UNIT- V
Cluster Analysis

Scalable Data Clustering

- Many clustering algorithms work well on small data sets containing fewer than several hundred data objects;

- A large database may contain millions of objects. Clustering on a *sample* of a given large data set may lead to biased results.

- Highly scalable clustering algorithms are needed.

General strategies for scalability

- Reducing the number of proximity calculations
- Sampling the data
- Partitioning the data
- Clustering a summarized representation of the data

Algorithm: *k*-means

- The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.
- Input:
  - *k: the number of clusters,*
  - *D: a data set containing n objects.*
- Output:
  - *A set of k clusters*
- It is relatively scalable and efficient in processing large data sets because the computational complexity $O(nkt)$.

Algorithm: *k*-means

"How can we make the *k*-means algorithm more scalable?"

- A recent approach to scaling the *k*-means algorithm is based on the idea of identifying three kinds of regions in data:
  – Regions that are discardable.
  – Regions that are compressible,
  – Regions that must be maintained in main memory,

- Scaling the *k*-means algorithm by exploring the micro-clustering idea

Scalable clustering algorithms

- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

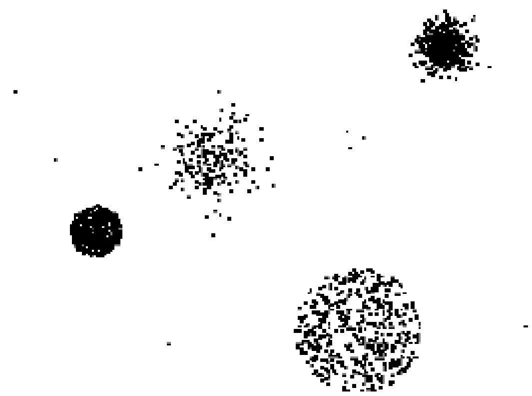
- CURE (Clustering Using Representatives).

Questions

UNIT- VI
Anomaly Detection

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

Anomaly Detection Applications

- Fraud Detection
  - purchasing behaviour of someone who steals a credit card information
- Intrusion Detection
  - attacks on computer systems to steal information
- Ecosystem Disturbance
  - floods, droughts, heat waves,and fires
- Medicine
  - For a particular patient, unusual symptoms or test results may indicate potential health problems

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 the 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

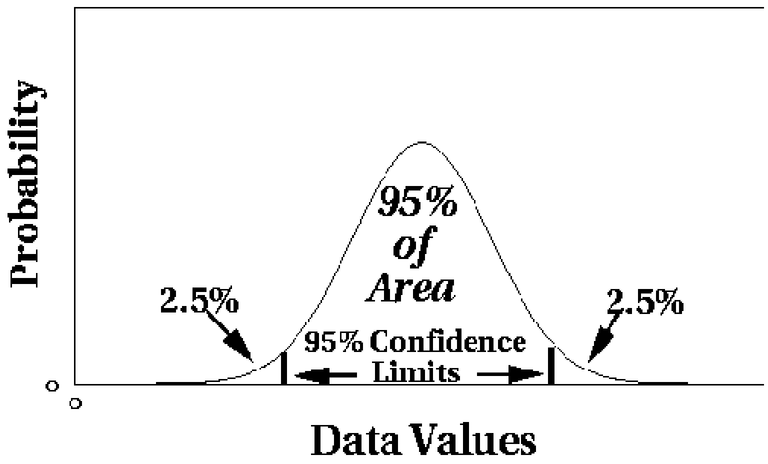
- Density-Based

Statistical Approaches

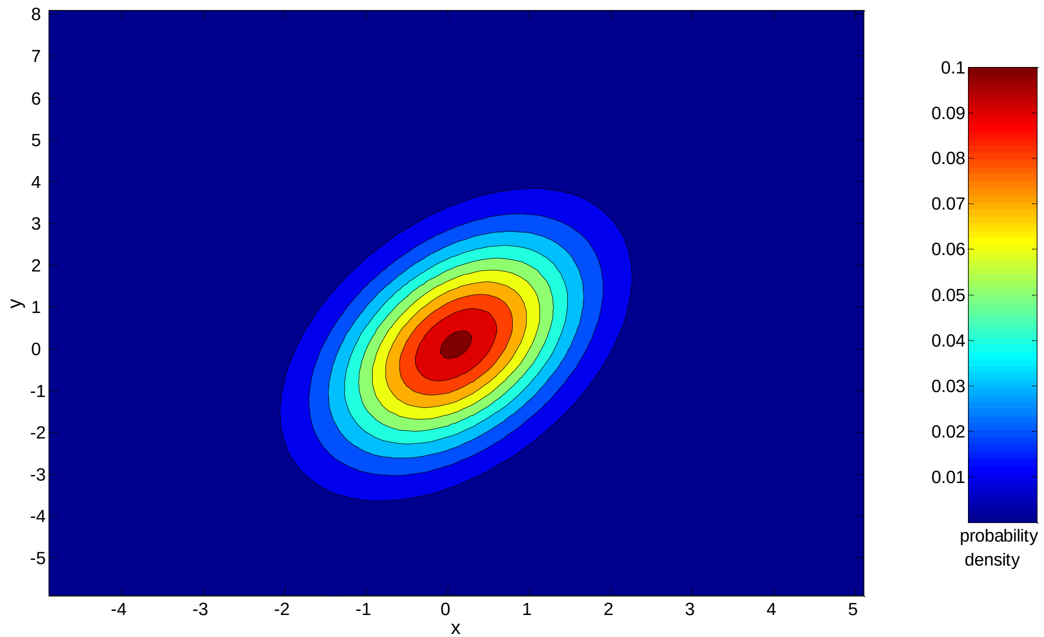
Probabilistic definition of an outlier: An outlier is an object that has a low probability with respect to a probability distribution model of the data.

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

Normal Distributions

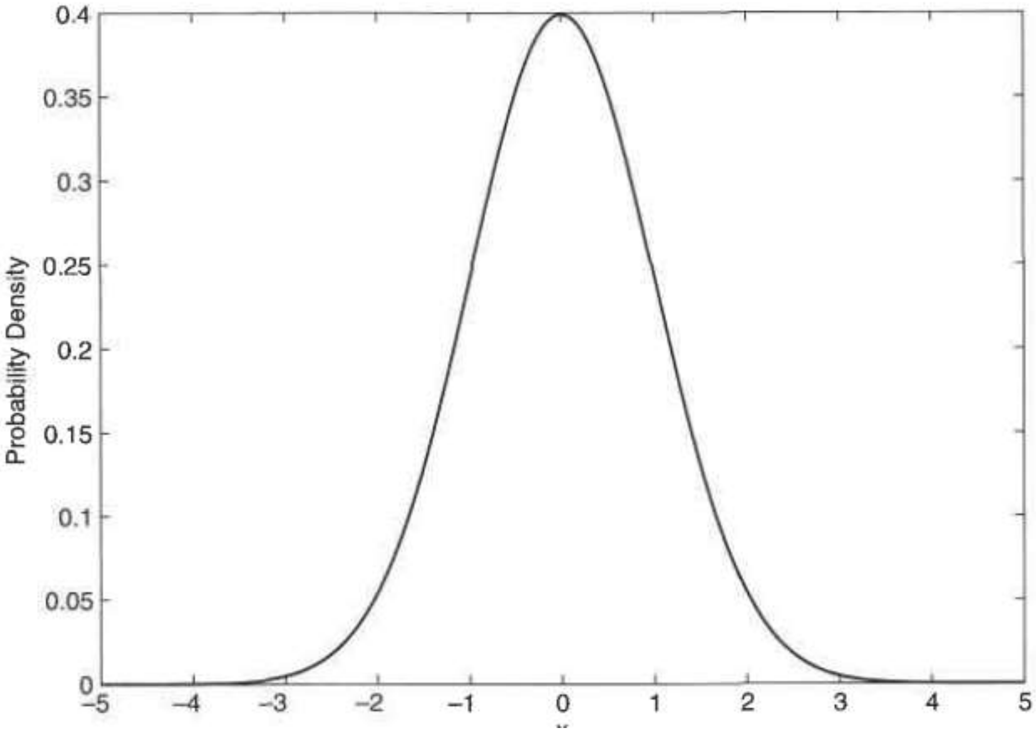


One-dimensional Gaussian



Two-dimensional Gaussian

Outlier for a Single N(0,1) Gaussian Attribute



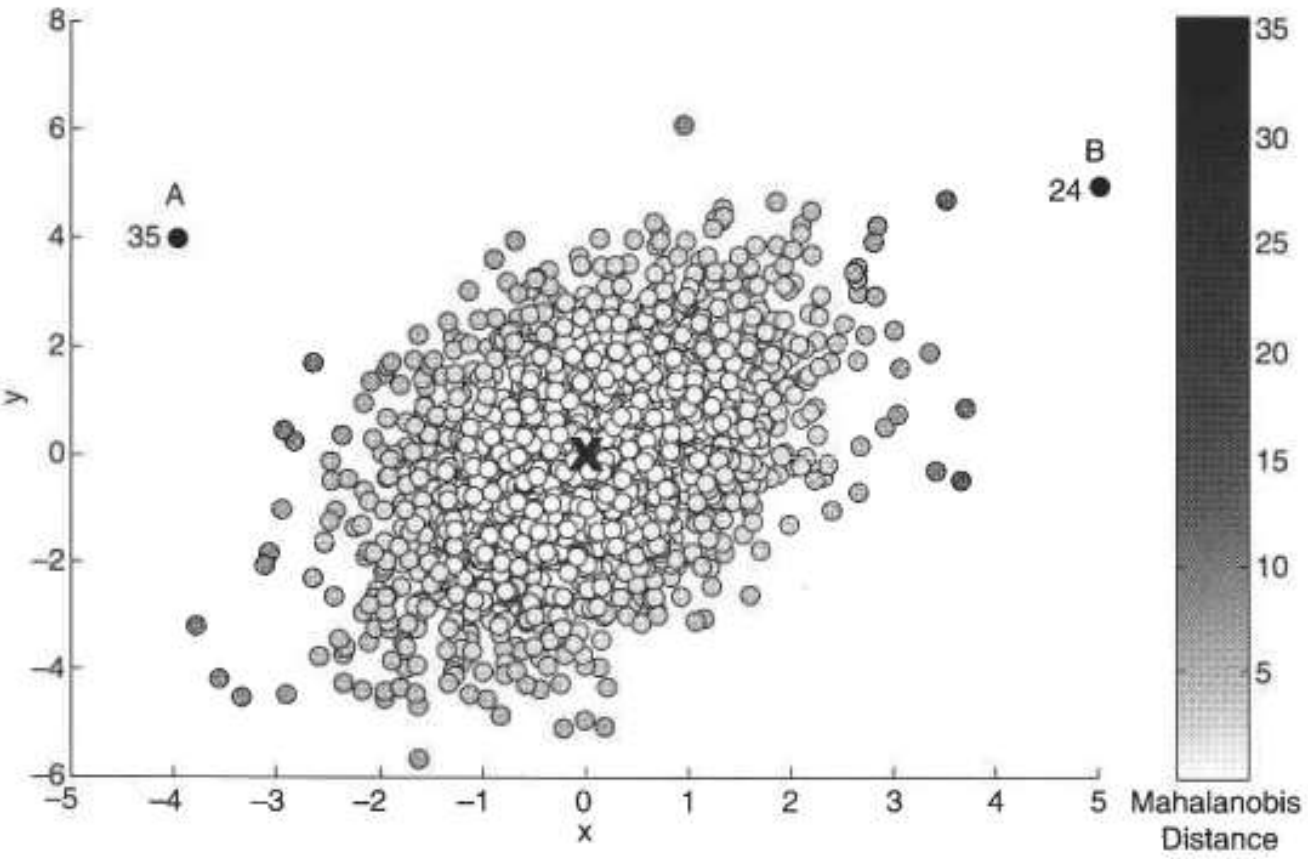
(c, α), $\alpha = prob(|x| \geq c)$

c	α for $N(0,1)$
1.00	0.3173
1.50	0.1336
2.00	0.0455
2.50	0.0124
3.00	0.0027
3.50	0.0005
4.00	0.0001

$$|x| \geq c,$$

Outliers in a Multivariate Normal Distribution

$$mahalanobis(\mathbf{x}, \overline{\mathbf{x}}) = (\mathbf{x} - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}})^T,$$



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 – λ) M + λ 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:

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

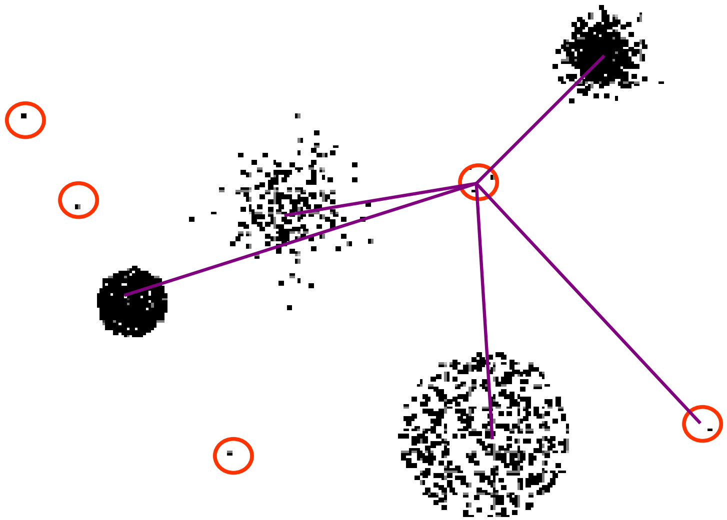
$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

Strengths/Weaknesses of Statistical Approaches

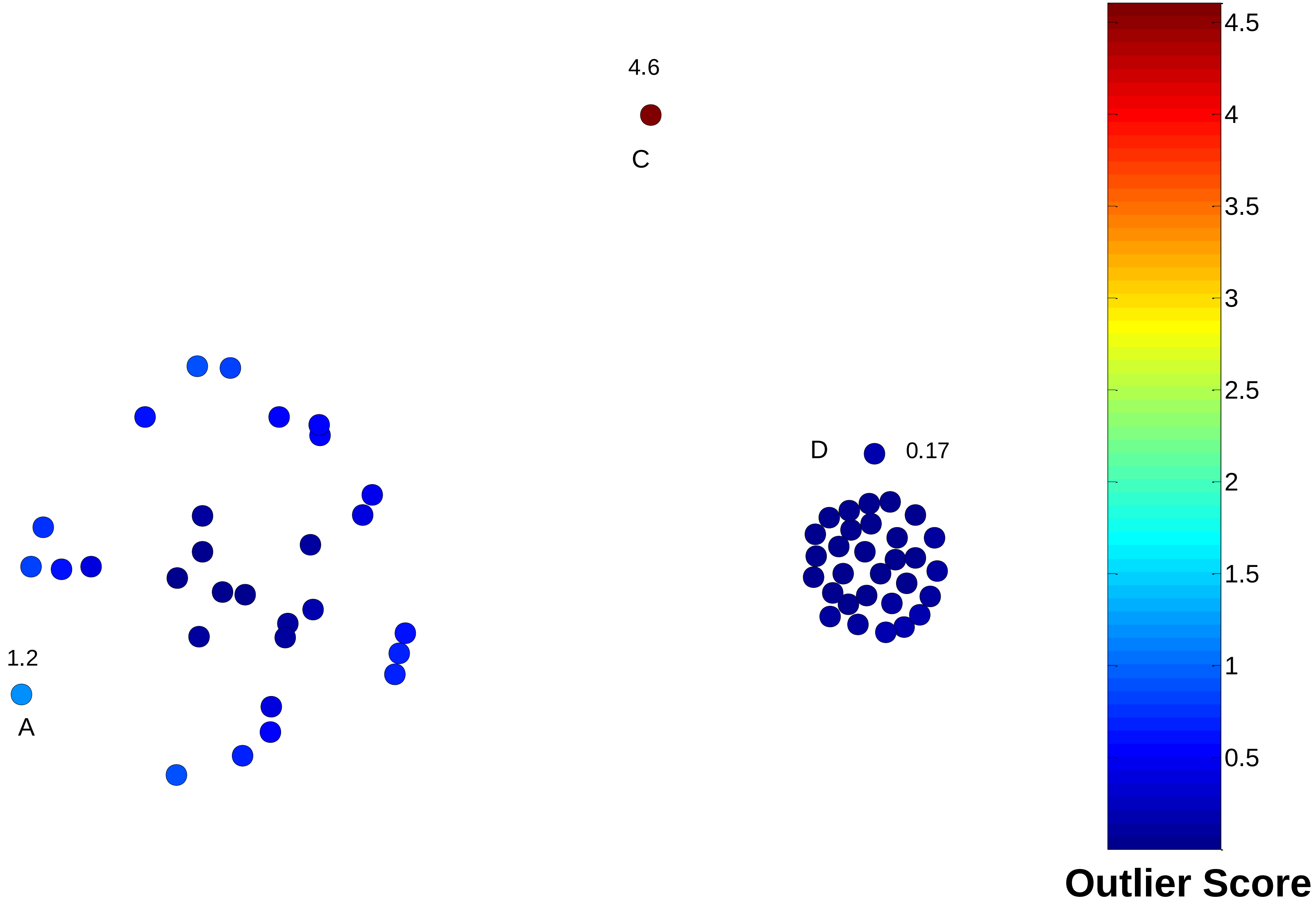
- Firm mathematical foundation
- Can be very efficient
- Good results if distribution is known
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
- Anomalies can distort the parameters of the distribution

Clustering-Based Approaches

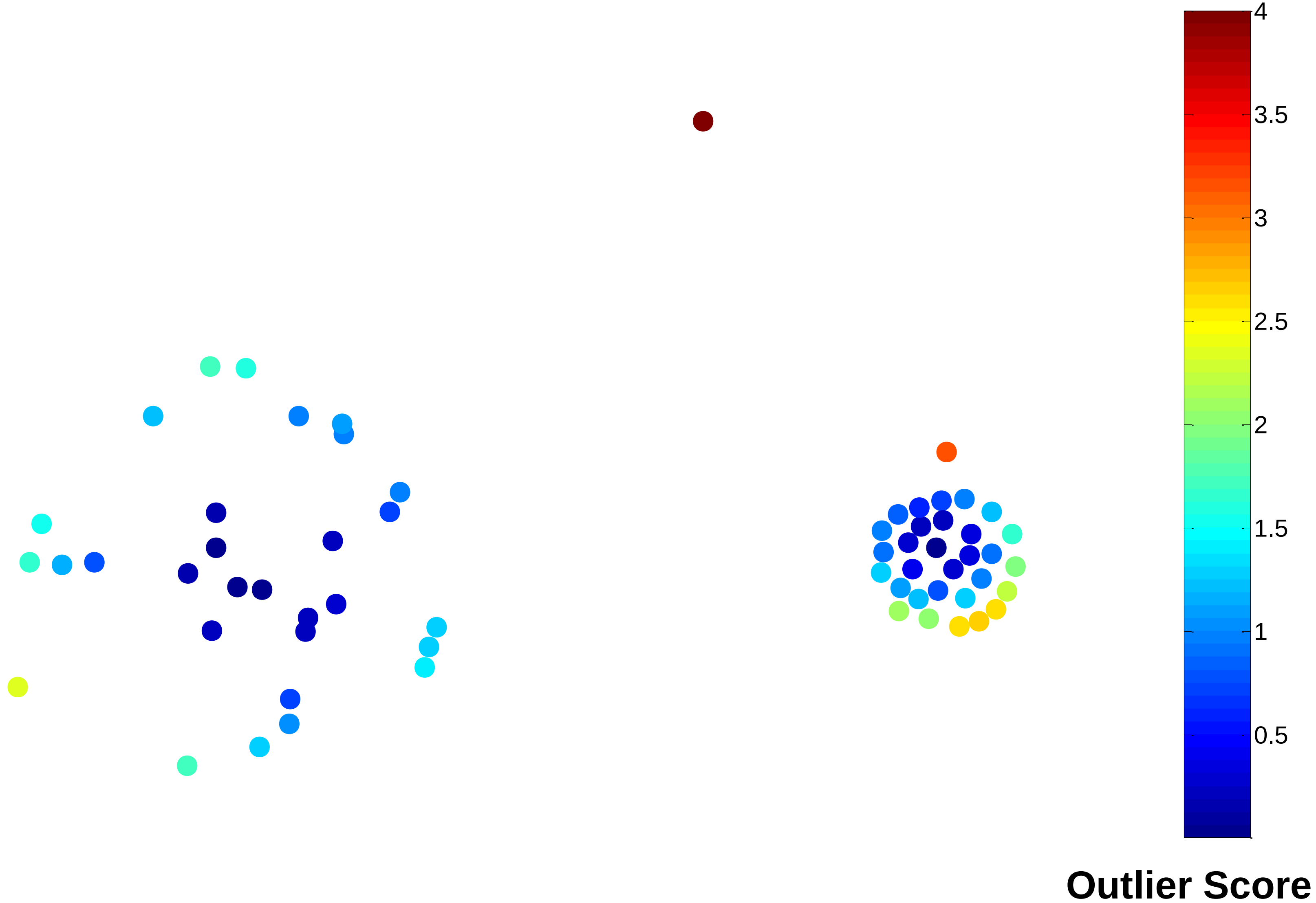
- An object is a cluster-based outlier if it does not strongly belong to any cluster
  - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
    - Outliers can impact the clustering produced
  - For density-based clusters, an object is an outlier if its 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

Evaluation of Anomaly Detection

- If class labels are present, then use standard evaluation approaches for rare class such as precision, recall, or false positive rate
  - FPR is also know as false alarm rate

- For unsupervised anomaly detection use measures provided by the anomaly method
  - E.g. reconstruction error or gain

- Can also look at histograms of anomaly scores.

Questions