DBSCAN

Density-based spatial clustering of applications with noise

Why DBSCAN?

- 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 nice about DBSCAN is that you don’t have to specify the number of clusters to use it.

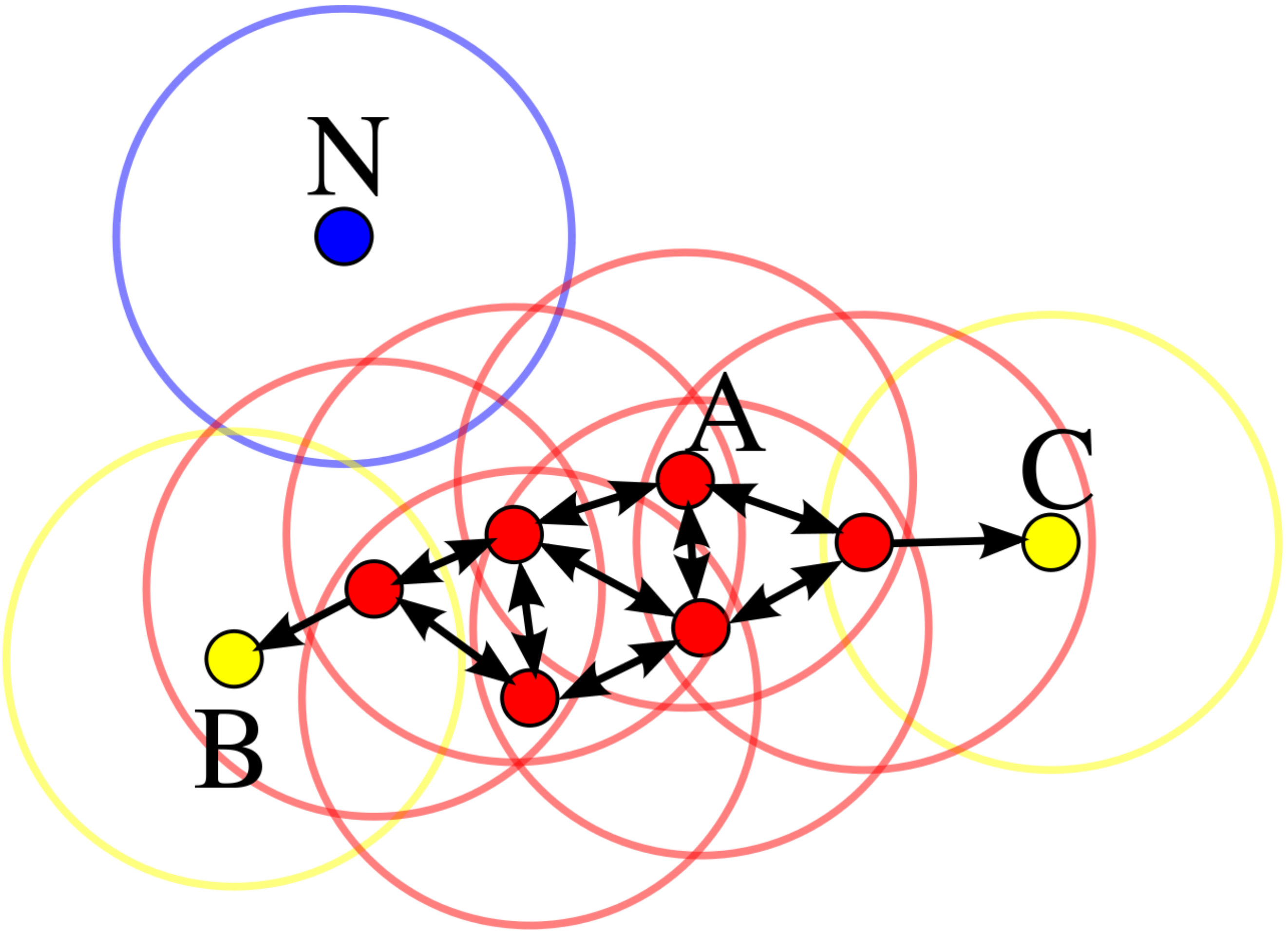
DBSCAN parameters

- Epsilon: Radius for the neighbourhood of points
- Minimum Points: Number of points required to be in the epsilon radius

DBSCAN Terms

- Reachability: The point B is said to be density reachable for point A if it lies within epsilon radius of point A
- Connectivity: Transitivity based chaining-approach to determine whether points are located in a particular cluster. For example, p and q points could be connected if p->r->s->t->q, where a->b means b is in the neighbourhood of a.

DBSCAN Algorithm



Working of DBSCAN

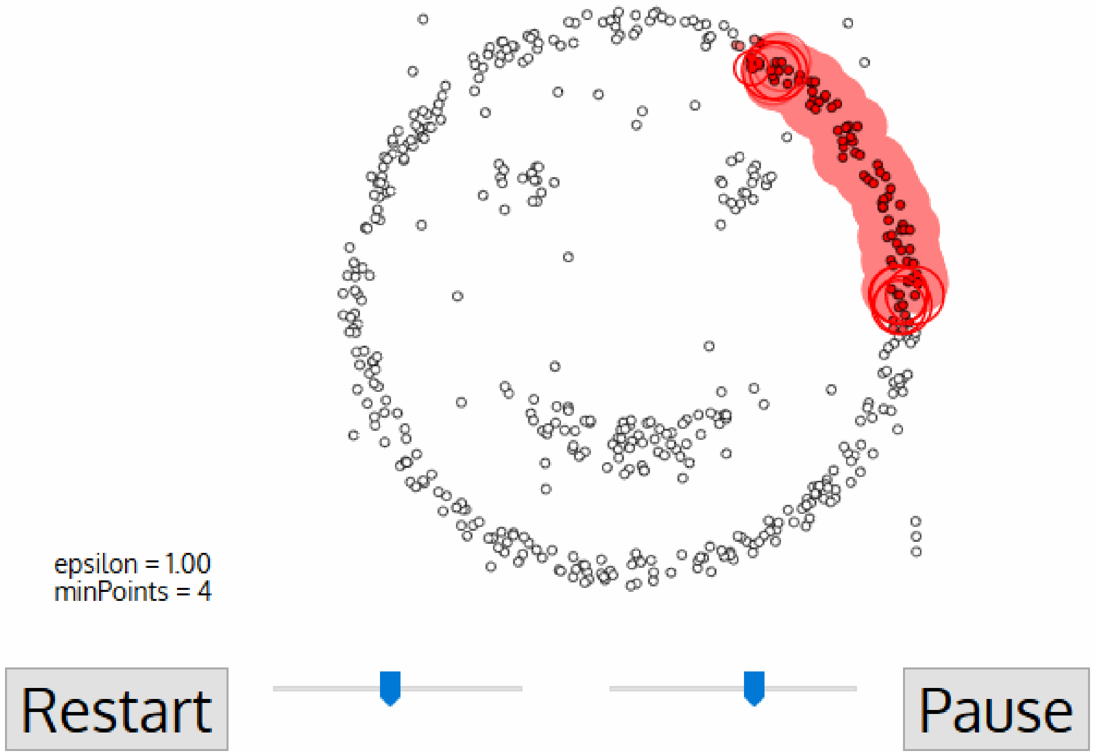


Image Courtesy: https://www.digitalvidya.com/blog/the-top-5-clustering-algorithms-data-scientists-should-know/

Silhouette Score

- The Silhouette Score is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each observation.
- The Silhouette Coefficient for any observations is (b - a) / max(a, b).
- In other words, b is the distance between that observation and the nearest cluster that the observation is not a part of.
- Note that Silhouette Coefficient is only defined if number of labels is 2 <= n_labels <= n_observations - 1

Good Explanation: https://www.youtube.com/watch?v=_jg1UFoef1c&t=140s