

Anomaly Detection:

UNIT 5

CONTENT

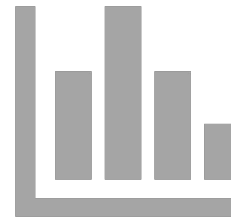


Anomaly Detection

Types of Anomalies

Challenges of Anomaly Detection

Type of Anomaly Detection



Methods

Statistical Methods


Proximity-based Methods

Clustering-based Methods

Classification- based Methods



WHAT IS ANOMALY DETECTION

- An anomaly is an observation that doesn't fit the distribution of the data for normal instances.
 - i.e., is unlikely under the distribution of the majority of instances.
- 

APPLICATION



Fraud Detection.

Intrusion Detection

Ecosystem Disturbances

Medicine and Public Health

Aviation Safety

Anomaly/Outlier Detection



What are anomalies/outliers?

The set of data points that are considerably different than the remainder of the data



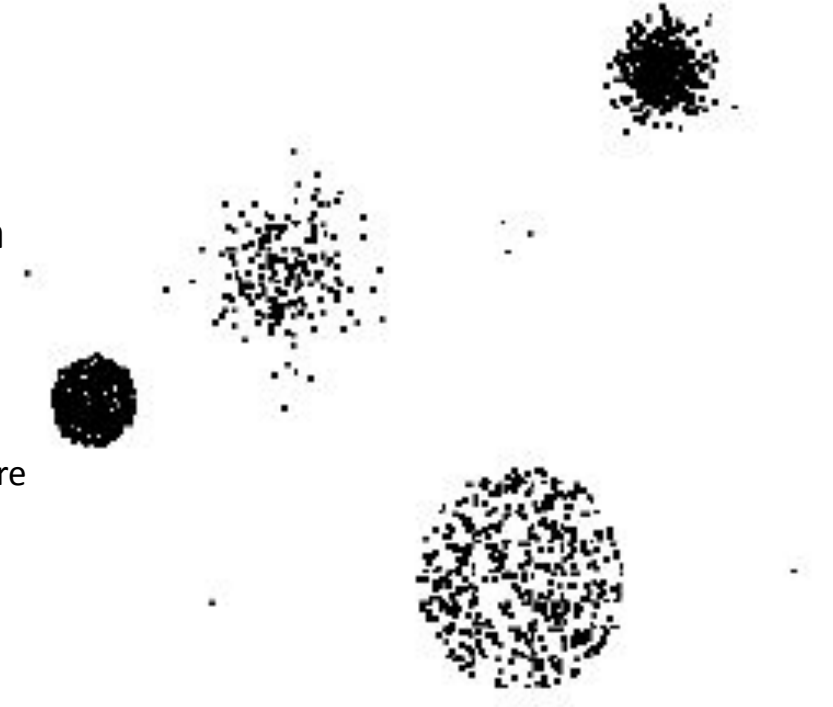
Natural implication is that anomalies are relatively rare

One in a thousand occurs often if you have lots of data
Context is important, e.g., freezing temps in July



Can be important or a nuisance

Unusually high blood pressure
200 pound, 2 year old



Importance of Anomaly Detection

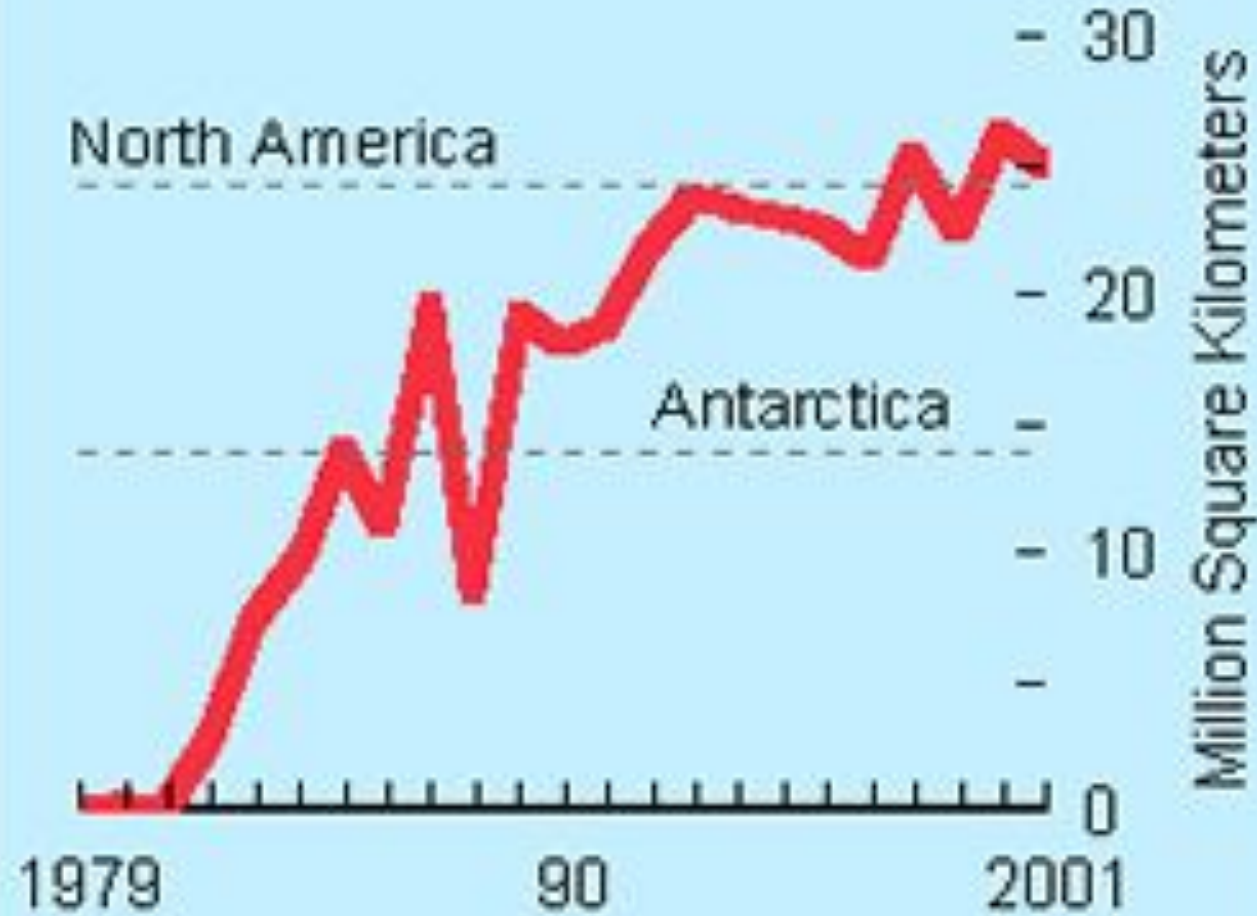
Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels
- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?
- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!

Antarctic Ozone Hole

Average Area

Source: <http://www.epa.gov/ozone/science/hole/size.html>



Hole defined as area < 220 Dobson Units

Source: NASA Goddard Space Flight Center

Causes of Anomalies



Data from different classes

- Measuring the weights of oranges, but a few grapefruit are mixed in

Natural variation

- Unusually tall people

Data errors

- 200 pound 2-year-old

Distinction Between Noise and Anomalies



Noise doesn't necessarily produce unusual values or objects



Noise is not interesting



Noise and anomalies are related but distinct concepts

Model-based vs Model-free



Model-based Approaches

Model can be parametric or non-parametric
Anomalies are those points that don't fit well
Anomalies are those points that distort the model



Model-free Approaches

Anomalies are identified directly from the data without building a model



Often the underlying assumption is that most of the points in the data are normal

General Issues: Label vs Score



Some anomaly detection techniques provide only a binary categorization



Other approaches measure the degree to which an object is an anomaly

This allows objects to be ranked
Scores can also have associated meaning (e.g., statistical significance)

Anomaly Detection Techniques



Statistical Approaches



Proximity-based

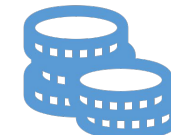
Anomalies are points far away from other points



Clustering-based

Points far away from cluster centers are outliers

Small clusters are outliers



Reconstruction Based

Statistical Approaches- Probabilistic definition of an outlier

01

An outlier is an object that has a low probability with respect to a probability distribution model of the data.

02

Usually assume a parametric model describing the distribution of the data (e.g., normal distribution)

03

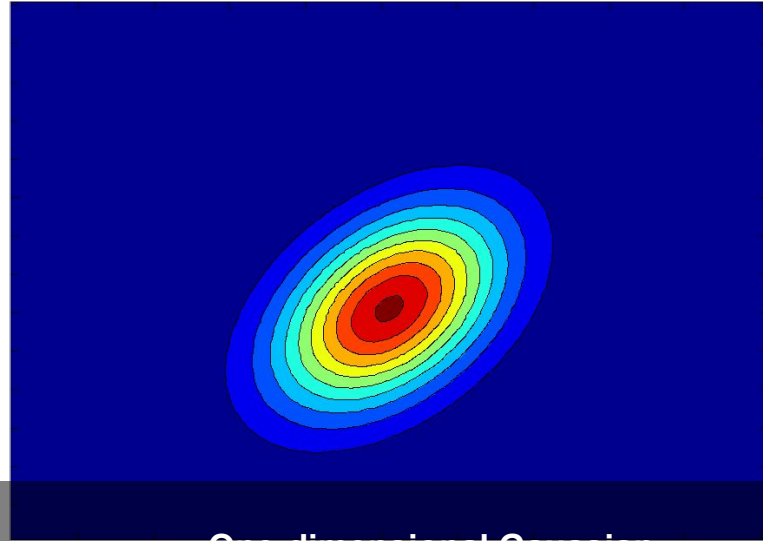
Apply a statistical test that depends on

- Data distribution
- Parameters of distribution (e.g., mean, variance)
- Number of expected outliers (confidence limit)

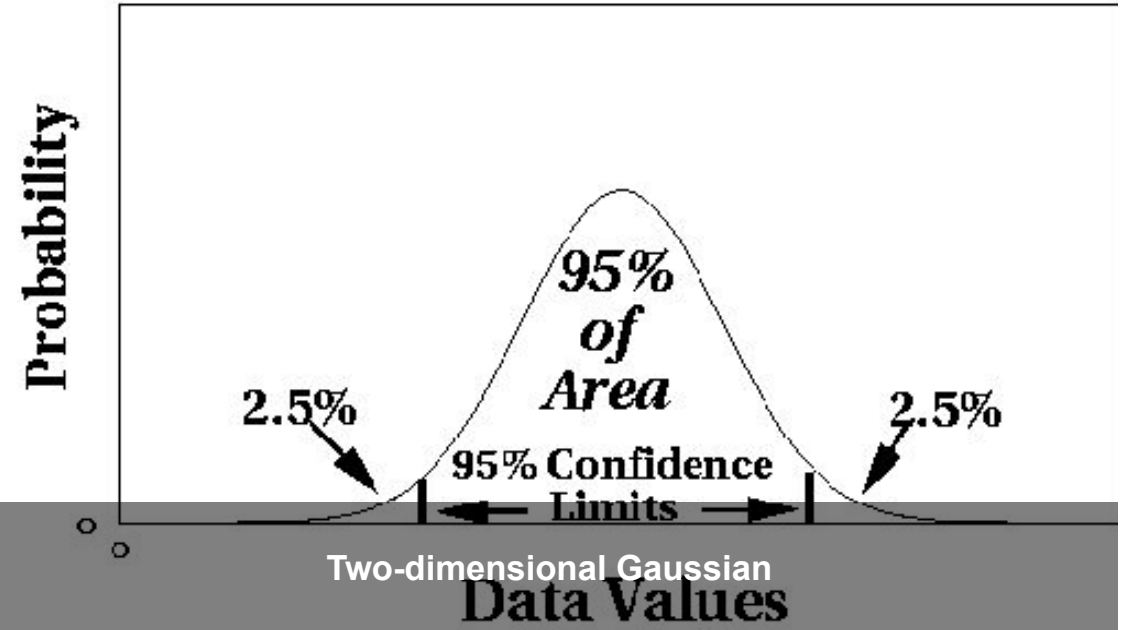
04

Issues

- Identifying the distribution of a data set
 - Heavy tailed distribution
- Number of attributes
- Is the data a mixture of distributions?



One-dimensional Gaussian



Normal Distributions

Grubbs' Test

- Detect outliers in univariate data
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
 - H_0 : There is no outlier in data
 - H_A : There is at least one outlier
- Grubbs' test statistic:
 - Reject H_0 if:

$$G = \frac{\max |X - \bar{X}|}{s}$$

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

Statistically-based – Likelihood Approach

Assume the data set D contains samples from a mixture of two probability distributions:

- M (majority distribution)
- A (anomalous distribution)

General Approach:

- Initially, assume all the data points belong to M
- Let $L_t(D)$ be the log likelihood of D at time t
- For each point x_t that belongs to M , move it to A
 - Let $L_{t+1}(D)$ be the new log likelihood.
 - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
 - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistically- based – Likelihood Approach

- Data distribution, $D = (1 - \lambda) M + \lambda A$
- M is a probability distribution estimated from data
 - Can be based on any modeling method (naïve Bayes, maximum entropy, etc.)
- A is initially assumed to be uniform distribution
- Likelihood at time t :

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

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

Strengths/Weaknesses of Statistical Approaches

Firm mathematical foundation

Can be very efficient

Good results if distribution is known

In many cases, data distribution may not be known

For high dimensional data, it may be difficult to estimate the true distribution

Anomalies can distort the parameters of the distribution

Proximity-based Approaches

Conceptual Basis:

- identify anomalies based on the distance of instances from others.
- If normal instances exhibit proximity to each other
- while anomalies deviate and appear relatively distant from normal instances.

Distance-Based Techniques:

- Proximity-based methods often use distances as a fundamental metric.
- They're also known as distance-based outlier detection techniques due to their reliance on calculating distances between data points.

Model-Free Approach:

- Proximity-based techniques are categorized as model-free anomaly detection methods.
- They don't create an explicit model of the normal class to determine anomaly scores.

Proximity-based Approaches

Local Perspective:

- They calculate anomaly scores based on the local perspective of each data instance.
- Instead of considering global properties or statistical distributions, they focus on the relationships among neighboring instances.

Advantages over Statistical Approaches:

- Proximity-based methods are more versatile than statistical approaches.
- Determining a meaningful proximity measure for a dataset is often easier than defining its statistical distribution.

Variations in Approaches:

- Different proximity-based techniques exist, primarily differing in how they assess the locality of a data instance.

Local Analysis of Data Instances:

- These methods analyze the neighborhood or local environment of each instance to gauge its anomaly score.
- The focus is on understanding the relationships and patterns among nearby instances to identify anomalies effectively.

Distance-Based Approaches

Definition:

Proximity-based anomaly scores for a data instance x can be computed using its distance to the k^{th} nearest neighbour, known as $\text{dist}(x, k)$.

This metric helps differentiate between normal and anomalous instances based on their proximity to neighbouring data points.

Relation to Normalcy:

Normal instances, being close to other instances, tend to have a low value of $\text{dist}(x, k)$ since they are surrounded by neighbouring data points.

Anomalies, being isolated or dissimilar, have a high value of $\text{dist}(x, k)$ as they are distant from their k -nearest neighbors.

Sensitivity to k :

The choice of k significantly impacts the anomaly scores.

A small k , like 1, might misrepresent outliers as normal if they are located close to each other, leading to low anomaly scores for anomalies in proximity.

A large k may label entire clusters with fewer than k objects as anomalies, thus increasing the anomaly scores of those points.

Distance-Based Approaches

Effects on Anomaly Identification:

Figure 9.5 illustrates that a small k can misclassify anomalies located in close proximity to each other, resulting in falsely low anomaly scores for these outliers.

Conversely, a large k might mark smaller clusters as anomalies, as depicted in Figure 9.6, causing higher anomaly scores for points in smaller clusters.

Robust Alternative Metric:

To address the sensitivity issue, the average distance to the first k -nearest neighbors, $\text{avg.dist}(x, k)$, serves as a more robust alternative to $\text{dist}(x, k)$.

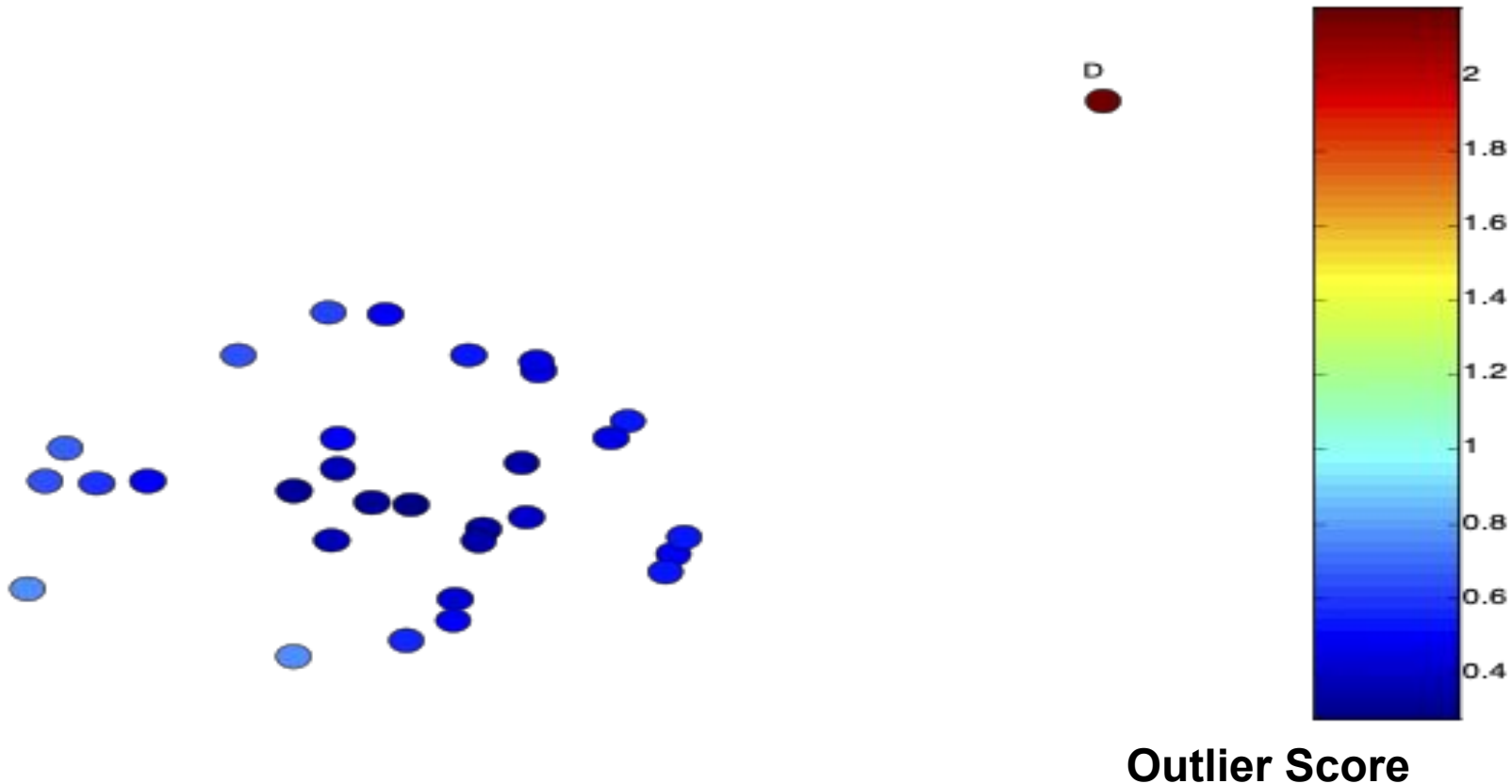
$\text{avg.dist}(x, k)$ is commonly used in various applications due to its reliability in determining proximity-based anomaly scores, being less influenced by the choice of k .

Robustness and Applicability:

$\text{avg.dist}(x, k)$ offers greater stability against variations in k compared to the direct distance metric, making it a preferred choice in practical applications for anomaly detection.

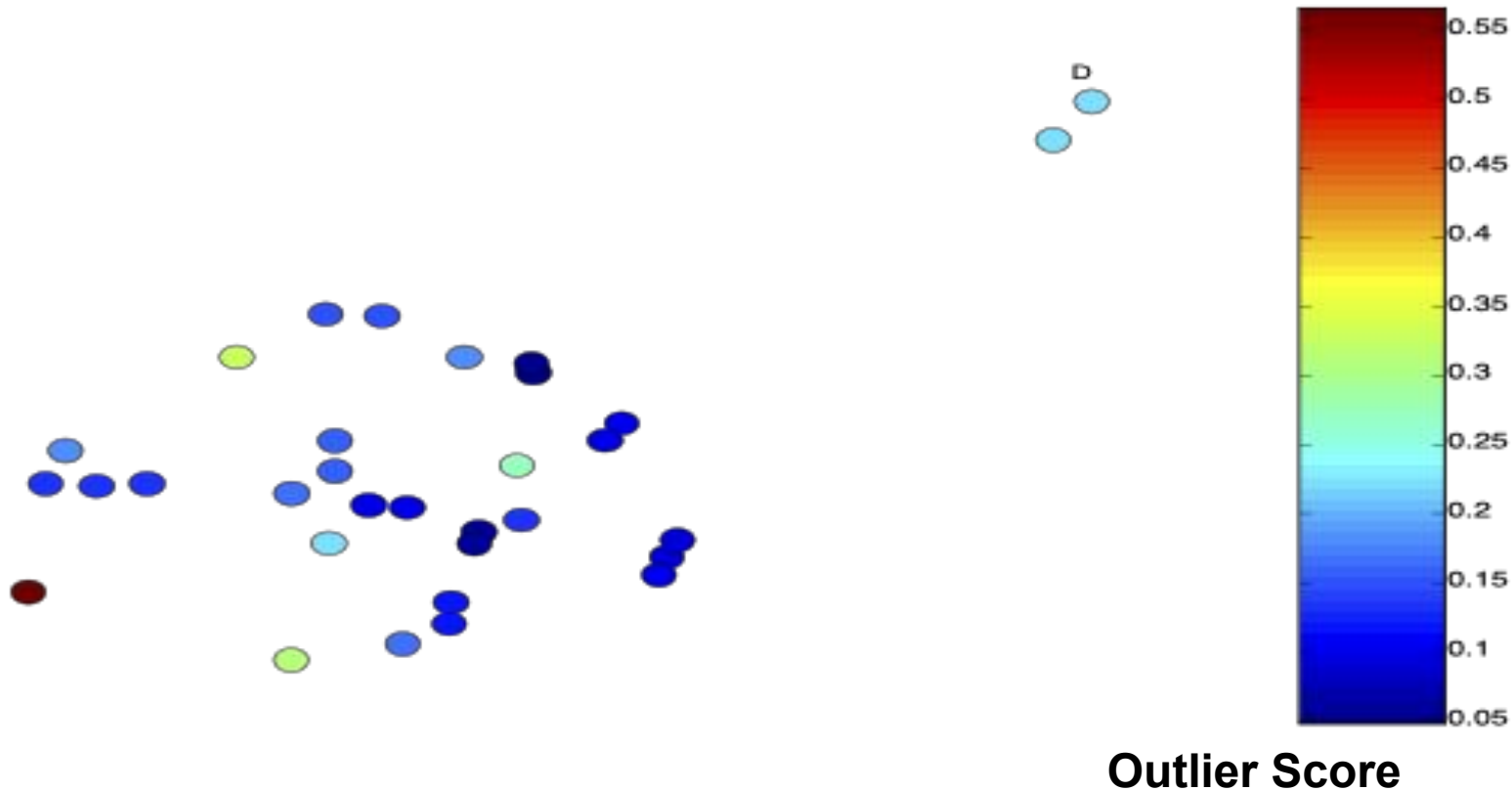
One Nearest Neighbor - One Outlier

- Anomaly score based on the distance to fifth nearest neighbour.
- A small k can misclassify anomalies located in close proximity to each other, resulting in falsely low anomaly scores for these outliers.



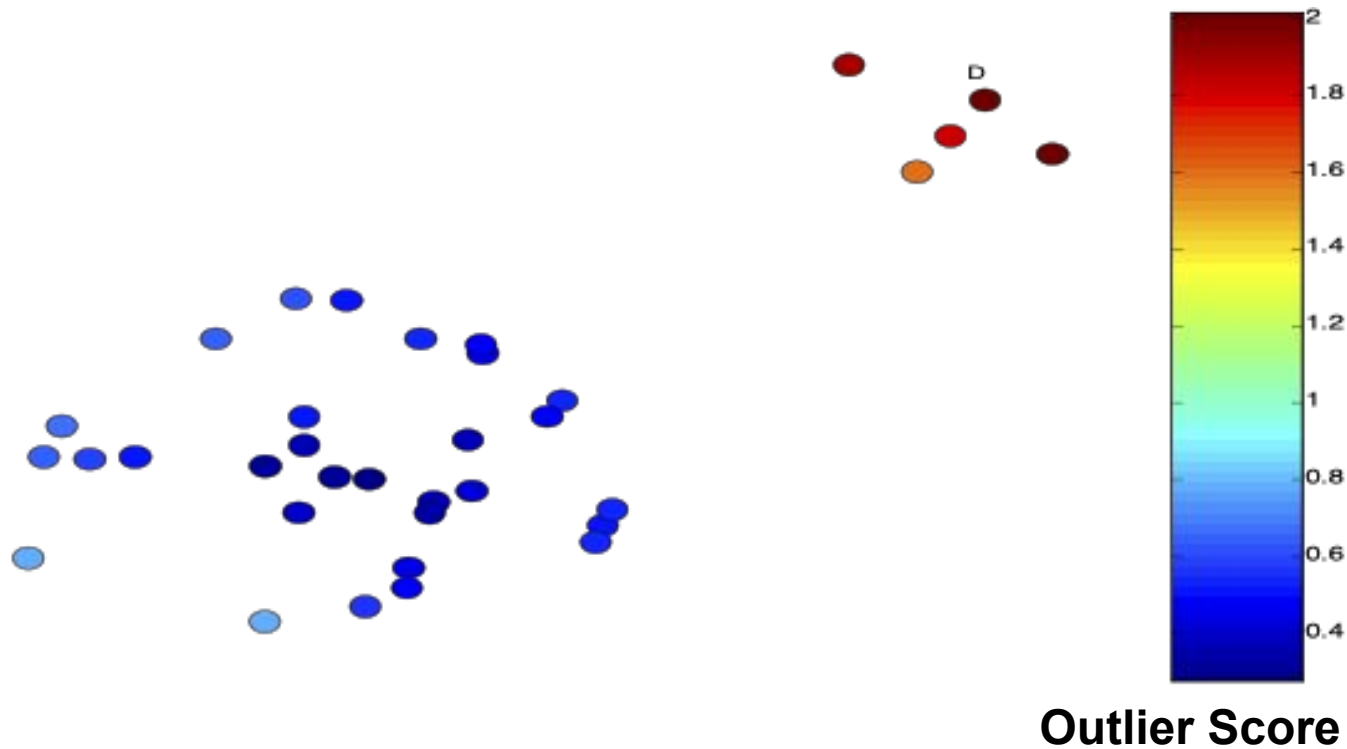
One Nearest Neighbor - Two Outliers

- Anomaly score based on the distance to the first nearest neighbour.
- Nearby outliers have low anomaly scores.
- a large k might mark smaller clusters as anomalies, causing higher anomaly scores for points in smaller clusters.



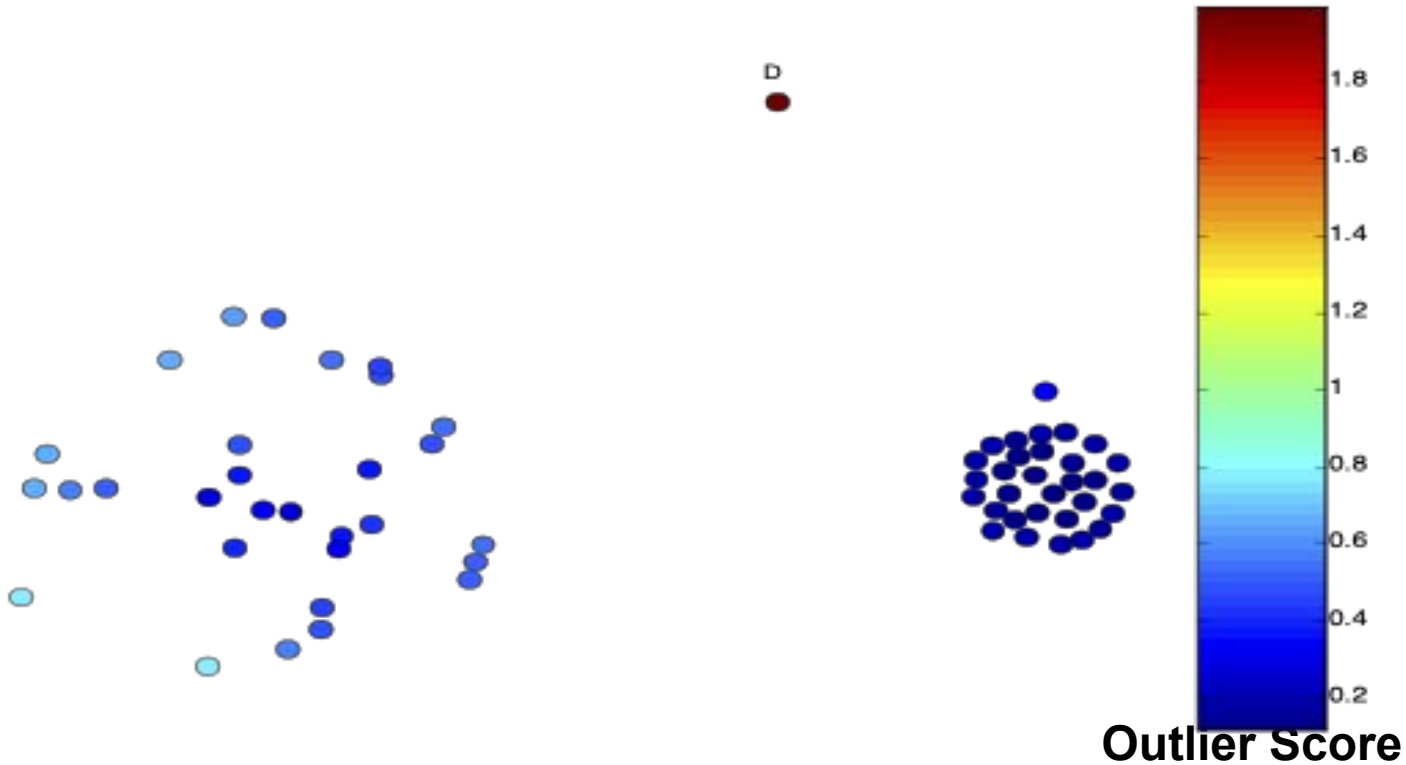
Five Nearest Neighbors - Small Cluster

- Anomaly score based on distance to the fifth nearest neighbour.
- A small cluster becomes an outlier.



Five Nearest Neighbors - Differing Density

- Anomaly score based on the distance to the fifth nearest neighbour, when there are clusters of varying densities.



Strengths/Weaknesses of Distance-Based Approaches

Simple

Expensive – $O(n^2)$

Sensitive to
parameters

Sensitive to variations
in density

Distance becomes less
meaningful in
high-dimensional
space

Density-based Anomaly Score

Density Calculation:

- Density around a data instance is calculated as the ratio $n/V(d)$, where:
 - n represents the number of instances within a specified distance d from the instance.
 - $V(d)$ denotes the volume of the neighborhood within the specified distance.
- For simplicity, the density is often represented by the count of instances (n) within a fixed distance (d).

Similarity to DBSCAN Algorithm:

- The density-based definition of anomalies aligns with the DBSCAN clustering algorithm's concept, where anomalies occur in regions of low density.
- Anomalies exhibit lower instance counts within a defined distance (d) compared to normal instances, indicating lower neighborhood density.

Parameter Selection Challenge:

- Similar to distance-based measures where selecting the parameter k poses challenges, in density-based measures, choosing the parameter d is equally challenging.
- A small d might incorrectly label many normal instances as anomalies due to their low-density values, while a large d might treat anomalies as similar to normal instances due to comparable densities.

Density-based Anomaly Score

Relationship between Distance and Density Views:

- The distance-based and density-based views of proximity exhibit similarity.
- Considering the k -nearest neighbors of a data instance x , where $\text{dist}(x, k)$ represents the distance to the k th nearest neighbor:
 - Larger $\text{dist}(x, k)$ signifies lower density around x , and vice versa.
 - Distance-based and density-based anomaly scores show an inverse relationship, allowing the definition of density measures based on distance measures.

Density Measures based on Distance Metrics:

- Utilizing distance measures such as $\text{dist}(x, k)$ and $\text{avg.dist}(x, k)$:
 - $\text{density}(x, k) = 1/\text{dist}(x, k)$
 - $\text{avg.density}(x, k) = 1/\text{avg.dist}(x, k)$

Inverse Relationship Utilization:

- The reciprocal relationship between distance and density measures helps in defining density-based measures, leveraging inverse relationships to identify anomalies based on density criteria.

Relative Density

- Consider the density of a point relative to that of its k nearest neighbors
- Let y_1, \dots, y_k be the k nearest neighbors of x

$$density(x, k) = \frac{1}{dist(x, k)} = \frac{1}{dist(x, y_k)}$$

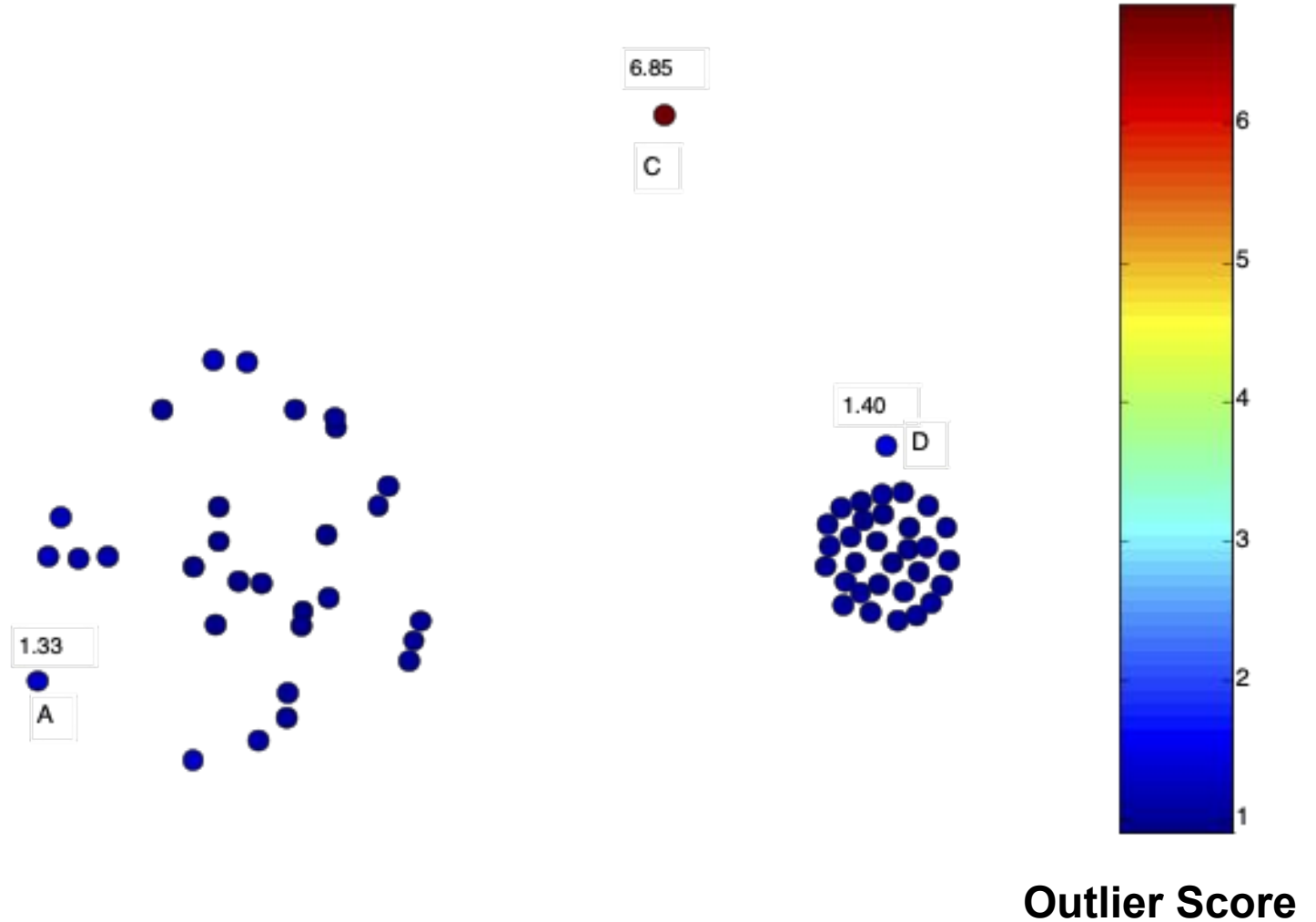
$$relative\ density(x, k) = \frac{\sum_{i=1}^k density(y_i, k)/k}{density(x, k)}$$

$$= \frac{dist(x, k)}{\sum_{i=1}^k dist(y_i, k)/k} = \frac{dist(x, y)}{\sum_{i=1}^k dist(y_i, k)/k}$$

- Can use average distance instead

Relative Density Outlier Scores

Relative density (LOF) outlier scores
for two-dimensional points

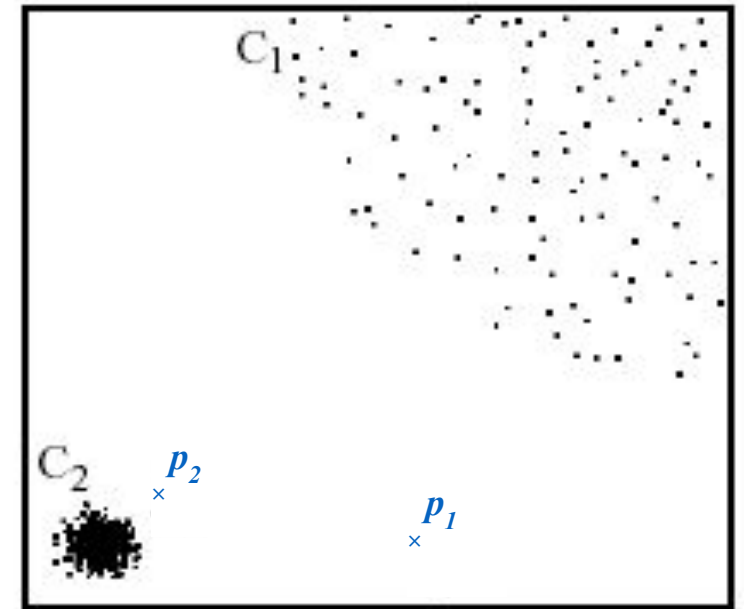


Relative Density-based: LOF approach

For each point, compute the density of its local neighborhood

Compute local outlier factor (LOF) of a sample p as the average of the ratios of the density of sample p and the density of its nearest neighbors

Outliers are points with largest LOF value



In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Strengths/Weaknesses of Density-Based Approaches

Simple

Expensive –
 $O(n^2)$

Sensitive to
parameters

Density
becomes less
meaningful in
high-dimension
al space

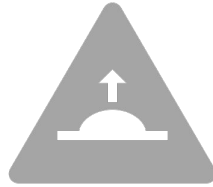
Clustering-Based Approaches



Fundamental Principle:

Clustering-based methods for anomaly detection utilize clusters to represent the normal class.

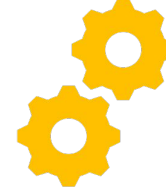
The foundational assumption is that normal instances tend to cluster together, forming distinct groups or clusters in the data.



Normal vs. Anomalous Instances:

Normal instances are expected to be closely grouped within clusters, exhibiting proximity to each other.

Anomalies are identified as instances that either don't fit well within the clusters of the normal class or appear in small, isolated clusters that are distinct from the main clusters representing the normal instances.



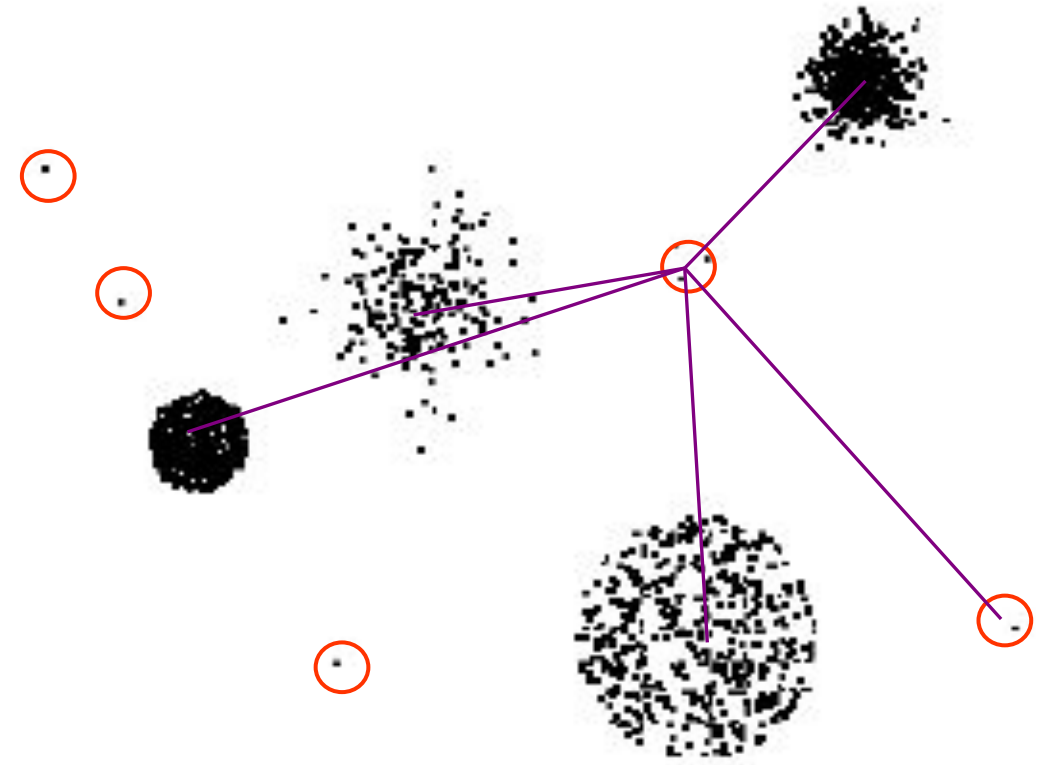
Categorization of Clustering Methods:

Clustering-based methods are categorized into two main types based on their anomaly detection approach:

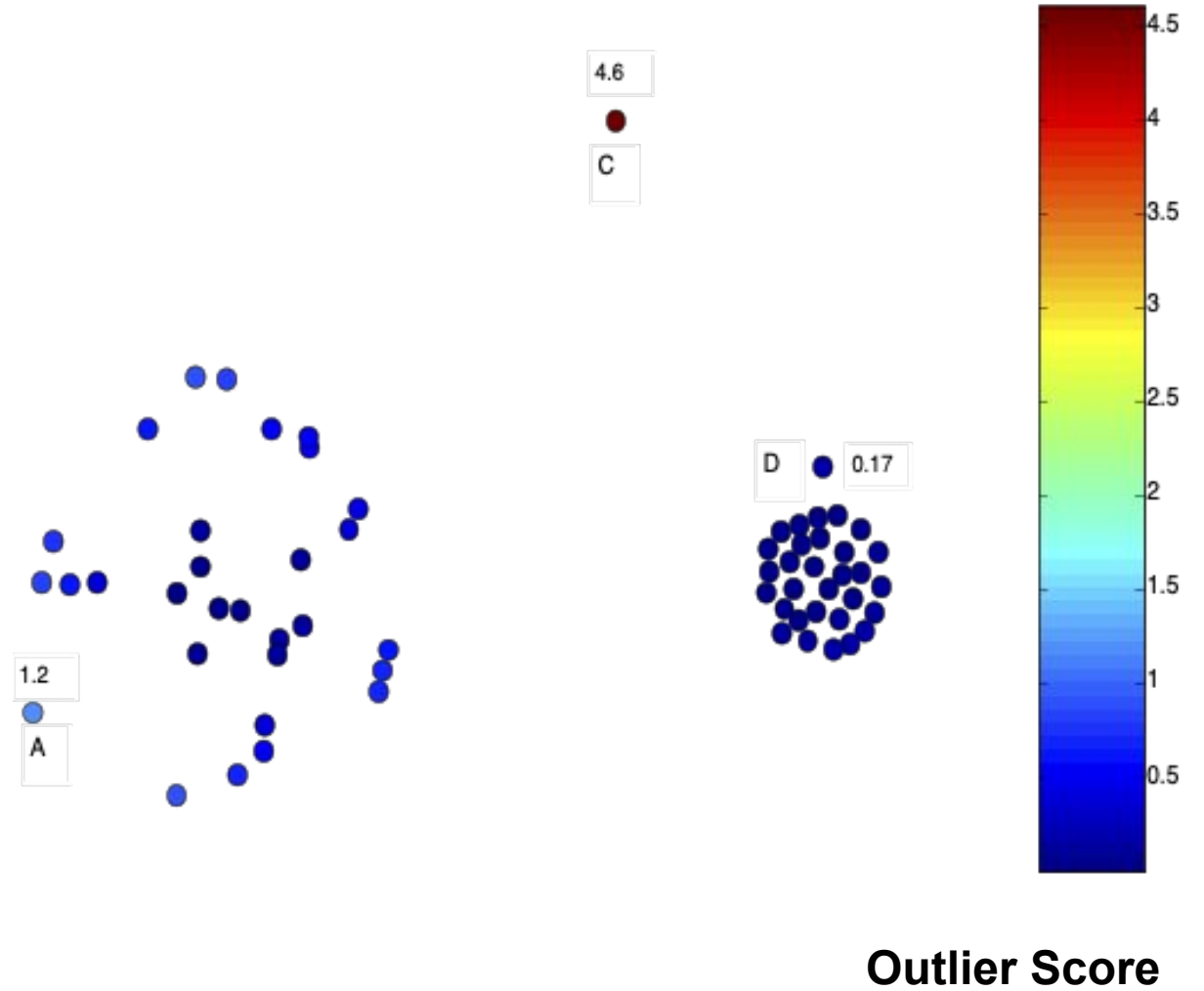
- Methods identifying small clusters as anomalies.
- Methods defining anomalies as instances that poorly fit within the established clustering, often measured by their distance from a cluster centre.

Clustering-Based Approaches

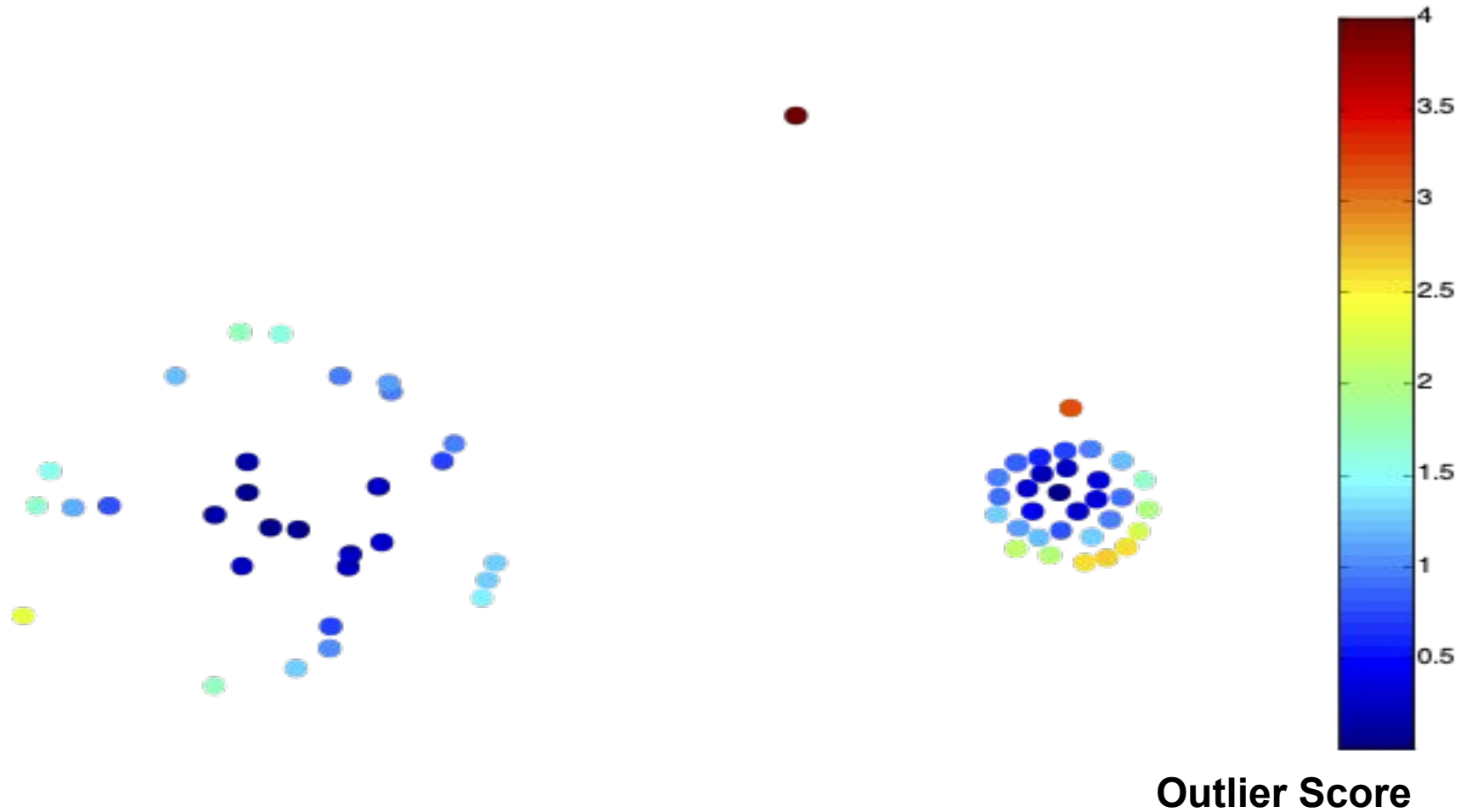
- An object is a cluster-based outlier if it does not strongly belong to any cluster
- For prototype-based clusters
 - an object is an outlier if it is not close enough to a cluster center
 - Outliers can impact the clustering produced
- For density-based clusters, an object is an outlier if its density is too low
 - Can't distinguish between noise and outliers
 - For graph-based clusters, an object is an outlier if it is not well connected



Distance of Points from Closest Centroids



Relative Distance of Points from Closest Centroid



Strengths/Weaknesses of Clustering-Based Approaches

Simple

Many clustering techniques can be used

Can be difficult to decide on a clustering technique

Can be difficult to decide on number of clusters

Outliers can distort the clusters

Reconstruction-Based Approaches

Assumption:

- Reconstruction-based techniques assume that the normal class exists in a lower-dimensional space compared to the original attribute space.
- Patterns within the distribution of the normal class can be captured using lower-dimensional representations via dimensionality reduction techniques.

Dimensionality Reduction and PCA:

- Principal Components Analysis (PCA) is utilized to derive useful features from the data, representing the normal class in a reduced dimensionality.
- PCA yields principal components that capture the maximum data variance, allowing approximation of the data with fewer derived features.

Reconstruction Error Computation:

- Data instances are projected onto their k-dimensional representation using PCA.
- The re-projection of these representations back to the original attribute space creates a reconstruction (\hat{x}).
- The squared Euclidean distance between the original instance (x) and its reconstruction (\hat{x}) serves as the reconstruction error: $\text{Reconstruction Error}(x) = \|x - \hat{x}\|^2$.

Reconstruction-Based Approaches

Utilization of Reconstruction Error:

- Low reconstruction error is anticipated for normal instances due to their adherence to the learned structure.
- Anomalies, not conforming to the hidden structure of the normal class, exhibit high reconstruction errors.
- Reconstruction error effectively serves as an anomaly detection score.

Illustration via PCA:

- The concept is demonstrated in a two-dimensional dataset where circles represent normal instances and squares represent anomalies.
- PCA helps in capturing the hidden structure by projecting instances onto a line (first principal component) as a lower-dimensional representation.
- Reconstruction errors are calculated based on the deviation of instances from this line, effectively identifying anomalies.

Limitations of PCA:

- PCA can only derive linear combinations of attributes and struggles with nonlinear patterns in the normal class.
- In scenarios with nonlinear patterns, autoencoders are introduced as a possible solution for nonlinear dimensionality reduction and reconstruction.

Reconstruction-Based Approaches

Autoencoders for Nonlinear Dimensionality Reduction:

- Autoencoders, multi-layer neural networks, facilitate nonlinear transformation for dimensionality reduction and reconstruction.
- They involve encoding data into a low-dimensional representation and decoding back to the original attribute space, allowing the calculation of reconstruction errors as anomaly scores.

Learning Complex Representations:

- Autoencoders, especially in the presence of primarily normal instances, enable the learning of complex and nonlinear representations using backpropagation techniques.
- Variants like denoising autoencoders robustly learn representations even in noisy datasets.

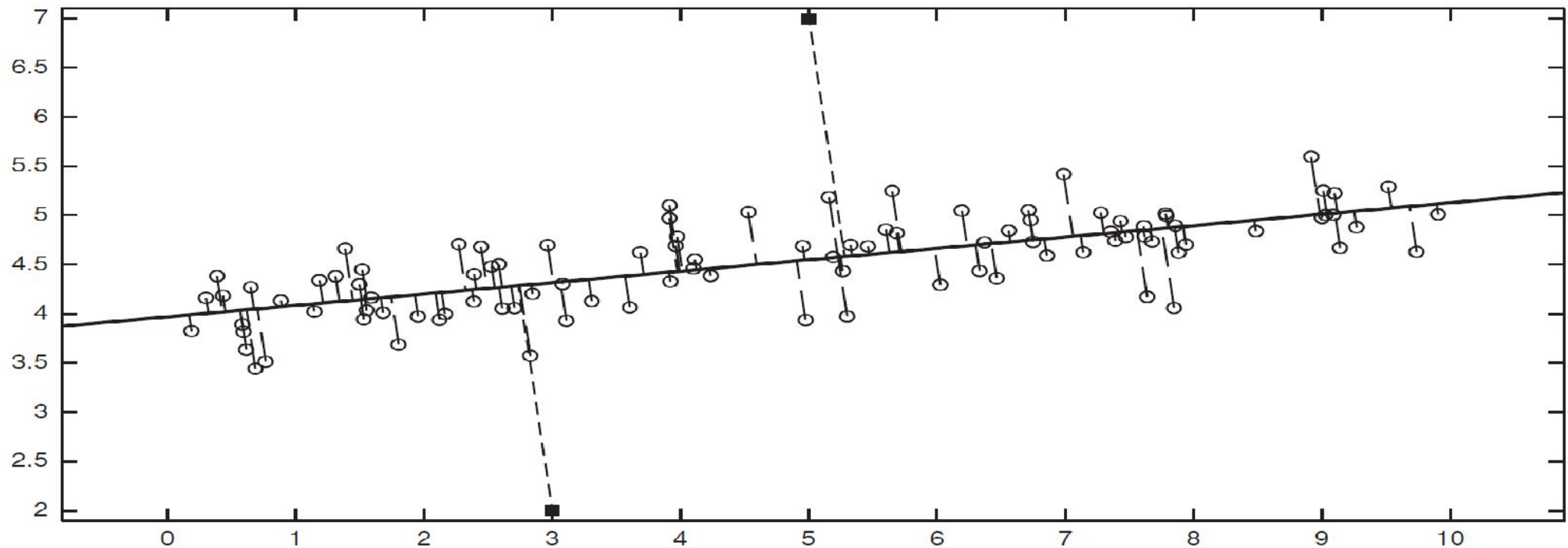
Reconstruction Error

-
- Let \mathbf{x} be the original data object
- Find the representation of the object in a lower dimensional space
- Project the object back to the original space
- Call this object $\hat{\mathbf{x}}$

$$\text{Reconstruction Error}(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|$$

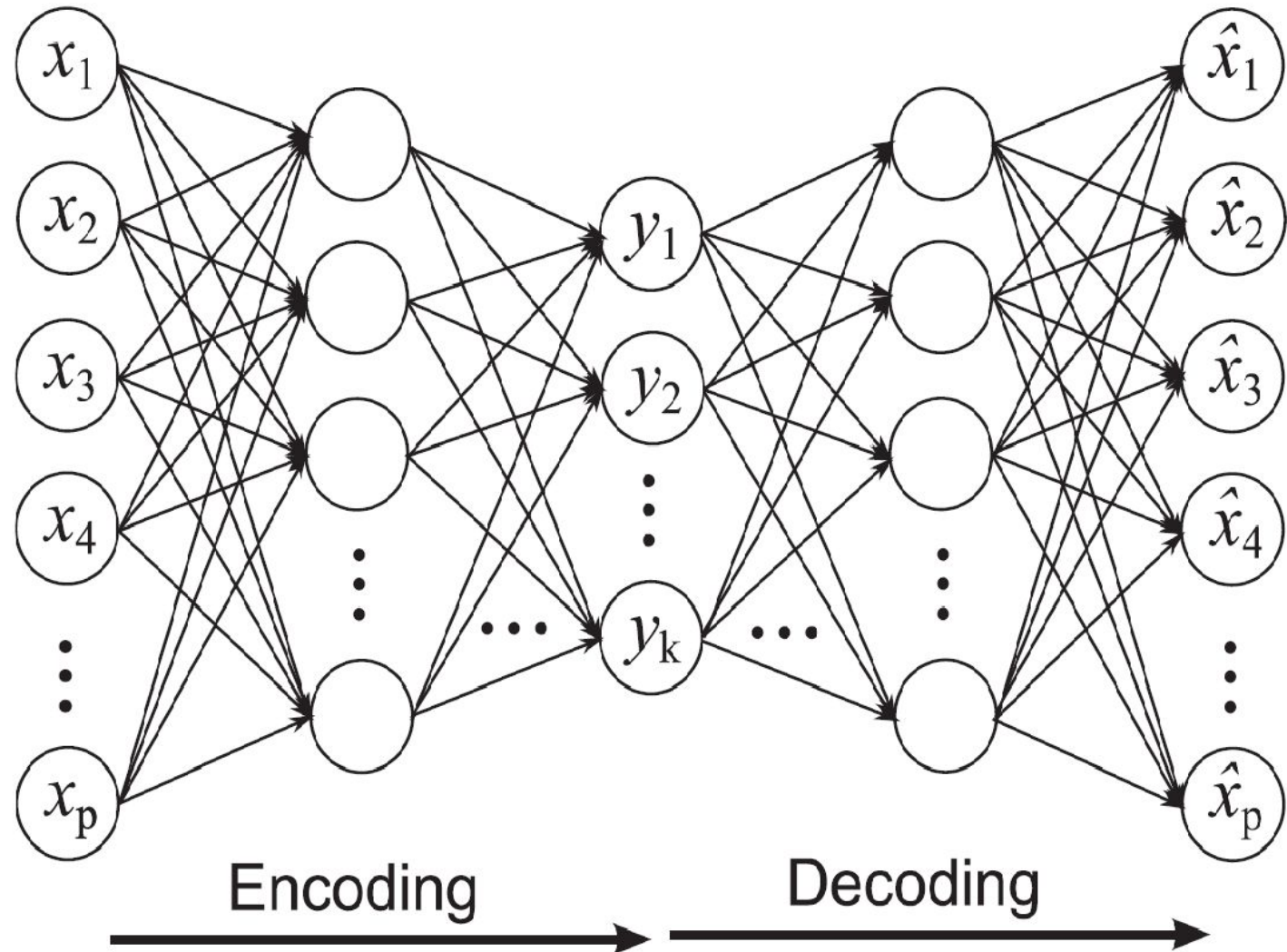
- Objects with large reconstruction errors are anomalies

Reconstruction of two-dimensional data



Basic Architecture of an Auto-encoder

- An autoencoder is a multi-layer neural network
- The number of input and output neurons is equal to the number of original attributes.



Strengths and Weaknesses

01

Does not require assumptions about distribution of normal class

02

Can use many dimensionality reduction approaches

03

The reconstruction error is computed in the original space

- This can be a problem if dimensionality is high

One-Class Classification

- Differs from binary classification by learning a boundary that encloses all normal objects in the attribute space.
- Focuses on modelling the boundary of the normal class for anomaly detection.
- Contrasts binary classification where boundaries separate two distinct classes.
- **One-Class SVM**
 - One-class SVM utilizes only normal class instances to learn a decision boundary.
 - Involves transforming data to a higher-dimensional space using a function ϕ for linear separation.
 - Kernels, like the Gaussian kernel, are used for learning nonlinear boundaries in the transformed space.

Use of Kernels

Nonlinear Boundary with Linear Separation:

To achieve a nonlinear boundary enclosing the normal class, data is transformed into a higher-dimensional space.

This transformation allows the use of a linear hyperplane for separating the normal class instances.

A function ϕ maps original attribute space instances x to points $\phi(x)$ in the transformed space.

Linear Hyperplane in Transformed Space:

The goal is to create a linear hyperplane, represented by parameters (w, ρ) , that ideally separates normal instances from anomalies.

The hyperplane equation is defined as $w \cdot \phi(x) = \rho$, using inner products between vectors x and y .

Representation of Separating Hyperplane:

The separating hyperplane is represented using α_i 's (weights) and ρ derived from the linear combination of $\phi(x_i)$'s.

This representation enables the description of the hyperplane based on inner products of $\phi(x)$ in the transformed space.



Use of Kernels

Utilizing Kernel Functions:

Kernel functions like $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$ are employed for computing inner products of $\phi(x)$ in the transformed space.

These kernel functions are extensively used for learning nonlinear boundaries, such as kernel-SVMs in binary classification problems.

Challenges in One-Class Learning:

Learning nonlinear boundaries in the one-class setting is challenging due to the absence of information about the anomaly class during training.

One-class SVM employs the "origin trick" to overcome this challenge, particularly effective with specific types of kernel functions.

Use of Kernels and Gaussian Kernel Properties

- Uses the “origin” trick

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right)$$

- Gaussian kernel maps points to a hypersphere of unit radius and the same orthant in the transformed space.
- Helps visualize the boundary for one-class SVM.
- Use a Gaussian kernel
 - Every point mapped to a unit hypersphere

$$\kappa(\mathbf{x}, \mathbf{x}) = \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle = \|\phi(\mathbf{x})\|^2 = 1$$


- Every point in the same orthant (quadrant)

$$\kappa(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \geq 0$$

- The "origin trick" serves as an approach to learning the separating hyperplane in one-class SVMs, tackling the difficulty of handling nonlinear boundaries without information about the anomaly class during training.

Requirements for Optimal Hyperplane:

Structural risk minimization principle guides the selection of the best hyperplane.



Three primary requirements:

A large margin or a small value of $\|w\|^2$ to avoid overfitting.

Hyperplane should be maximally distant from the origin, ensuring a tight representation of normal class points.

Minimization of distances of potential anomalies lying on the wrong side of the hyperplane.

Hyper-Parameter v

Role of Hyper-Parameter v :

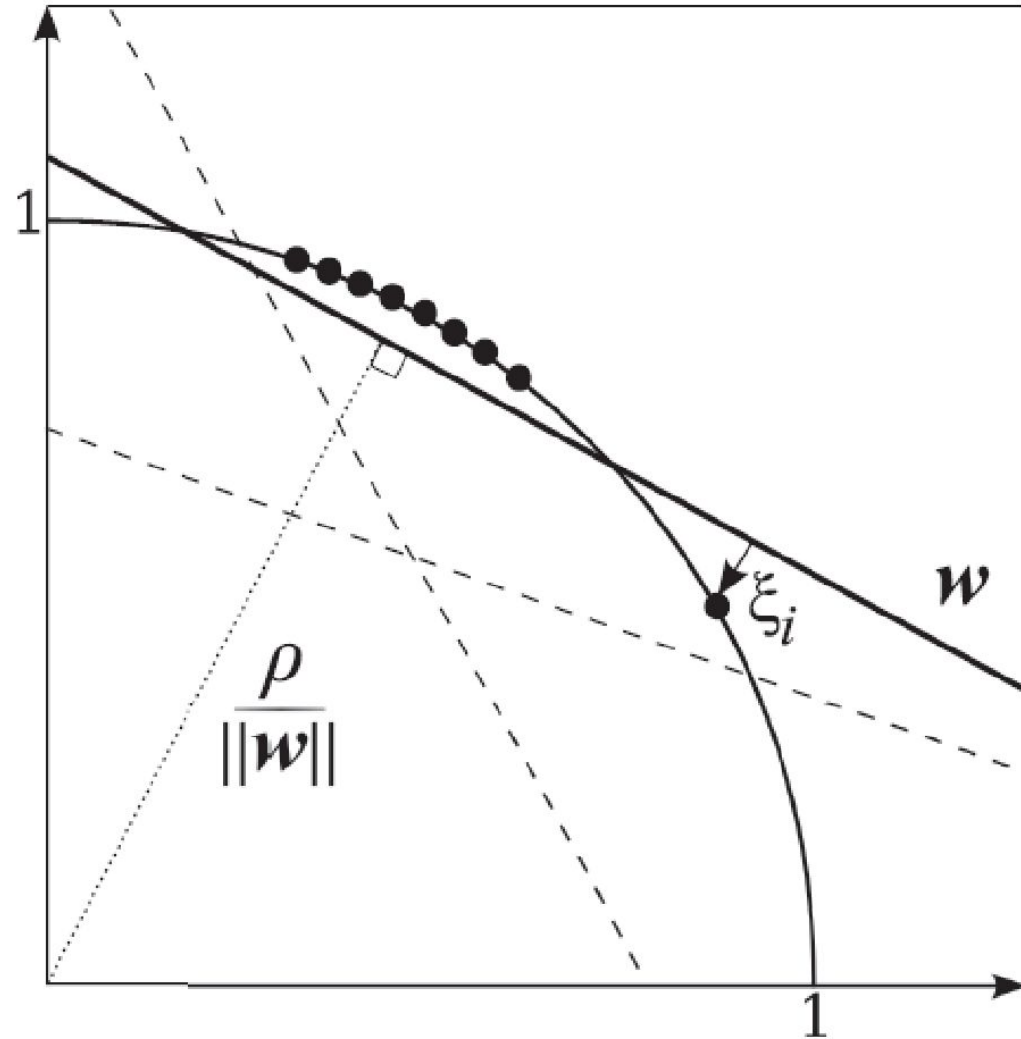
- $v \in (0, 1]$ represents an upper bound on tolerated anomalies during hyperplane learning.
- Controls the maximum number of training instances that can be considered as anomalies.
- Crucial in adapting the decision boundary to handle different numbers of outliers.

Effect of v on Decision Boundary:

- Examples with varying v values (e.g., $v = 0.1, 0.05, 0.2$) showcase how the decision boundary changes concerning outlier tolerance.
- Lower v allows more outliers, resulting in decision boundaries encompassing larger regions for the normal class.
- Higher v leads to more compact decision boundaries, as the model is less tolerant of outliers.

Two-dimensional One Class SVM

- Illustrating the concept of one-class SVM in the transformed space.
- an anomalous training instance is shown in Figure as the lower-most black dot on the quarter arc.
- If a training instance x_i lies on the opposite side of the hyperplane (corresponding to the anomaly class), its distance from the hyperplane, as measured by its slack variable ξ_i , should be kept small.
- If x_i lies on the side corresponding to the normal class, then $\xi_i = 0$.



Equations for One-Class SVM

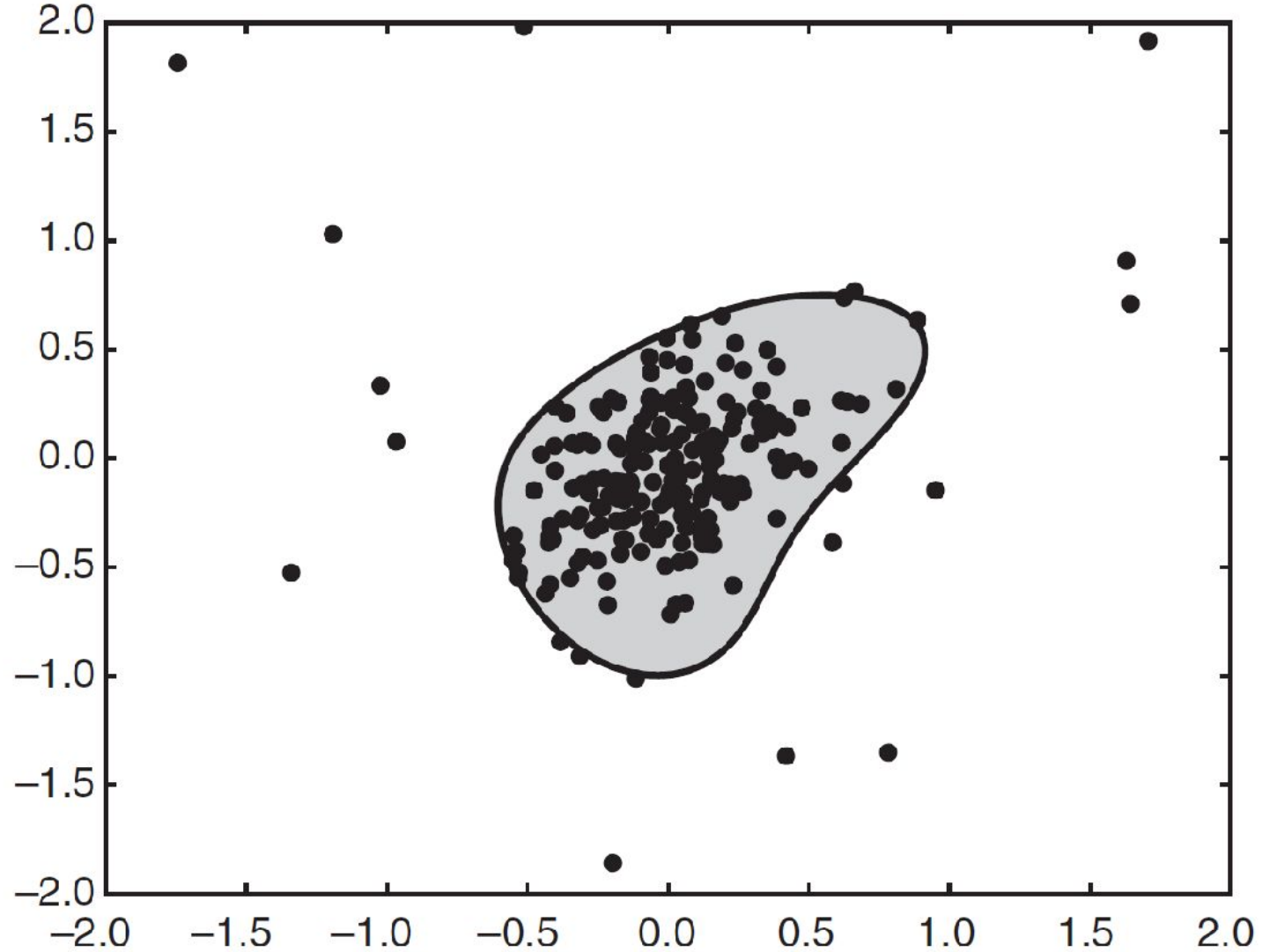
- Equation of hyperplane $\langle \mathbf{w}, \phi(\mathbf{x}) \rangle = \rho$
- ϕ is the mapping to high dimensional space
- Weight vector is $\mathbf{w} = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i)$
- ν is fraction of outliers
- Optimization condition is the following

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

$$\text{subject to: } \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0$$

Finding Outliers with a One-Class SVM

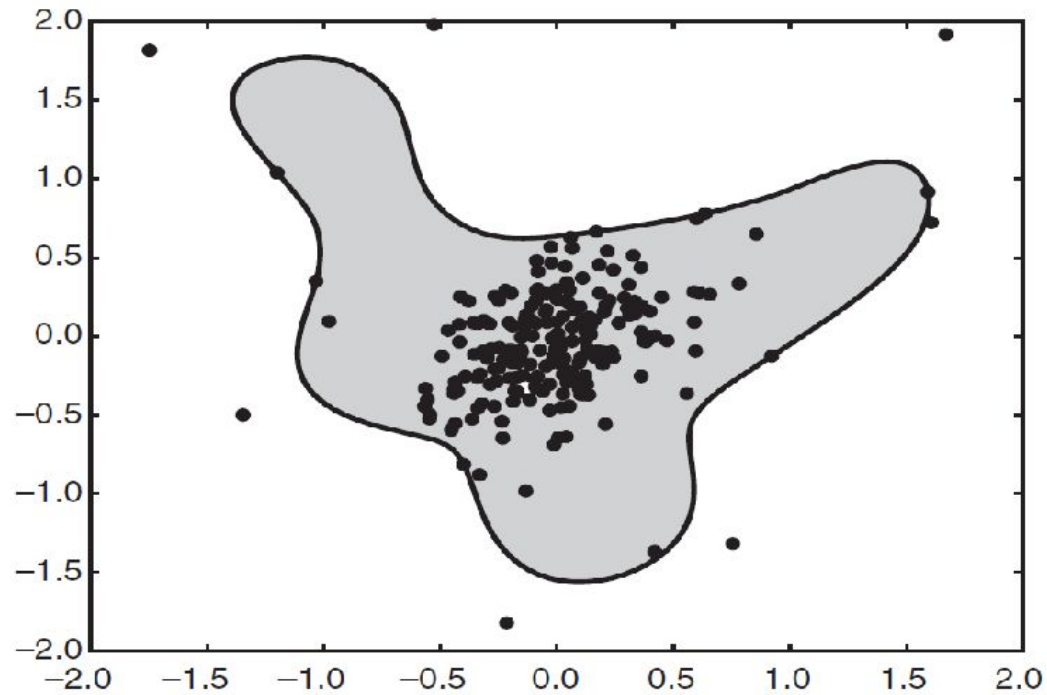
- The decision boundary of a one-class classification problem attempts to enclose the normal instances on the same side of the boundary.
- Decision boundary with $\nu = 0.1$



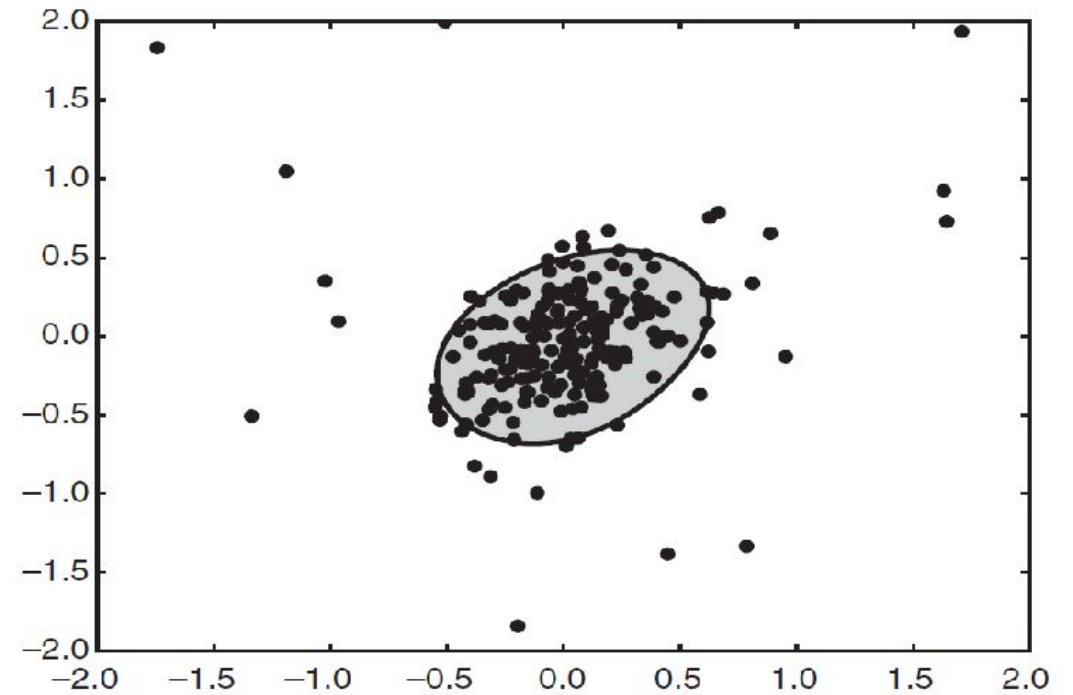
Finding Outliers with a One-Class SVM

• Decision boundaries of one-class SVM for varying values of ν

Decision boundary with $\nu = 0.05$ and $\nu = 0.2$



(a) $\nu = 0.05$.



(b) $\nu = 0.2$.

Strengths and Weaknesses

-
- Strong theoretical foundation
- Choice of v is difficult
- Computationally expensive