Nearest Neighbor Based Anomaly Detection

국가수리과학연구소 산업수학혁신센터

최 동 헌 (dhchoe@nims.re.kr)

2020.11.05

- Ph.D. in Mathematics (Topology)
- Post Doctoral Researcher in NIMS
- Research interests
  - Low dimensional topology, Geometric topology
  - Topological data analysis
  - Time series analysis, Anomaly detection

Review

- Anomaly detection?
    - Identification of unobserved pattern in the data
    - Describe outlier as a data point that is dissimilar to other point
- Challenges in Anomaly detection
    - Inaccurate boundaries between outlier and normal behavior
    - Noise in the data which mimics real outlier
    - Labeled data might be hard to obtain
    - Highly imbalanced classification problem
    - Context dependent and so hindering the use of one model for multiple problems

Review

- Outlier detection
    - Polluted training data
    - Unsupervised learning
- Novelty detection
    - Training data consisting only of normal data
    - Semi-supervised learning
- Methods
    - Probabilistic methods, Distance-based , Neighbor-based, Domain Based, Isolation methods, Neural Networks

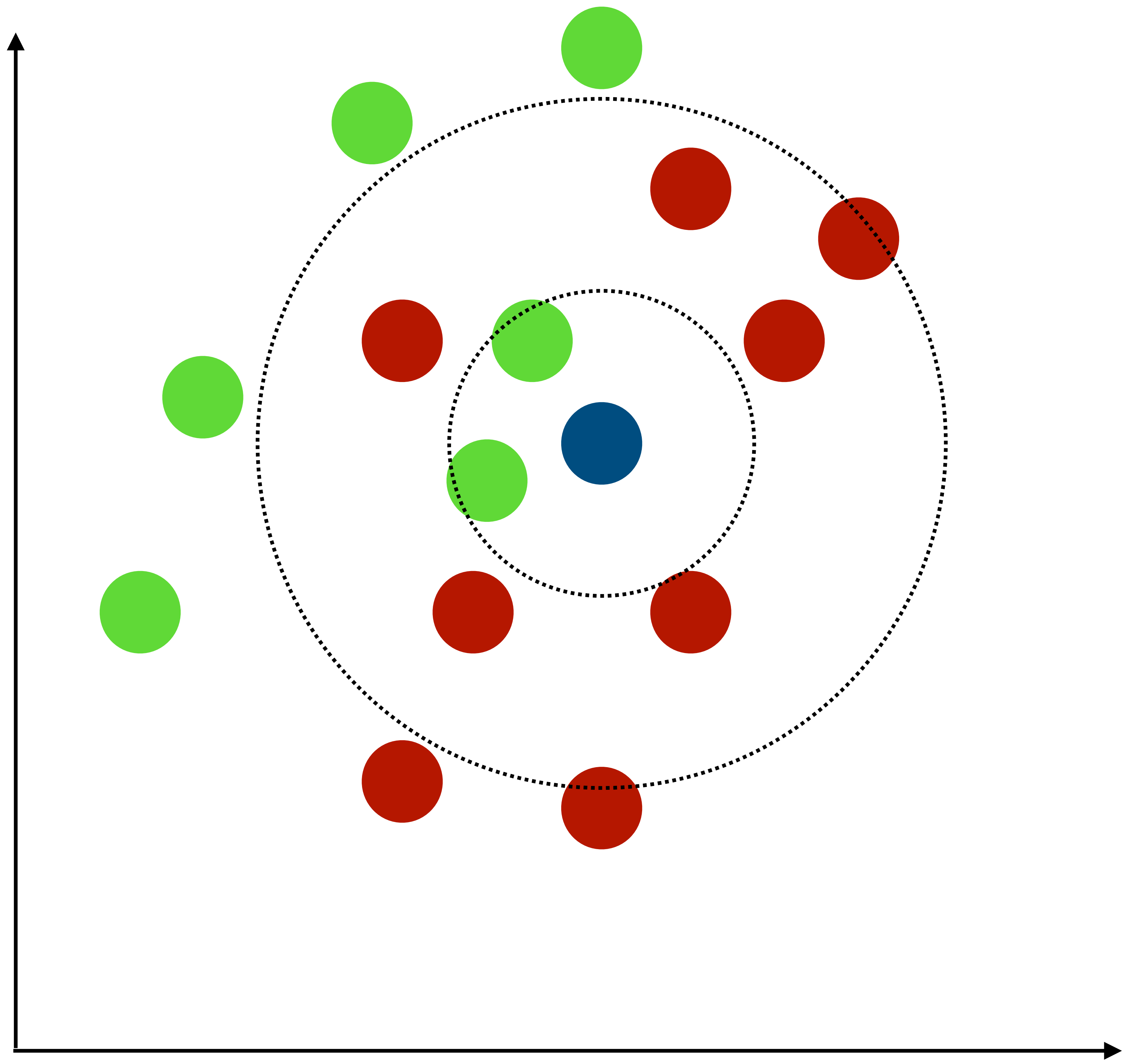
Review

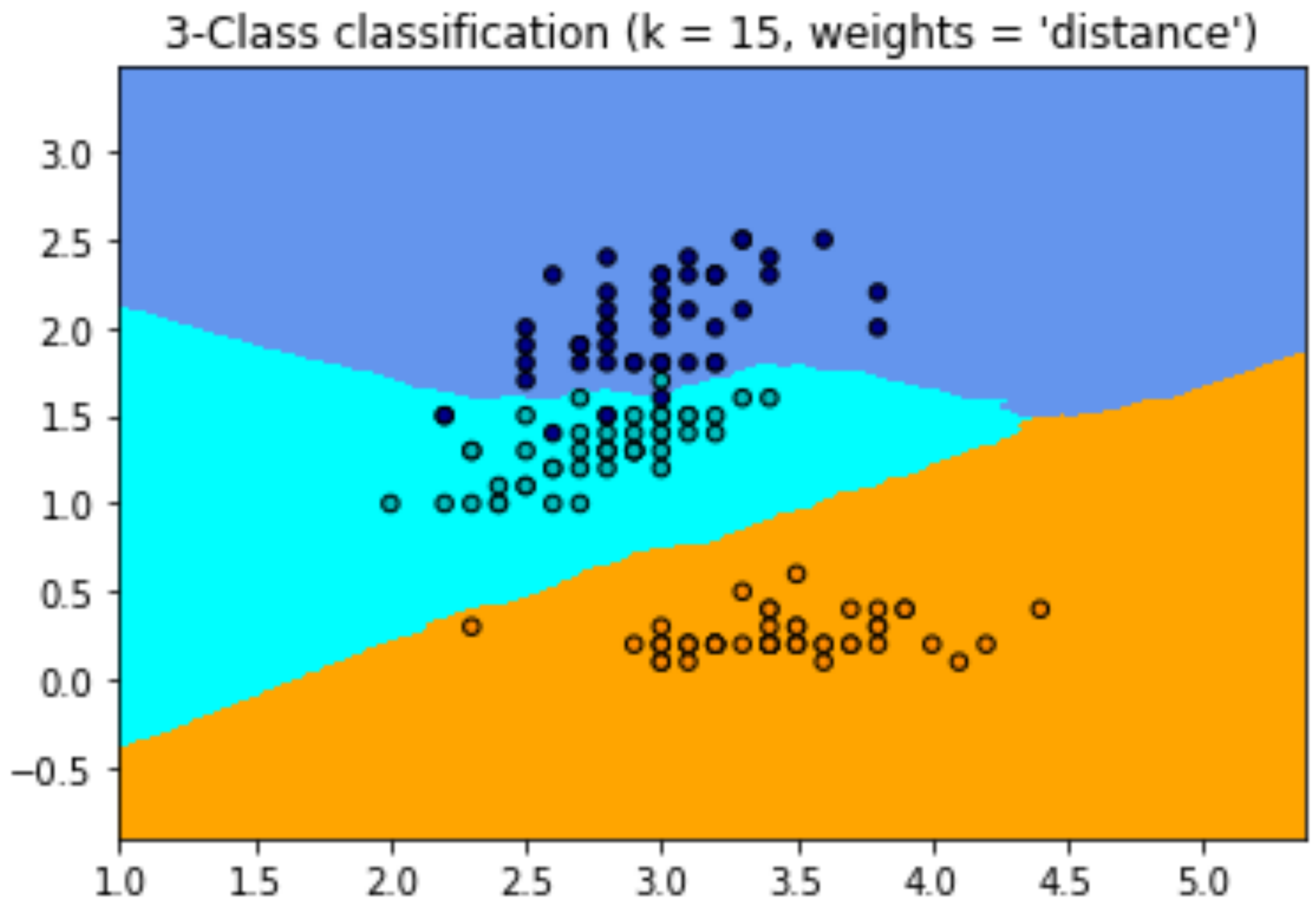
- Classifier performance
    - To evaluate, we need labeled test data
    - Binary classifier, score function
- Evaluation
    - Confusion Matrix, Recall, Precision, AUROC
- Robust Covariance, One Class SVM

k-Nearest Neighbor(kNN) algorithm

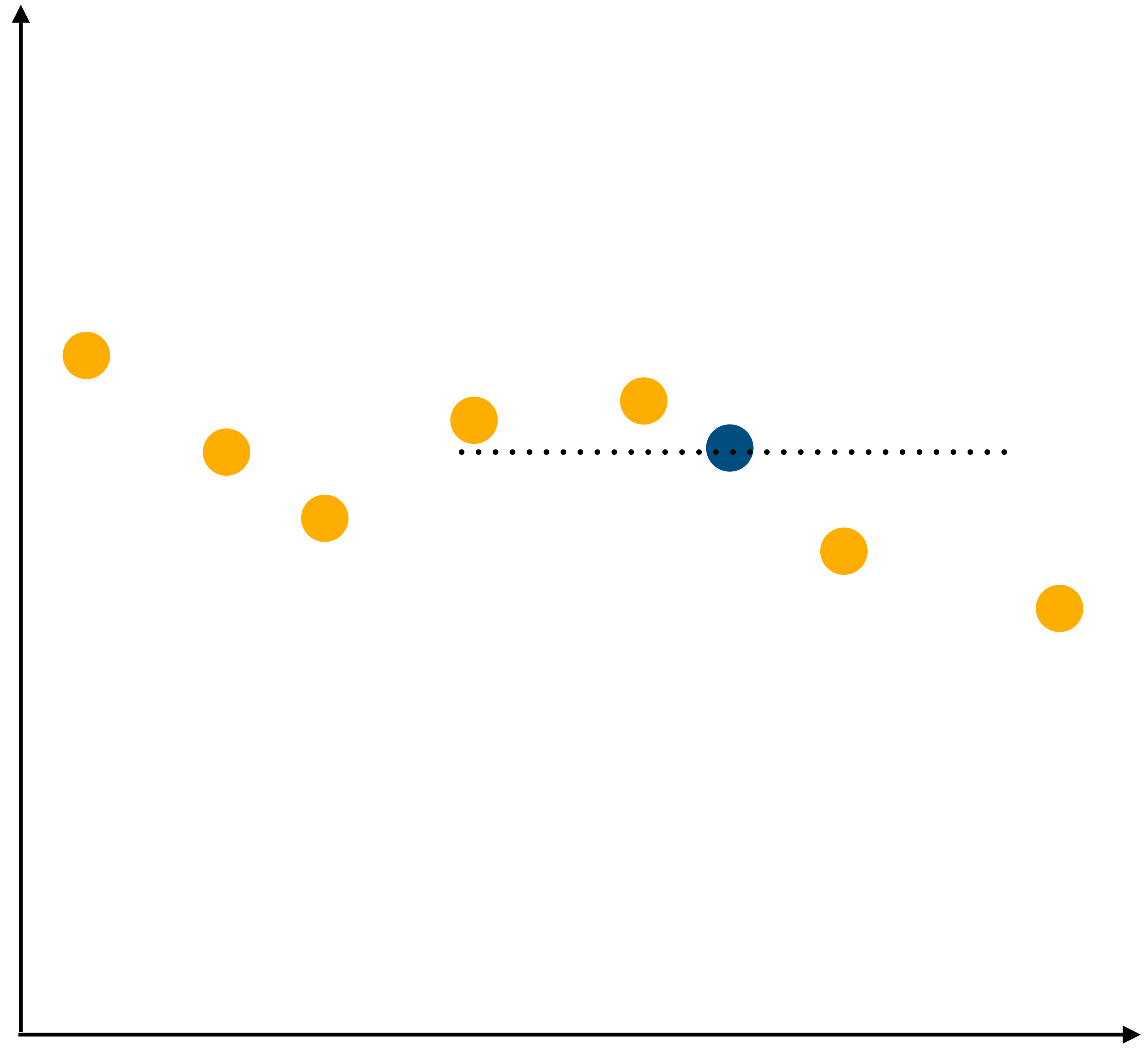
- Simplest machine learning algorithm
- Sensitive for local structure
- Instance based learning (without model generating)
- Hyperparameters : number of neighbors (k), metric
- Optimal k depend on data (effect on decision boundary)

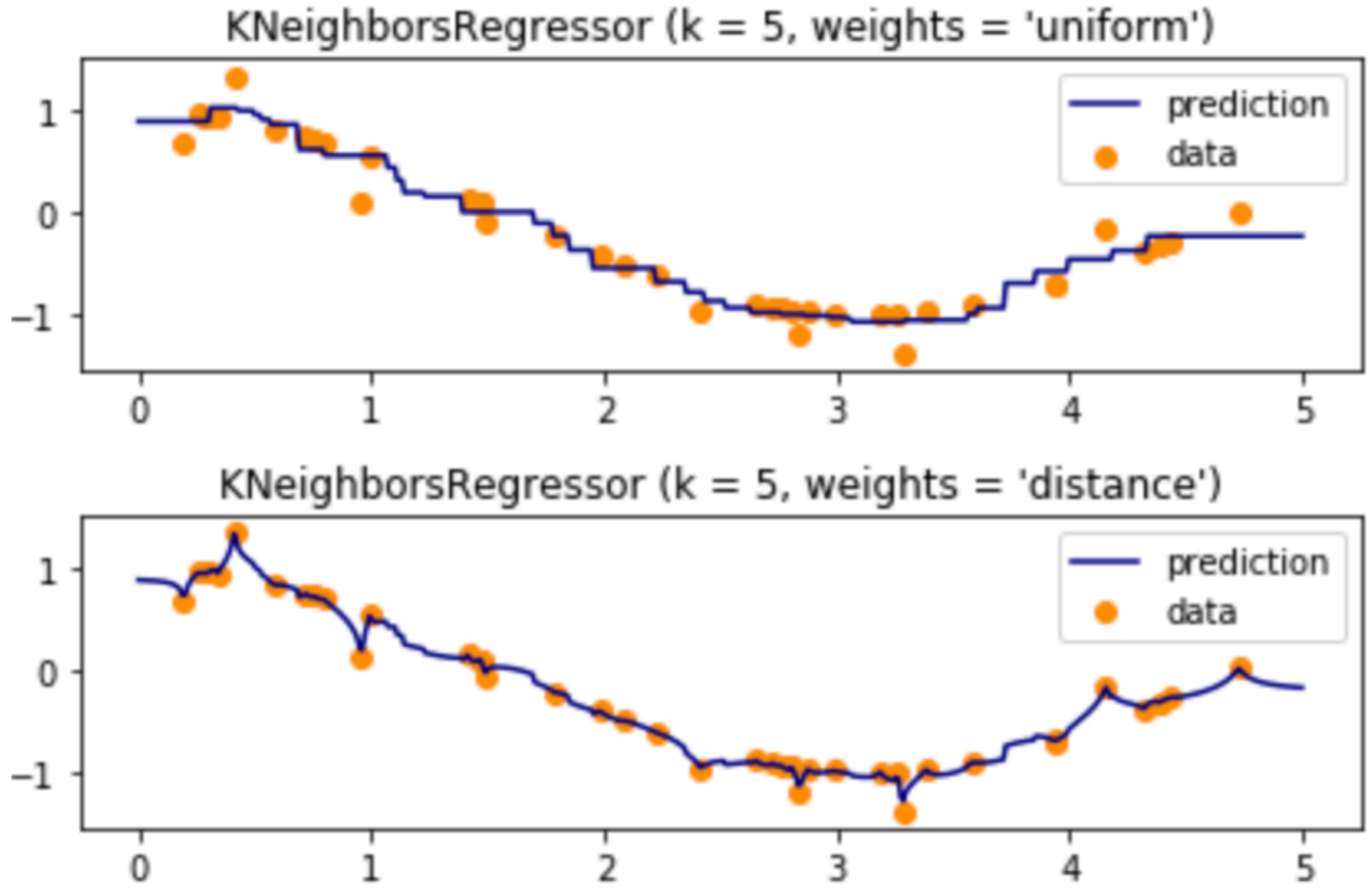
k-NN Classification





k-NN Regression

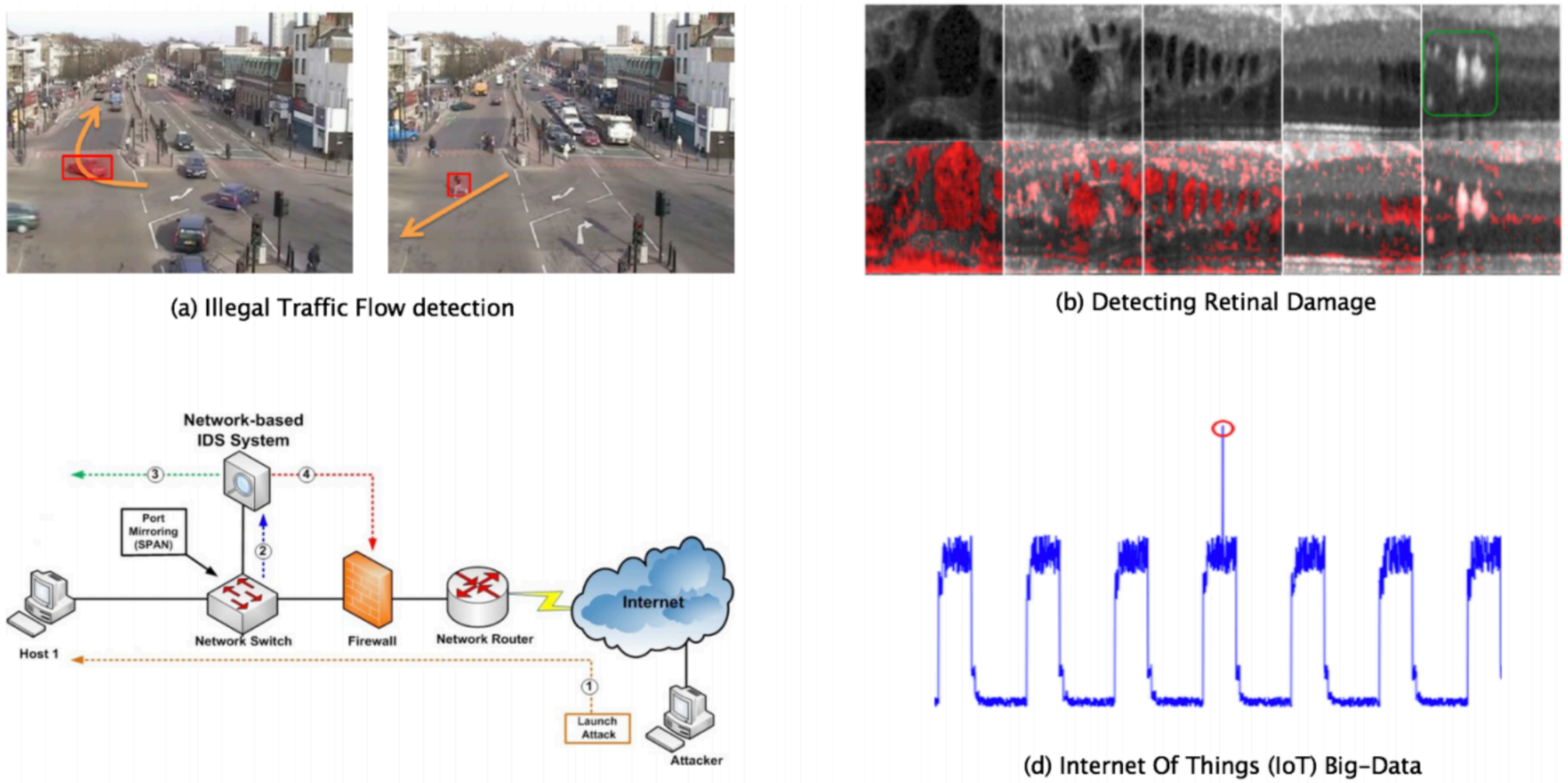




Outlier?

Definition 1: (Hawkins-Outlier)

An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.



(a) Illegal Traffic Flow detection

(b) Detecting Retinal Damage

(c) Cyber-Network Intrusion detection

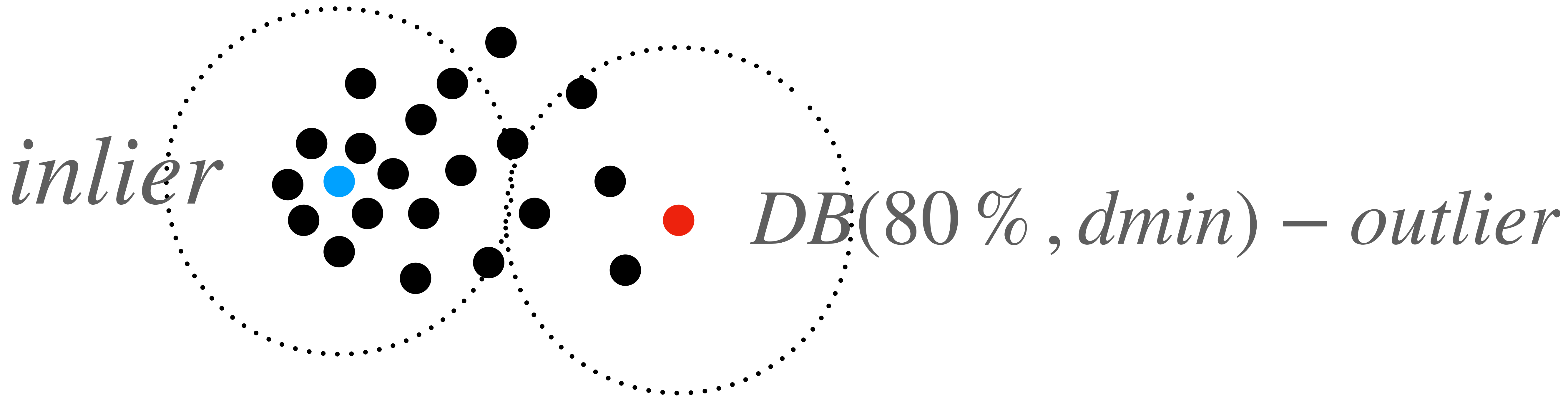
(d) Internet Of Things (IoT) Big-Data Anomaly detection

Reference : R.Chalapathy, S.Chawla, Deep learning for anomaly detection: A survey, 2019 Jan

Distance based approach

Definition 2: (DB(*pct, dmin)*-Outlier)

An object *p* in a dataset *D* is a *DB(pct, dmin)*-outlier if at least percentage *pct* of the objects in *D* lies greater than distance *dmin* from *p*



Distance based approach

- DB-method takes a global view of data
- o_1, o_2 are outliers according to Hawkins' definition
- C_1, C_2 are the set of inliers.
- In right side dataset, DB-method cannot detect outliers.

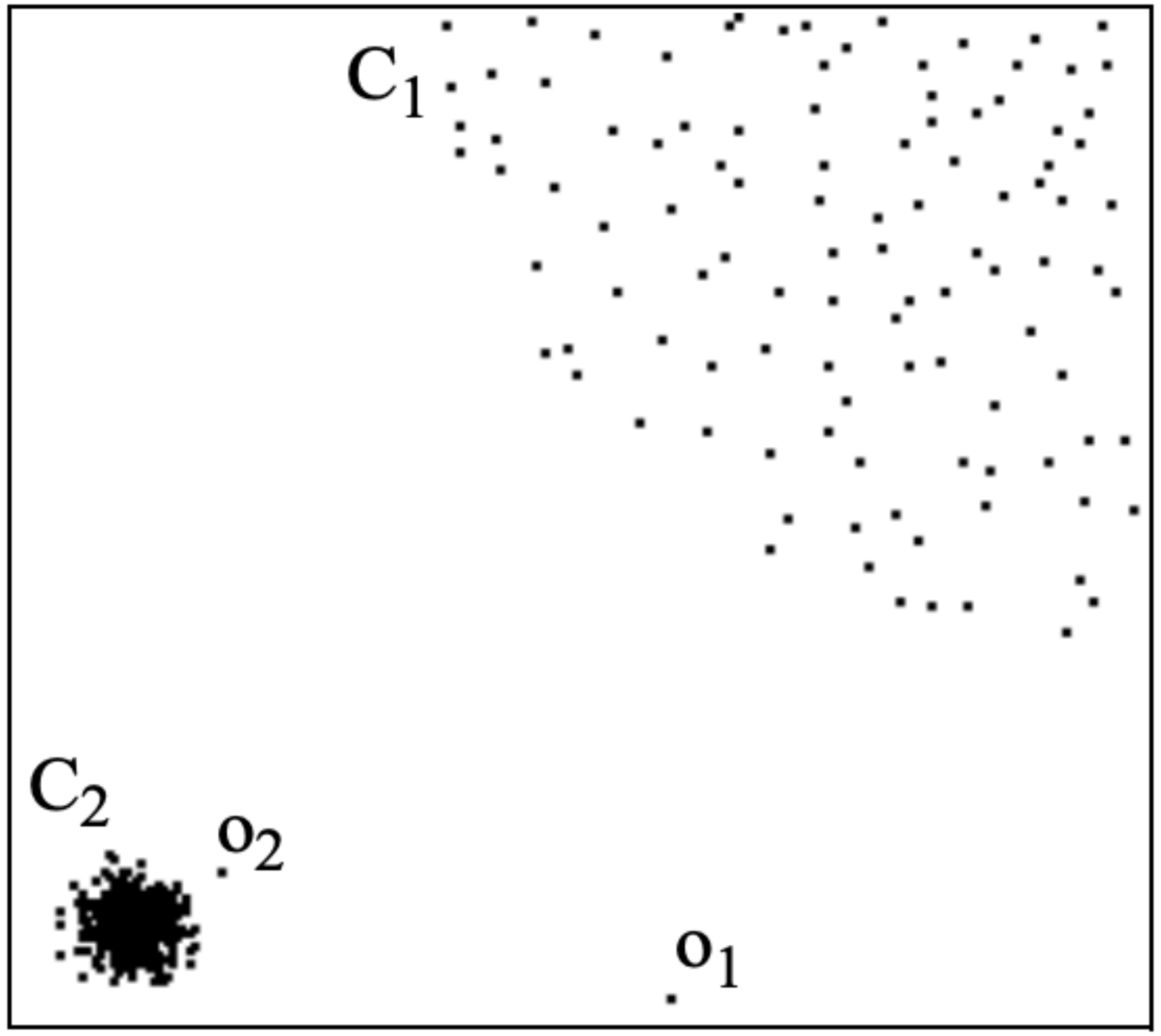


Figure 1: 2-*d* dataset DS1

Local Outlier Factor

Definition 3: (k-distance of p)

For any positive integer k, the k-distance of object p, denoted as k-distance(p), is defined as the distance d(p,o) between p and an object o $\in$ D such that:

(i) for at least k objects o'$\in$D \ {p} it holds that d(p,o') ≤ d(p,o), and

(ii) for at most k-1 objects o'$\in$D \ {p} it holds that d(p,o') < d(p,o).

Definition 4: (k-distance neighborhood of p)

Given the k-distance of p, the k-distance neighborhood of p contains every object whose distance from p is not greater than the k-distance,

i.e. $N_{k-distance(p)}$(p) = { q $\in$ D\{p} | d(p, q) ≤ k-distance(p) }.

Local Outlier Factor

Definition 5: (reachability distance of p w.r.t o)

Let k be a natural number. The reachability distance of object p with respect to object o is defined as

$$reach\_dist_k(p, o) = \max\{k - distance(o), d(p, o)\}$$

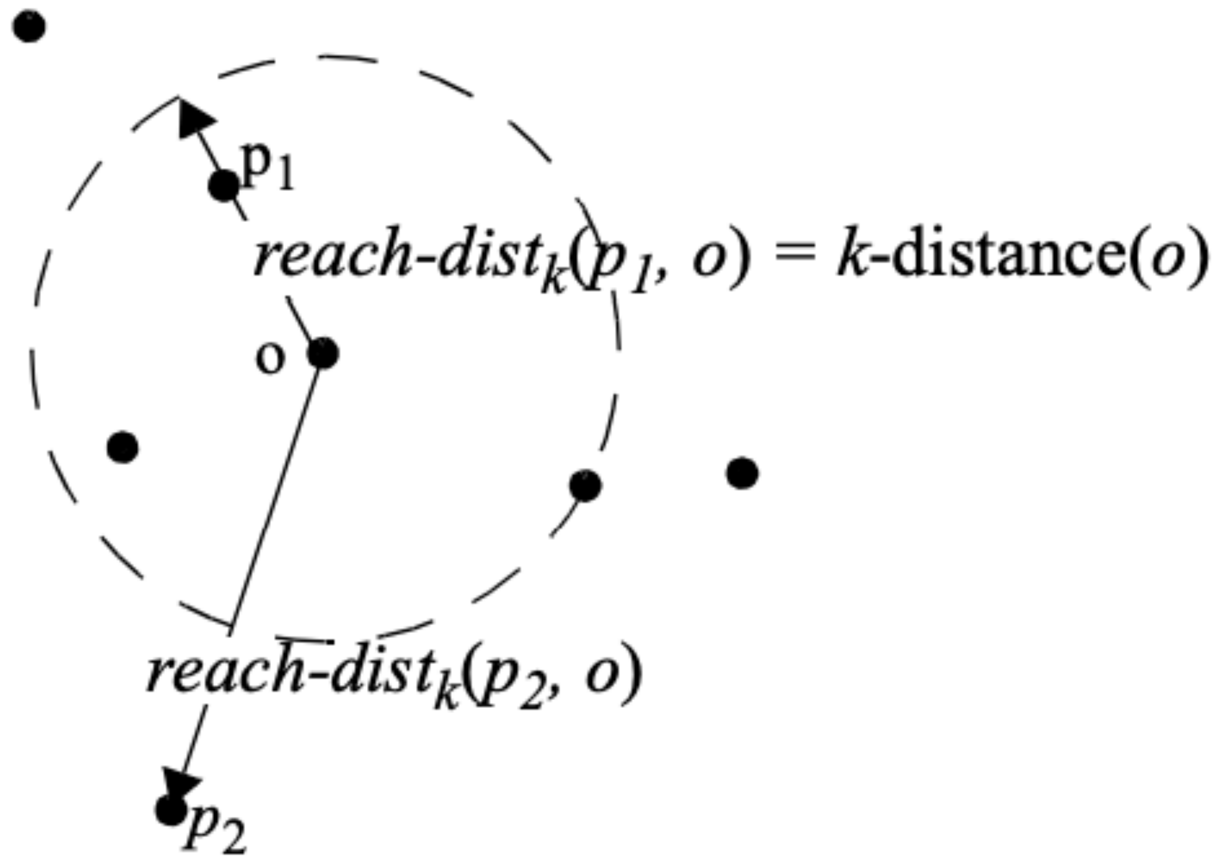


Figure 2: ***reach-dist***(p_1,o) **and** ***reach-dist***(p_2,o)**, for** k**=4**

Local Outlier Factor

Definition 6: (local reachability density of p)

The local reachability density of p is defined as

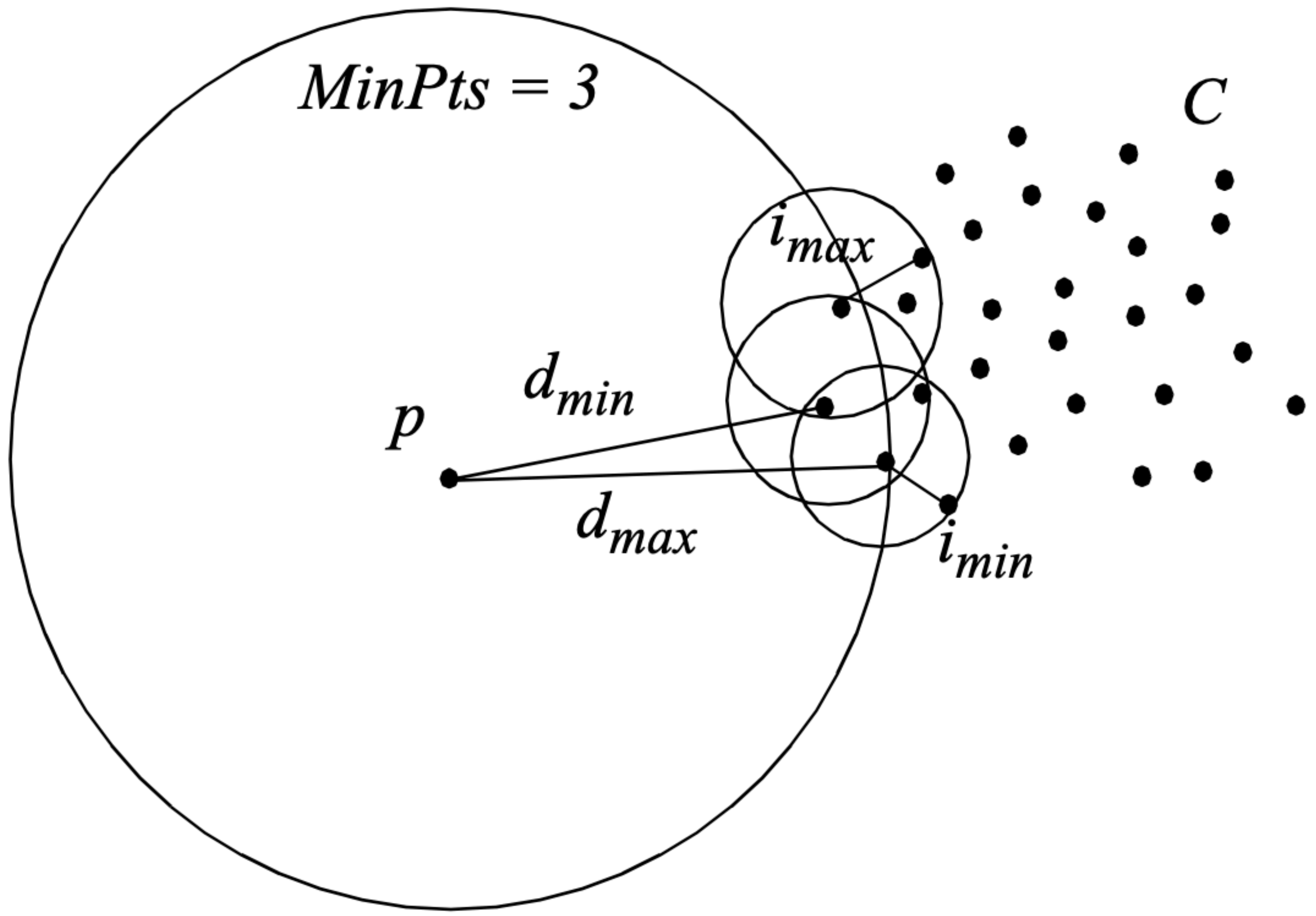
$$lrd_{MinPts}(p) = 1/\left(\frac{\Sigma_{o \in N_{MinPts}} reach\_dist_{MinPts}(p, o)}{|N_{MinPts}(p)|}\right)$$

Local Outlier Factor

Definition 6: ((local) outlier factor of p)

The (local) outlier factor of p is defined as

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}|}$$



A upper and lower bound of LOF

$direct_{min}(p) = min\{reach\_dist(p, q) \,|\, q \in N_{MinPts}(p)\}$

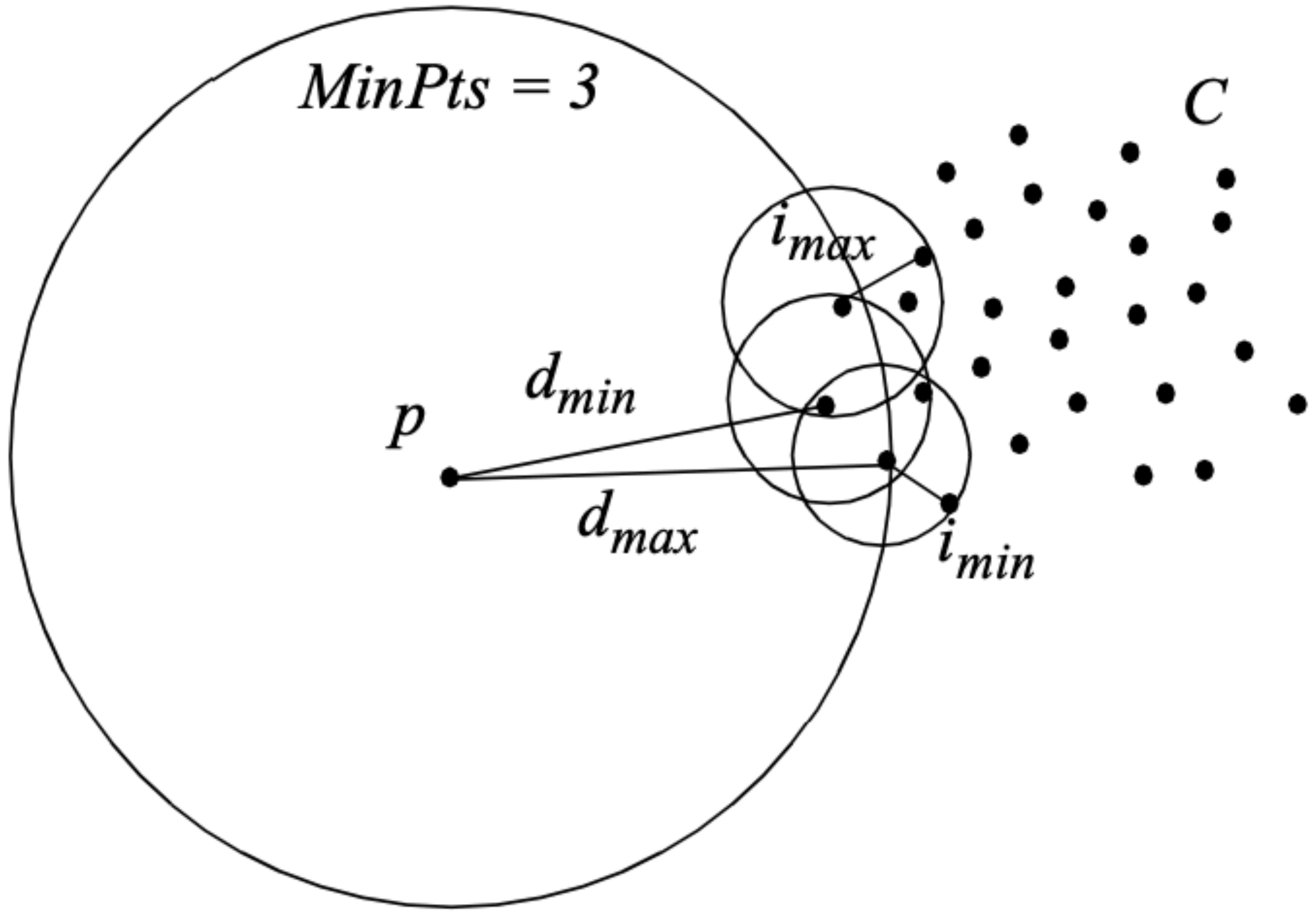
$indirect_{min}(p) = min\{reach\_dist(q, o) \,|\, q \in N_{MinPts}(p)$ and $o \in N_{MinPts}(q)\}$

Theorem 1: Let p be an object from D, and $1 \leq MinPts \leq |D|$.

Then, it is the case that

$$\frac{direct_{min}(p)}{indirect_{max}(p)} \leq LOF(p) \leq \frac{direct_{max}(p)}{indirect_{min}(p)}$$

A upper and lower bound of LOF



$$d_{min} = 4 * i_{max} \quad \Rightarrow LOF_{MinPts}(p) \geq 4$$

$$d_{max} = 6 * i_{min} \quad \Rightarrow LOF_{MinPts}(p) \leq 6$$

Figure 3: Illustration of theorem 1

Impact of the parameter MinPts

- How LOF varies according to changing MinPts values?

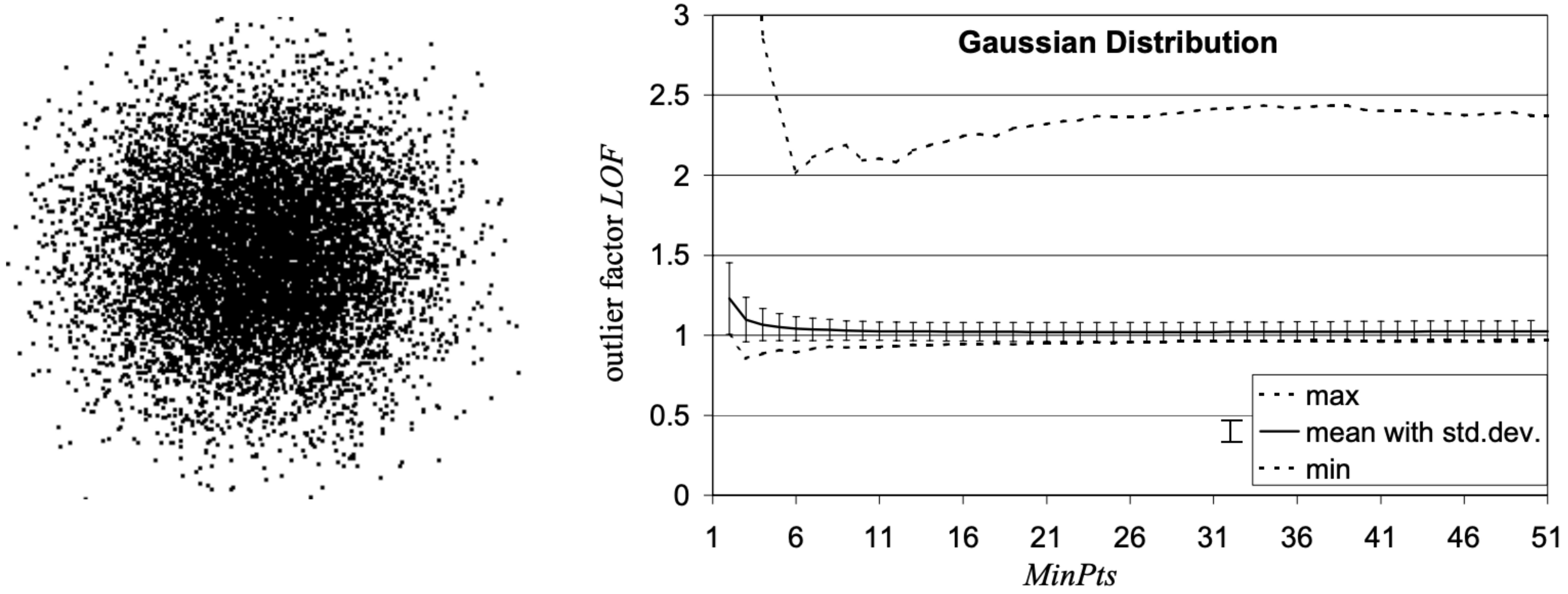


Figure 7: Fluctuation of the outlier-factors within a Gaussian cluster

Impact of the parameter MinPts

- Determining a Range of MinPts Values
    - Too small MinPts cause unwanted statistical fluctuations.
    - MinPtsLB can be regarded as the minimum number of objects a “cluster” has contain
    - MinPtsUB is the maximum number of “close by” objects that can potentially be local outliers.

Impact of the parameter MinPts

S_1 consist of 10 objects
S_2 consist of 35 objects
S_3 consist of 500 objects

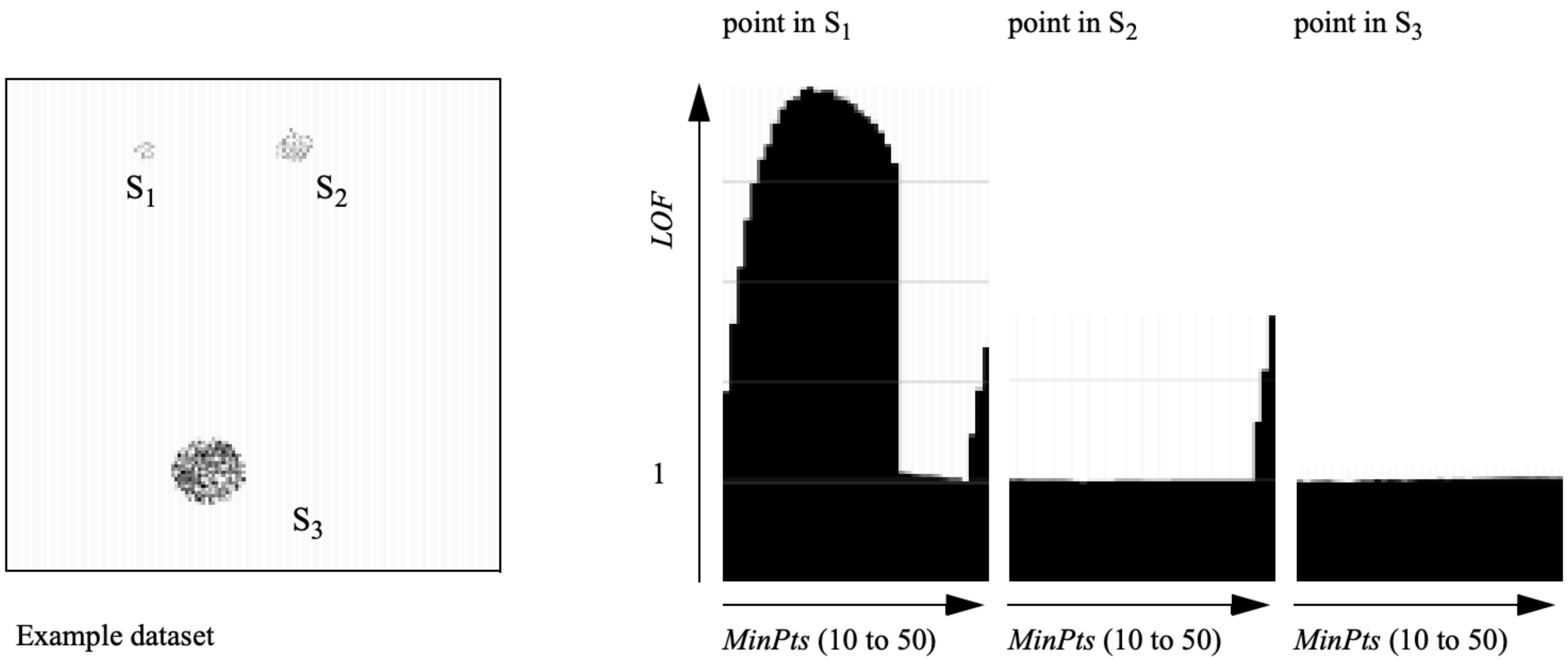


Figure 8: Ranges of *LOF* values for different objects in a sample dataset