FIAP

MBA+ | Artificial Intelligence & Machine Learning


AI

Aprendizado não supervisionado

K-means / EM / Regras de Associação

Prof. Dr. Vinicius F. Caridá

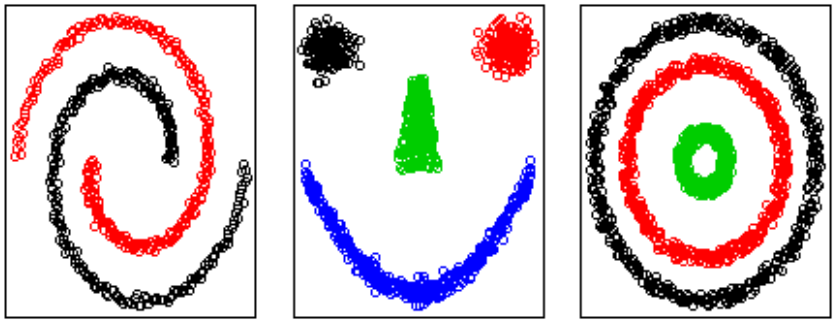
Algoritmos Baseados em Densidade

Algoritmos Baseados em Densidade

Paradigma de Agrupamento por Densidade

- Clusters como regiões de alta concentração de objetos separadas por regiões de baixa concentração de objetos
- Paradigma alternativo àquele baseado em protótipos: K-means e variantes, EM, etc

Existem vários algoritmos, veremos a seguir um dos mais conhecidos: **DBSCAN**

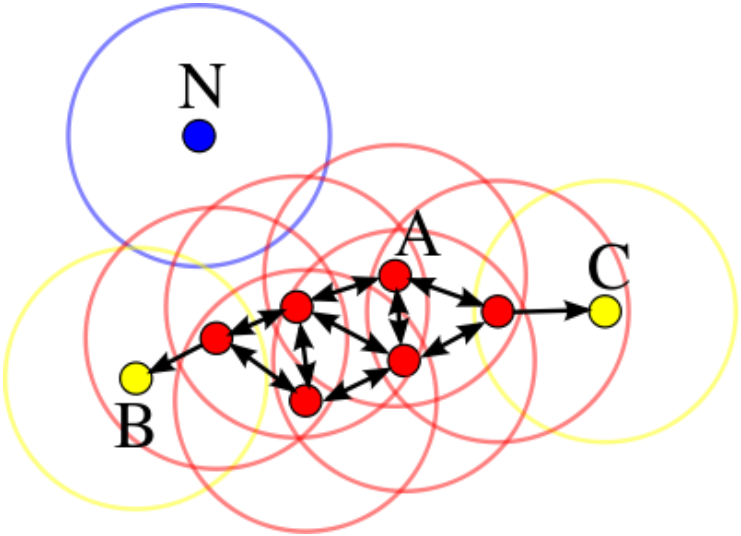


DBScan: definições

- A point is a **core** point if it has at least a specified number of points (MinPts) within the radius Eps (including the point itself)
  - These are points that are in the interior of a cluster

- A **border** point has fewer than MinPts within Eps, but is in the neighborhood (within the radius) of at least 1 core point

- A **noise** point is neither a core point nor a border point



DBScan: algoritmo

Algoritmo Conceitual:

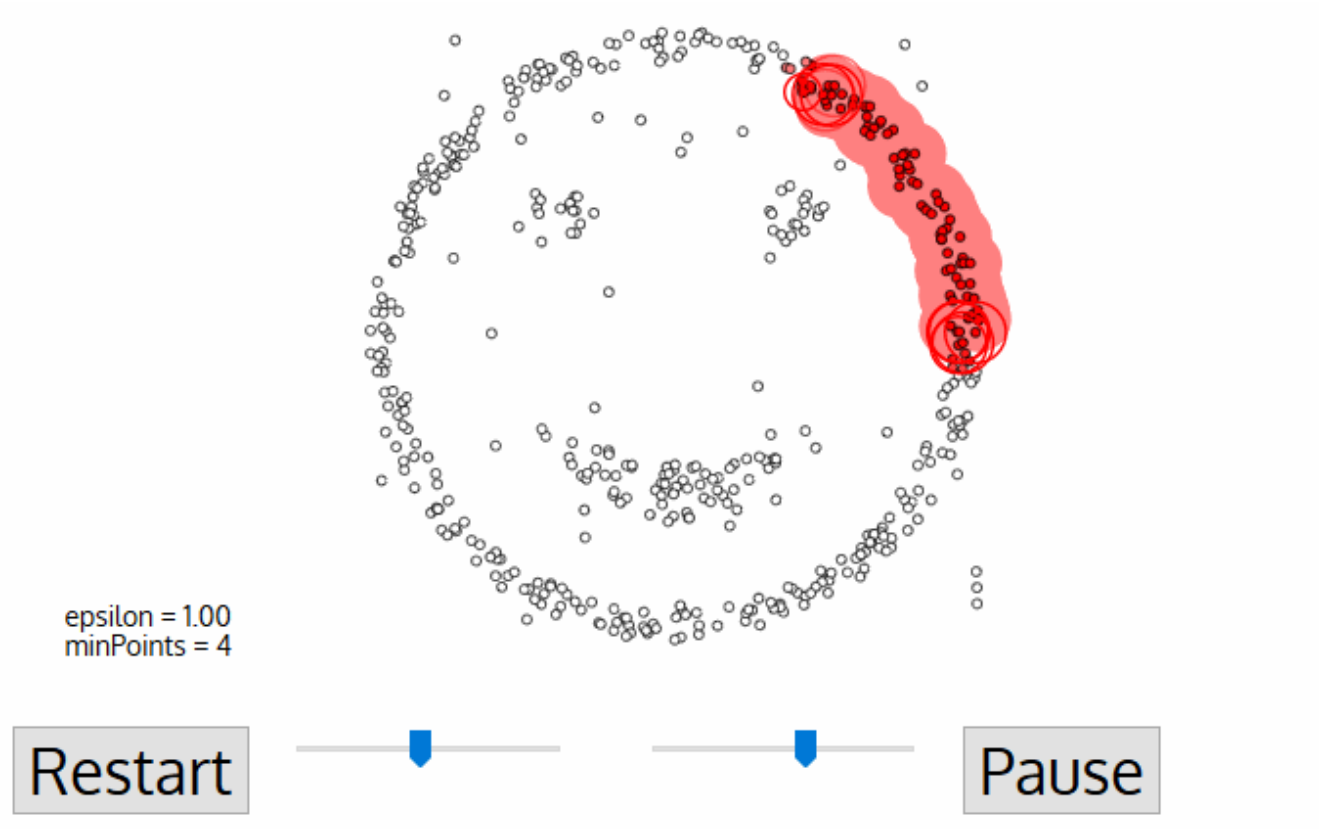
1. Percorra a BD e rotule os objetos como core, border ou noise
2. Elimine aqueles objetos rotulados como **noise**
3. Insira uma aresta entre cada par de objetos **core** vizinhos

- 2 objetos são vizinhos se um estiver dentro do raio Eps do outro

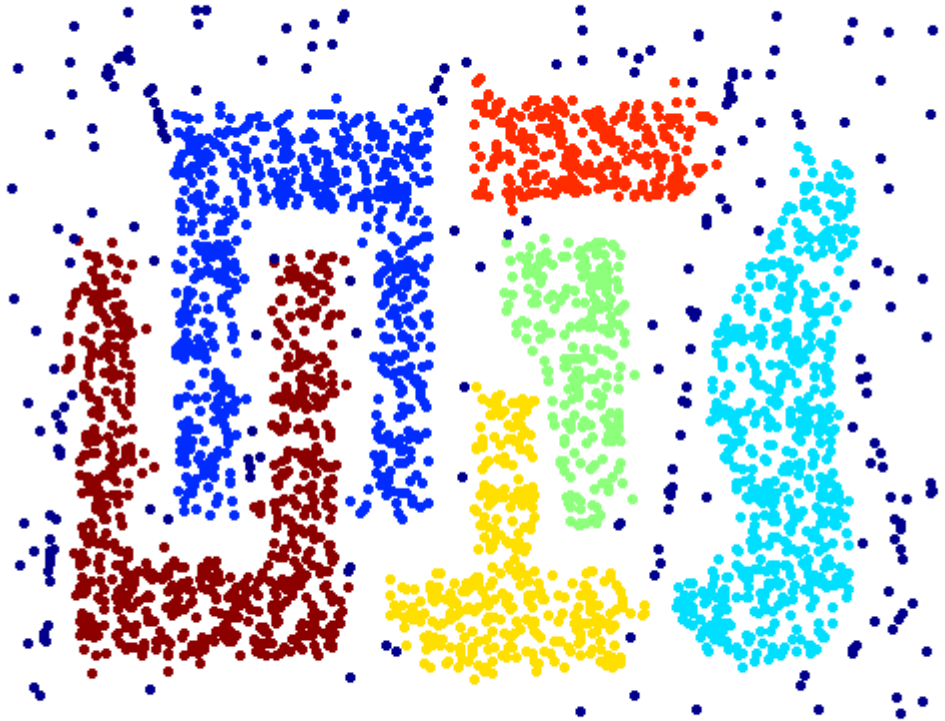
4. Faça cada componente conexo resultante ser um cluster
5. Atribua cada **border** ao cluster de um de seus core associados

- Resolva empates se houver objetos core associados de diferentes clusters

DBScan: algoritmo



DBScan: Exemplo



Point types: **core**, **border** and **noise**

Resumo das (des)vantagens do DBScan

Vantagens

- Não necessita do número de clusters a priori
- Consegue encontrar clusters com formatos arbitrários
- Tem uma definição de ruído e é robusto a outliers
- Necessita de apenas dois parametros:
  - Raio
  - Número de vizinhos para virar core (minpts)

Desvantagens

- Extremamente sensível aos parametros Raio e minPts
- Depende da distância utilizada para determiner se um ponto está ou nao presente dentro do raio. (tipicamente se utiliza euclidiana)
- Não consegue clusterizar dados com grupos com grandes diferenças de densidades
- Se a escala dos dados não for conhecida, desterminar o raio pode ser difícil

Questions and Feedback

Thank you!

Obrigado !

Vinicius Fernandes Caridá
vfcarida@gmail.com

@Vinicius Caridá

@Vinicius Caridá
@machine learning Brasil

@vfcarida

Copyright © **2018 Prof. Vinicius Fernandes Caridá**
Todos direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proíbido sem o consentimento formal, por escrito, do Professor (autor).