# DBSCAN

Um estudo sobre o algoritmo de agrupamento por densidade.

Brainer Sueverti de Campos

# O que é o DBSCAN?

- É uma algoritmo de aprendizado de máquina (não supervisionado) e de agrupamento.
- Baseado em Densidade.
- DBSCAN - Density-Based Spatial Clustering of Applications of Noise.
- A principal ideia do algoritmo é separar regiões de alta densidade por regiões de baixa densidade.

# Parâmetros

- O algoritmo possui dois parâmetros, sendo: **epsilon** (raio da hiperesfera) e a **quantidade mínima de pontos dentro da região**.
- Para um epsilon muito grande haverá junção de clusters e para um epsilon muito pequeno boa parte dos dados não serão agrupados. ***Geralmente***, **valores pequenos** geram melhores resultados.
- Há uma regra para determinar a quantidade mínima de pontos, sendo: **num_min >= dimensão + 1**. Obs.: num_min podem ajudar com dados com ruídos.

# Como funciona? - Definições

- **Elementos de Núcleos** - São os elementos que possuem a quantidade mínima ou mais de elementos dentro da sua hiperesfera de raio epsilon.
- **Elementos de Bordas** - São os elementos que não possuem a quantidade mínima de elementos dentro da sua hiperesfera de raio epsilon, *porém possui um elemento de núcleo dentro da hiperesfera***.**
- **Outliers (ou Ruídos)** - São os elementos que não possuem ligação com nenhum elemento de núcleo e não possui a quantidade mínima de elementos dentro de sua hiperesfera.

# Como funciona? - Passos

O Algoritmo possui alguns passos, sendo eles:

- Escolha de um elemento arbitrário (aleatoriamente).
- Classificar o elemento como Núcleo, Borda ou Outlier. Se for núcleo, o elementos e os elementos dentro da hiperesfera tornam-se um cluster
- Depois de classificar todos os pontos da região do núcleo, de forma recursiva, volta-se ao passo 1. (chain effect).
- O algoritmo acaba quando todos os pontos foram visitados.

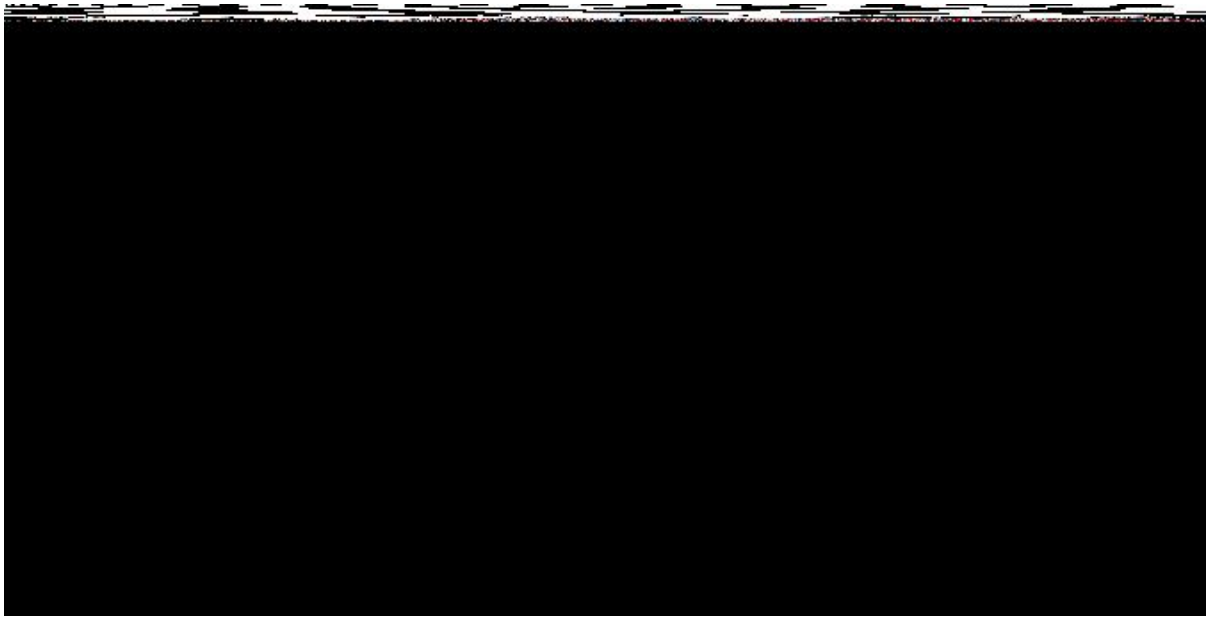
# Como funciona? - Classificação

- Feito a seleção do elemento a ser analisado, é necessário verificar quantos pontos estão dentro da hiperesfera, através da comparação com os outros elementos usando uma função de distância (geralmente é a euclidiana).
- Assim, é feita a classificação,  de acordo com as definições já abordadas.
- Pode haver reclassificações.

# Como funciona? - Recursão

- Se o ponto classificado for de núcleo, é recuperado todos os pontos dentro da sua hiperesfera e classificado. Caso seja núcleo, é feita a recursão. Esse efeito é chamado de chain effect.
- Finalizado a recursão, é escolhido outro ponto aleatório e feito os mesmos passos.
- Assim, o algoritmo acaba quando nenhum elemento mudar de grupo.

# Como funciona? -  Demonstração



# Eficiência

O DBSCAN tem complexidade **O($n^2$)**. Se utilizar R-Tree, a complexidade é **O(nlogn)**.

# Vantagens

- Não é necessário o número de classes, como o K-Means.
- Caso haja um bom reconhecimento dos dados, a escolha dos parâmetros torna-se fácil.
- A classificação de Outliers é bem feita.
- Os dados podem ser esféricos ou não.

# Desvantagens

- Caso não haja um bom conhecimento prévio dos dados, a determinação dos parâmetros torna-se um pouco mais difícil.
- Quando a função de distância é a euclidiana e o espaço tem alta dimensionalidade, o algoritmo não funciona bem.
- Quando os agrupamentos possuem densidades variadas, o algoritmo não funciona bem.

# Dúvidas

- Como funciona a função que retorna um valor heurístico para o valor do epsilon?
- Quais outras funções de distância existem?
- Quando a complexidade é O(n log n)? (R-Tree)
- O que seria densidades variadas?

# Referências


- DBSCAN - Clusterização. Disponível em: <https://pt.coursera.org/lecture/machine-learning-with-python-pt/dbscan-B8ctK>. Acesso em: 9 dez. 2022.

- SALTON, K. How DBSCAN works and why should we use it? Disponível em: <https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80>.

- MONTEIRO, G. Entendendo DBSCAN. Disponível em: <https://gabriellm.medium.com/entendendo-dbscan-770f680d9160>. Acesso em: 9 dez. 2022.

- NALDI, M. C. Técnicas de combinação para agrupamento centralizado e distribuído de dados. [s.d.].

- Ester, M., H.-P. Kriegel, J. Sander, & X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), pp. 226–231.