7.02 DBSCAN Clustering

DBSCAN Clustering

- Density-based spatial clustering of applications with noise (DBSCAN) clustering method

- It is an unsupervised method that:
  - clusters core samples (dense areas of data set)
  - denotes non-core samples (sparse portions of data set)

- Commonly used for outlier detection

- Outliers should make up <= 5% of the total observations
  - Enabled by adjusting model parameters accordingly

DBSCAN Clustering

- Key Model Parameters

  - EPS
    - maximum distance between 2 samples for them to be clustered in the same neighbourhood
    - Start at a value of 0.1

  - min_samples
    - minimum number of samples in a neighbourhood for a data point to be classified as a core point
    - Start with a very low sample size

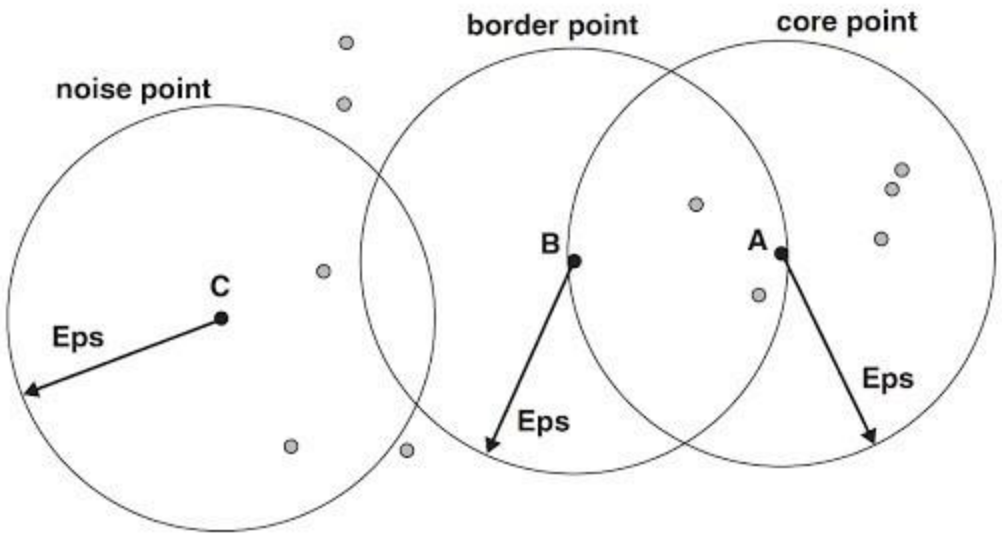
DBSCAN Clustering – Data Classification

- Based on these two parameters i.e., epsilon and min_samples, we are first going to classify every point in our dataset into three categories:

    1. Core points
    2. Boundary points
    3. Noise points

DBSCAN Clustering – **Core points**

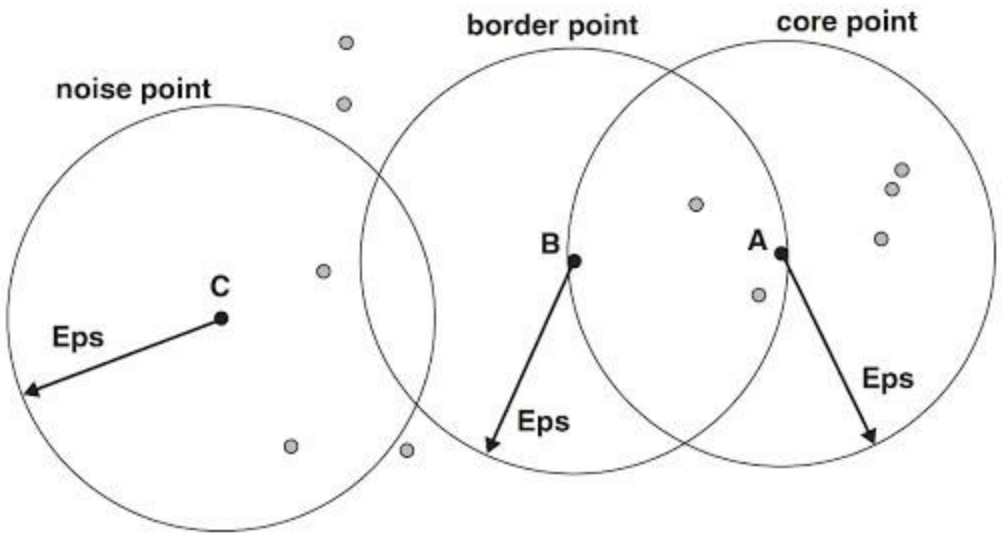
- A point is a core point when:
  - number of neighbours is greater than or equal to min_samples.



Example: If we set min_samples = 3, then Point A satisfies this condition.

DBSCAN Clustering – **Boundary points**

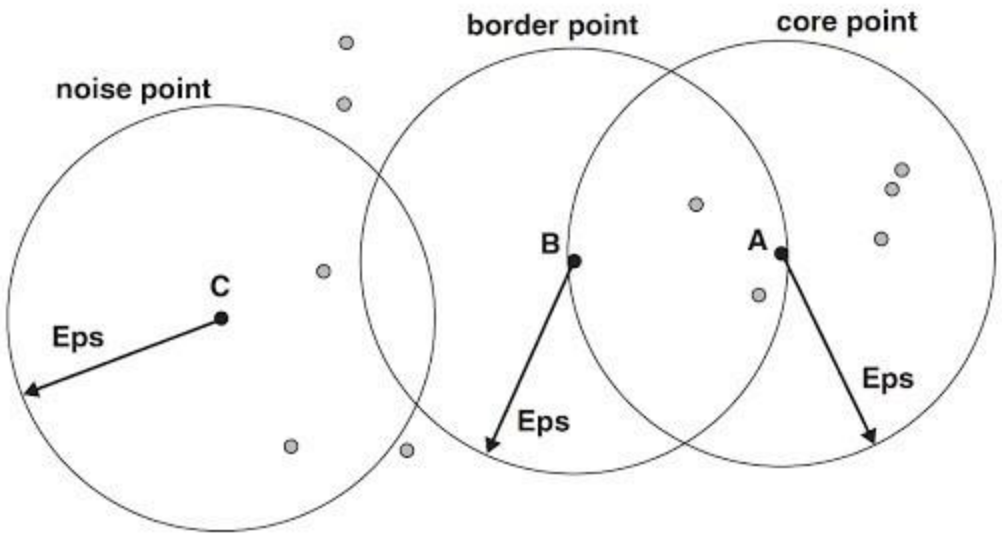
- A point is a boundary point when:
  - The number of neighbors is less than min_samples
  - The point is in neighbourhood of a core point



Example: Point B has less than min_samples=3 in its neighbourhood and it is in the neighbourhood of another core point.

DBSCAN Clustering – **Noise points**

- A point is a noise point when:
  - it is neither a core point nor a boundary point



Example: Point C is neither a core nor a boundary point

Why do we need DBSCAN when we already have K-means clustering?

- K-Means clustering may cluster loosely related observations together. Every observation becomes a part of some cluster eventually, even if the observations are scattered far away in the vector space. Since clusters depend on the mean value of cluster elements, each data point plays a role in forming the clusters. A slight change in data points might affect the clustering outcome. This problem is greatly reduced in DBSCAN due to the way clusters are formed. This is usually not a big problem unless we come across some odd shape data.

- What's good about DBSCAN is that you don't have to specify the number of clusters to use it. All you need is a function to calculate the distance between values and some guidance for what amount of distance is considered "close". DBSCAN also produces more reasonable results than k-means across a variety of different distributions.

Algorithmic steps for DBSCAN clustering

- The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited)
- If there are at least ‘min_sample’ points within a radius of ‘ε’ to the point then we consider all these points to be part of the same cluster
- The clusters are then expanded by recursively repeating the neighborhood calculation for each neighboring point

Why should we use DBSCAN?

- The DBSCAN algorithm should be used to find associations and structures in data that are hard to find manually but that can be relevant and useful to find patterns and predict trends.

- Clustering methods are usually used in biology, medicine, social sciences, archaeology, marketing, characters recognition, management systems and so on.

Why should we use DBSCAN?

- Example: Suppose we have an online store, and we want to improve our sales by recommending relevant products to our customers.
  - We don’t know exactly what our customers are looking for but based on a data set we can predict and recommend a relevant product to a specific customer.
  - We can apply the DBSCAN to our data set (based on the e-commerce database) and find clusters based on the products that the users have bought.
  - Using this clusters we can find similarities between customers, for example, the customer A have bought 1 pen, 1 book and 1 scissors and the customer B have bought 1 book and 1 scissors, then we can recommend 1 pen to the customer B.