DBSCAN Clustering

Amanda Landi

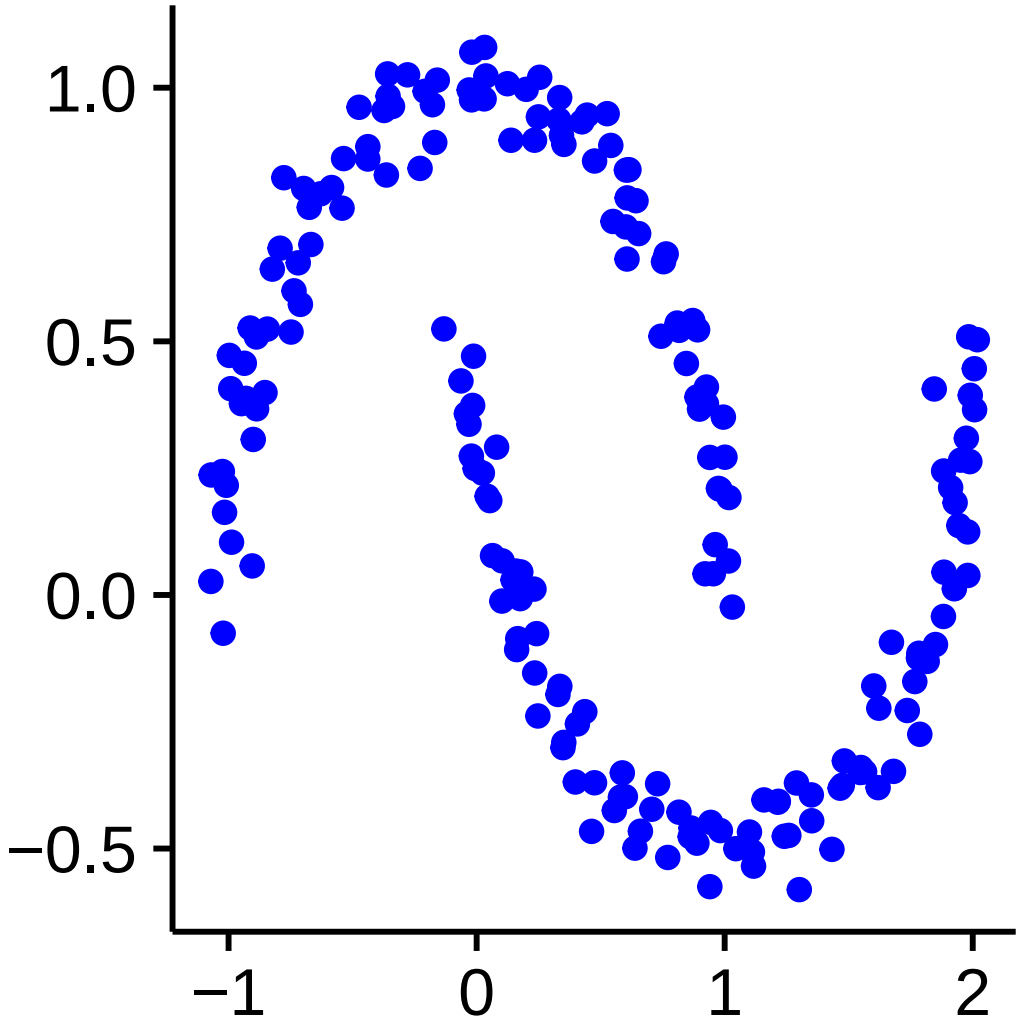
April 11, 2017

Outline

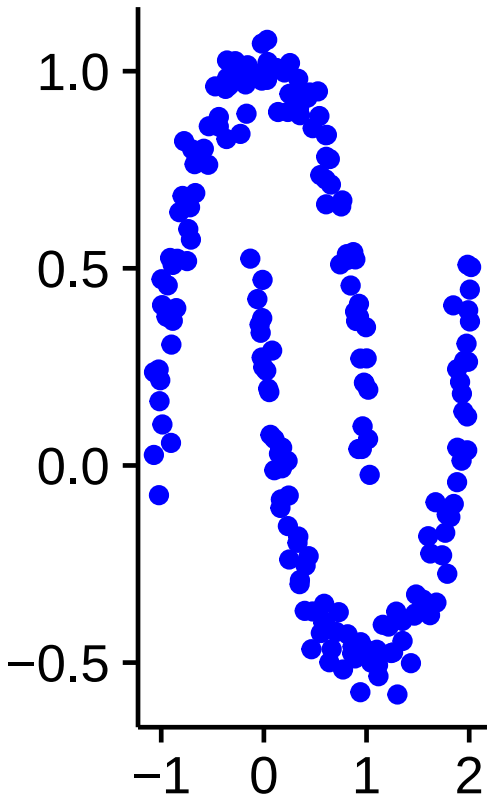
1. Why we need an alternative?
2. General idea behind DBSCAN
3. Necessary Definitions
4. Results

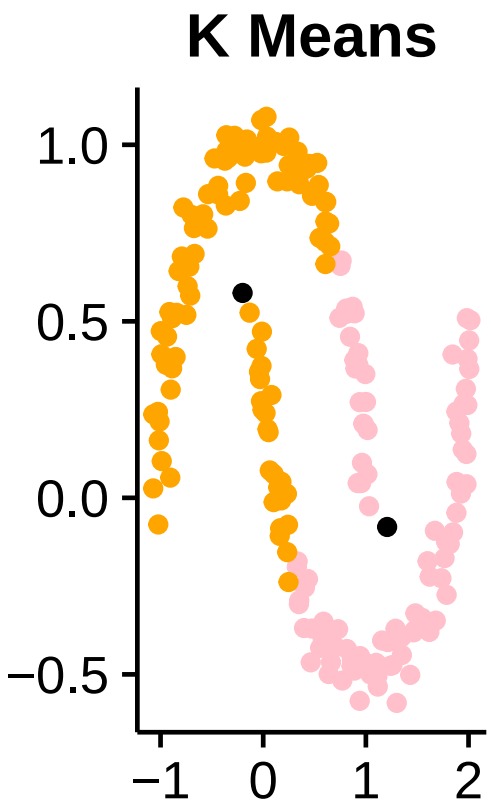
Why

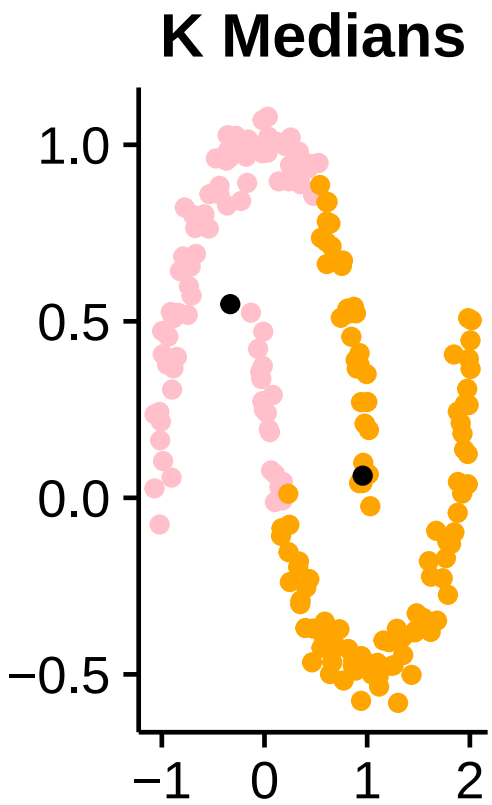
Consider the data



Why


1.0
0.5
0.0
−0.5
−1
0
1
2


K Means
1.0
0.5
0.0
−0.5
−1
0
1
2


K Medians
1.0
0.5
0.0
−0.5
−1
0
1
2

General Idea - DBSCAN

Background:

- Density based spatial clustering algorithm
- Inspired by spatial data (data from space)
- Introduced in 1996 by Ester et. al. and given an award because of the substantial attention it receives in theory and practice at the leading data mining conference, KDD, in 2014

Advantages:

- Rather than assuming a circular region for the cluster, DBSCAN finds clusters of arbitrary shape.
- Does not require the number of clusters specified *a priori*
- Can find clusters completely surrounded by a different cluster, though NOT CONNECTED
- Robust to outliers
- Converges O(n log(n))

DBSCAN - The Method

- Given a set of points $\{x_1, x_2, \ldots, x_n\}$, it groups together points that are closely packed together, marking outlier points that lie alone in low-density regions.
- Definitions
  - A point p is a **core point** if at least r points are within ϵ of it (one has to specify ϵ, can influence the types of clusters we get)
  - The r points are **directly reachable** from p. No points are directly reachable from a non-core point.
  - A point q is **reachable** from p if there is a path of J points $\{p_i\}_{i=1}^{j}$ between p and q such that p_{i+1} is directly reachable from p_i and $p_1 = p$ and $p_J = q$.
  - If p is a core point, then it forms a **cluster** together with all points that are reachable from it.
  - Points not reachable from any other point are **outliers**.
  - Two points are **density-connected** if there exists a point t such that p and q are directly reachable from t. Symmetric!
  - Clusters satisfy
    - All points within a cluster are mutually density-connected
    - If a point is directly reachable from any point of the cluster, it is part of the cluster as well.

DBSCAN In-Class Activity

- Get into groups of 4 students.
- Discuss how to turn the definitions into a method for clustering.
- Begin writing pseudo-code, you will turn this in by end of class.
- IF you have pseudo-code you are happy with, begin writing your own implementation for DBSCAN.
- You can create your own half moons with
  - *from sklearn.datasets import make_moons*
  - *X,y = make_moons(n_samples = 200, noise = 0.05, random_state = 0)*

Note, for your projects, there exists a dbscan function in R in the library dbscan and a dbscan function in Python in the sklearn.cluster package.

DBSCAN Results - Moons

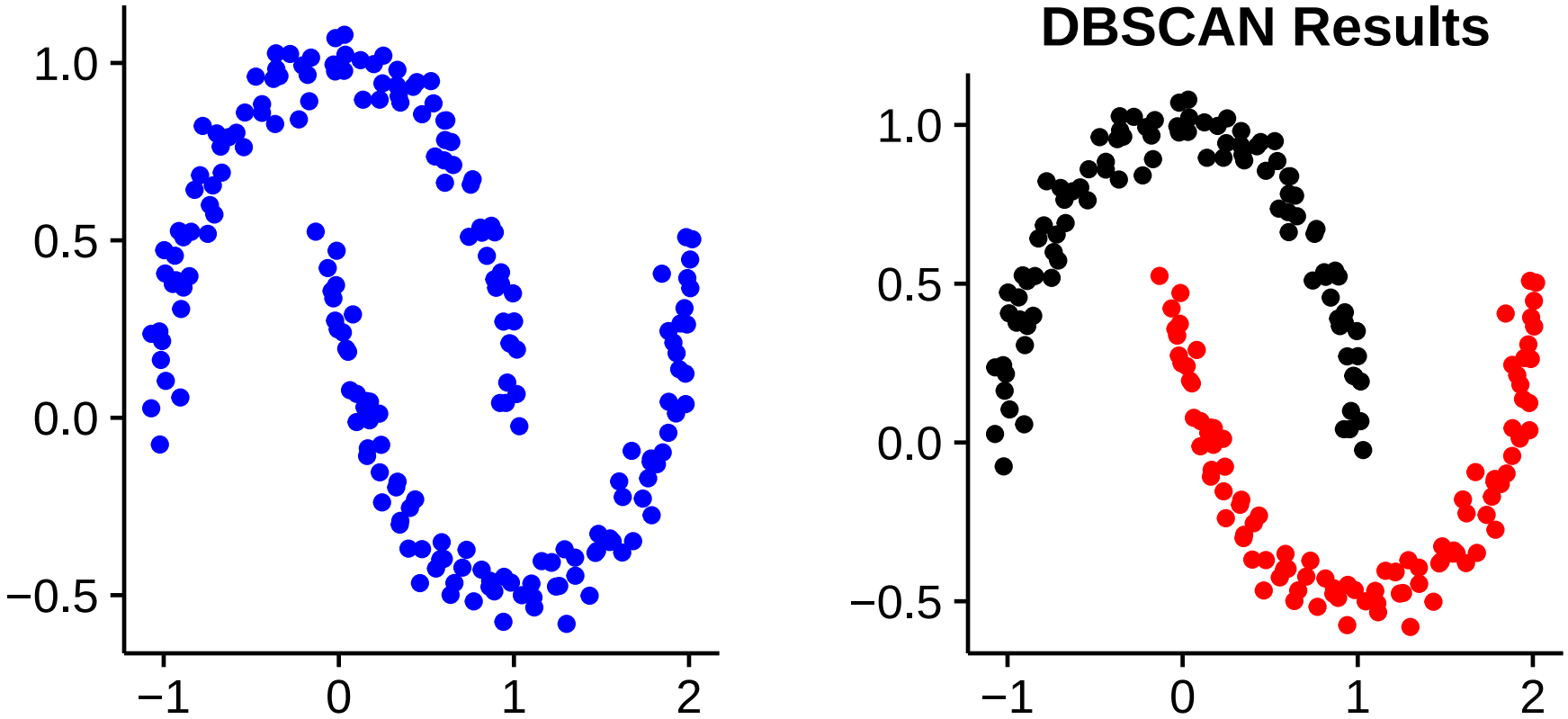


Figure 1: Results by DBSCAN for eps = 0.2 and r = 5

References

- Raschka, S. **Python Machine Learning**. 2015
- Ester, Martin, et al. **A density-based algorithm for discovering clusters in large spatial databases with noise.** 1996.