**DBSCAN: Density-Based Spatial Clustering of Applications with Noise**

**Introduction**

DBSCAN is a popular clustering algorithm which is used for finding clusters of arbitrary shapes and sizes in a  database. Opposite to traditional clustering techniques like K-means which presuppose spherical clusters DBSCAN can effectively find clusters in datasets having irregular shapes. This nature of the algorithm makes it special and important to analyze irregular datasets like. geographical data and other complicated datasets,  
  
  
The logic behind DBSCAN  focuses on  the idea of density. It clusters points which are closely packed and marks those that lie in a low-density region alone as an outlier or noise. It works mainly on two parameters:  
  
Epsilon (ε): This defines the radius around a point(any given arbitrary point) to consider in clustering. It is used to decide the neighborhood around a point.

MinPts: Minimum number of points required to form a dense region. A point is treated as a core point if it has at least MinPts neighbors within the radius of ε.

**Steps of DBSCAN Algorithm**

Classify Points:

For each point in the data-set, you first count the number of points lying in the ε neighborhood, meaning inside the radius specified.  
Tag the point as a core point if it has at least MinPts neighbors. I.e, if MinPts is set to 4, the points picked as a core point would be the once that have at least 4 points inside the radius specified for them.  
If a point is surrounded by less than MinPts neighbors it is set as a non core point and it becomes a noise point.

Forming the Clusters:

Choose any core point arbitrarily. Then create a cluster of that point with its neighbors. Then, choose any non-visited point, If this point is a core point, then create a cluster. Recursively add all directly density-reachable points (points within ε) to this cluster. Keep in processing until all the points in the cluster are processed. Points which are not a part of any cluster are labeled as noise.

**Note that:** If the point is a non-core point and is a density-reachable point it is added to the cluster but it is not used to extend the cluster even if it is a non-visited point.

Example  
Consider a data-set with points scattered in space. Using DBSCAN, ε = 0.5 and MinPts = 3, the following can be obtained:  
  
Points that have enough neighbors with in a radius of 0.5 form clusters. Points that do not meet this criteria are considered noise(i.e, if there is only one neighbor of a point within this radius, then it is excluded as noise). If a point does not meet the criteria and is found in the radius of another core point it is included in the cluster but does not contribute to the expansion of the cluster and stays idle.

**Advantages of DBSCAN**  
1. No need for a predefined number of clusters. Unlike K-means, DBSCAN does not need to presuppose any number of clusters since it finds clusters with any form of shape from data depending on the varying density of data.  
2. Noised Robust: Identifies outliers impressively.  
3. Ability to Identify Arbitrarily Shaped Clusters: DBSCAN can identify a cluster of any shape,and it does not presuppose the clusters to be spherical in shape.

**Limitations of DBSCAN**

1. Parameter sensitive: The choice of ε and MinPts highly affects the performance of this algorithm. Hence poor selection of this parameters can cause bad clustering.  
   Sensitive to   
   2. Difficulty with Varying Densities: DBSCAN does not work well with datasets containing clusters of widely varying densities since it uses a single ε for the entire data-set.  
   3. High-Dimensional Data: It does not work well on high-dimensional data due to the curse of dimensionality since the notion of density loses meaning in high-dimensional spaces.

**Applications**  
DBSCAN can be used in many applications such as:  
  
Geo-spatial Data Analysis: Identification of cluster of geographical location, for example, Crime hot-spots and Customer Distribution.

Market Basket Analysis: Patterns in transaction data show customer behavior.

**Conclusion**

DBSCAN is one of the powerful clustering techniques which perform well on the identification of clusters with arbitrary shapes and can handle noise effectively. Its density-based approach provides many advantages over traditional clustering methods, hence making it good for a lot of different applications in different domains. However, careful consideration of its parameters and limitations is required to achieve optimal performance. As data and different problems grow, this algorithm provides a suitable way to handle this data analyse is and recognize patter while picking out out-liner's making it very effective.