UNIVERSITÀ DEGLI STUDI DI TRIESTE

MACHINE LEARNING LAB

PoAC for Outlier Detection

Introduction

Simone Cappiello

Literature Review on Denoising in Clustering:

- What is noise?
- What types are there?
- Which metrics are used?
- Are there solutions to remove noise?

What is noise?

- **Noise** refers to **unwanted** or **irrelevant information** within a dataset, including **duplicates**, **outliers**, and **random or systematic errors**.

- It complicates the discernment of true underlying trends or relationships, leading to potential inaccuracies in data interpretation and analysis outcomes.

- **Data cleaning** deals with removing noise from data.

Types of data cleaning tasks

- **Deduplication:** refers to the process of identifying tuples in one or more relations that refer to the same world entity.
- **Data Transformation**: refers to the task of transforming data from one format to another.
- **Outlier Detection**: it consists in detecting “outlying” values (definition in next slide)

Clustering algorithms excel at outlier detection by effectively identifying data points that don't align with any cluster, making them standalone solutions for this task.
For deduplication and data transformation, though, clustering aids in the process but requires additional methods for comprehensive solutions. Given the PoAC framework's focus on leveraging clustering for specific challenges, our analysis will focus on outlier detection.

Outliers

“an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”

Hawkins, 1980

“an outlier observation is one that appears to deviate markedly from other members of the sample in which it occurs”

Barnett and Lewis, 1994

Types of Outliers

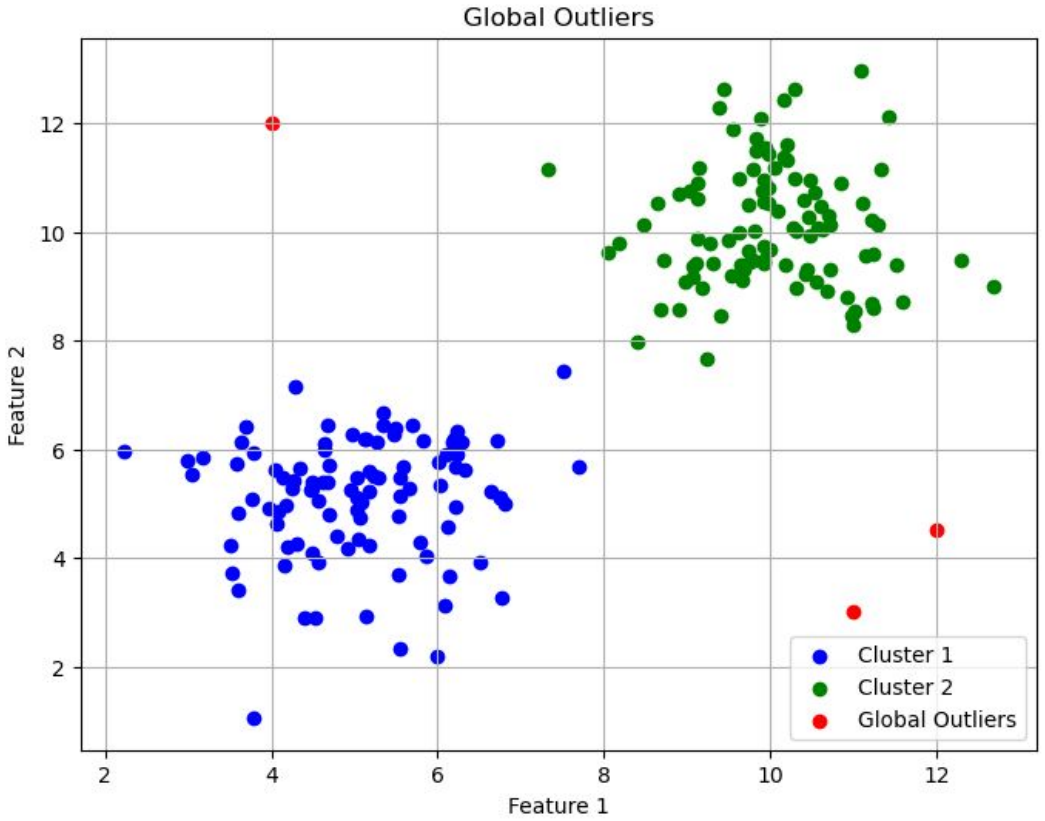
- **Global Outliers (Point Anomalies):** Individual data points that significantly deviate from the rest of the dataset. Easily identifiable as they don't fit the pattern of any data cluster.

- **Contextual Outliers (Conditional Anomalies):** Data points that deviate significantly based on a specific context or condition. Unusual for a particular time or location, but might not be outliers in a different context.

- **Collective Outliers:** Subsets of data points that collectively deviate from the dataset's overall pattern. Individual points might not be outliers, but their anomalous nature is noticeable when evaluated together.

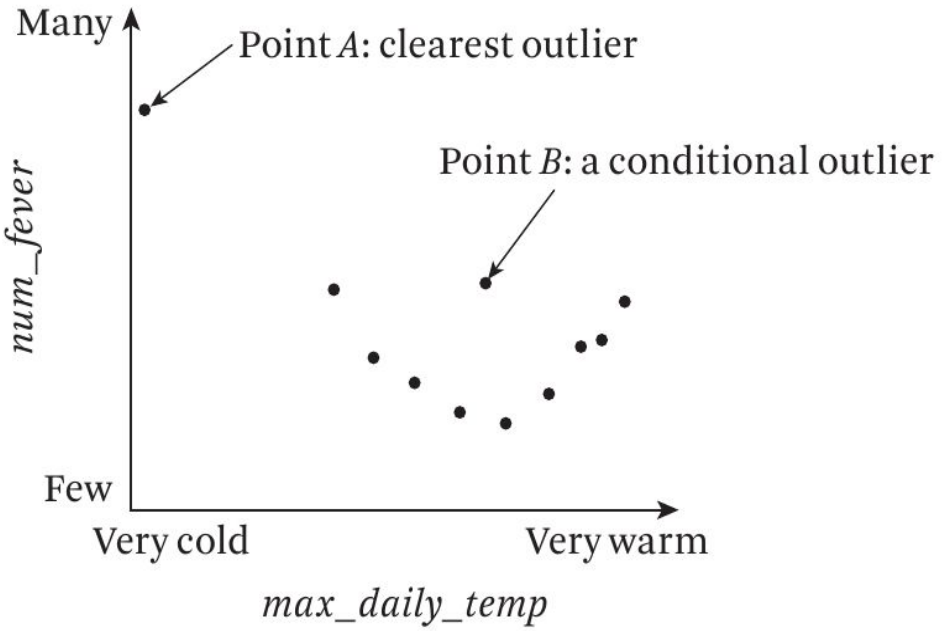
Global Outliers

They individually stand out from the main groups of the entire dataset. They don't belong to any specific subgroup and are easy to identify because they don't match the common patterns found in the data.



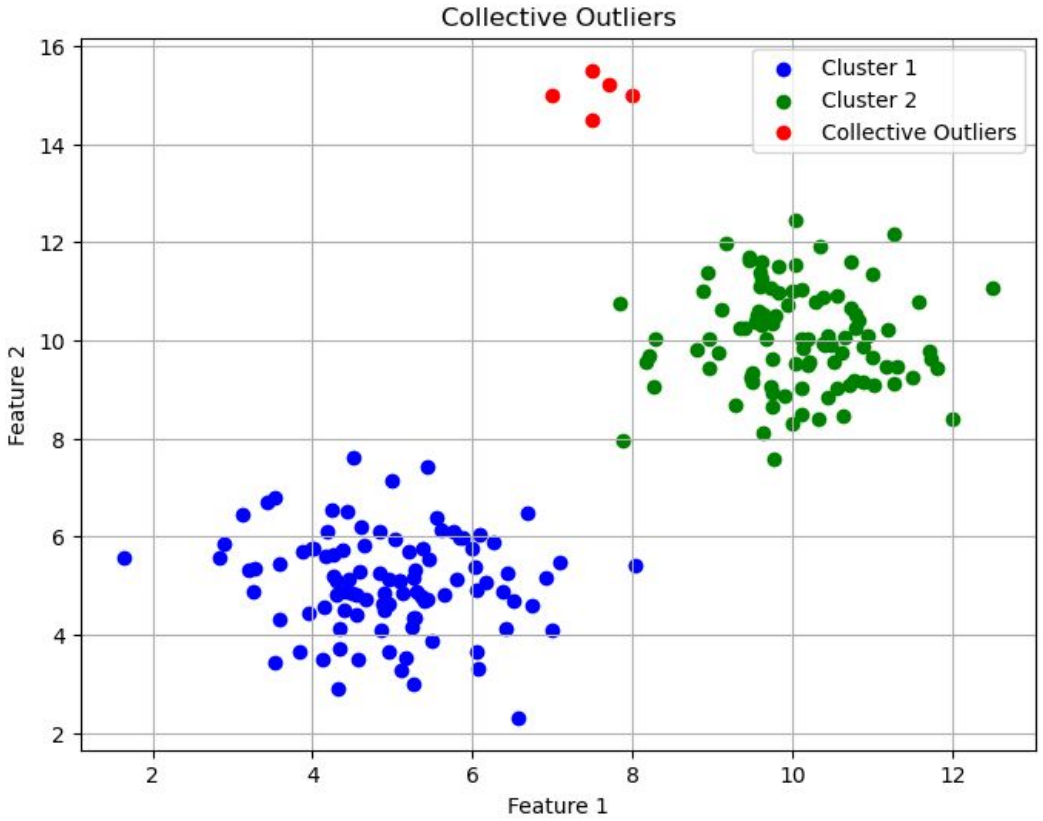
Contextual Outliers

They are the most difficult to detect, especially in high dimensions. Determining the context and behavioral variables often necessitates insights from domain experts, who can provide critical knowledge about which features are relevant and how they interact to define normal behavior within specific conditions.



Collective Outliers

Identifying collective outliers can be tricky because it requires recognizing groups of data points that only appear unusual when viewed together, understanding the relationships between these points, and differentiating these groups from normal patterns.



Challenges of Outlier Detection

1. Defining a normal data pattern and identifying outliers can be challenging, as what constitutes as "normal" varies across different datasets and applications.

2. As the number of dimensions (attributes) in a dataset grows, many techniques for detecting outliers tend to become less effective, an issue commonly known as the curse of dimensionality.

Performance Metrics

Outlier detection can be cast as a binary classification task.
Precision and **recall** are most appropriate when there is a minority class.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Algorithms for OD

Algorithm	Type	Sklearn package	Computational Complexity	Memory Complexity
DBSCAN	Density-based flat clustering	`sklearn.cluster.DBSCAN`	$O(n^2)$ *[n is the size of the dataset]*	$O(n)$
OPTICS	Density-based hierarchical clustering	`sklearn.cluster.OPTICS`	$O(n^2)$	$O(n)$
HDBSCAN	Density-based hierarchical clustering	`sklearn.cluster.HDBSCAN`	$O(n^2)$	$O(n^2)$
ISOLATION FOREST	Ensemble method	`sklearn.ensemble.IsolationForest`	$O(t \cdot \psi \cdot \log n)$ *[t is the number of trees in the forest]*	$O(t \cdot \psi)$ [ψ *represents the number of samples selected by each tree]*

DBSCAN

- Groups closely packed points together by identifying the number of points within a specified radius (ε) and labels points as core points, border points, or outliers. Core points have at least a minimum min_samples of points within their ε-neighborhood. Border points have fewer than min_samples within their ε-neighborhood but are in the neighborhood of a core point. Outliers are points that are not core points nor border points.

- DBSCAN is good mostly for data with clusters of similar density.

- **Main hyperparameters**

  -`eps`: maximum distance between two samples for one to be considered as in the neighborhood of the other.
  -`min_samples`: number of samples in a neighborhood for a point to be considered as a core point. This includes the point itself. It reflects the minimum density required to form a cluster.

OPTICS

- ➔ creates a density-based ordering of points in a dataset, facilitating the identification of clusters of varying densities without a preset distance threshold.

- ➔ Outputs a reachability plot, which provides an ordered list of points based on their density connectivity. This list is accompanied by each point's reachability distance, effectively encoding the clustering structure of the dataset across different density levels.

- ➔ **Main hyperparameters**

-`min_cluster_size` : the smallest size a cluster may be. Any cluster with fewer points than this threshold will be considered noise.
-`min_samples`: Similar to DBSCAN, it defines the number of samples in a neighborhood for a point to be considered as a core point. It influences the density required to start a cluster.
-`max_eps`: The maximum distance between two points to be considered as neighbors. Unlike DBSCAN's eps, this does not limit the neighborhood search, but it is used for reachability distance calculation.
-`xi`: Determines the minimum steepness on the reachability plot that constitutes a cluster boundary. Higher values mean fewer, but more significant clusters.

HDBSCAN

- Extends DBSCAN by converting it into a hierarchical clustering algorithm, and then uses a technique to extract flat clusters out of the hierarchical tree of clusters. It does not require a preset ε distance to be defined.

- Produces a set of clusters that have been extracted based on the stability of clusters across different densities. The algorithm directly outputs clusters and their membership, including outlier points that are not assigned to any cluster.

- **Main hyperparameters**

  -`min_cluster_size`: the smallest size a cluster may be. Any cluster with fewer points than this threshold will be considered noise.

  -`min_samples` : used similarly to DBSCAN, it defines the number of samples in a neighborhood for a point to be considered as a core point. It affects how conservatively the algorithm declares points as noise.

Isolation Forest

- In order to isolate a data point, the algorithm recursively generates partitions on the sample by randomly selecting an attribute and then randomly selecting a split value between the minimum and maximum values allowed for that attribute.

- Particularly effective in detecting global outliers (also in high dimensions). Less suited for the identification of collective outliers (unless they are substantially different in density or distribution from the rest of the data.).

- **Main hyperparameters**

  -`n_estimators`: number of base estimators (trees).
  -`max_samples`: number of samples in a neighborhood for a point to be considered as a core point. This includes the point itself. It reflects the minimum density required to form a cluster.
  -`contamination`: proportion of outliers in the dataset.
  -`max_features`: number of features to draw from the total features to train each base estimator.