CS5228 – Tutorial 2

Clustering: K-Means & DBSCAN

K-Means and DBSCAN are two very popular clustering methods. Since clustering is typically used to find patterns in unlabeled data, it is generally very difficult to reliably assess if a resulting clustering is "good". This makes it even more important to properly understand the underlying principles, as well as the pros and cons of the different methods to better interpret the results.

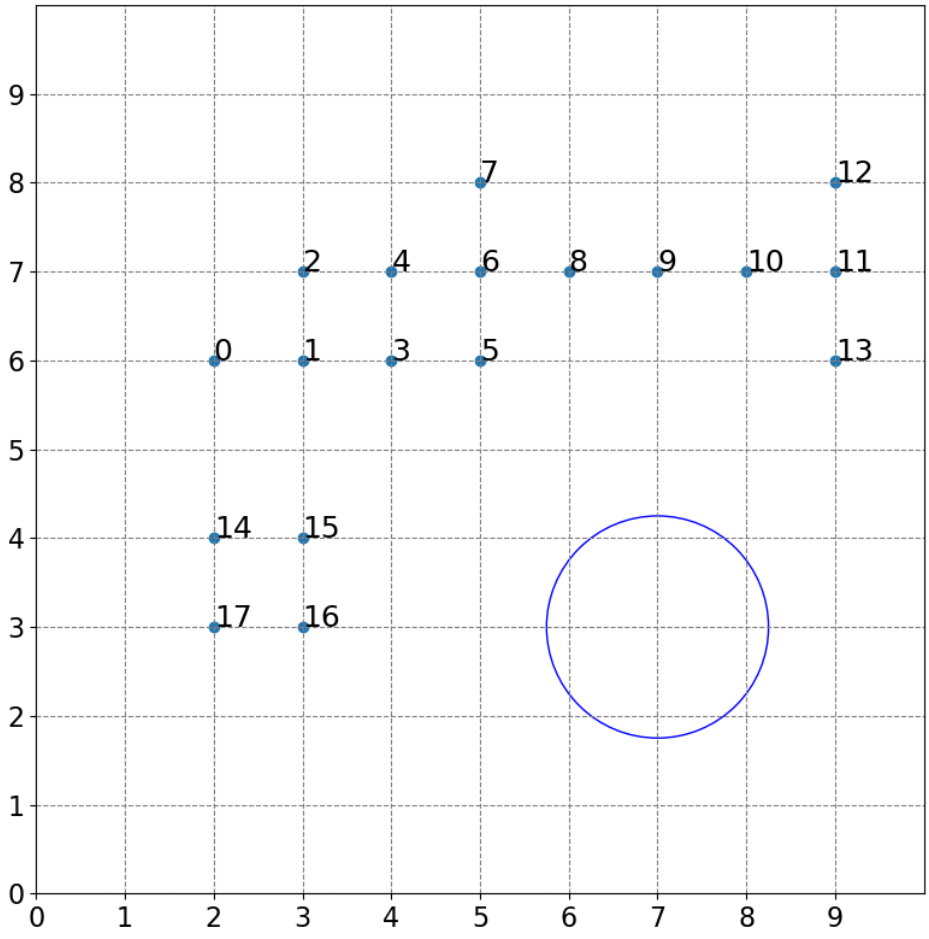

Figure 1: Toy dataset for "manually" performing DBSCAN.

1. **Performing DBSCAN "by hand".** Figure 1 shows a toy dataset with 18 data points. Let's assume we run DBSCAN over this data with $\epsilon = 1.25$ and $MinPts = 4$.
   (a) What will be the result of DBSCAN? Describe the ouput by listing all
      - core points
      - border points
      - noise points

   (b) How many clusters are there, and what are their data points?

   (c) Can you add 2 data points such that the resulting clustering contains only 1 cluster and no noise? If so, give the coordinates for both points!

   The simplicity of the toy dataset should allow you to answer all three tasks by just looking at the plot in Figure 1. There should be no need to actually calculate any distances. The blue circle reflects the chosen radius of $\epsilon = 1.25$ to make it easier for you.

2. **K-Means.** For the following questions, assume that we now want to run K-Means over the toy dataset shown in Figure 1.
   (a) For $K = 3$, can you find locations for the initial centroids so that the resulting clustering will contain 0, 1, 2, or 3 non-empty clusters? You can answer this question qualitatively; there is no need to list any exact coordinates for the initial centroids.

   (b) Now we assume that K-Means++ initialization is used. For $K = 3$, what is the minimum and maximum number of clusters? (Comment: For this question you may need to consider arbitrary datasets and not just the toy dataset!)

3. **K-Means vs. DBSCAN**
   (a) Apart from their implementation, what is the fundamental difference between K-Means and DBSCAN?

   (b) What are meaningful criteria to decide whether K-Means or DBSCAN is the preferable clustering method for a certain task?

   (c) Come up with 5 example tasks and discuss why K-Means or DBSCAN would be your method of choice!

   (d) Is there any example where fundamentally only K-Means is applicable but not DBSCAN, or vice versa?

   (e) Is it possible that both K-Means and DBSCAN return the same clustering for a dataset or will the clusterings always be different?