Anomaly Detection

Business Analytics (IME654)

2022. 11. 06

Team: 동기사랑

Member: 김창현, 정진용

해당 발표자료는

고려대학교 산업경영공학과

강필성 교수님: 비즈니스 애널리틱스(IME654)
김성범 교수님: 다변량 통계분석 및 데이터 마이닝(IME567)

의 강의자료를 사용했음을 미리 밝힙니다.

Anomaly Detection

❖ Anomaly detection 방법론

1. Density/Distance-based methods
   - Gaussian Mixture Model
   - K-Nearest Neighbors (KNN) method
   - LOF(Local Outlier Factors) : 데이터 밀도 또는 거리 척도를 통해, majority 군집과 minority 군집을 생성하여 이상치를 탐지

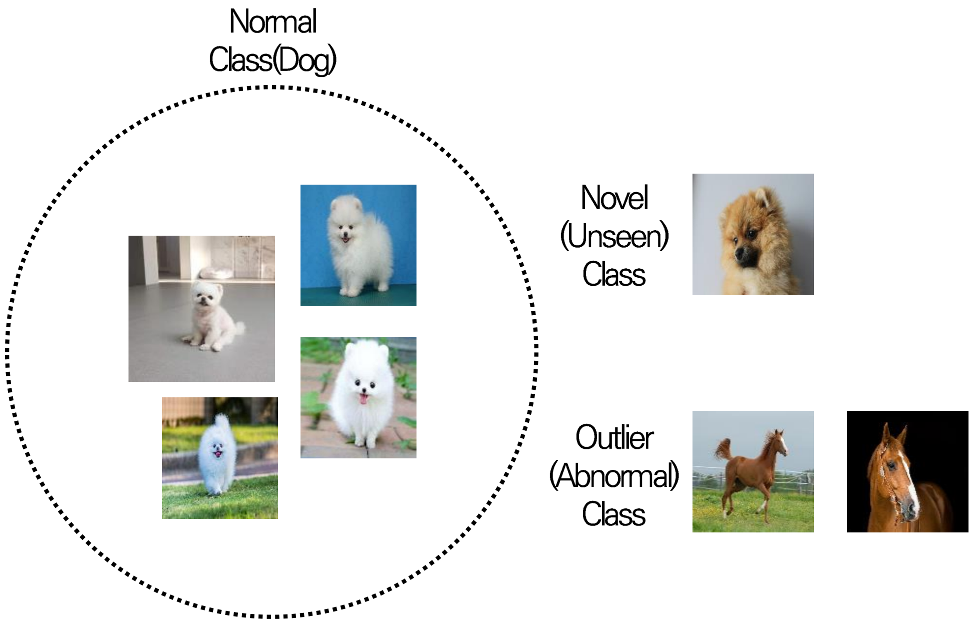
2. Model-based methods
   - Isolation Forest : Tree based method로써 데이터를 분할 및 고립시켜 이상치를 탐지
   - 1-class SVM : 데이터가 존재하는 영역을 정의하여, 영역 밖의 데이터들은 이상치로 간주

3. Reconstruction-based methods
   - PCA(Principal Component Analysis) method
   - Auto-Encoder based method : 고차원 데이터에서 주로 사용하는 방법론으로서 데이터를 압축/복원하여 복원된 정도로 이상치를 판단

Anomaly Detection

❖ Anomaly detection과 novelty detection의 차이?

- 비정상 sample의 정의하는 방식에 따른 분류 차이임
- Anomaly를 정의하는 방식을 잘 살펴보고 접근해야 함



학습시 비정상 sample의 사용 여부 및 label 유무에 따른 분류

용어	비정상 sample
Novelty Detection	지금까지 등장하지 않았지만 분포 내에 존재할 수 있는 sample
Outlier Detection	지금까지 등장하지 않았지만 분포 내에 존재할 수 없는 sample 이며, 동시에 데이터에 오염이 발생했을 수도 있는 sample

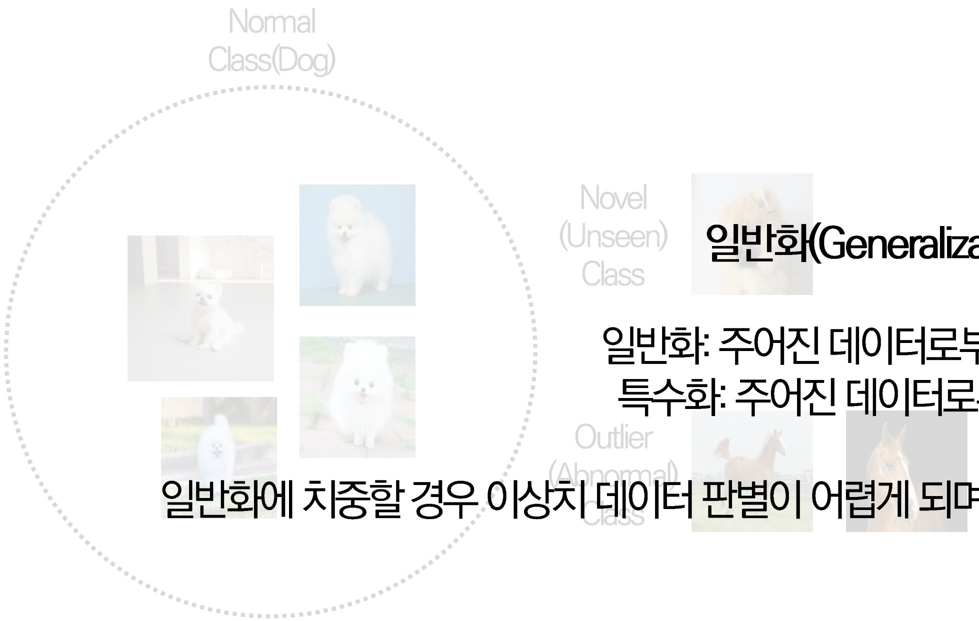
비정상 sample 정의에 따른 분류

용어	정상 sample	비정상 sample
Supervised Anomaly Detection	학습 사용	학습 사용
Semi-Supervised (one-class) Anormaly Detection	학습 사용	학습에 사용 안함
Unsupervised Anomaly Detection	Label이 없어서 모름 학습에 사용하는 데이터의 대다수가 정상 sample일 것이라고 가정함	

Anomaly Detection

❖ Anomaly detection과 novelty detection의 차이?

- 비정상 sample의 정의하는 방식에 따른 분류 차이임
- Anomaly를 정의하는 방식을 잘 살펴보고 접근해야 함



일반화(Generalization)와 특수화(Specialization)

일반화: 주어진 데이터로부터 정상 범주의 개념을 확장해 가는 것
특수화: 주어진 데이터로부터 정상 범주의 개념을 좁혀 가는 것

일반화에 치중할 경우 이상치 데이터 판별이 어렵게 되며, 특수화에 치중할 경우 과적합 위험(잦은 false alarm)에 빠질 수 있음

학습시 비정상 sample의 사용 여부 및 label 유무에 따른 분류

용어	비정상 sample
Novelty Detection	지금까지 등장하지 않았지만 분포 내에 존재할 수 있는 sample
Outlier Detection	지금까지 등장하지 않았지만 분포 내에 존재할 수 없는 sample 이며, 동시에 데이터에 오염이 발생했을 수도 있는 sample

비정상 sample 정의에 따른 분류

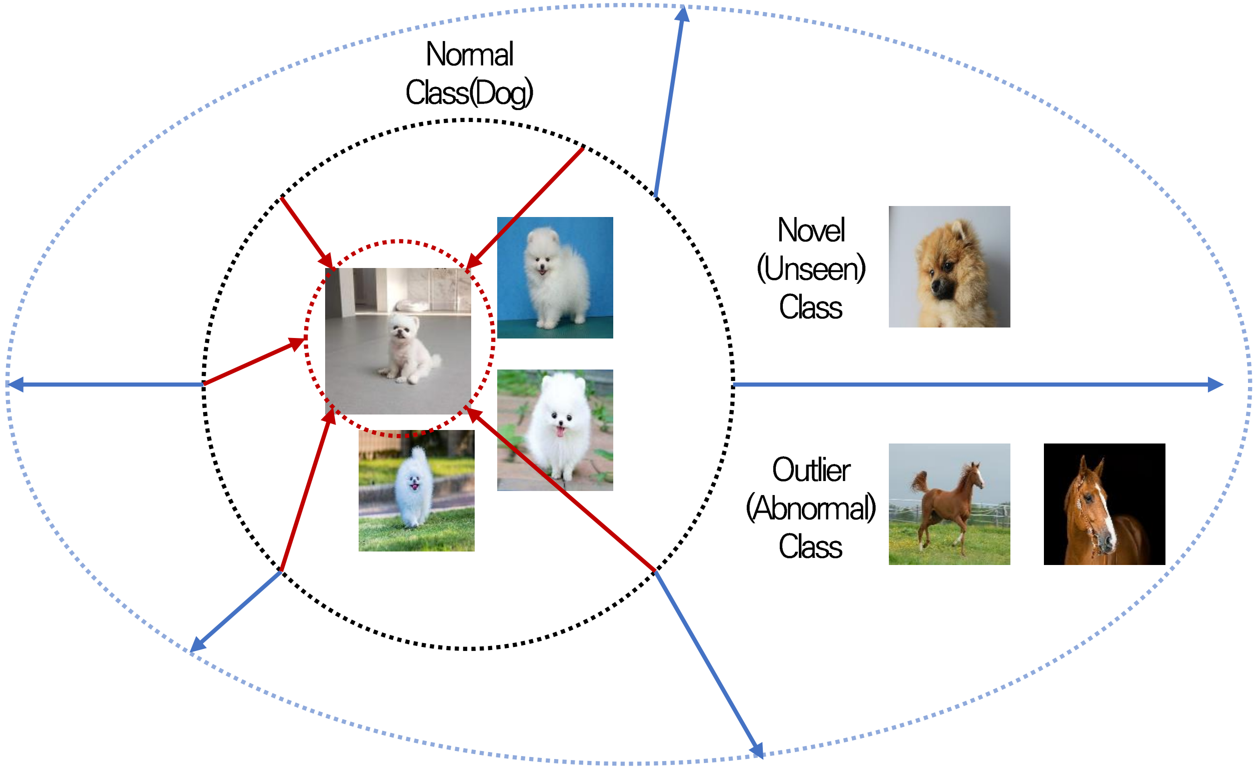
용어	정상 sample	비정상 sample
Supervised Anomaly Detection	학습 사용	학습 사용
Semi-supervised (one-class) Anormaly Detection	학습 사용	학습에 사용 안함
Unsupervised Anomaly Detection	Label이 없어서 모름 학습에 사용하는 데이터의 대다수가 정상 sample일 것이라고 가정함	

Anomaly Detection

일반화(Generalization)와 특수화(Specialization)

일반화: 주어진 데이터로부터 정상 범주의 개념을 확장해 가는 것
특수화: 주어진 데이터로부터 정상 범주의 개념을 좁혀 가는 것

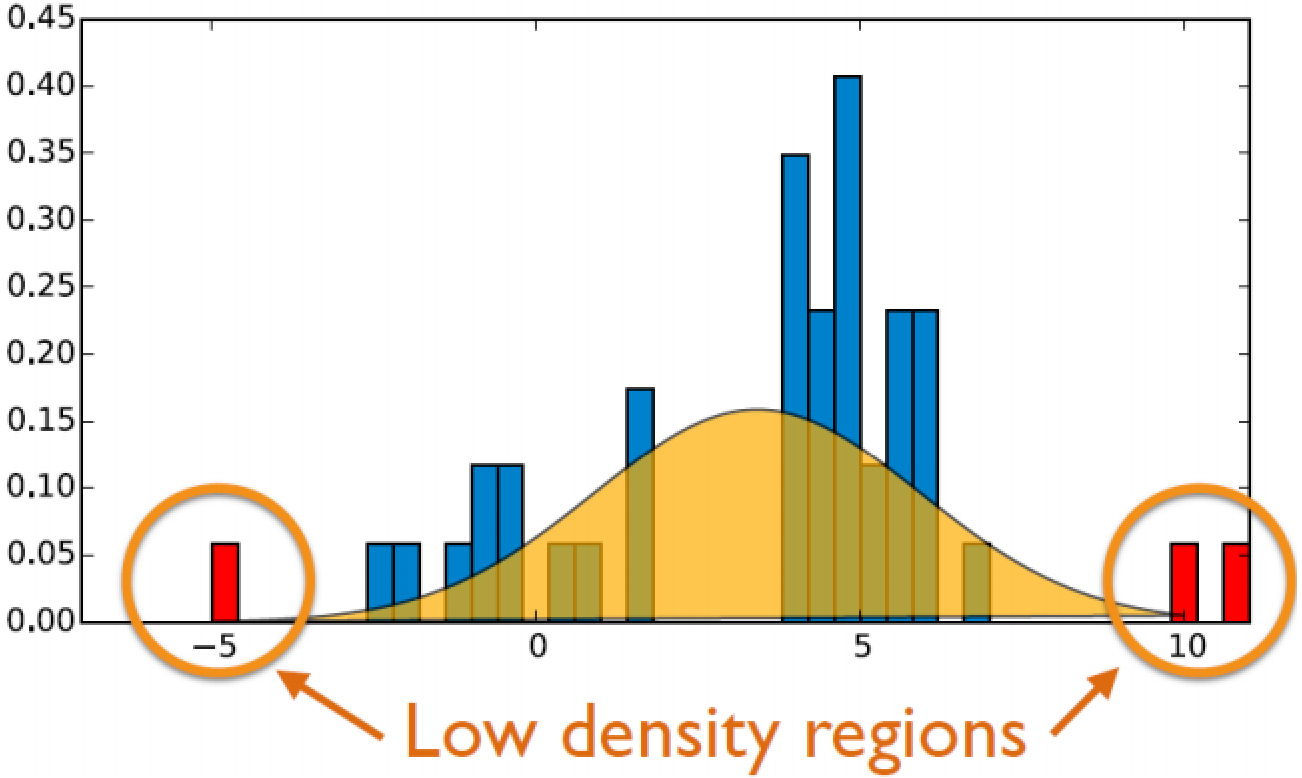
일반화에 치중할 경우 이상치 데이터 판별이 어렵게 되며, 특수화에 치중할 경우 과적합 위험(잦은 false alarm)에 빠질 수 있음



Density-based Anomaly Detection

❖ Density-based anomaly detection이란?

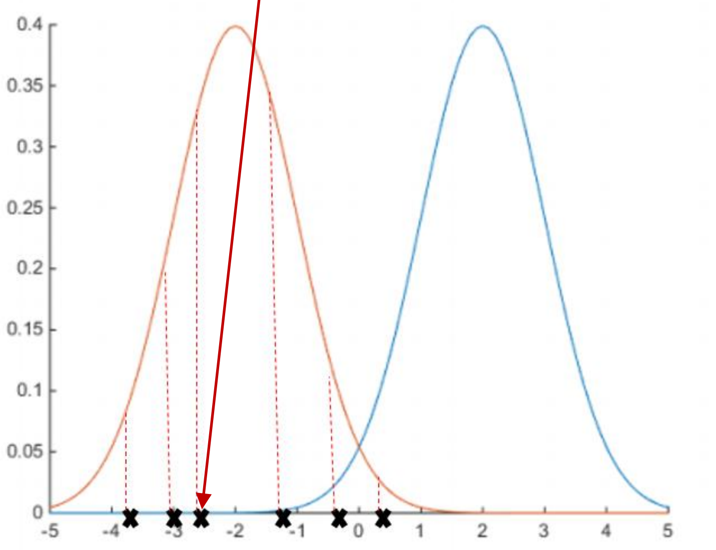
- 주어진 데이터를 바탕으로 각 객체들이 생성될 확률을 추정한 후, 추정된 확률이 낮을 경우 이상치로 판단
- 이를 통해서 기존 데이터 중에서도 이상치를 탐지할 수 있고 새로운 데이터가 들어와도 이상치인지 판단할 수 있음

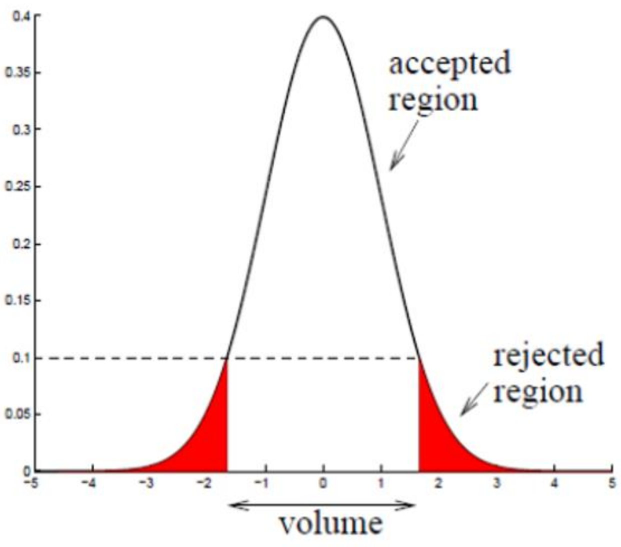


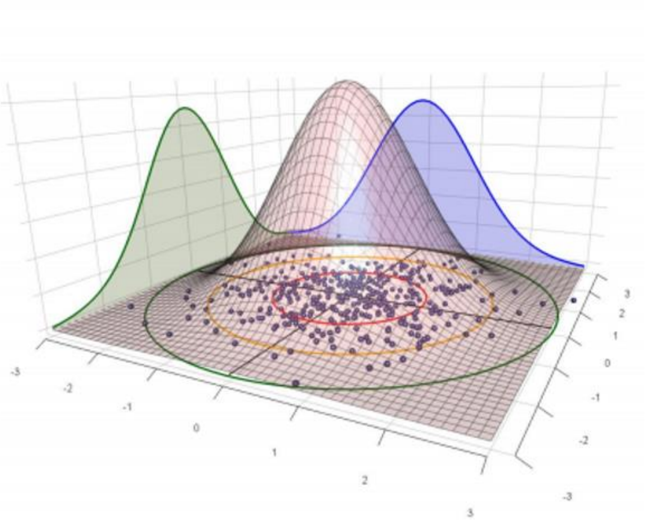
Density-based Anomaly Detection

❖ Gaussian density estimation

- 가우시안 밀도 추정 방법은 데이터가 하나의 정규 분포를 따른다고 가정하는 방법론임
- 주어진 sample들을 정규분포 식에 대입했을 때 가장 최대가 되는 분포가 가장 알맞은 정규분포가 됨
  - ✓ Log-likelihood가 convex 형태이므로 1차 도함수로 최적값을 찾을 수 있음



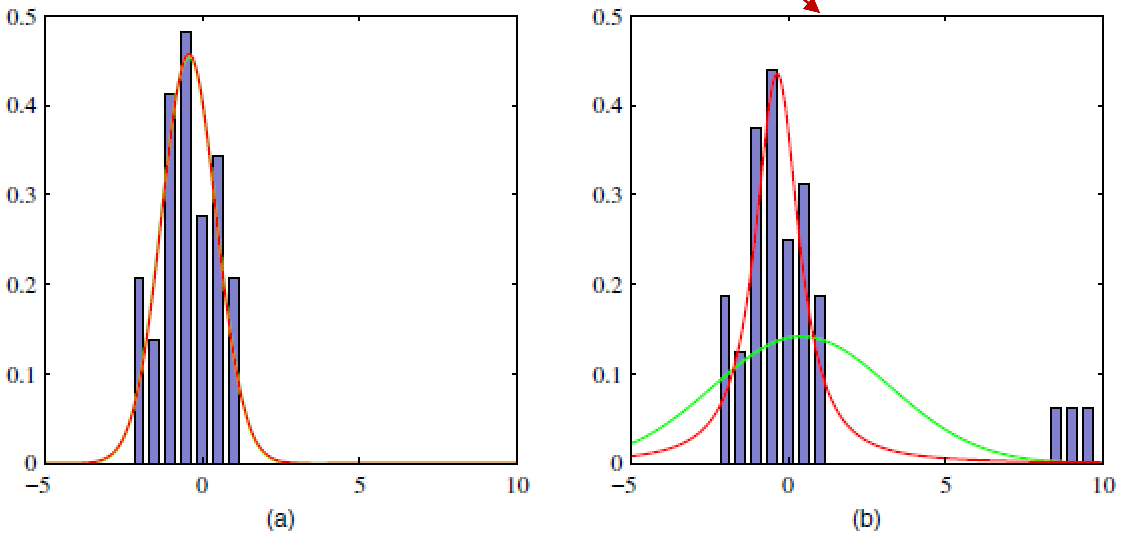




Density-based Anomaly Detection

❖ Gaussian density estimation

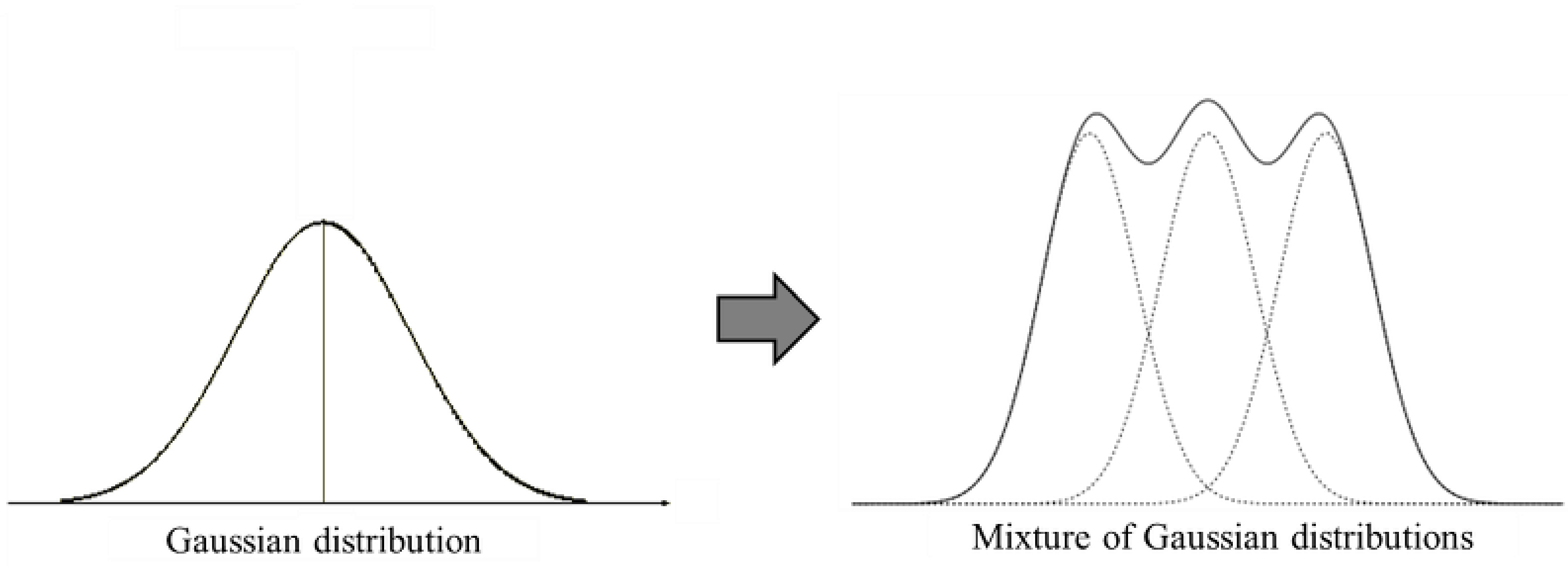
- 대부분의 데이터셋은 단일 가우시안 분포로 표현할 수 없음
- 단일 가우시안 밀도 추정으로 표현하기 어려운 데이터는 어떻게 처리를 해야될까?



Density-based Anomaly Detection

❖ Gaussian mixture model(GMM)

- Gaussian 분포가 여러 개 혼합된 clustering 알고리즘
- 이 알고리즘의 아이디어는 현실에 존재하는 복잡한 확률 분포의 형태를 k개의 gaussian 분포를 혼합하여 표현하는 것임



여러 gaussian distribution의 혼합 분포 예시

Density-based Anomaly Detection

❖ Gaussian mixture model(GMM)

- 주어진 데이터 x에 대해 GMM은 x가 발생할 확률을 식 (1)과 같이 여러 가우시안 확률 분포 함수 합으로 표현함
- 이 알고리즘의 아이디어는 현실에 존재하는 복잡한 확률 분포의 형태를 k개의 gaussian 분포를 혼합하여 표현하는 것임

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \qquad (1)$$

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$0 \le \pi_k \le 1 \qquad (2)$$

$$\sum_{k=1}^{K} \pi_k = 1 \qquad (3)$$

- (1)식은 k개의 가우시안 밀도의 superposition 표현식
- (2), (3)식은 mixing coefficient이며 각각의 가우시안 밀도의 비율을 의미
- 아래 파라미터를 적절히 추정하는 것을 통해 학습 진행

$$\theta = \{\pi, \mu, \Sigma\}$$

$$\pi \equiv \{\pi_1, \ldots, \pi_K\}, \mu \equiv \{\mu_1, \ldots, \mu_K\}\ \Sigma \equiv \{\Sigma_1, \ldots \Sigma_K\}.$$

Density-based Anomaly Detection

$$\theta = \{\pi, \mu, \Sigma\}$$

$$\pi \equiv \{\pi_1, \ldots, \pi_K\}, \mu \equiv \{\mu_1, \ldots, \mu_K\} \ \Sigma \equiv \{\Sigma_1, \ldots \Sigma_K\}.$$

❖ Gaussian mixture model(GMM)

- E step : 초기 파라미터를 고정해주고 responsibility를 구함
- M step : 현재 responsibility를 앞서 구한 세가지 파라미터의 해(MLE solution)에 대입하여 업데이트 함

EM for Gaussian Mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.
2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{9.23}$$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{9.24}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \tag{9.25}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{9.26}$$

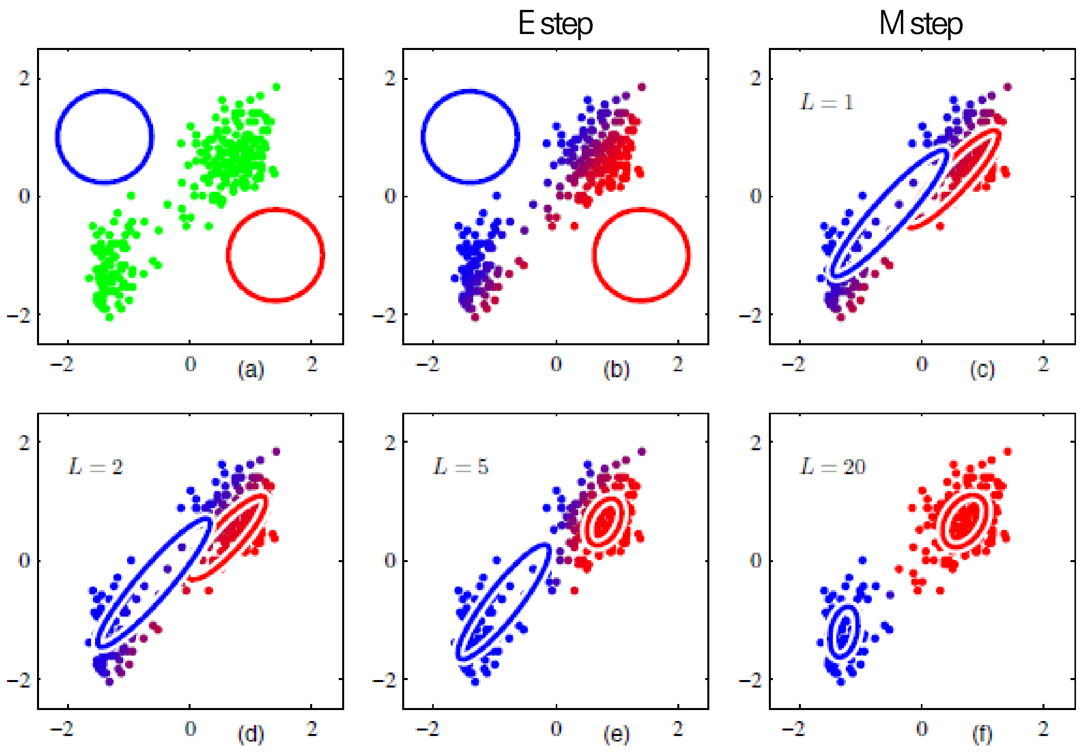
where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{9.27}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{9.28}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

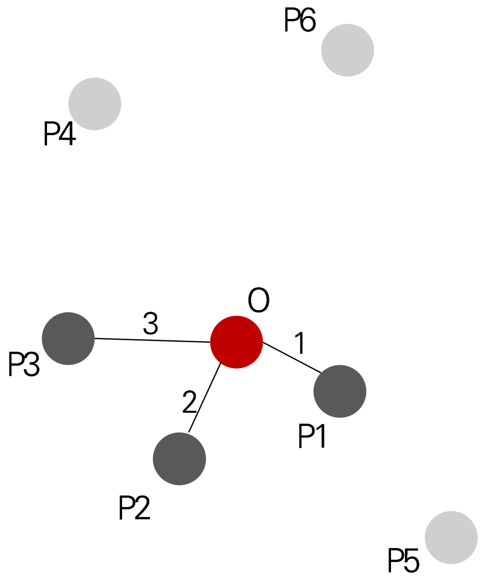


K=2일때 20번 반복한 EM알고리즘 적용 예시

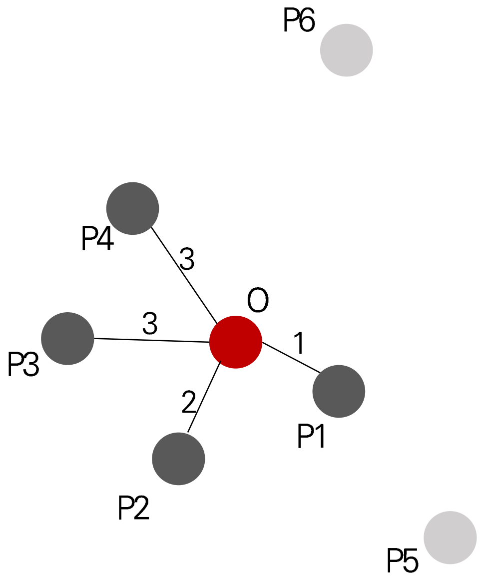
Density-based Anomaly Detection

❖ LOF(Local Outlier Factor)

- 각각의 관측치가 데이터 안에서 얼마나 벗어나 있는가에 대한 이상치 정도를 나타냄
- 가장 중요한 특징은 모든 데이터를 전체적으로 고려하는 것이 아닌, 국소적(local) 관점으로 주변 데이터(neighbor)를 이용하여 이상치를 파악하는 것



예시 1



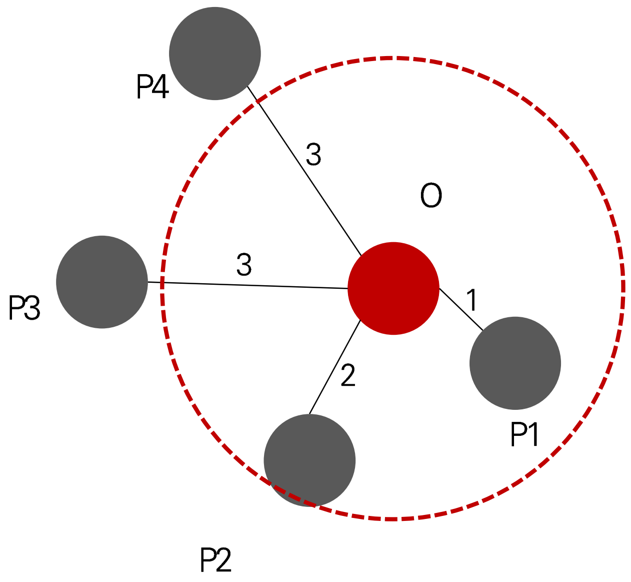
예시 2

	3-distance(O)	N_3(O) 개수
예시 1	2	3
예시 2	2.25	4

Density-based Anomaly Detection

❖ LOF(Local Outlier Factor)

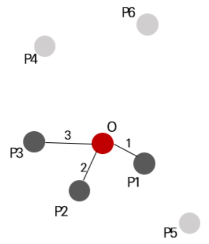
- Reachability distance : $reach-dist_k(p,o) = \max\{k-distance(o), d(o,p)\}$
  - ✓ 관측치 p가 o에서 멀다면 관측치 p와 o의 실제 유클리디안 거리
  - ✓ 관측치 p가 o에서 가깝다면 관측치 o의 k-distance



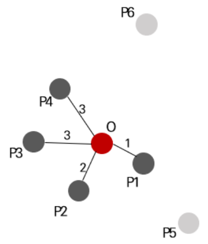
	P_1	P_2	P_3	P_4
$d(O,P_i)$	1	2	3	3
$reach-dist_3(P_i,O)$	2.25	2.25	3	3

$3-distance(o) = 2.25$

Density-based Anomaly Detection



예시1

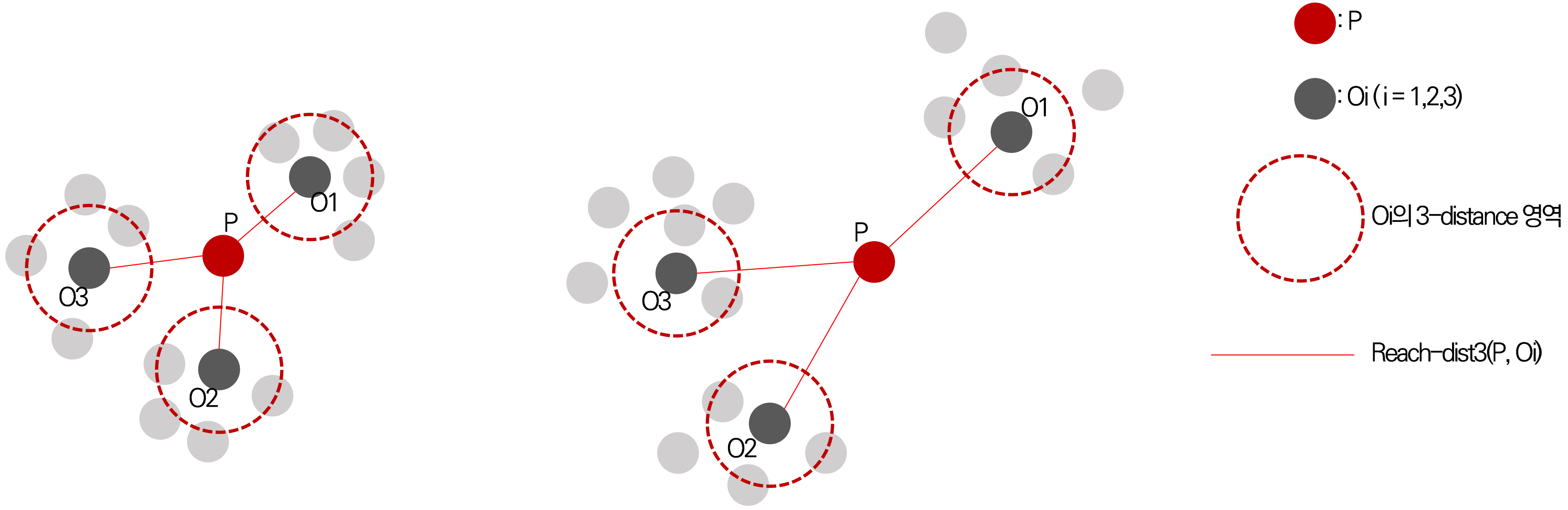


예시2

	3-distance(O)	N_3(O) 개수
예시1	2	3
예시2	2.25	4

❖ LOF(Local Outlier Factor)

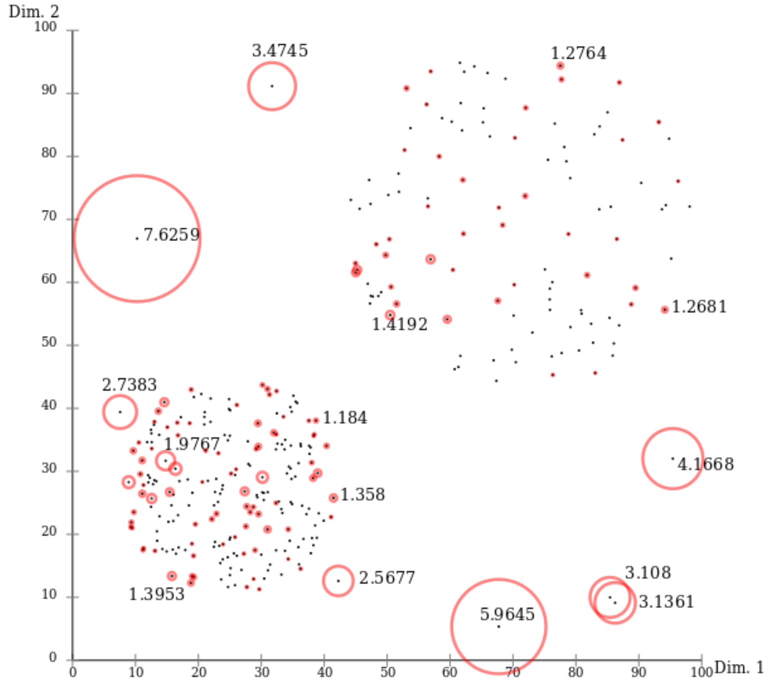
- Object p에 대한 local reachability density(lrd)를 산출
- 관측치 p주변에 이웃이 얼마나 밀도 있게 있는가를 대변함
- $lrd_k(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} reach\text{-}dist_k(p,o)}$



Density-based Anomaly Detection

❖ LOF(Local Outlier Factor)

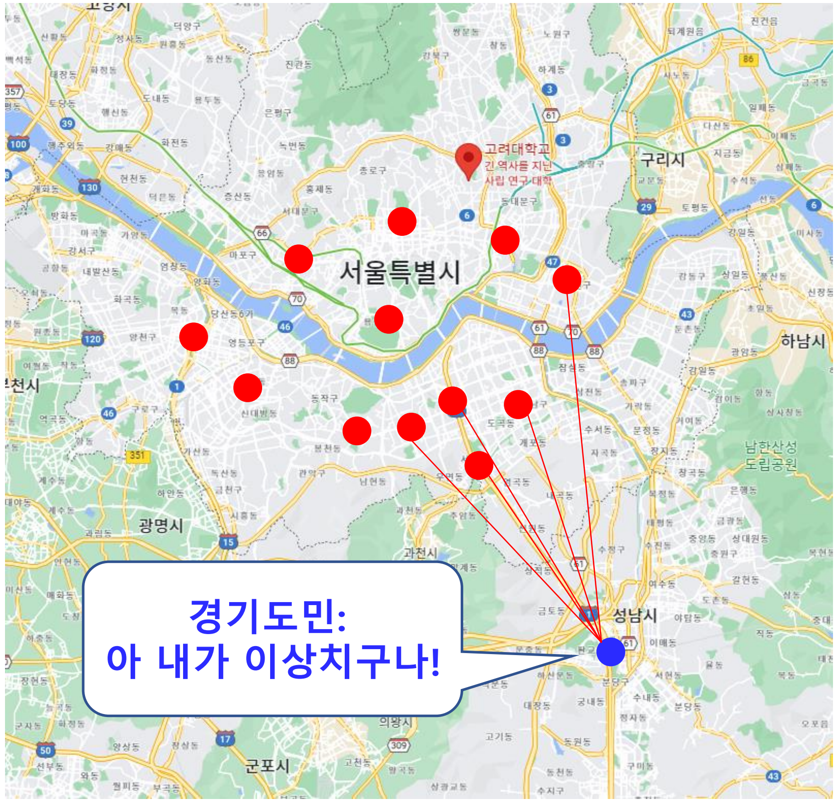
- 최종적으로 관측치 p의 이상치 정도를 나타내는 LOF는 아래 수식과 같이 정의됨
- 관측치 p의 밀도($lrd_k(p)$)와 이웃 o의 밀도 ($lrd_k(o)$)의 비율을 평균한 것
- $\mathrm{LOF}_k(p) = \dfrac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} = \dfrac{\frac{1}{lrd_k(p)} \sum_{o \in N_k(p)} lrd_k(o)}{|N_k(p)|}$



- LOF < 1 : 밀도가 높은 분포
- LOF ≈ 1 : 이웃 관측치와 비슷한 분포
- LOF > 1 : 밀도가 낮은 분포, 크면 클수록 이상치 정도가 큼

Distance-based Anomaly Detection

❖ k-Nearest Neighbor-based Anomaly Detection

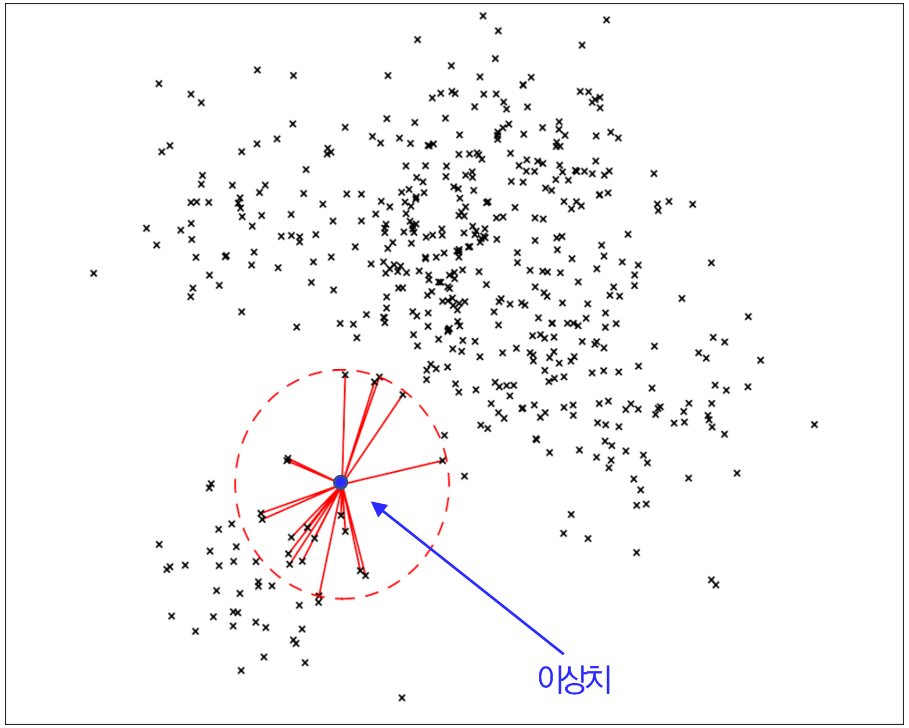


Kang, Pilsung, and Sungzoon Cho. "A hybrid novelty score and its use in keystroke dynamics-based user authentication." *Pattern recognition* 42.11 (2009): 3115-3127.

Distance-based Anomaly Detection

❖ k-Nearest Neighbor-based Anomaly Detection

- Supervised learning 기반의 k-nearest neighbors에서 발전된 알고리즘
- 정상 데이터에 대해서 어떤 사전 분포(prior probability)도 가정하지 않음



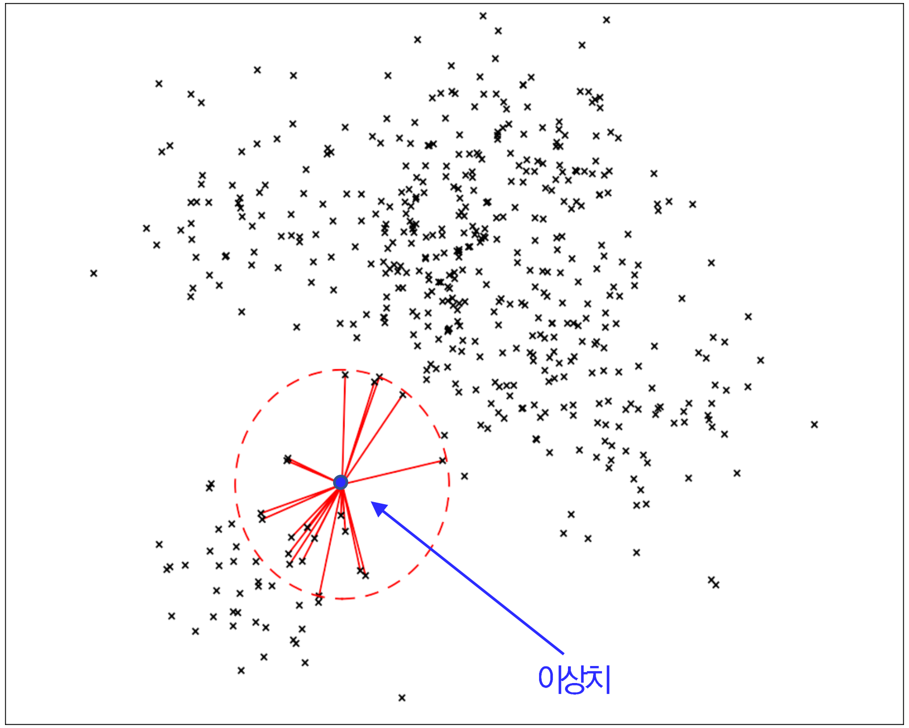
Kang, Pilsung, and Sungzoon Cho. "A hybrid novelty score and its use in keystroke dynamics-based user authentication." *Pattern recognition* 42.11 (2009): 3115-3127.

Distance-based Anomaly Detection

❖ k-Nearest Neighbor-based Anomaly Detection

- Supervised learning 기반의 k-nearest neighbors에서 발전된 알고리즘
- 정상 데이터에 대해서 어떤 사전 분포(prior probability)도 가정하지 않음!

거리 (Distance)



Kang, Pilsung, and Sungzoon Cho. "A hybrid novelty score and its use in keystroke dynamics-based user authentication." *Pattern recognition* 42.11 (2009): 3115-3127.

Distance-based Anomaly Detection

❖ 그렇다면 거리는 어떻게 측정?

1. Maximum distance to the k-th nearest neighbor (주변에서 가장 먼 이웃과의 거리)

$$d_{max}^{k} = \kappa(x) = \|x - z_k(x)\|$$

2. Average distance to the k-th nearest neighbor (거리의 평균)

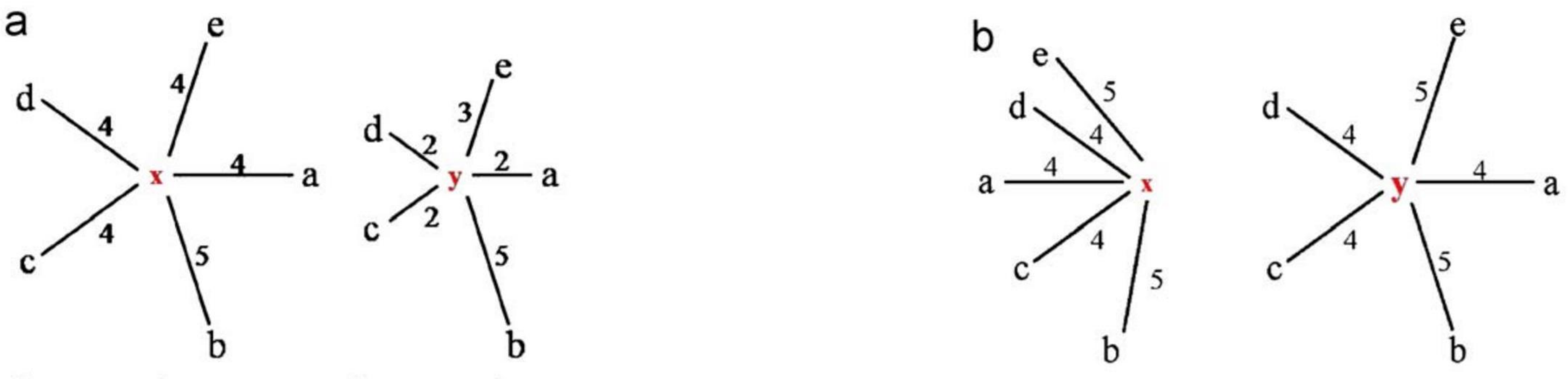
$$d_{avg}^{k} = \gamma(x) = \frac{1}{k}\sum_{j=1}^{k}\|x - z_j(x)\|$$

3. Distance to the mean of the k-nearest neighbor (이웃들 간 중심을 구해서 거리 구하기)

$$d_{mean}^{k} = \delta(x) = \left\| x - \frac{1}{k}\sum_{j=1}^{k} z_j(x) \right\|$$

Distance-based Anomaly Detection

❖ 그렇다면 거리는 어떻게 측정?



A (max vs average)

d^k_{max}	5.0	5.0
d^k_{avg}	4.2	2.8

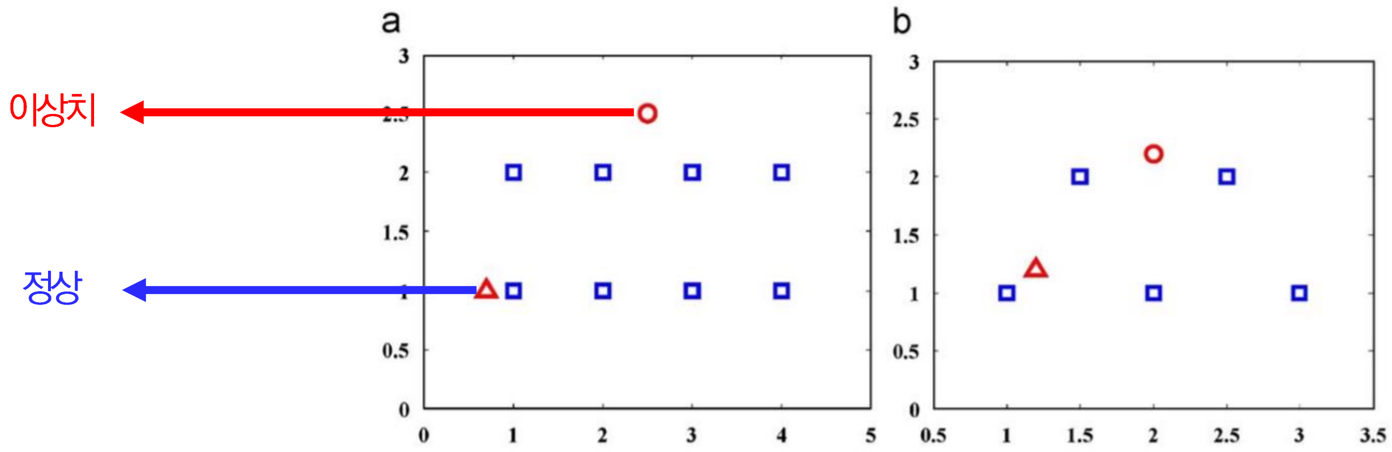
B (mean vs average)

d^k_{mean}	3.3	2.1
d^k_{avg}	4.4	4.4

Kang, Pilsung, and Sungzoon Cho. "A hybrid novelty score and its use in keystroke dynamics-based user authentication." *Pattern recognition* 42.11 (2009): 3115-3127.

Distance-based Anomaly Detection

❖ 그렇다면 거리는 어떻게 측정?



		d^k_{max}	d^k_{avg}	d^k_{mean}
A (k=4)	Circle	1.58	1.14	0.50
	Triangle	1.64	1.07	0.94
B (k=5)	Circle	1.56	1.08	0.80
	Triangle	1.86	1.09	0.88

Wrong!!!

Kang, Pilsung, and Sungzoon Cho. "A hybrid novelty score and its use in keystroke dynamics-based user authentication." *Pattern recognition* 42.11 (2009): 3115-3127.

Distance-based Anomaly Detection

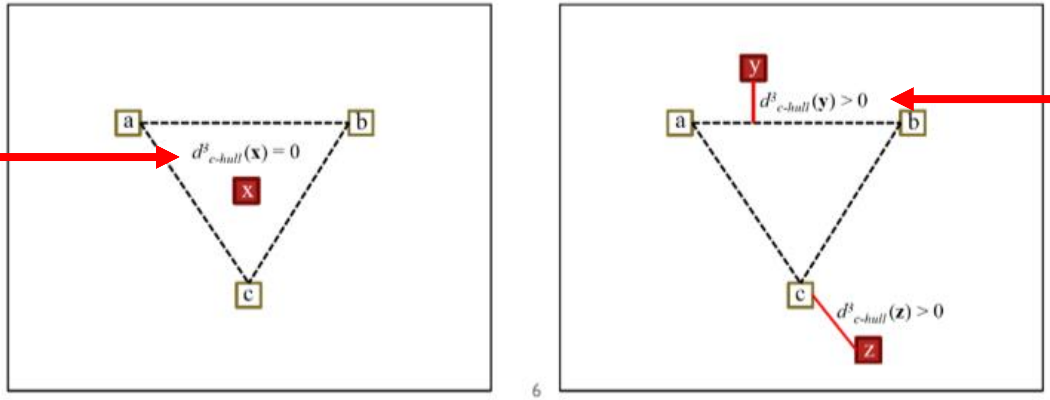
❖ Consider additional factor

- 한 가지 항을 추가 → 이웃들의 convex hull 까지의 거리를 고려
- Convex hull: 이웃들끼리 연결했을 때 그 안에 있으면 거리가 0, 밖에 있으면 0 이상

$$\min_{\mathbf{w}} \left(d^k_{c-hull}(\mathbf{x})\right)^2 = \left|\left|\mathbf{x}_{new} - \sum_{j=1}^{k} \mathbf{w}_i z_j(\mathbf{x})\right|\right|^2$$

$$s.t. \sum_{i=1}^{k} \mathbf{w}_i = 1, \quad \mathbf{w}_i \geq 0, \ \forall i.$$

안에 있으면 0



밖에 있으면 0 이상

Kang, Pilsung, and Sungzoon Cho. "A hybrid novelty score and its use in keystroke dynamics-based user authentication." *Pattern recognition* 42.11 (2009): 3115-3127.

Distance-based Anomaly Detection

❖ Hybrid Distance (Convex hull)

1. Average distance to the k-th nearest neighbor (거리의 평균)

$$d_{avg}^{k} = \frac{1}{k}\sum_{j=1}^{k} ||\mathbf{x} - z_j(\mathbf{x})||$$

2. Convex distance to its k-nearest neighbors

$$d_{c-hull}^{k} = \left|\left|\mathbf{x} - \sum_{j=1}^{k} \mathbf{w}_i z_j(\mathbf{x})\right|\right|$$

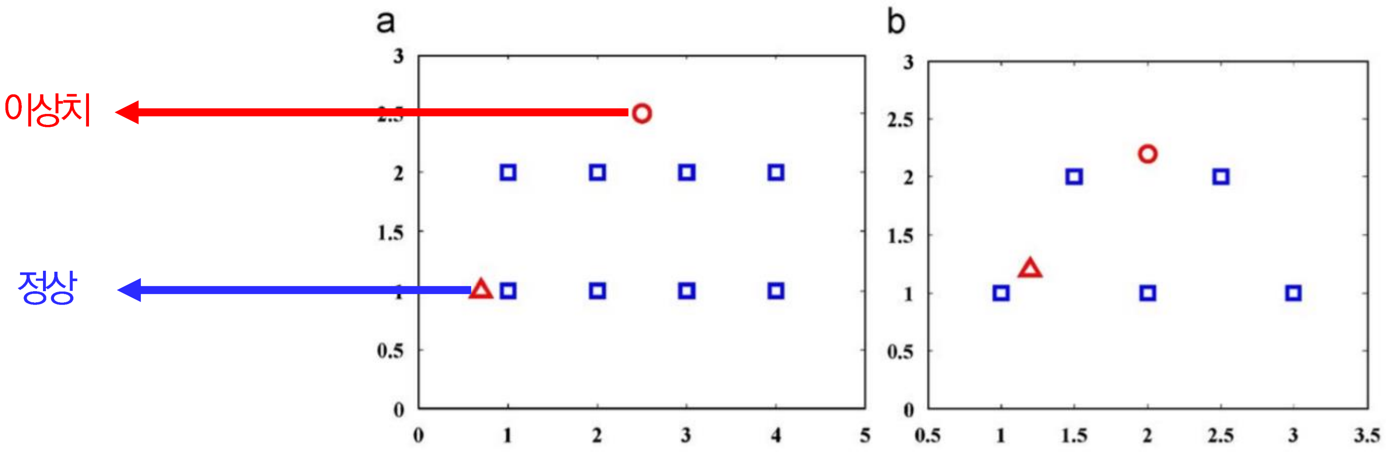
3. Hybrid Distance

$$d_{hybrid}^{k} = d_{avg}^{k} \times \left(\frac{2}{1 + exp(-d_{c-hull}^{k})}\right)$$

Kang, Pilsung, and Sungzoon Cho. "A hybrid novelty score and its use in keystroke dynamics-based user authentication." *Pattern recognition* 42.11 (2009): 3115-3127.

Distance-based Anomaly Detection

❖ 그렇다면 거리는 어떻게 측정?



		d^k_{max}	d^k_{avg}	d^k_{mean}	d^k_{hybrid}
A (k=4)	Circle	1.58	1.14	0.50	1.42
	Triangle	1.64	1.07	0.94	1.18
B (k=5)	Circle	1.56	1.08	0.80	1.18
	Triangle	1.86	1.09	0.88	1.09

Kang, Pilsung, and Sungzoon Cho. "A hybrid novelty score and its use in keystroke dynamics-based user authentication." *Pattern recognition* 42.11 (2009): 3115-3127.

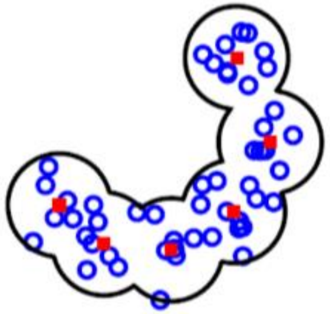
Distance-based Anomaly Detection

❖ Clustering-based Approach

- 가장 근처 군집의 centroid와의 거리를 계산
- 정상 데이터에 대해서 어떤 사전 분포(prior probability)도 가정하지 않음

$$\mathcal{X} = C_1 \cup C_2 \ldots \cup C_K, \quad C_i \cap C_j = \phi, \quad i \neq j.$$

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in C_i} ||\mathbf{x}_j - \mathbf{c}_i||^2$$

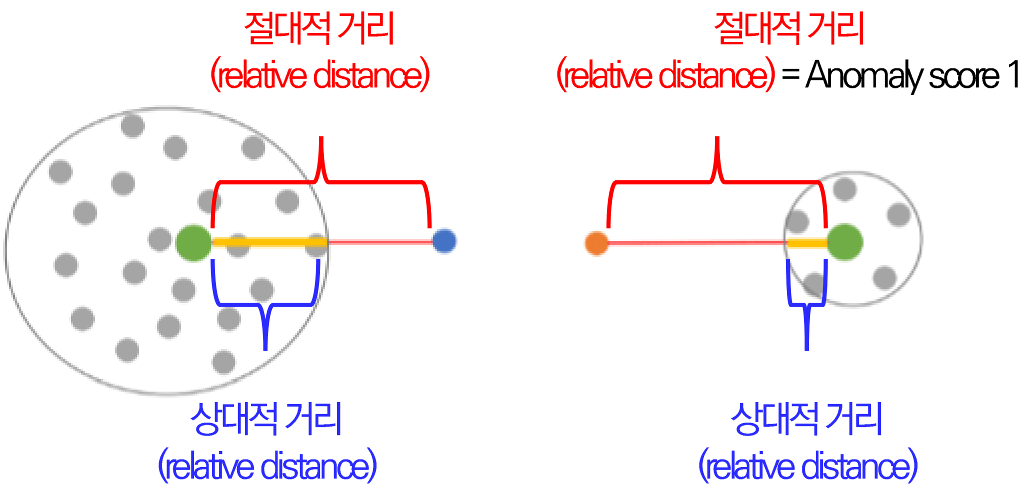


1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Distance-based Anomaly Detection

❖ Clustering-based Approach

- K-means clustering의 두 가지 anomaly score(distance) 존재
- 정상 데이터에 대해서 어떤 사전 분포(prior probability)도 가정하지 않음

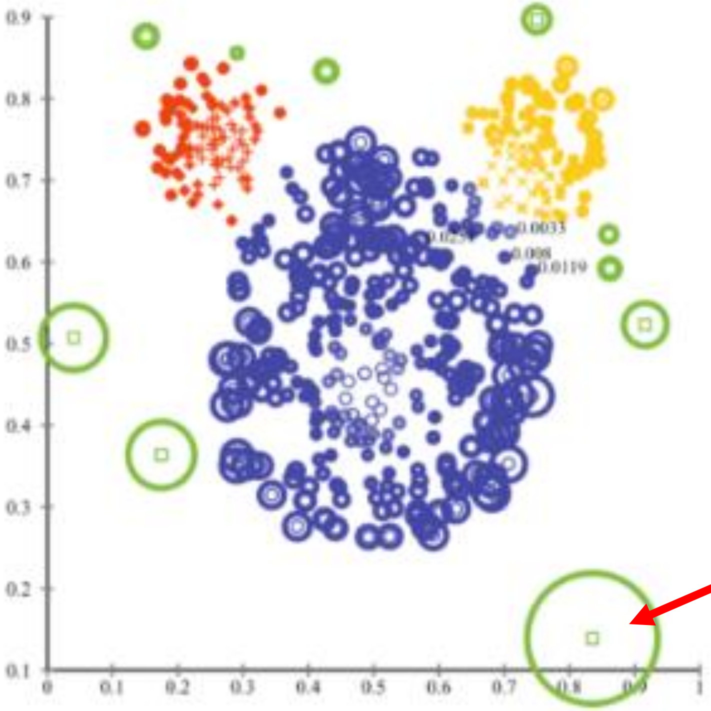


$$\frac{\text{절대적 거리 (relative distance)}}{\text{상대적 거리 (relative distance)}} = \text{Anomaly score 2}$$

Distance-based Anomaly Detection

❖ Clustering-based Approach

- K-means clustering의 두 가지 anomaly score(distance) 존재
- 정상 데이터에 대해서 어떤 사전 분포(prior probability)도 가정하지 않음

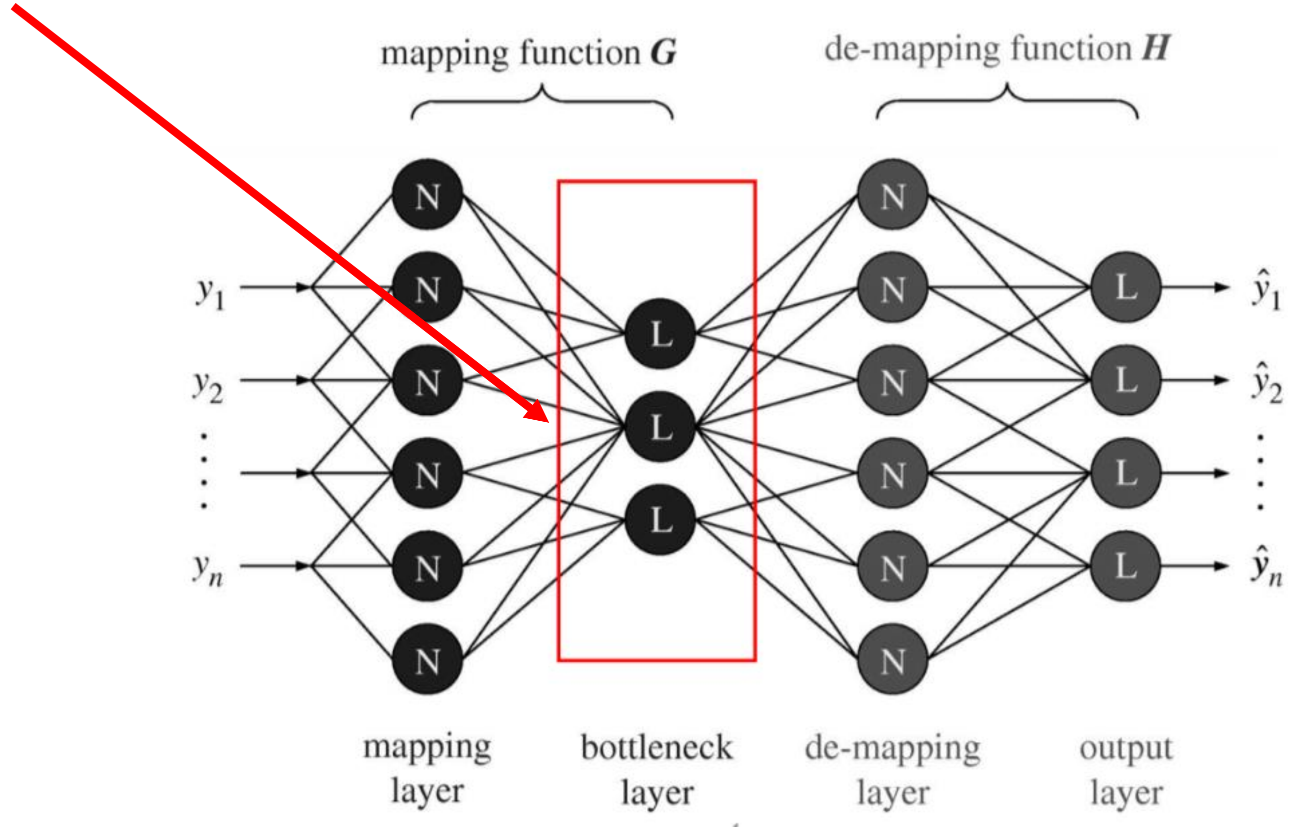


군집에서 멀리 떨어질수록
Anomaly score up!

Model-based Anomaly Detection

❖ Auto-Encoder for Anomaly Detection

- 반드시 입력 변수의 수보다 은닉 노드의 수가 더 적은 은닉 층이 있어야 함!
- 은닉 층에서 정보의 축약이 이루어짐

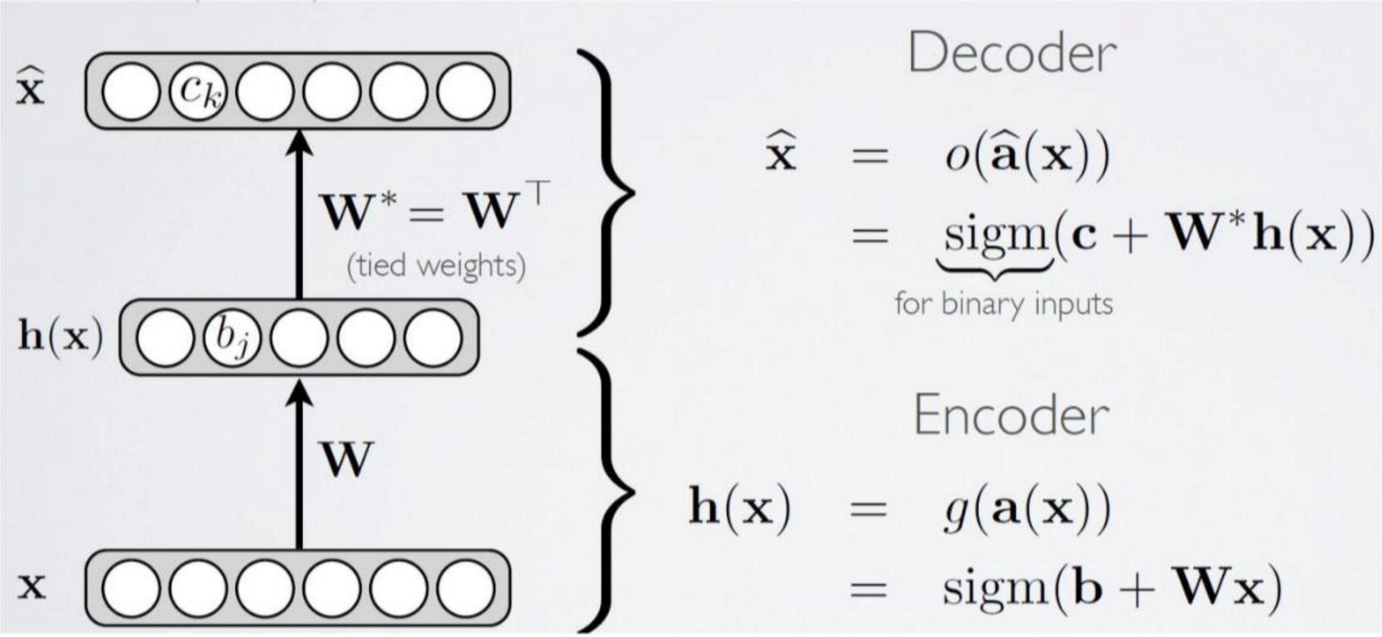


Model-based Anomaly Detection

- Auto-Encoder for Anomaly Detection
  - Auto-Encoder: 입력과 출력이 동일한 인공 신경망 구조

Loss = Anomaly Score ←

$$l(f(\mathbf{x})) = \frac{1}{2}\sum_k (\widehat{x}_k - x_k)^2$$

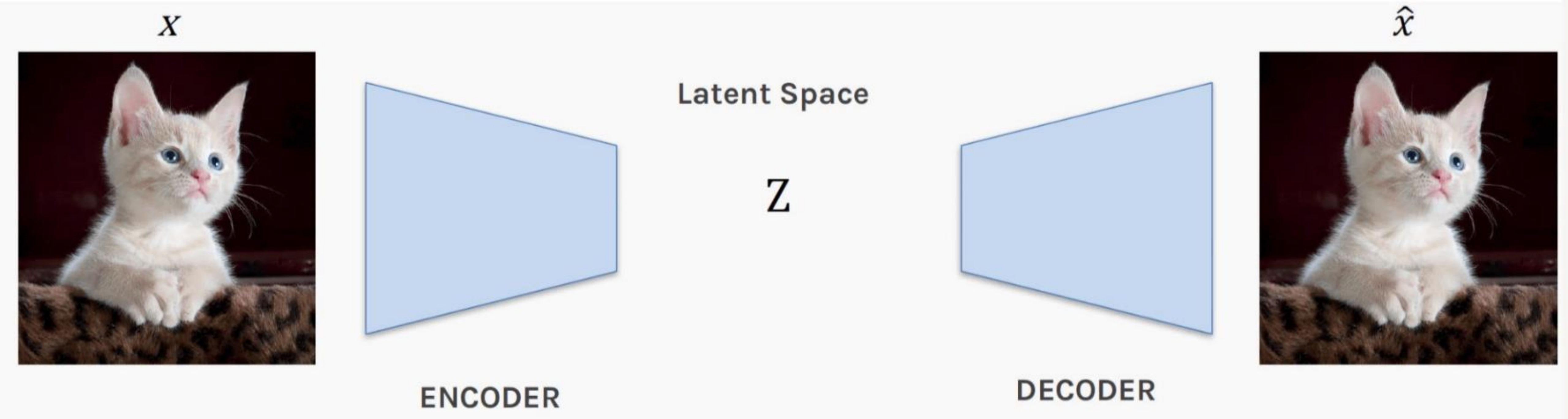


Model-based Anomaly Detection

❖ Auto-Encoder for Anomaly Detection

- 정상데이터들에 대한 학습이 충분히 되어 있으면

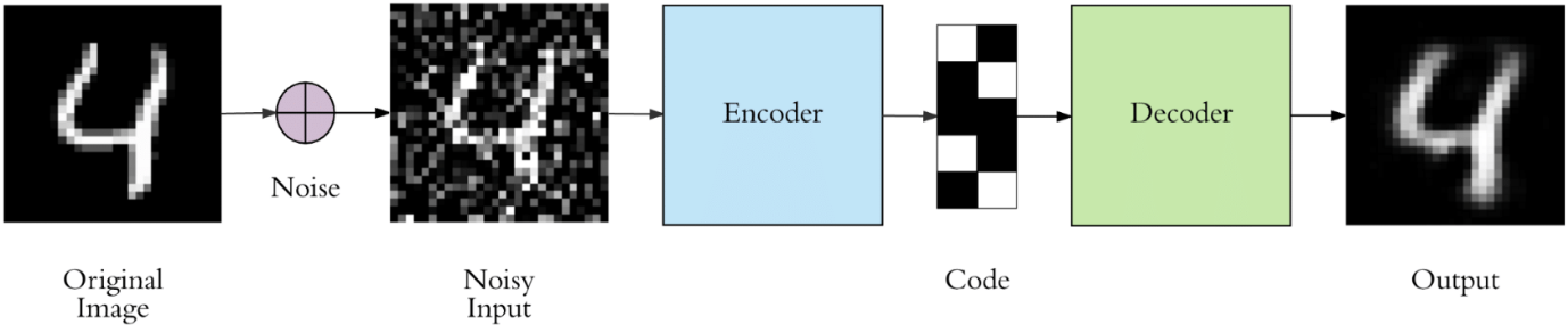
1. 정상 데이터는 자기 자신을 잘 복제할 수 있는 신경망이 됨
2. 이상치 데이터는 학습 기회가 적어서 상대적으로 복제를 잘 못하는 것을 가정함!



Model-based Anomaly Detection

❖ Auto-Encoder for Anomaly Detection

- Auto-Encoder를 포함한 인공신경망의 단점 → 입력에 대한 약간의 변형(small perturbations)에도 모델이 민감하게 반응!
- 학습 과정에서 입력에 일부러 noise를 첨가하자!



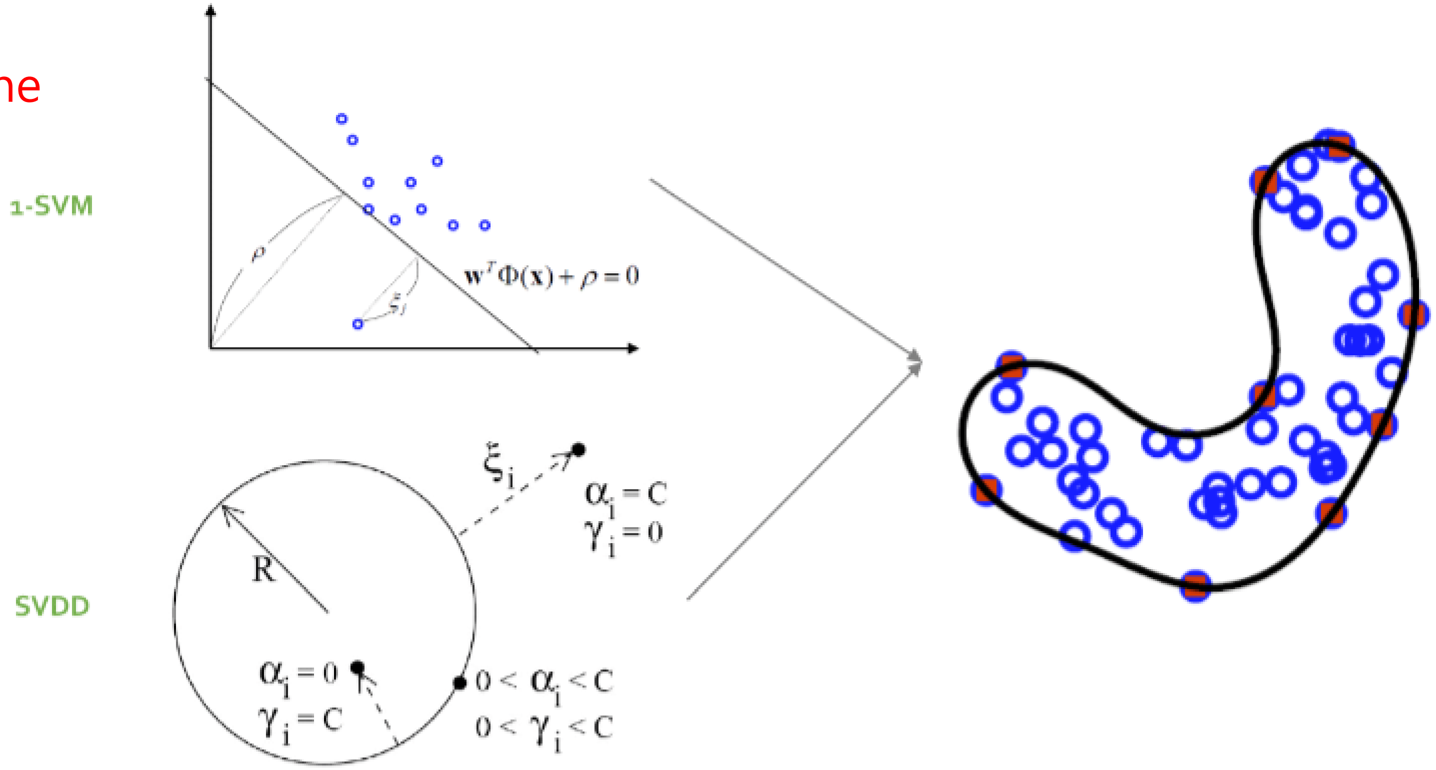
Model-based Anomaly Detection

❖ Support Vector-based Novelty Detection

- 정상과 이상치를 구별하는 boundary를 찾는 것!

One class
support vector machine

Support vector data
description



Model-based Anomaly Detection

❖ Support Vector-based Novelty Detection

- One-Class Support Vector Machine

원점에서부터 최대한 멀리 떨어진

hyperplane과의 거리

▪ Optimization problem

Hyperparameter

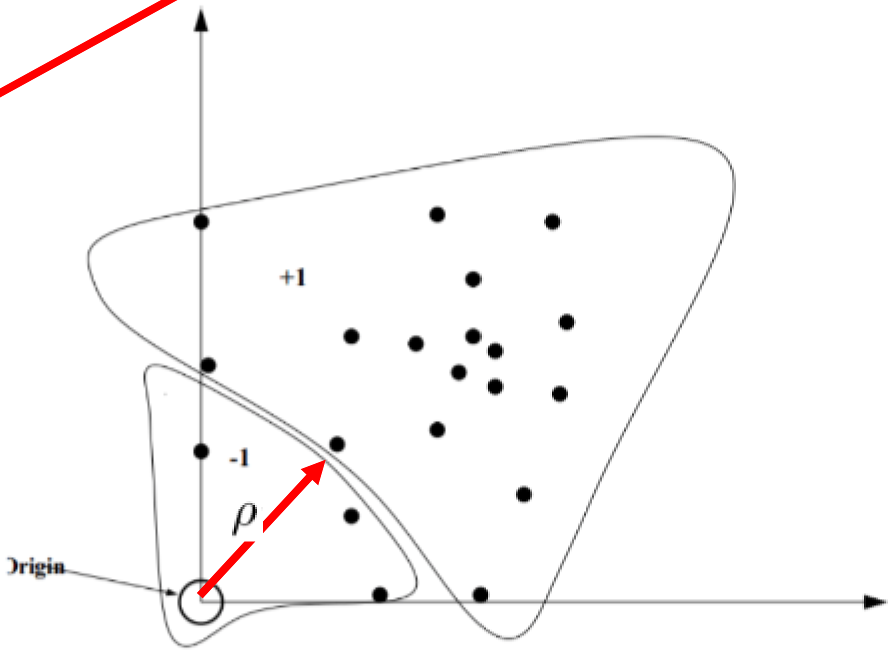
$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i - \rho$$

$$s.t. \quad \mathbf{w}\cdot\Phi(\mathbf{x}_i) \geq \rho - \xi_i$$

$$i = 1, 2, \cdots, l, \quad \xi_i \geq 0$$

▪ Decision function

$$f(\mathbf{x}_i) = sign\big(\mathbf{w}\cdot\Phi(\mathbf{x}_i) - \rho\big)$$



Model-based Anomaly Detection

❖ Support Vector-based Novelty Detection

- One-Class Support Vector Machine

원점에서부터 최대한 멀리 떨어진

hyperplane과의 거리

V: Hyperparameter

l: 정상 데이터 수

- Optimization problem

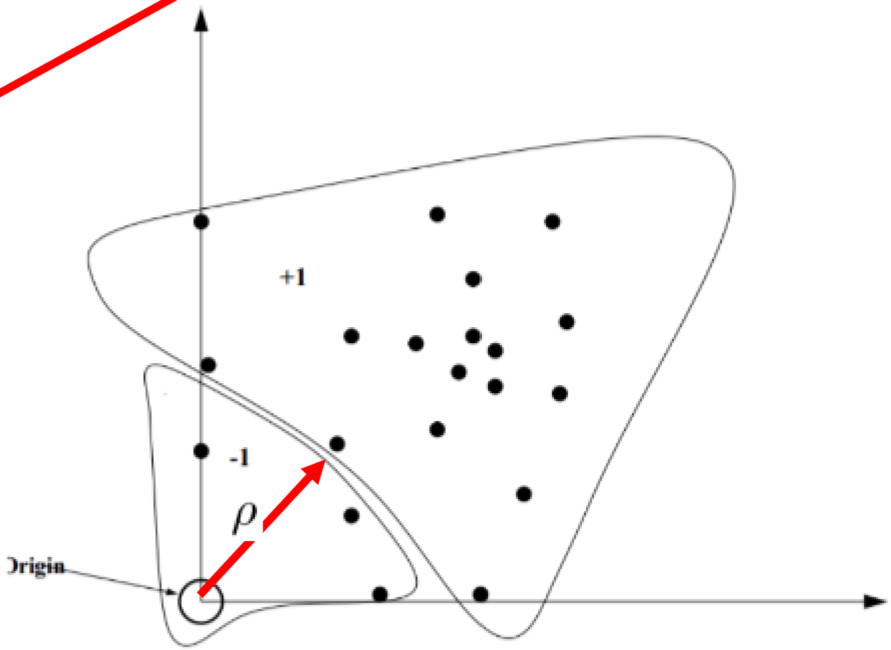
$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i - \rho$$

$$s.t. \quad \mathbf{w}\cdot\Phi(\mathbf{x}_i) \geq \rho - \xi_i$$

$$i = 1, 2, \cdots, l, \quad \xi_i \geq 0$$

- Decision function

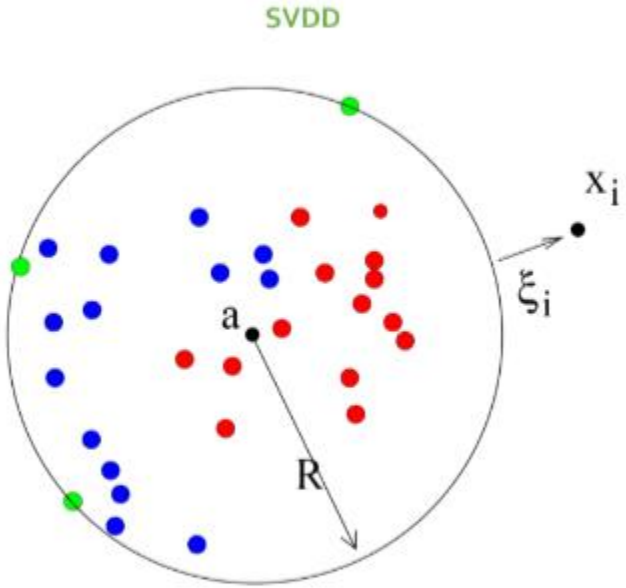
$$f(\mathbf{x}_i) = sign\big(\mathbf{w}\cdot\Phi(\mathbf{x}_i) - \rho\big)$$

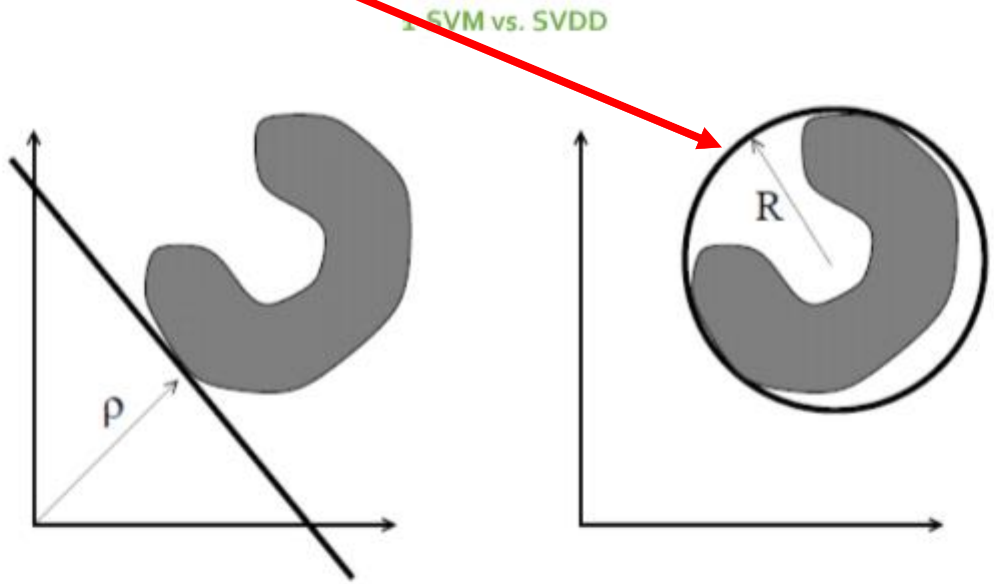


Model-based Anomaly Detection

❖ Support Vector Data Description (SVDD)

- 정상 데이터를 모두 포함하는 가장 작은 초구를 찾고 싶어함!





Model-based Anomaly Detection

❖ Support Vector Data Description (SVDD)

- 정상 데이터를 모두 포함하는 가장 작은 초구를 찾고 싶어함!

▪ Optimization function

$$\min_{R,\mathbf{a},\xi_i} R^2 + C\sum_{i=1}^{l}\xi_i$$

객체와 초구 사이의 거리

$$s.t. \quad ||\Phi(\mathbf{x}_i) - \mathbf{a}||^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad \forall i.$$

▪ Decision function

$$f(\mathbf{x}) = sign\left(R^2 - ||\Phi(\mathbf{x}_i) - \mathbf{a}||^2\right)$$

내부(normal) +

외부 (abnormal) −

Model-based Anomaly Detection

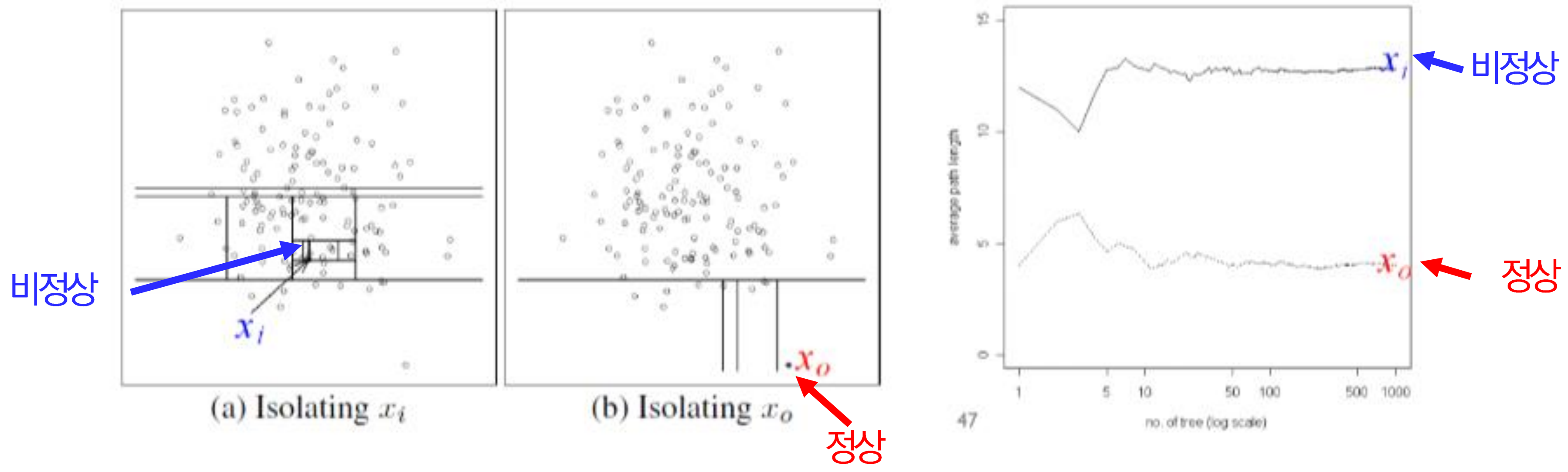
❖ Isolation Forest

- 소수 범주(이상치)는 개체수가 적음
- 소수 범주데이터는 정상 범주 데이터와는 특정 속성 값이 많이 다름!

→ 하나의 객체(이상치)를 고립(isolation)시키는 tree를 생성해보자!

→ 정상 데이터라면 고립시키는데 많은 split

→ 이상치 데이터라면 상대적으로 적은 split만으로 고립 가능

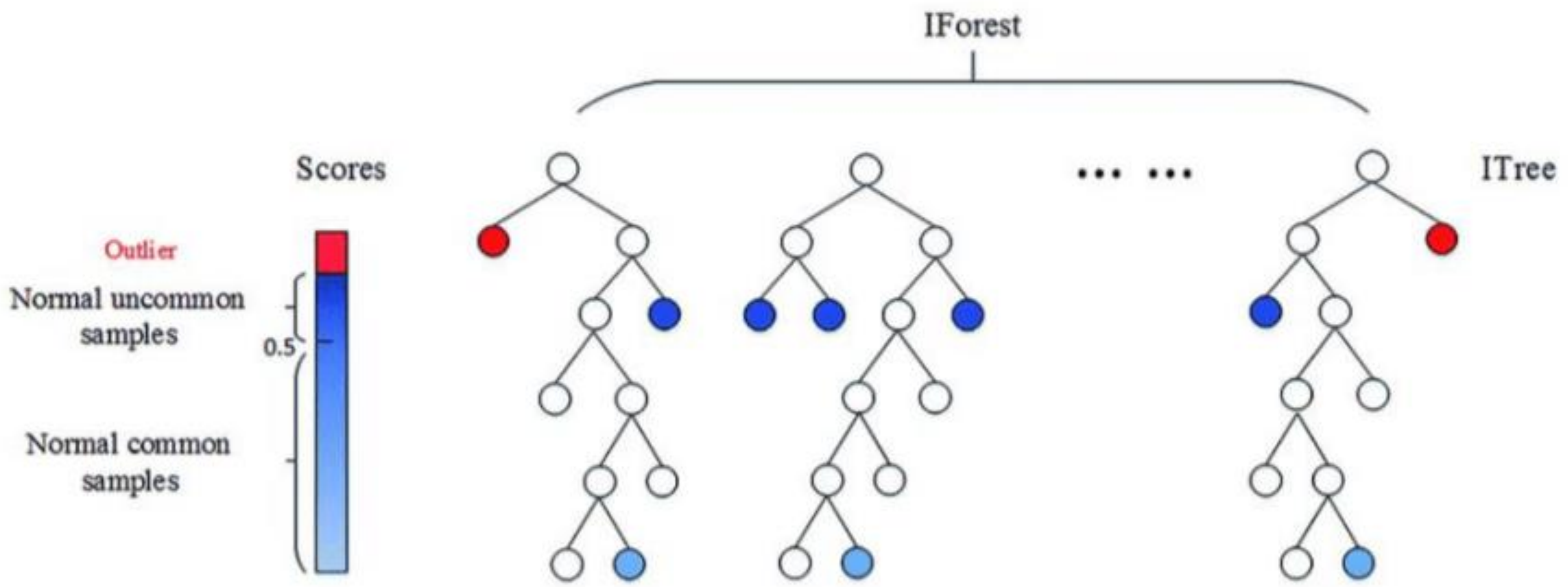


Model-based Anomaly Detection

❖ Isolation Forest

- I-Forest

✓ 객체를 고립시킬 때까지 몇 번이나 분기(split)를 했는지에 대한 정보로 이상치 점수를 부여할 수 있지 않을까?



Model-based Anomaly Detection

❖ Isolation Forest

- Path Length (경로 길이)
  - ✓ 객체 x의 경로 길이 h(x)는 Root node로부터 x가 속한 말단 노드까지 도달하기 위해 거쳐간 edge의 수로 정의됨
    - ▪ h(x)는 평균 기대 path length c(n)을 사용하여 다음과 같이 정규화 가능

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}, \quad H(i) = ln(i) + 0.5772156649 \ (\text{Euler's constant})$$

1개의 tree에 대해 객체를 isolation 시키기 위한 path length

  - ✓ 객체 x의 이상치 스코어 s는 다음과 같이 정의됨

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

평균 path length

- ▪ When $E(h(x)) \rightarrow c(n), \quad s \rightarrow 0.5$
- ▪ When $E(h(x)) \rightarrow 0, \quad s \rightarrow 1$ Split 쉬운 경우 = 정상 = 2^0
- ▪ When $E(h(x)) \rightarrow n-1, \quad s \rightarrow 0$ Split 쉬운 경우 = 비정상 = 2^{-inf}

✓ 즉, Tree에서 path length가 짧을수록 이상치 스코어는 1에 가까워지고, path length가 길수록 이상치 스코어는 0에 가까워짐

Thank you