UNIT IV CLASSIFICATION AND CLUSTERING 10

Decision Tree Induction - Bayesian Classification – Rule Based Classification – Classification by Back propagation – Support Vector Machines – Associative Classification – Lazy Learners – Other Classification Methods – Clustering techniques – , Partitioning methods- k-means- Hierarchical Methods – distance based agglomerative and divisible clustering, Density-Based Methods – expectation maximization -Grid Based Methods – Model-Based Clustering Methods – Constraint – Based Cluster Analysis – Outlier Analysis

Clustering techniques - Outlier Analysis

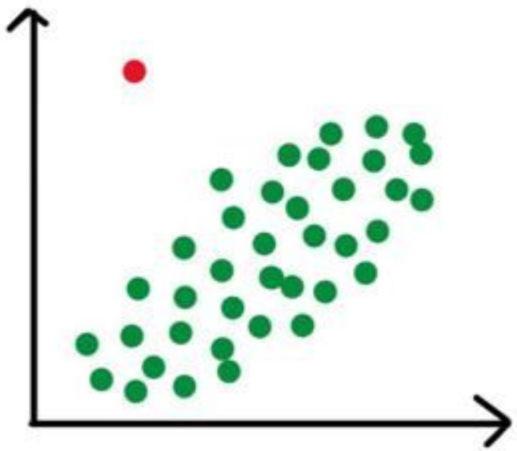
- Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner. An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution errors. The analysis of outlier data is referred to as outlier analysis or outlier mining.
- An outlier cannot be termed as a noise or error. Instead, they are suspected of not being generated by the same method as the rest of the data objects.

- Outliers are of three types, namely –

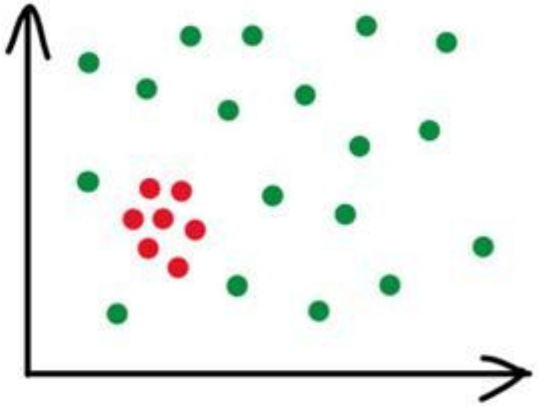
1. Global (or Point) Outliers
2. Collective Outliers
3. Contextual (or Conditional) Outliers

- clustering is a try to advance the fit between the given data and some mathematical model and is based on the assumption that data are created by a combination of a basic probability distribution.

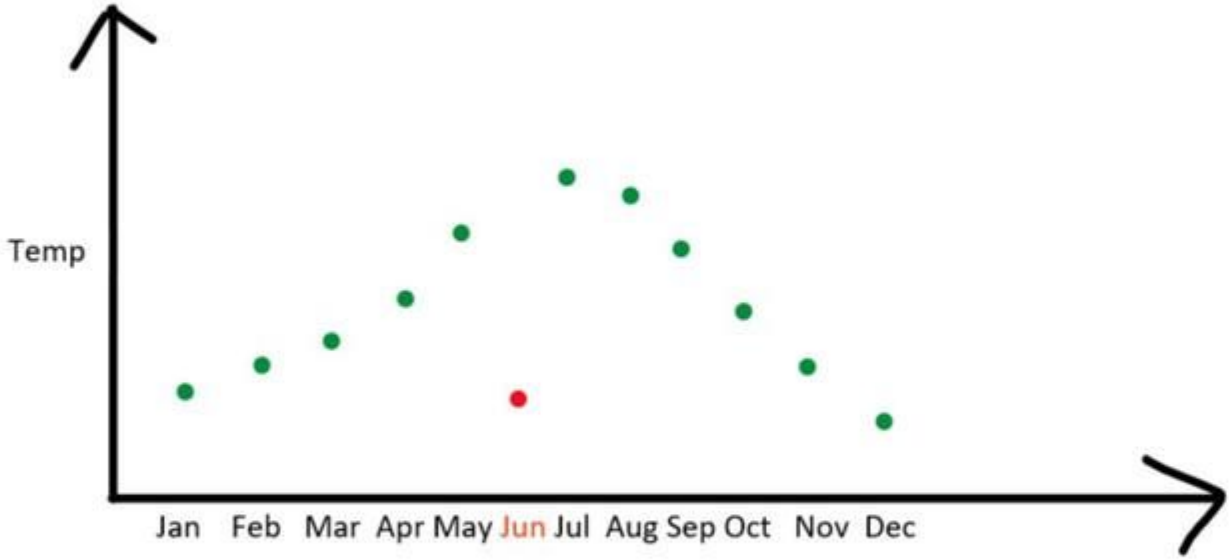
- **1. Global Outliers**
- They are also known as *Point Outliers*. These are the simplest form of outliers. If, in a given dataset, a data point strongly deviates from all the rest of the data points, it is known as a global outlier. Mostly, all of the outlier detection methods are aimed at finding global outliers.
- *For example,* In Intrusion Detection System, if a large number of packages are broadcast in a very short span of time, then this may be considered as a global outlier and we can say that that particular system has been potentially hacked.



- **2. Collective Outliers**
- As the name suggests, if in a given dataset, some of the data points, as a whole, deviate significantly from the rest of the dataset, they may be termed as collective outliers. Here, the individual data objects may not be outliers, but when seen as a whole, they may behave as outliers. To detect these types of outliers, we might need background information about the relationship between those data objects showing the behavior of outliers.
- *For example:* In an Intrusion Detection System, a DOS (denial-of-service) package from one computer to another may be considered as normal behavior. However, if this happens with several computers at the same time, then this may be considered as abnormal behavior and as a whole they can be termed as collective outliers.



- **3. Contextual Outliers**
- They are also known as *Conditional Outliers.* Here, if in a given dataset, a data object deviates significantly from the other data points based on a specific context or condition only. A data point may be an outlier due to a certain condition and may show normal behavior under another condition. Therefore, a context has to be specified as part of the problem statement in order to identify contextual outliers. Contextual outlier analysis provides flexibility for users where one can examine outliers in different contexts, which can be highly desirable in many applications. The attributes of the data point are decided on the basis of both contextual and behavioral attributes.
- *For example:* A temperature reading of 40°C may behave as an outlier in the context of a “winter season” but will behave like a normal data point in the context of a “summer season”.


Temp
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec