L.Sanchirbold (23M1NUM0353)

ANOMALY DETECTION

**DEF:**

**ANOMALY DETECTION IS A TECHNIQUE USED IN DATA ANALYSIS TO IDENTIFY PATTERNS OR INSTANCES THAT DO NOT CONFORM TO EXPECTED BEHAVIOR OR A NORM.**

**Here are some key concepts and techniques related to anomaly detection:**

**Anomaly detection techniques :**

1. Model-Based Techniques

Many anomaly detection techniques first build a model of the data. Anomalies are objects that do not fit the model very well. For example, a model of the distribution of the data can be created by using the data to estimate the parameters of a probability distribution. An object does not fit the model very well; i.e., it is an anomaly, if it is not very likely under the distribution

2. Proximity-Based Techniques

Many of the techniques in this area are based on distances and are referred to as **distance-based outlier detection techniques**. When the data can be displayed as a two- or three-dimensional scatter plot, distancebased outliers can be detected visually, by looking for points that are separated from most other points

3. Density-Based Techniques

Estimates of the density of objects are relatively straightforward to compute, especially if a proximity measure between objects is available. Objects that are in regions of low density are relatively distant from their neighbors, and can be considered anomalous.

**Statistical Approaches**

Statistical approaches are **model-based approaches**; i.e., a model is created for the data, and objects are evaluated with respect to how well they fit the model. Most statistical approaches to outlier detection are based on building a probability distribution model and considering how likely objects are under that model.

**Definition** (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.

**Proximity-Based Outlier Detection**

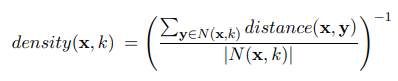
One of the simplest ways to measure whether an object is distant from most points is to use the distance to the k-nearest neighbor. This is captured by definition. The lowest value of the outlier score is 0, while the highest value is the maximum possible value of the distance function—usually infinity.

**Definition** (Distance to k-Nearest Neighbor). The outlier score of an object is given by the distance to its k-nearest neighbor.

**Density-Based Outlier Detection**

From a density-based viewpoint, outliers are objects that are in regions of low density. Density-based outlier detection is closely related to proximity-based outlier detection since density is usually defined in terms of proximity. One common approach is to define density as the reciprocal of the average distance to the k nearest neighbors. If this distance is small, the density is high, and vice versa.

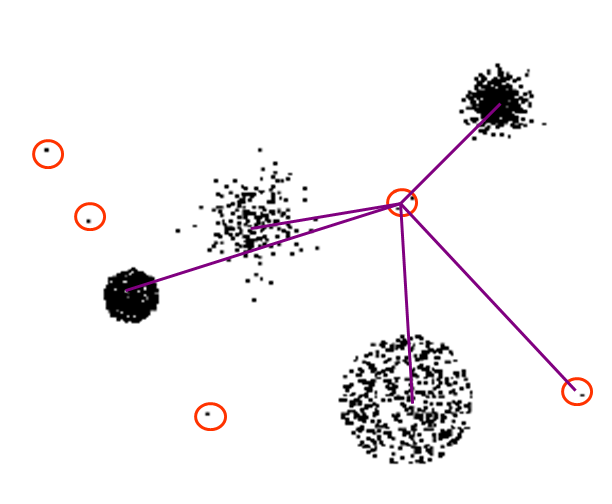
**Definition** (Inverse Distance).



where N(x, k) is the set containing the k-nearest neighbors of x, |N(x, k)| is the size of that set, and y is a nearest neighbor.

**Definition** (Count of Points within a Given Radius). The density around an object is equal to the number of objects that are within a specified distance d of the object.

**Clustering-Based Approaches**

 **Information Theoretic Approaches**

Key idea is to measure how much information decreases when you delete an observation.



Anomalies should show higher gain and normal points should have less gain.

**Example**

Survey of height and weight for 100 participants.



Our data for the weight and height information of 100 participants has an entropy of 2.08.



By eliminating last group, entropy equal to 1.89 and give a gain of 2.08 - 1.89 = 0.19.