

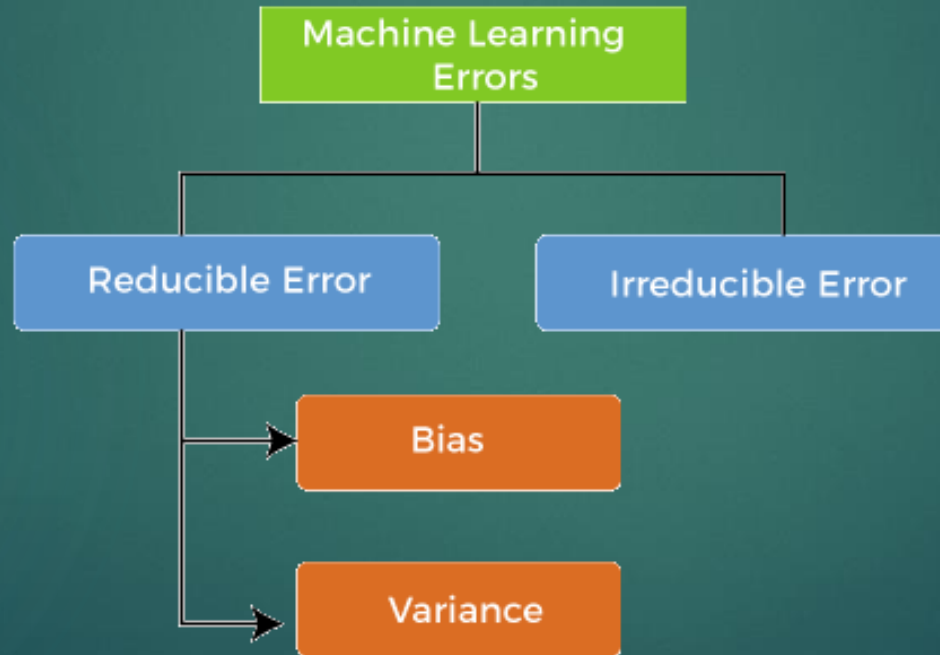


Bias Density Based Clustering

BY
ANAY SHUKLA

Errors in Machine Learning

- ▶ In machine learning, an error is a measure of how accurately an algorithm can make predictions for the previously unknown dataset. On the basis of these errors, the machine learning model is selected that can perform best on the particular dataset.



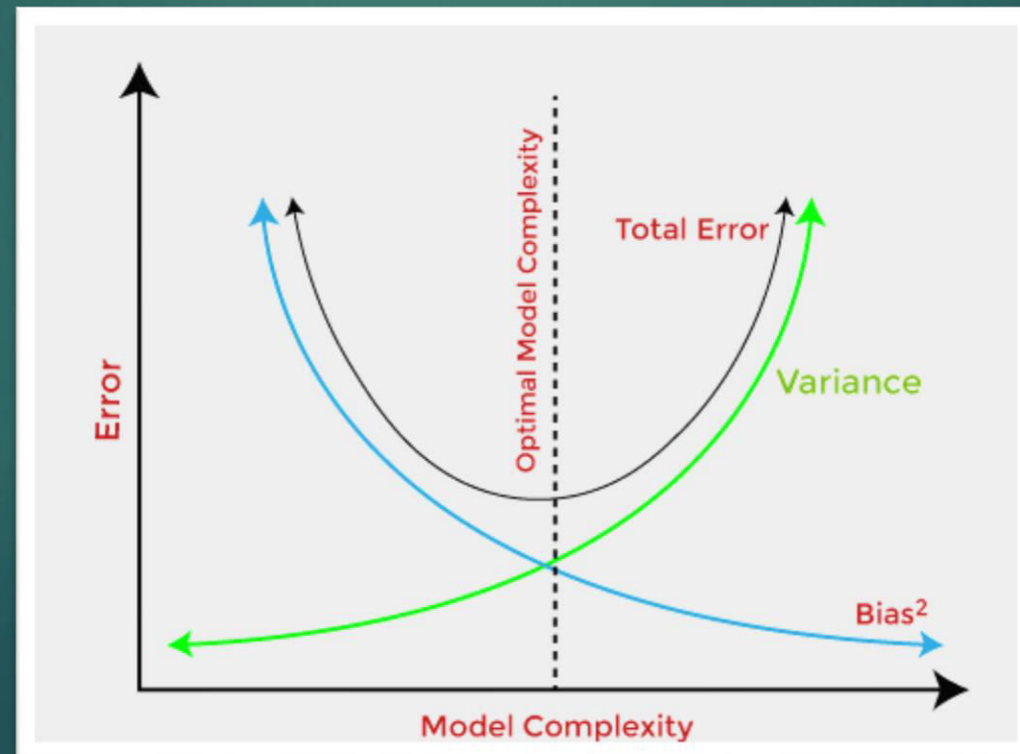
What is Biasing


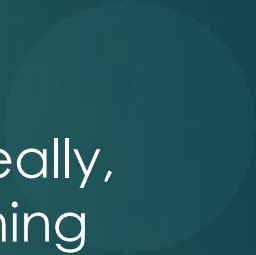
- ▶ While making predictions, a difference occurs between prediction values made by the model and actual values/expected values, and this difference is known as bias errors or Errors due to bias.
- ▶ Each algorithm begins with some amount of bias because bias occurs from assumptions in the model, which makes the target function simple to learn. A model has either:
 - ▶ Low Bias: A low bias model will make fewer assumptions about the form of the target function.
 - ▶ High Bias: A model with a high bias makes more assumptions, and the model becomes unable to capture the important features of our dataset. A high bias model also cannot perform well on new data.

- ▶ Some examples of machine learning algorithms with low bias are Decision Trees, k-Nearest Neighbours and Support Vector Machines. Algorithms with high bias is Linear Regression, Linear Discriminant Analysis and Logistic Regression.
- ▶ Ways to reduce High Bias:
 - ▶ Increase the input features as the model is underfitted.
 - ▶ Decrease the regularization term.
 - ▶ Use more complex models, such as including some polynomial features.

Bias-Variance Trade-Off

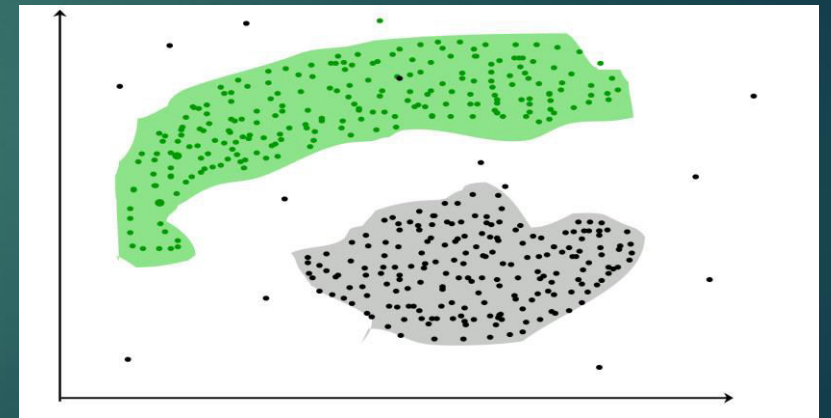
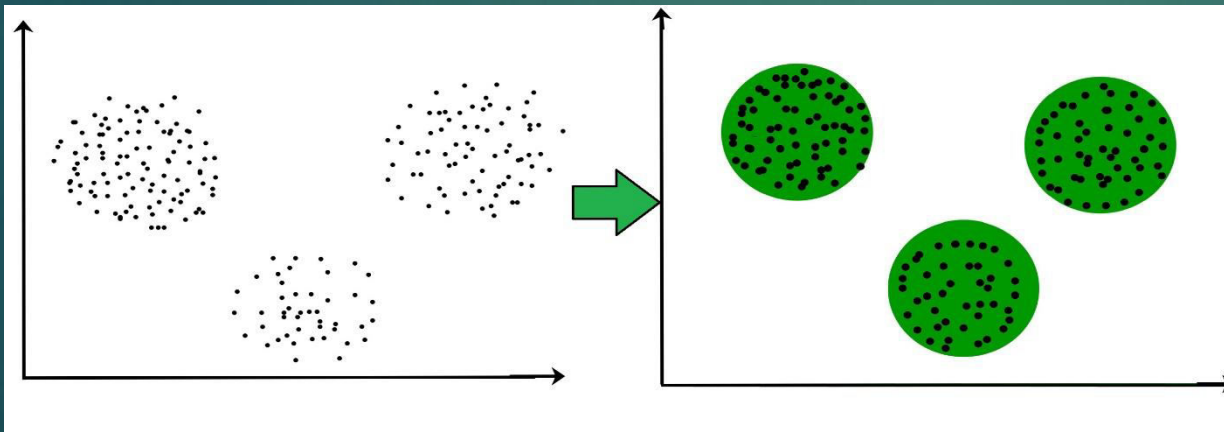
- ▶ In order to prevent overfitting and under fitting in the machine learning model, bias and variance must be carefully considered while the model is being built.



- 
- 
- ▶ For an accurate prediction of the model, algorithms need a low variance and low bias. But this is not possible because bias and variance are related to each other:
 - ▶ If we decrease the variance, it will increase the bias.
 - ▶ If we decrease the bias, it will increase the variance.
 - ▶ Bias-Variance trade-off is a central issue in supervised learning. Ideally, we need a model that accurately captures the regularities in training data and simultaneously generalizes well with the unseen dataset. Unfortunately, doing this is not possible simultaneously.
 - ▶ Hence, the Bias-Variance trade-off is about finding the sweet spot to make a balance between bias and variance errors

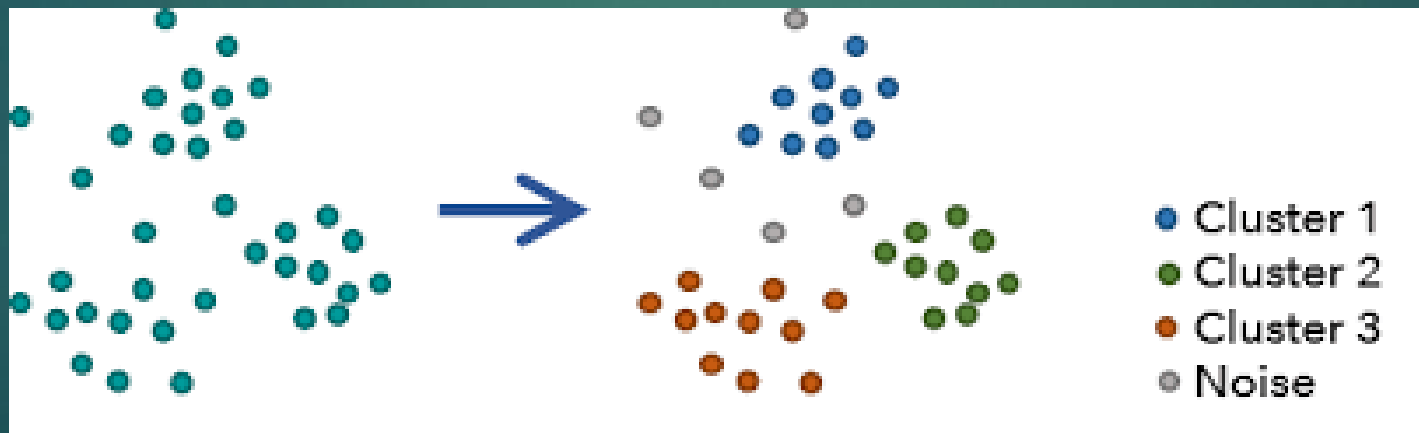
What is Clustering?

- ▶ Clustering analysis or simply Clustering is basically an Unsupervised learning method.
- ▶ Under this method that divides the data points into a number of specific batches or groups, such that the data points in the same groups have similar properties and data points in different groups have different properties in some sense.



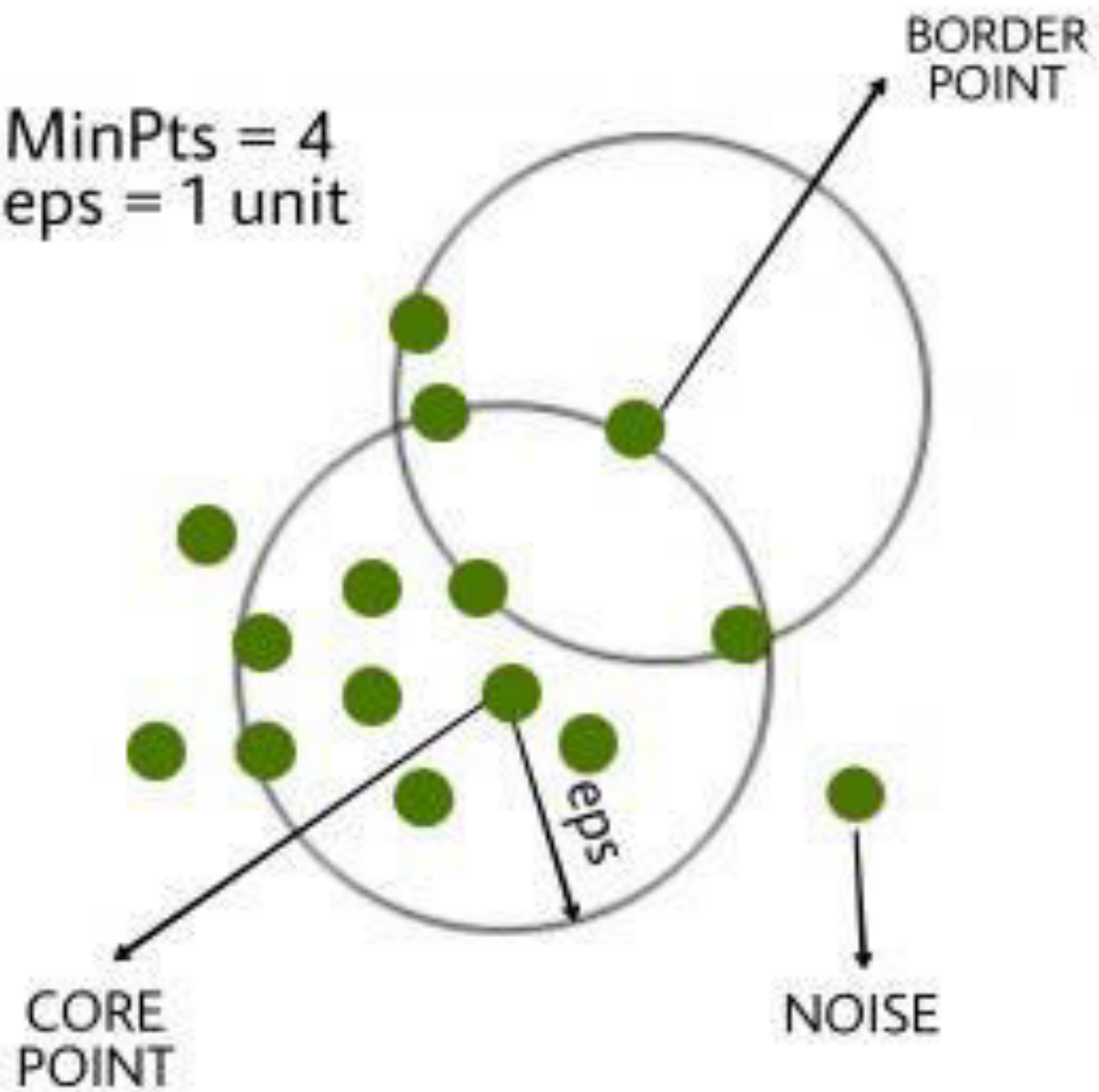
Density-based Clustering

- ▶ This method considers the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters.
- ▶ The data points in the region separated by two clusters of low point density are considered as noise or outliers.



- ▶ There are two different parameters to calculate the density-based clustering
- 1. **Eps:** It defines the neighbourhood around a data point i.e. if the distance between two points is lower or equal to 'eps' then they are considered neighbours. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters.
- 2. **MinPts:** Minimum number of neighbours (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, $\text{MinPts} \geq D+1$. The minimum value of MinPts must be chosen at least 3.
- ▶ **Core Point:** A point is a core point if it has more than MinPts points within eps.
- ▶ **Border Point:** A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.
- ▶ **Noise or outlier:** A point which is not a core point or border point.

MinPts = 4
eps = 1 unit

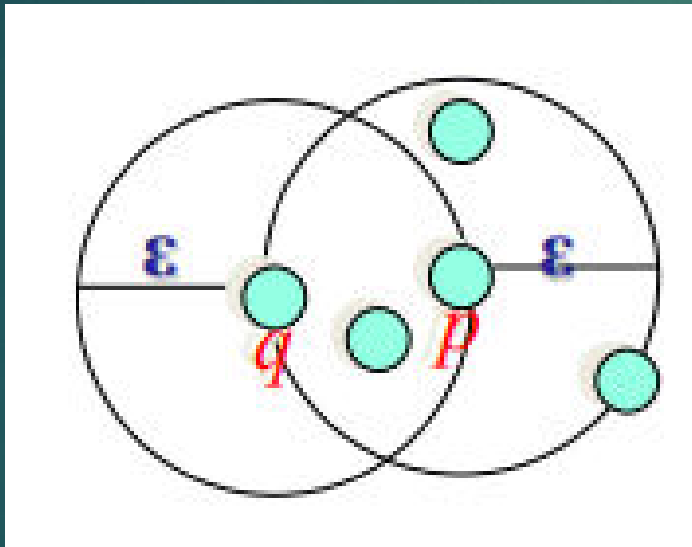


Working of Density-Based Clustering

► Density reachable:

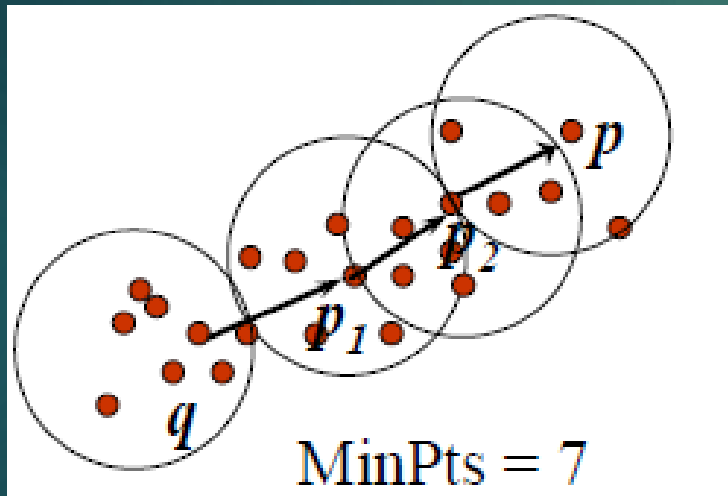
► Directly density-reachable

- An object q is directly density-reachable from object p if p is a core object and q is in p 's neighborhood.

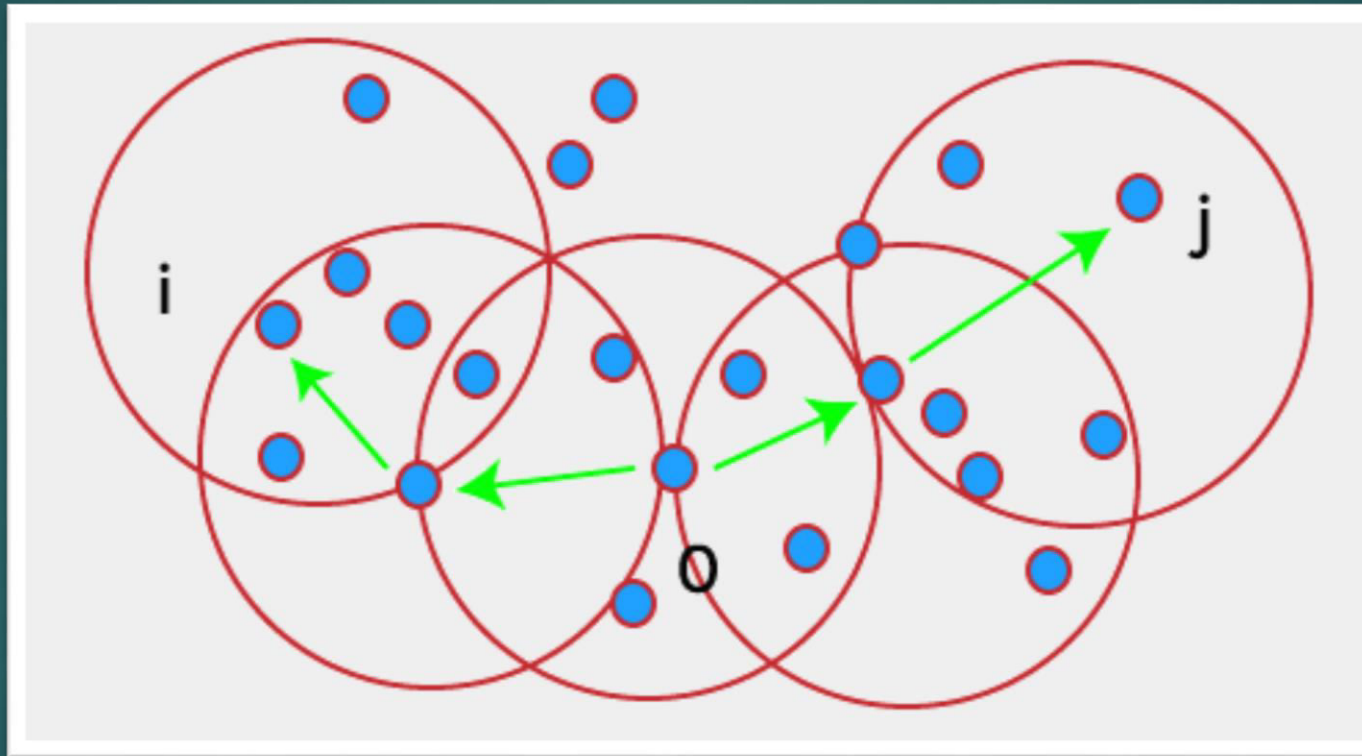


- q is directly density-reachable from p
- p is not directly density-reachable from q
- Density-reachability is asymmetric

- ▶ A point p is directly density-reachable from p_2
- ▶ p_2 is directly density-reachable from p_1
- ▶ p_1 is directly density-reachable from q
- ▶ $p \leftarrow p_2 \leftarrow p_1 \leftarrow q$ form a chain



- **Density connected:** A point i refers to density connected to a point j with respect to Eps , $MinPts$ if there is a point o such that both i and j are considered as density reachable from o with respect to Eps and $MinPts$.



Major Features of Density-Based Clustering

- ▶ It is a scan method.
- ▶ It requires density parameters as a termination condition.
- ▶ It is used to manage noise in data clusters.
- ▶ Density-based clustering is used to identify clusters of arbitrary size.

Density-Based Clustering Methods

- ▶ **DBSCAN**(Density-Based Spatial Clustering of Applications with Noise): It depends on a density-based notion of cluster. It also identifies clusters of arbitrary size in the spatial database with outliers.
 - ▶ **DBSCAN algorithm can be abstracted in the following steps:**
 - ▶ Find all the neighbor points within ϵ and identify the core points or visited with more than MinPts neighbors.
 - ▶ For each core point if it is not already assigned to a cluster, create a new cluster.
 - ▶ Find recursively all its density connected points and assign them to the same cluster as the core point.

A point a and b are said to be density connected if there exist a point c which has a sufficient number of points in its neighbors and both the points a and b are within the ϵ distance. This is a chaining process. So, if b is neighbor of c , c is neighbor of d , d is neighbor of e , which in turn is neighbor of a implies that b is neighbor of a .
 - ▶ Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

► **OPTICS**(Ordering Points To Identify the Clustering Structure):

It gives a significant order of database with respect to its density-based clustering structure. The order of the cluster comprises information equivalent to the density-based clustering related to a long range of parameter settings. OPTICS methods are beneficial for both automatic and interactive cluster analysis, including determining an intrinsic clustering structure.

► **DENCLUE**

Density-based clustering by Hinneburg and Kiem. It enables a compact mathematical description of arbitrarily shaped clusters in high dimension state of data, and it is good for data sets with a huge amount of noise.