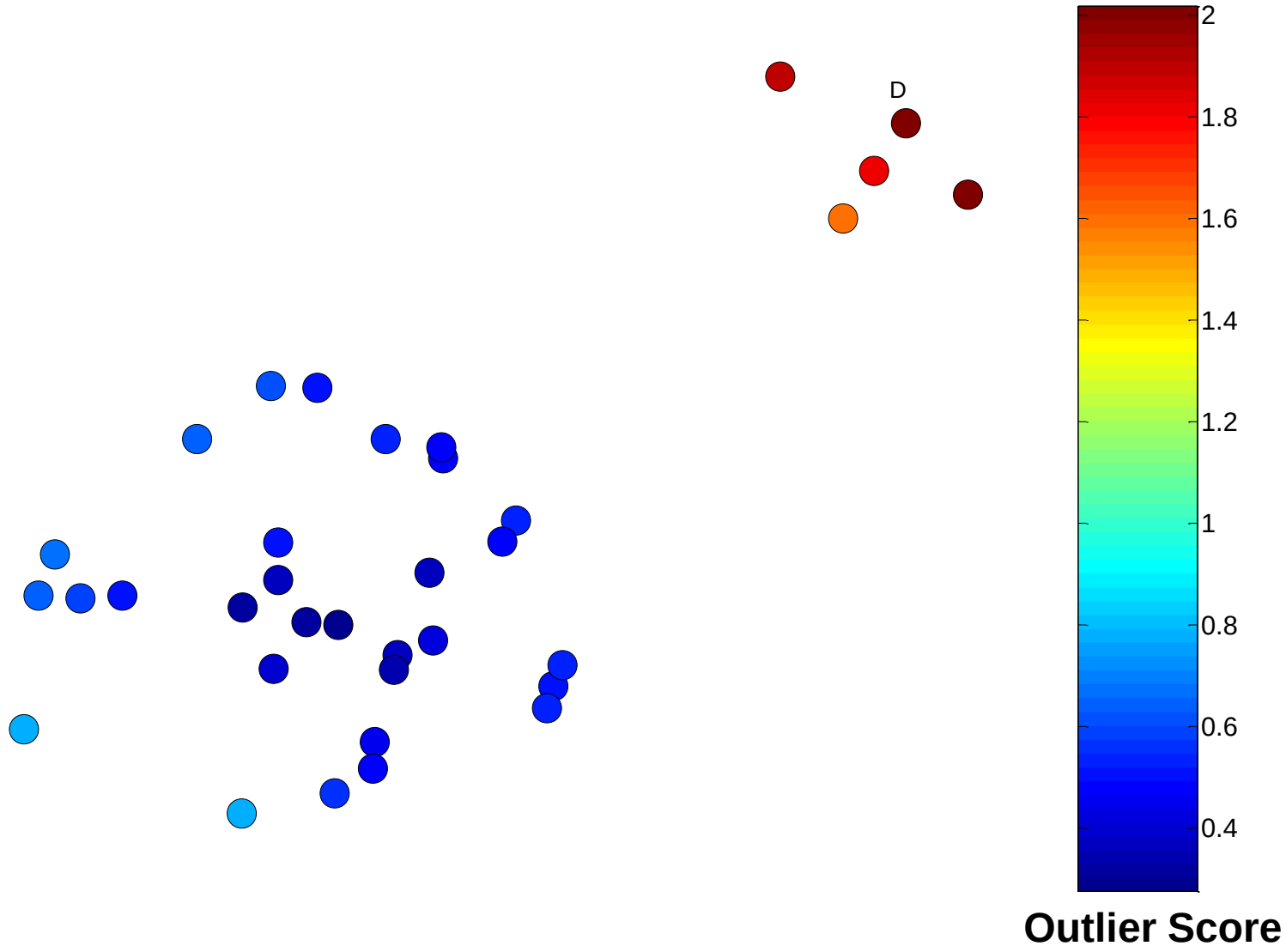
하나의 가장 가까운 이웃 – 하나의 이상치



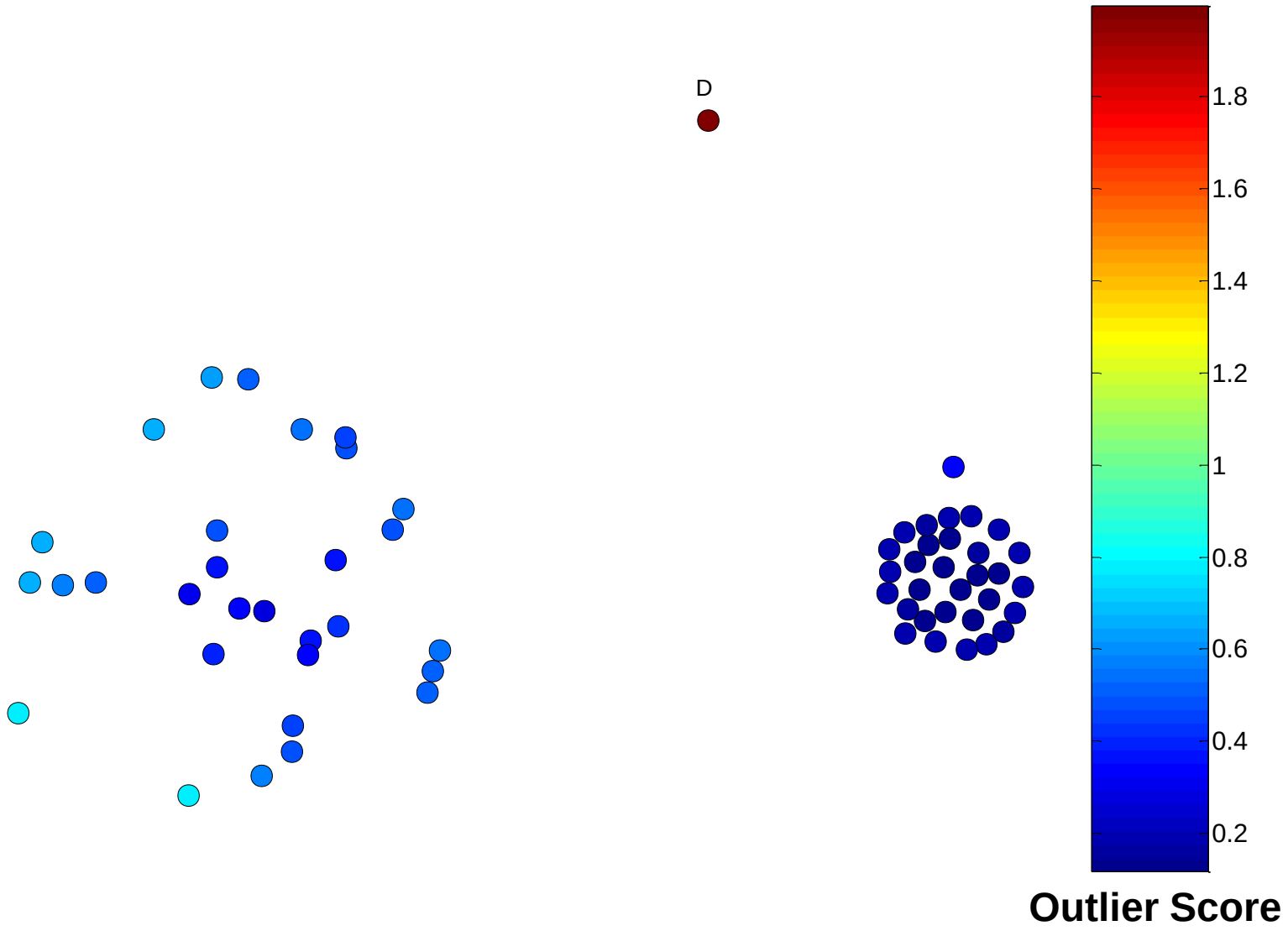
하나의 가장 가까운 이웃 – 두개의 이상치



가장 가까운 5명의 이웃 – 소규모 클러스터

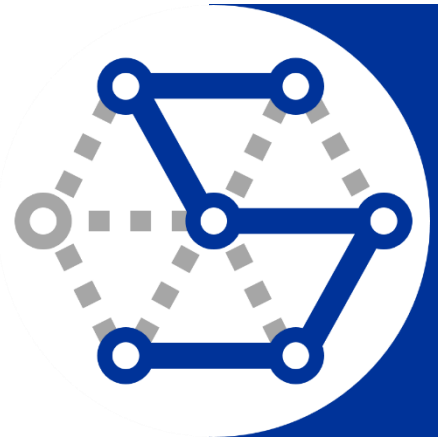


5명의 가장 가까운 이웃 – 다른 밀도



거리 기반 접근Distance-Based Approaches 의 강점과 단점

- 단순함Simple
- 비쌈 – O(n2)
- 매개 변수에 민감
- 밀도의 변화에 민감
- 고차원 공간에서 거리가 덜 의미가 있음



4. 밀도 기반 이상치 탐지

밀도 기반 접근Density-Based Approaches

- 밀도 기반 이상치Density-based Outlier: 객체의 이상치 점수는 객체 주위의 밀도의 역수
  - K개의 갖아 가까운 이웃들의 관점에서 정의 가능
  - 하나의 정의, k 번째 이웃과의 거리의 반전One definition: Inverse of distance to kth neighbor
  - 또 다른 정의: k거리에 평균 거리의 반전 예측

- 만약 밀도가 다른 영역이 있는 경우, 이 접근 방식에 문제가 있을 수 있음

밀도 기반 접근Density-Based Approaches

- 밀도 기반 이상치Density-based Outlier: 객체의 이상치 점수는 객체 주위의 밀도의 역수
  - K개의 가장 가까운 이웃들의 관점과 정의 될 수 있음
  - 하나의 정의: k 번째 이웃과의 거리를 반전
  - 또다른 정의: k 거리에 대한 평균 거리의 반전
  - DBSCAN 정의

- 밀도가 다른 영역이 있는 경우 이 접근 방식은 문제가 있을 수 있음

상대 밀도Relative Density

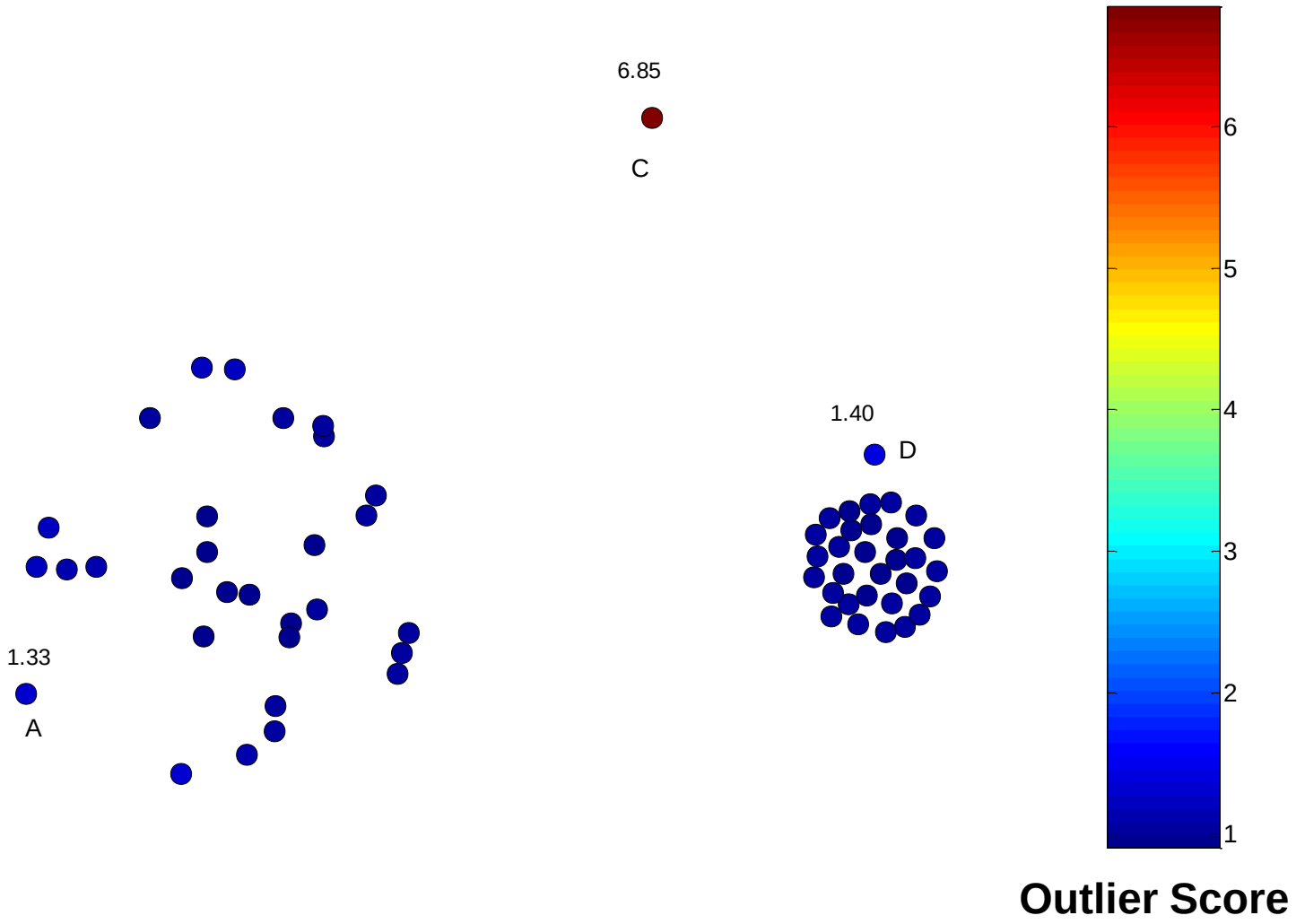
- K 개의 가장 가까운 이웃들에 대한 점의 밀도를 고려

$$average\ relative\ density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x},k)} density(\mathbf{y}, k)/|N(\mathbf{x}, k)|}. \quad (10.7)$$

Algorithm 10.2 Relative density outlier score algorithm.

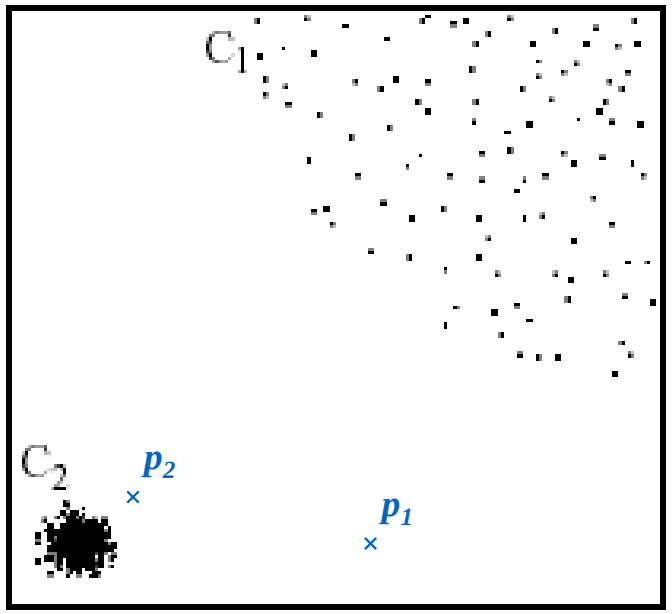
1: {k is the number of nearest neighbors}
2: **for all** objects $\mathbf{x}$ **do**
3: Determine $N(\mathbf{x}, k)$, the k-nearest neighbors of $\mathbf{x}$.
4: Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$, using its nearest neighbors, i.e., the objects in $N(\mathbf{x}, k)$.
5: **end for**
6: **for all** objects $\mathbf{x}$ **do**
7: Set the $outlier\ score(\mathbf{x}, k) = average\ relative\ density(\mathbf{x}, k)$ from Equation 10.7.
8: **end for**

상대 밀도 이상치 점수Relative Density Outlier Scores



밀도 기반Density-based: LOF 접근 LOFLocal Outlier Factor approach

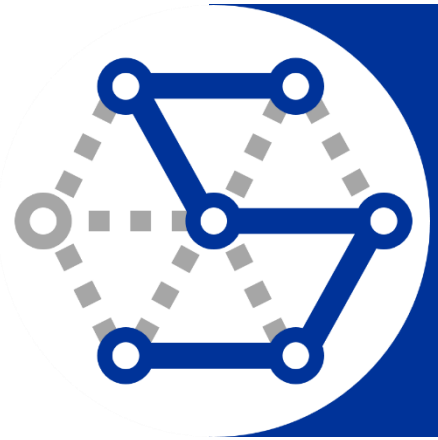
- 각 포인트마다, 해당 지역의 밀도를 계산
- 샘플 p의 밀도와 가장 가까운 이웃의 밀도의 비의 평균으로 샘플
- 이상치는 가장 큰 LOF 값을 갖는 포인트
- Sed: LOF 방식



NN 접근법에서 p_2 이상치로 간주하지 않지만 LOF 접근법은 1p와 2p 둘다 왜래키를 찾습니다.

밀도 기반 접근법 강정/약점

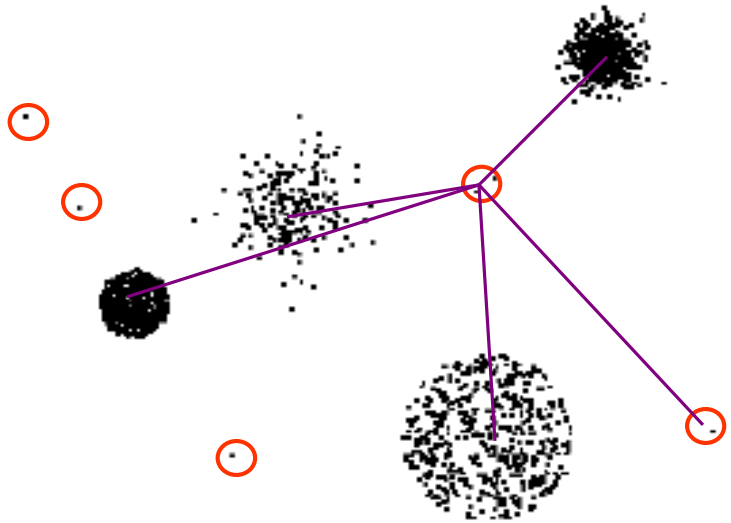
- 단순함
- 비쌈 – O(n2)
- 매개 변수는 민감
- 밀도는 고차원 공간에서 덜 의미있게 됨



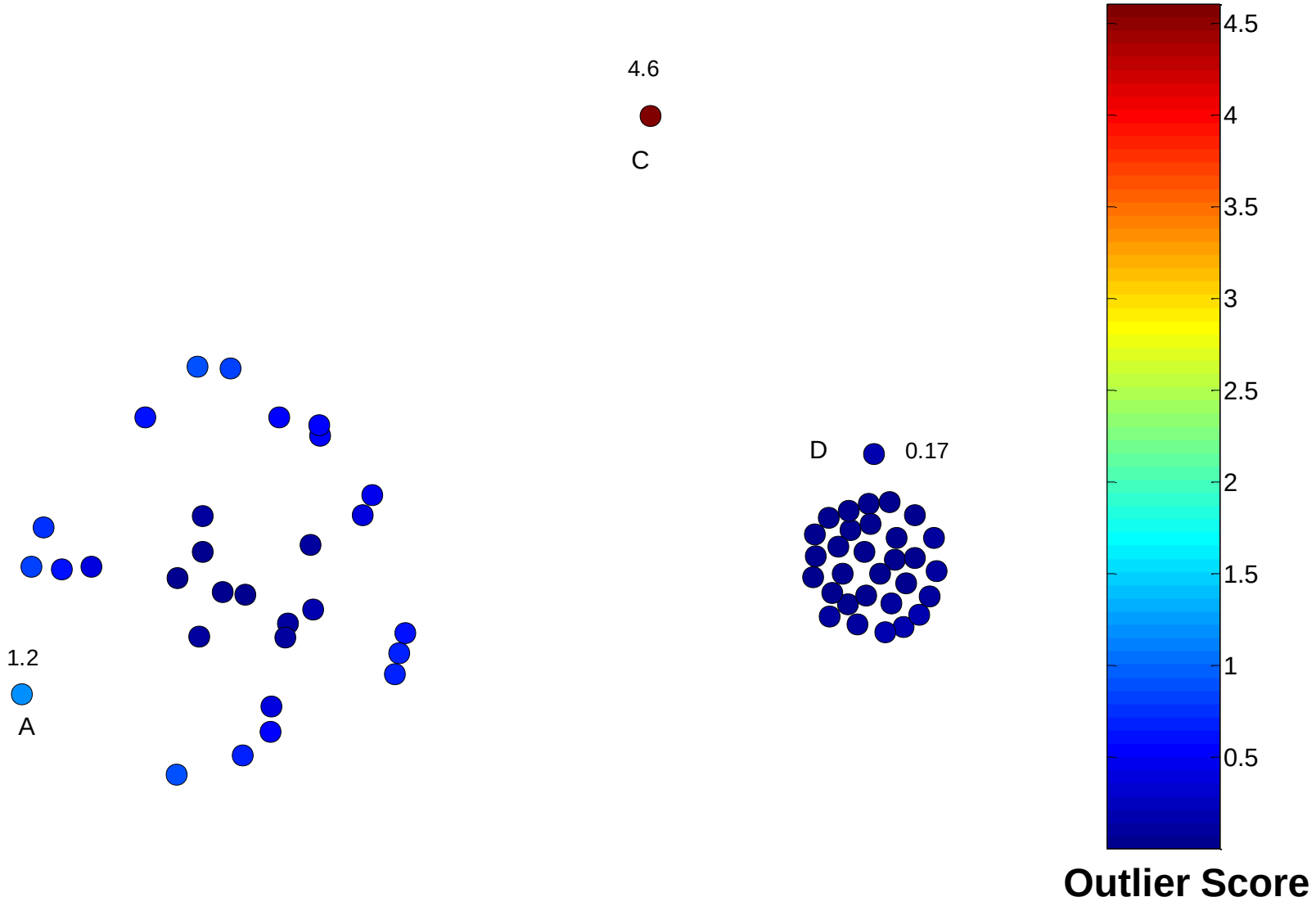
5. 군집 기반 이상치 탐지

군집 기반 접근Clustering-Based Approaches

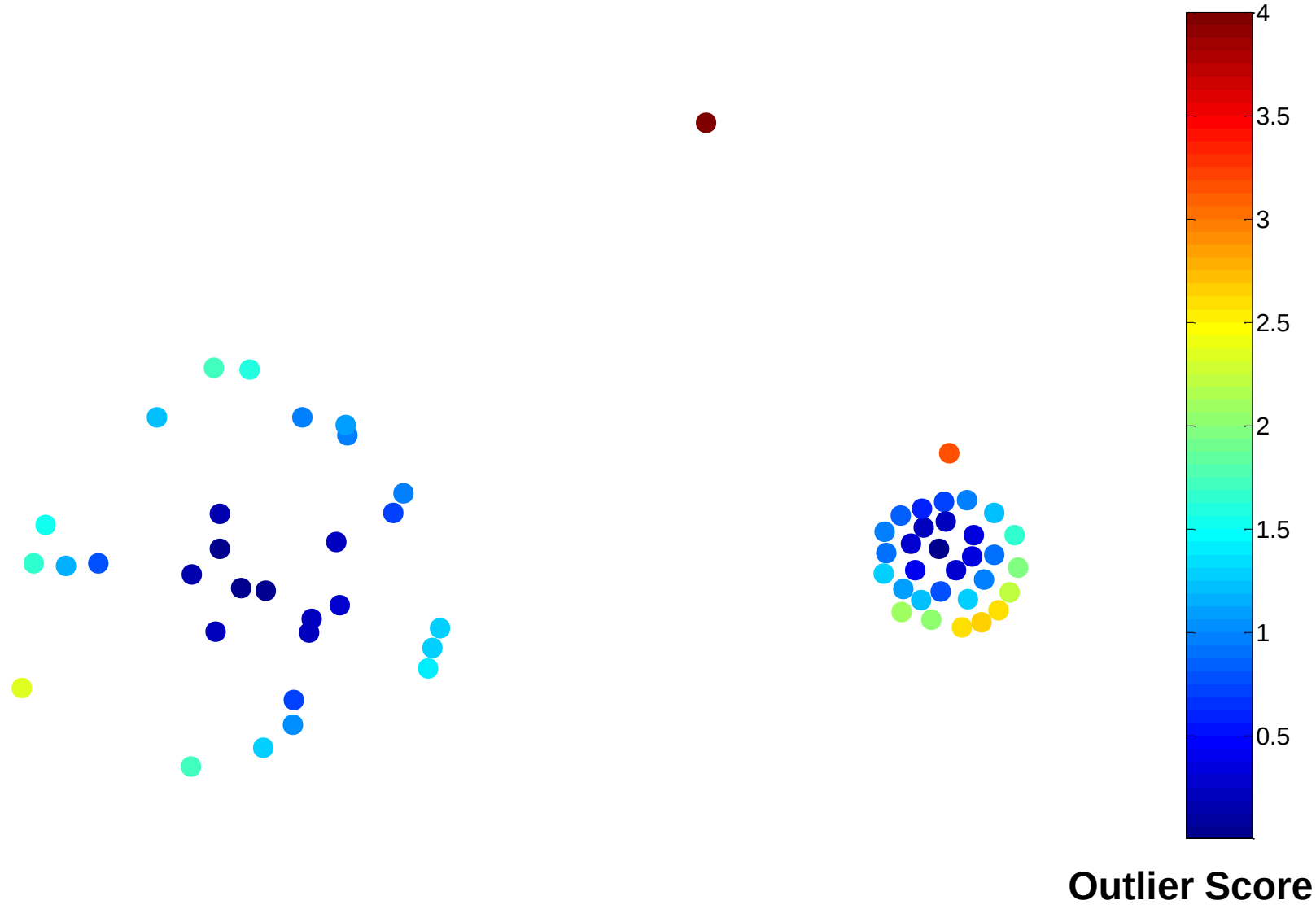
- 클러스터링 기반 이상치: 어떤 클러스터에도 강력하게 속하지 않으면 객체는 클래스터 기반 이상치
  - 프로토타입 클러스터prototype-based clusters, 클러스터 센터에 충분히 근접하지 않으면 아웃 라이어
  - 밀도 기반 클러스터의 경우 밀도가 너무 낮으면 객체는 이상치
  - 그래프 기반 클러스터의 경우 오브젝트가 제대로 연결되지 않을 경우 아웃 리어입니다.
- 다른 문제로는 클러스터에 이상치가 미치는 영향과 클러스터 수



가장 가까운 중심지에서의 거리



가장 가까운 중심지에서의 거리



거리 기반 접근법의 강정/약점

- 단순함Simple

- 많은 클러스터링 기술 사용 가능

- 클러스터링 기술을 결정하기 어려울 수 있음

- 클러스터의 수를 결정하기 어려움

- 이상치를 사용하면 클러스터가 왜곡될 수 있음

Q & A