Lecture 19: Anomaly Detection

COMP90049
Introduction to Machine Learning
Semester 2, 2020

Lida Rashidi, CIS

Acknowledgement: Jeremy Nicholson, Tim Baldwin & Karin Verspoor

Lecture Outline

- Anomaly Detection
  - Definition
  - Importance
  - Structure
- Anomaly Detection Algorithms
  - Statistical
  - Proximity-based
  - Density-based
  - Clustering-based
- Summary

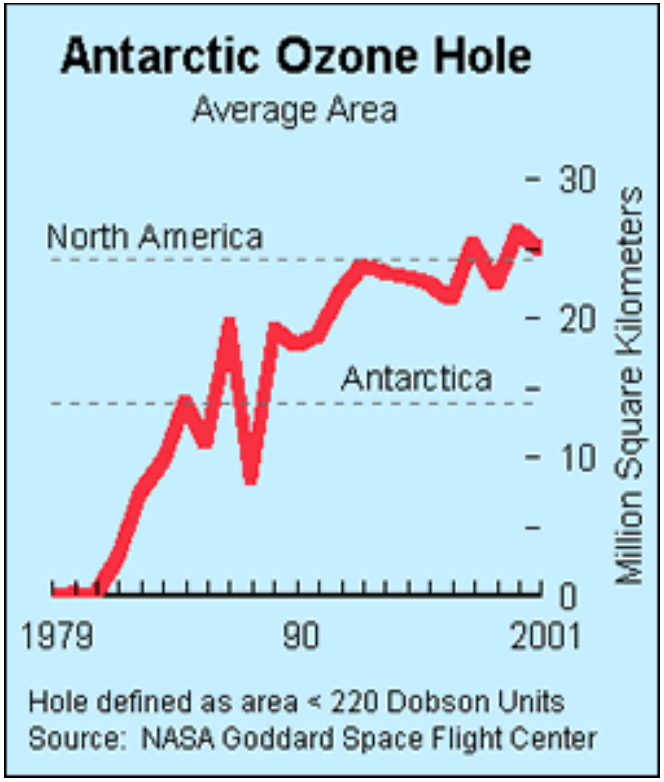
What are Outliers/Anomalies?

- **Anomaly**: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism**
  - Ex.: Unusual credit card purchase, sports: Michael Jordon, Wayne Gretzky, ...
- Anomalies are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before anomaly detection
- Anomalies are interesting:
  - They violate the mechanism that generates the normal data
  - translate to significant (often critical) real life entities
    - Cyber intrusions
    - Credit card fraud

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 noise by a computer program and discarded!



Variants of Anomaly Detection Problem

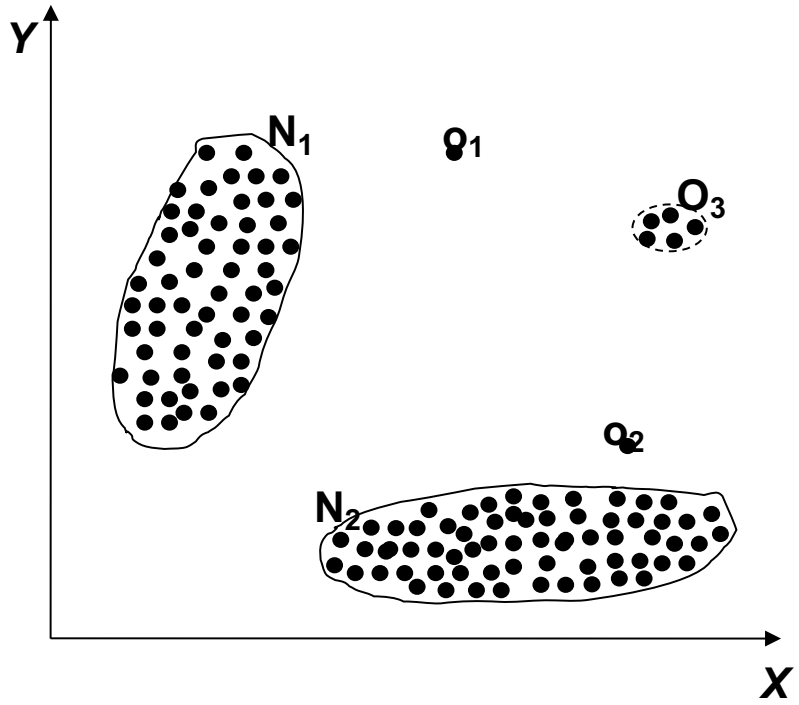
- Variants of Anomaly/Outlier Detection Problems
  - Given a database D, find all the data points $\mathbf{x} \in D$ with anomaly scores greater than some threshold t
  - Given a database D, find all the data points $\mathbf{x} \in D$ having the top-n largest anomaly scores f($\mathbf{x}$)
  - Given a database D, containing mostly normal (but unlabeled) data points, and a test point $\mathbf{x}$, compute the anomaly score of $\mathbf{x}$ with respect to D

Structure of Anomalies

- Global/Point anomalies
- Contextual/Conditional anomalies
- Collective anomalies

Global/Point anomalies

- **Global Anomaly** (or point)
  - Object is O_g if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation



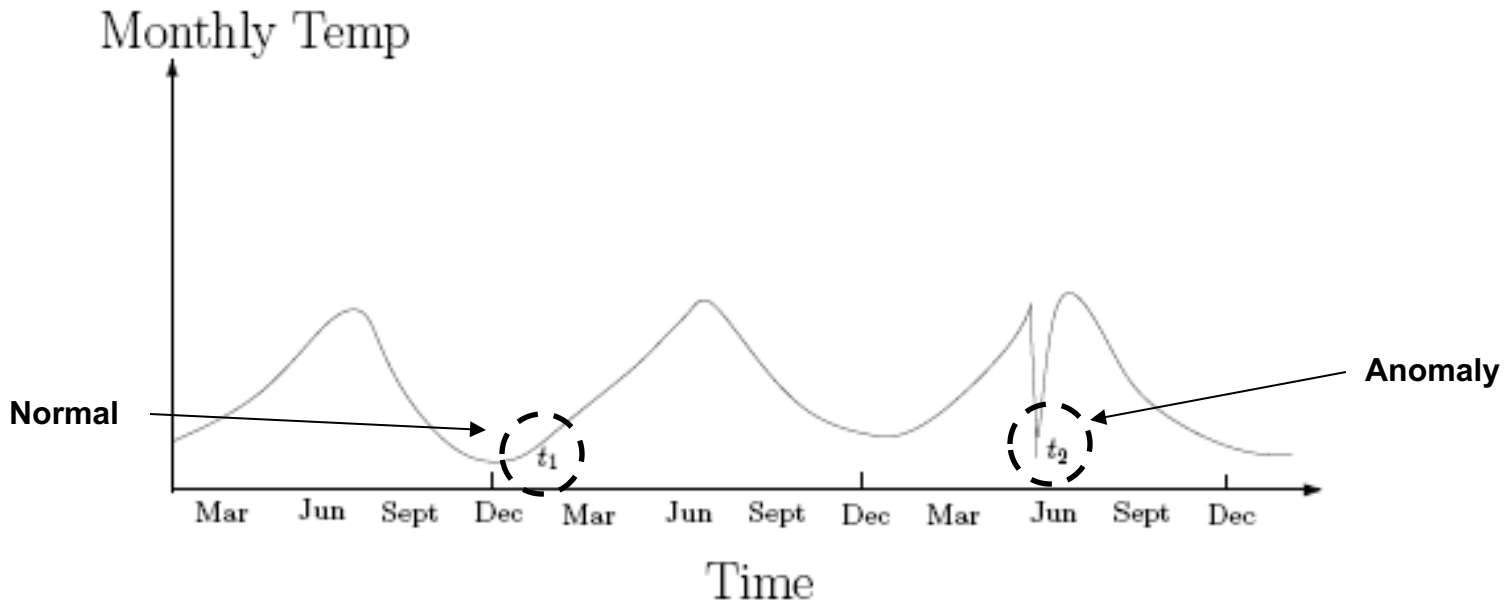
Contextual anomalies

- **Contextual Anomaly** (or *conditional*)
  - Object is O_c if it deviates significantly based on a selected context
  - Requires a notion of context
  - Attributes of data objects should be divided into two groups
    - Contextual attributes: defines the context, e.g., time & location
    - Behavioral attributes: characteristics of the object, used in anomaly evaluation, e.g., temperature
  - Can be viewed as a generalization of *local anomalies*—whose density significantly deviates from its local area
  - Issue: How to define or formulate meaningful context?

* Song, et al, “Conditional Anomaly Detection”, IEEE Transactions on Data and Knowledge Engineering, 2006.

Example of Contextual Anomalies

- Ex. 10^{o} C in Paris: anomaly? (depending on summer or winter?)

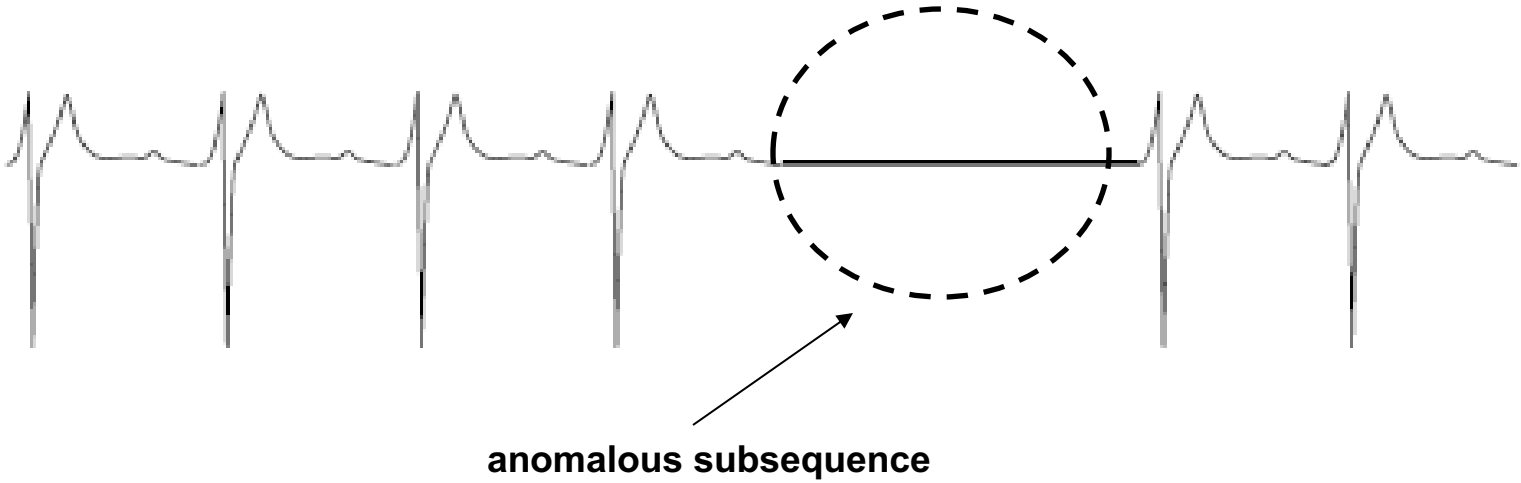


Collective anomalies

- **Collective Anomalies**
  - A subset of data objects that *collectively* deviate significantly from the whole data set, even if the individual data objects may not be anomalies
  - Applications: E.g., *intrusion detection*:
    - When a number of computers keep sending denial-of-service packages to each other
  - Detection of collective anomalies
    - Consider not only behavior of individual objects, but also that of groups of objects
    - Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects.

Example of Collective anomalies

- Requires a relationship among data instances
  - Sequential data
  - Spatial data
  - Graph data
- The individual instances within a collective anomaly are not anomalous by themselves



Supervised Anomaly Detection

- Supervised anomaly detection
  - Labels available for both normal data and anomalies
  - Samples examined by domain experts used for training & testing
  - Challenges
    - Require both labels from both normal and anomaly class
    - Imbalanced classes, i.e., anomalies are rare: Boost the anomaly class and make up some artificial anomalies
    - Cannot detect unknown and emerging anomalies
    - Catch as many outliers as possible, i.e., recall is more important than accuracy (i.e., not mislabeling normal objects as outliers)

Semi-supervised Anomaly Detection

- Semi-Supervised anomaly detection
  - Labels available only for normal data
  - Model normal objects & report those not matching the model as outliers
  - Challenges
    - Require labels from normal class
    - Possible high false alarm rate - previously unseen (yet legitimate) data records may be recognized as anomalies

Unsupervised Anomaly Detection I

- Unsupervised anomaly detection
  - Assume the normal objects are somewhat ``clustered'' into multiple groups, each having some distinct features
  - An outlier is expected to be far away from any groups of normal objects
- General steps
  - Build a profile of “normal” behavior
    - summary statistics for overall population
    - model of multivariate data distribution
  - Use the “normal” profile to detect anomalies
    - anomalies are observations whose characteristics differ significantly from the normal profile

Unsupervised Anomaly Detection II

- Unsupervised anomaly detection Challenges
  - Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area
  - Ex. In some intrusion or virus detection, normal activities are diverse
    - Unsupervised methods may have a high false positive rate but still miss many real outliers.
- Many clustering methods can be adapted for unsupervised methods
  - Find clusters, then outliers: not belonging to any cluster
  - Problem 1: Hard to distinguish noise from outliers
  - Problem 2: Costly since first clustering: but far less outliers than normal objects

Unsupervised point anomaly detection

- Statistical
  - Statistical methods (also known as model-based methods) assume that the normal data follow some statistical model (a stochastic model)
- Proximity-based
  - An object is an outlier if the nearest neighbors of the object are far away
- Density-based
  - Outliers are objects in regions of low density
- Clustering-based
  - Normal data belong to large and dense clusters

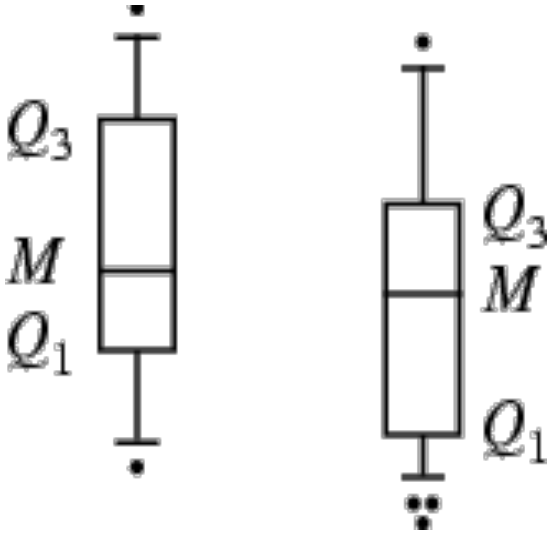
Statistical anomaly detection

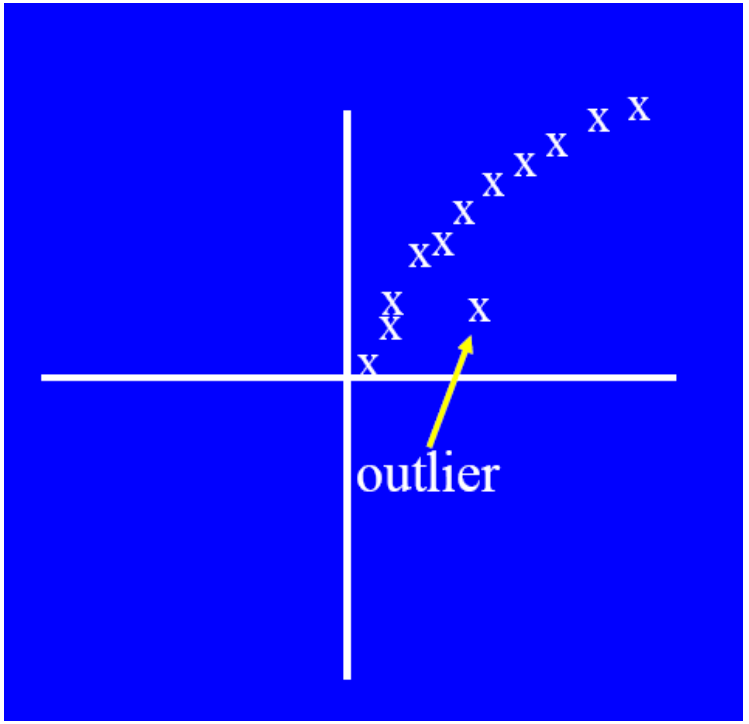
Anomalies are objects that are fit poorly by a statistical model.

- **Idea**: learn a model fitting the given data set, and then identify the objects in low probability regions of the model as anomalies
- **Assumption**: normal data is generated by a parametric distribution with parameter θ
  - The probability density function of the parametric distribution $f(x, \theta)$ gives the probability that object x is generated by the distribution
  - The smaller this value, the more likely x is an outlier
- Challenges of Statistical testing:
  - highly depends on whether the assumption of statistical model holds in the real data

Visualizing the data

- Graphical Approaches
  - Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)
  - Limitations
    - Time consuming
    - Subjective





Univariate Data – General Approach

- Avg. temp.: {24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4}
  - Use the maximum likelihood method to estimate μ and σ

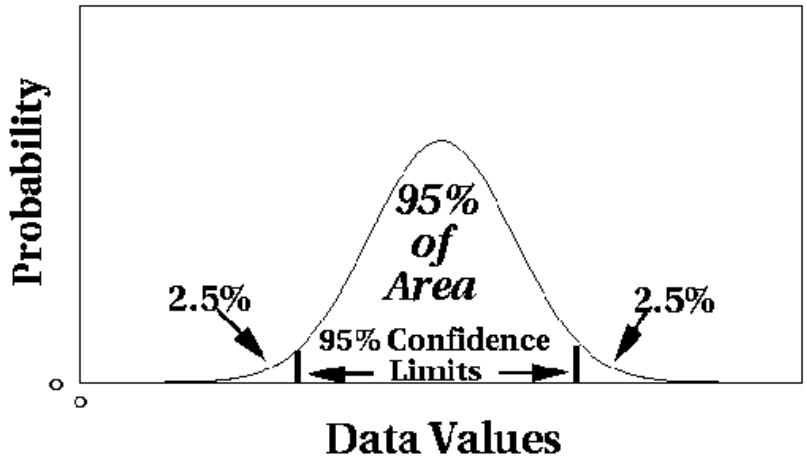
$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  - For the above data with n = 10, we have:

$$\hat{\mu} = 28.61 \;\; \hat{\sigma} = \sqrt{2.29} = 1.51$$

  - Then 24 is an outlier since:
    - (24 – 28.61) /1.51 = – 3.04 < –3

$\mu \pm 3\sigma$ region contains 99.7% data



Grubbs' test – Univariate Data

- Detect outliers in univariate data
  - Univariate data: A data set involving only one attribute or variable
- Assumption: data is generated from a **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:

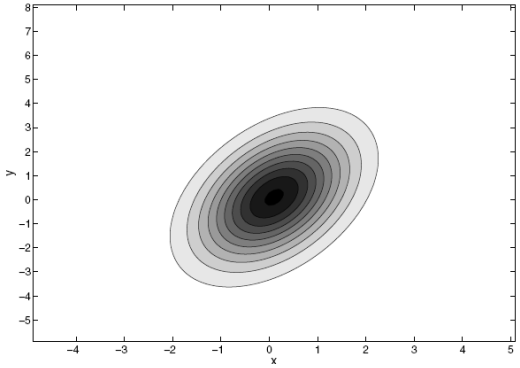
$$\mathbf{z} = \frac{|\mathbf{x} - \bar{\mathbf{x}}|}{s}$$

- Reject H_0 if:

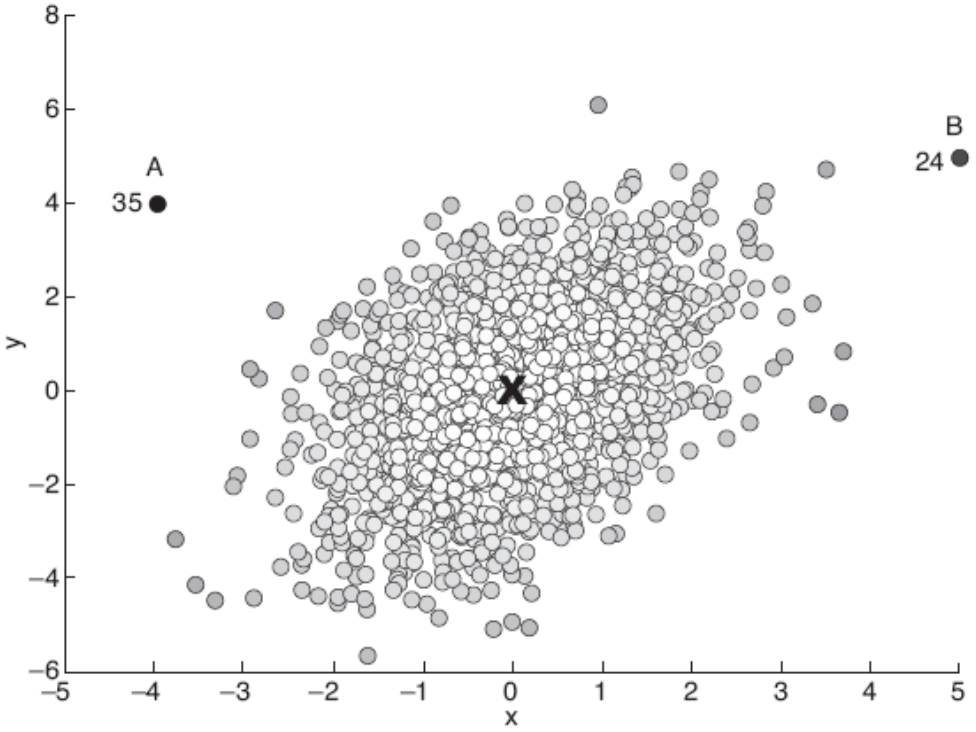
$$\mathbf{z} > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

Multivariate Data

- Multivariate Gaussian distribution
  - Outlier defined by Mahalanobis distance
  - Grubb's test on the distances



	Distance	
	Euclidean	Mahalanobis
A	5.7	35
B	7.1	24



Mahalanobis Distance

- Mahalanobis Distance

$$y^2 = (\mathbf{x} - \bar{\mathbf{x}})' S^{-1} (\mathbf{x} - \bar{\mathbf{x}})$$

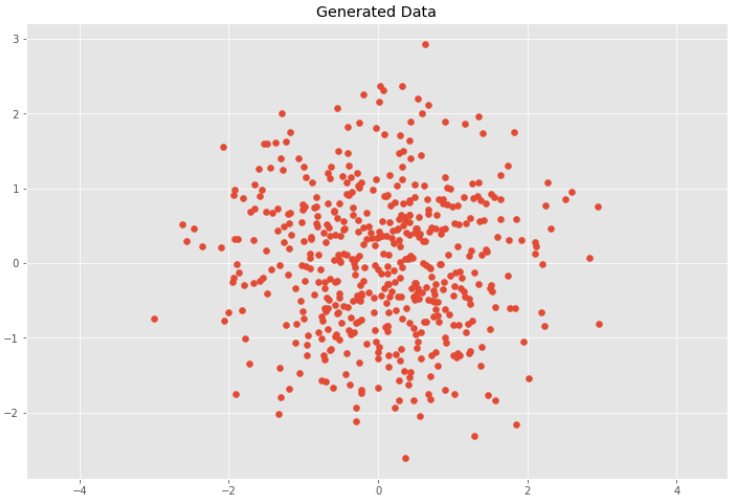
- S is the covariance matrix:

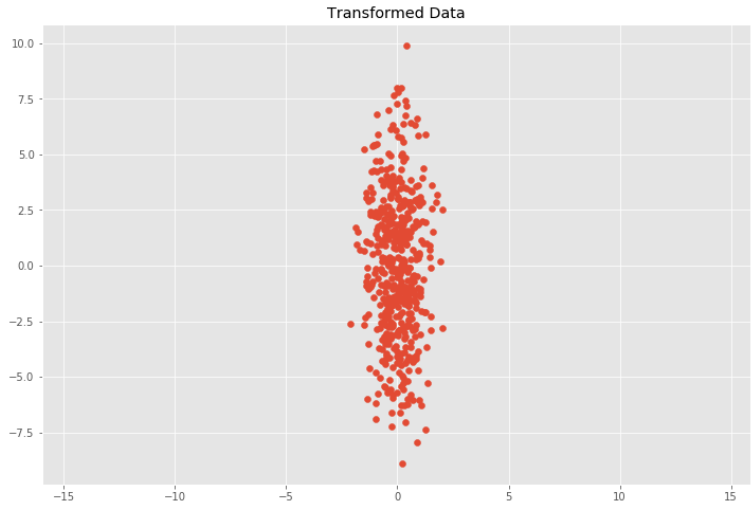
$$S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$$

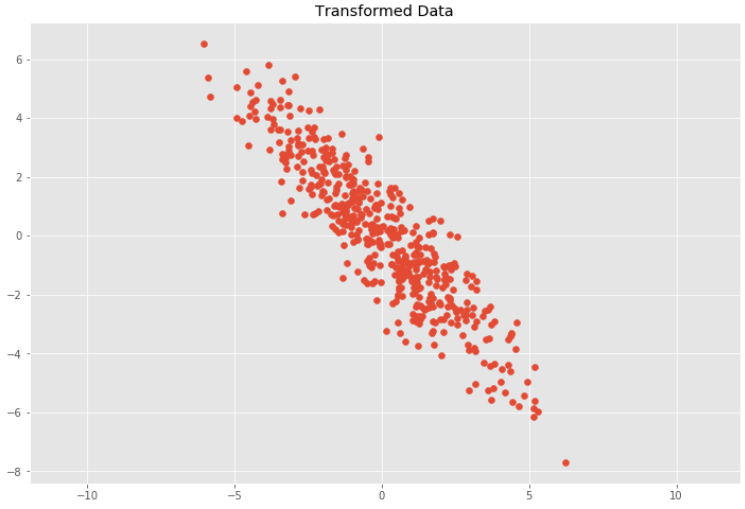
- For a 2-dimensional data:

$$\begin{pmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{pmatrix}$$

Mahalanobis Distance







Likelihood approach

- Assume the dataset 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

Likelihood approach

- Data distribution, D = (1 – λ) M + λ A
- M is a probability distribution estimated from data
- A is initially assumed to be uniform distribution
- Likelihood at time t:

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

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

Statistical Anomaly detection

- Pros
  - Statistical tests are well-understood and well-validated.
  - Quantitative measure of degree to which object is an outlier.
- Cons
  - Data may be hard to model parametrically.
    - multiple modes
    - variable density
  - In high dimensions, data may be insufficient to estimate true distribution.

Proximity-based Anomaly detection

Anomalies are objects far away from other objects.

- An object is an **anomaly** if the nearest neighbors of the object are **far** away, i.e., the **proximity** of the object significantly deviates from the proximity of most of the other objects in the same data set

- Common approach:
  - Outlier score is distance to k^{th} nearest neighbor.
  - Score sensitive to choice of k.

Nearest-Neighbor Based Approach

- Approach:
  - Compute the distance between every pair of data points
  - There are various ways to define outliers:
    - Data points for which there are fewer than *p* neighboring points within a distance *D*
    - The top n data points whose distance to the kth nearest neighbor is greatest
    - The top n data points whose average distance to the k nearest neighbors is greatest

Proximity-based anomaly detection

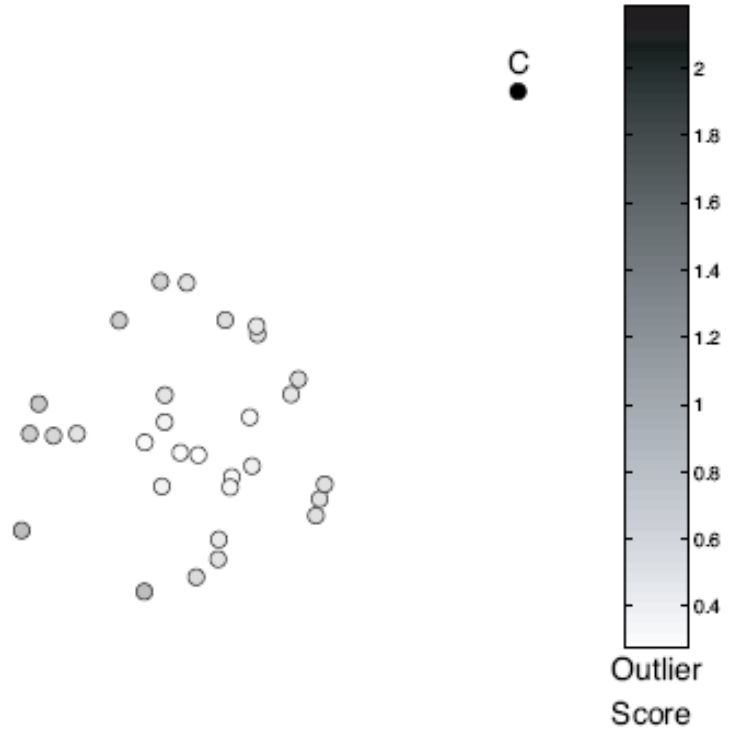


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

Proximity-based anomaly detection

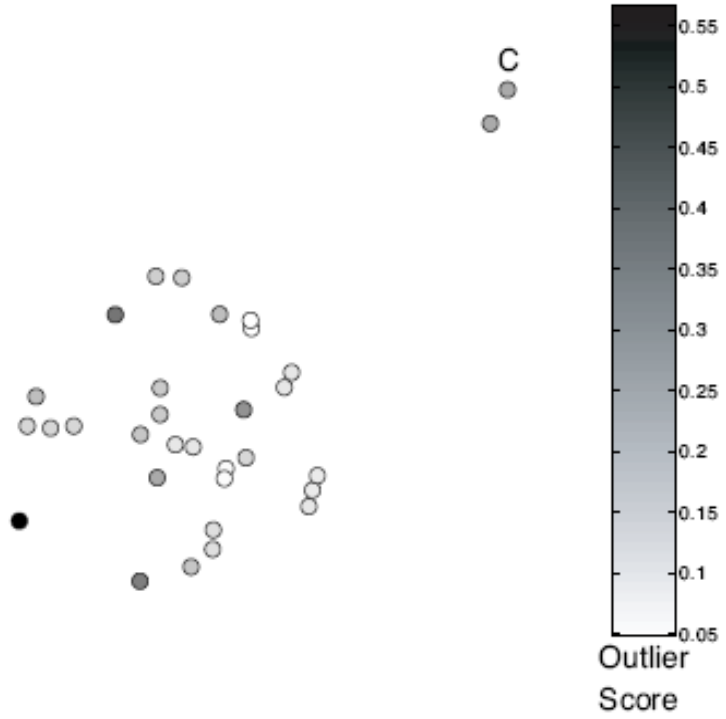


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

Proximity-based anomaly detection

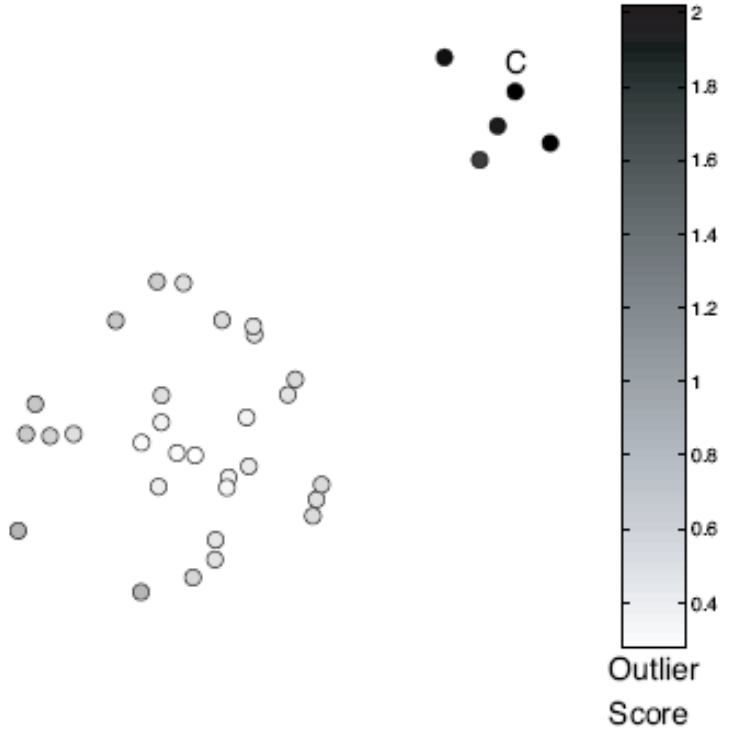


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

Proximity-based anomaly detection

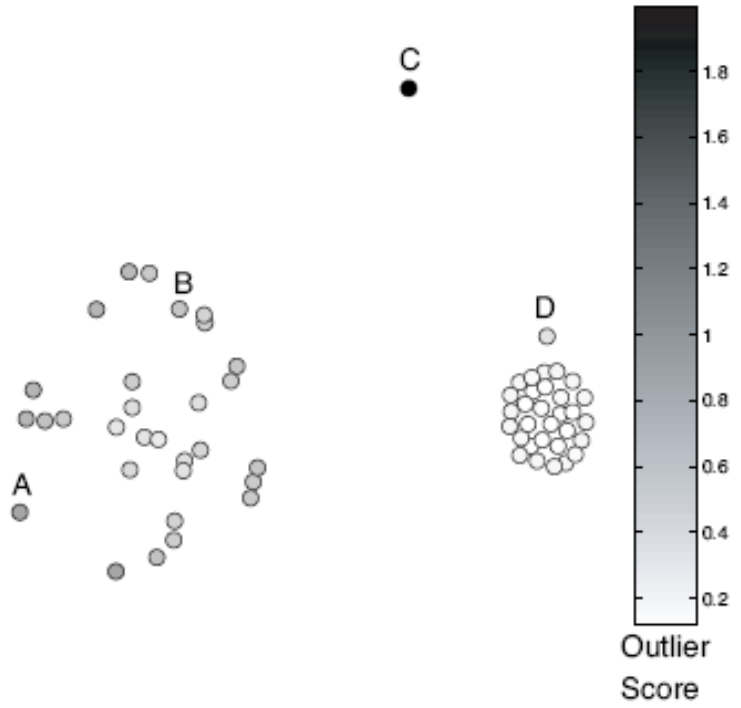


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

Proximity-based outlier detection

- Pros
  - Easier to define a proximity measure for a dataset than determine its statistical distribution.
  - Quantitative measure of degree to which object is an outlier.
  - Deals naturally with multiple modes.

- Cons
  - $O(n^2)$ complexity.
  - Score sensitive to choice of k.
  - Does not work well if data has widely variable density.

Density-based outlier detection

Outliers are objects in regions of low density.

- Outlier score is inverse of density around object.
- Scores usually based on proximities.
- Example scores:
  - Reciprocal of average distance to *k* nearest neighbors:

$$\text{density}(\mathbf{x}, k) = \left(\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x},k)} \text{distance}(\mathbf{x}, \mathbf{y}) \right)^{-1}$$

  - Number of objects within fixed radius *d*.
  - These two example scores work poorly if data has variable density.

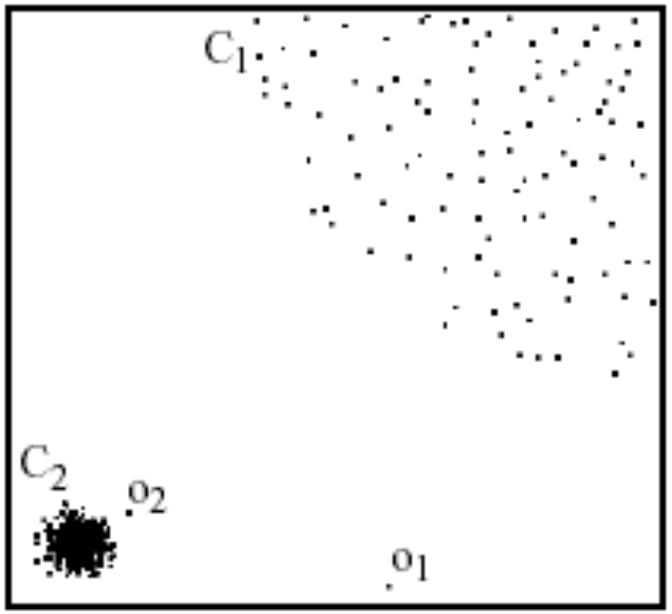
Density-based outlier detection

- Relative density outlier score (Local Outlier Factor, LOF)
  - Reciprocal of average distance to *k* nearest neighbors, relative to that of those *k* neighbors.

$$\text{relative density}(\mathbf{x}, k) = \frac{\text{density}(\mathbf{x}, k)}{\frac{1}{k}\sum_{\mathbf{y} \in N(\mathbf{x}, k)} \text{density}(\mathbf{y}, k)}$$

Density-based outlier detection

In the NN approach, o2 is not considered as outlier, while LOF approach find both o1 and o2 as outliers!



Density-based outlier detection

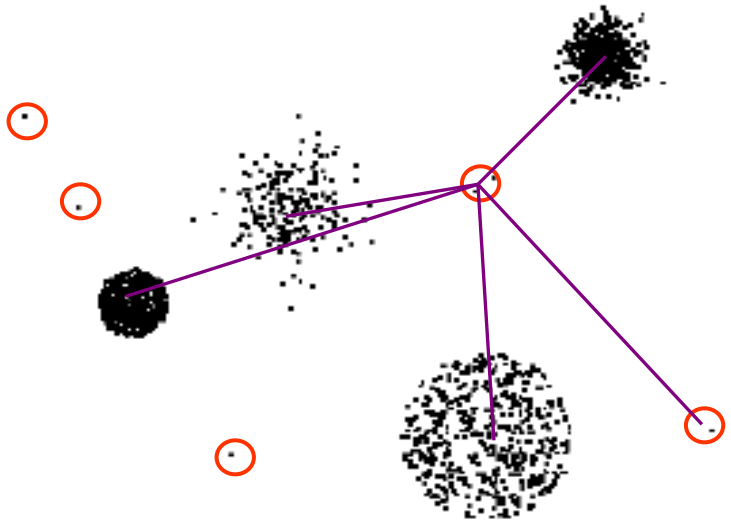
- Pros
  - Quantitative measure of degree to which object is an outlier.
  - Can work well even if data has variable density.

- Cons
  - $O(n^2)$ complexity
  - Must choose parameters
    - k for nearest neighbor
    - d for distance threshold

Cluster-based outlier detection

Outliers are objects that do not belong strongly to any cluster.

- Approaches:
  - Assess degree to which object belongs to any cluster.
  - Eliminate object(s) to improve objective function.
  - Discard small clusters far from other clusters.
- Issue:
  - Outliers may affect initial formation of clusters.



Cluster-based outlier detection

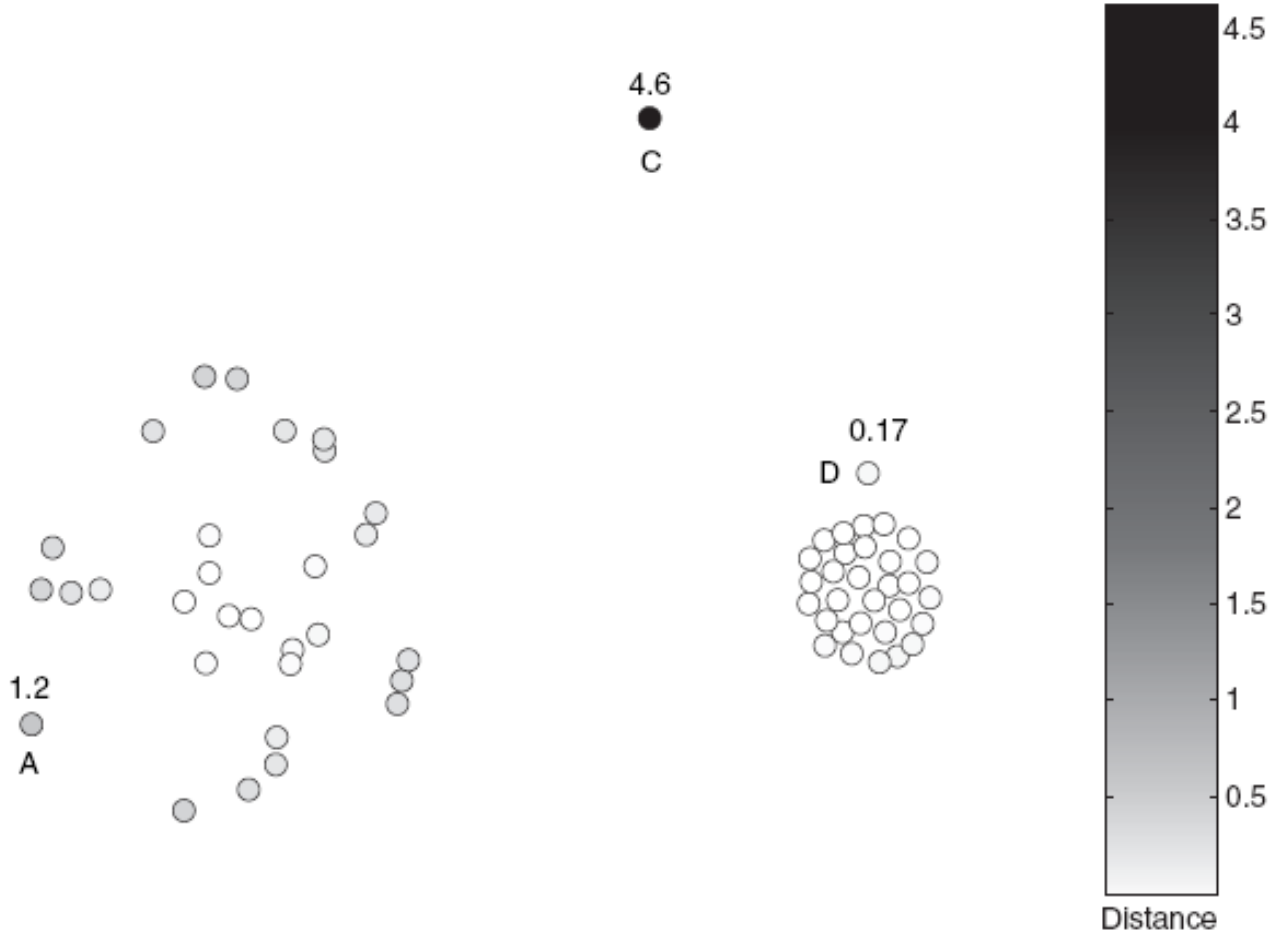
Assess degree to which object belongs to any cluster.

- For prototype-based clustering (e.g. k-means), use distance to cluster centers.
  - To deal with variable density clusters, use relative distance:

$$\frac{\text{distance}(\mathbf{x}, centroid_C)}{\text{median}\left(\left\{\forall_{x' \in C}\ \text{distance}(\mathbf{x}', centroid_C)\right\}\right)}$$

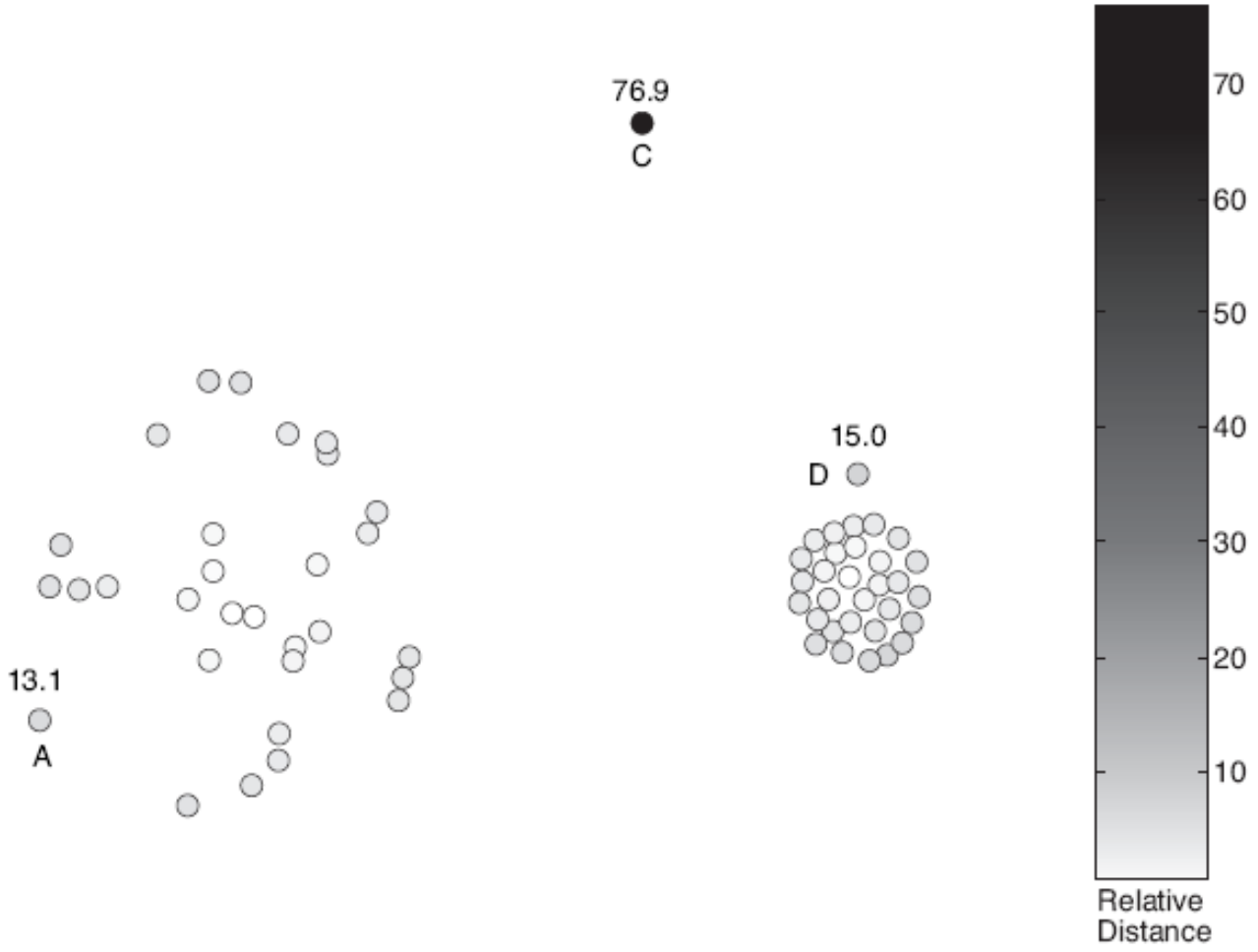
- Similar concepts for density-based or connectivity-based clusters.

Cluster-based outlier detection



distance of points from nearest centroid

Cluster-based outlier detection



relative distance of points from nearest centroid

Cluster-based outlier detection

Eliminate object(s) to improve objective function.

1) Form initial set of clusters.
2) Remove the object which most improves objective function.
3) Repeat step 2) until …

Discard small clusters far from other clusters.

- Need to define thresholds for “small” and “far”.

Cluster-based outlier detection

- Pro:
  - Some clustering techniques have $O(n)$ complexity.
  - Extends concept of outlier from single objects to groups of objects.

- Cons:
  - Requires thresholds for minimum size and distance.
  - Sensitive to number of clusters chosen.
  - Hard to associate outlier score with objects.
  - Outliers may affect initial formation of clusters.

Summary

- Types of outliers
  - global, contextual & collective outliers
- Outlier detection
  - supervised, semi-supervised, or unsupervised
- Statistical (or model-based) approaches
- Proximity-base approaches
- Clustering-base approaches

References

- *Tan et al (2006) Introduction to Data Mining. Section 4.3, pp 150-171. (Chapter 10)*
- V. Chandola, A. Banerjee, and V. Kumar, (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1-58.
- A. Banerjee, et al (2008). Tutorial session on anomaly detection. The SIAM Data Mining Conference (SDM08)