### **Módulo 7 - Aprendizagem não supervisionada** Aula 52b - Introdução a Density Based Spatial Clustering of Application with Noise (DBSCAN)

**Objetivos de aprendizagem**

1. Apresentar o conceito de DBSCAN
2. Fazer uma análise de clustering com DBSCAN
3. Avançar na interpretação dos resultados da análise
4. Conceituar e compreender as diferenças em relação aos métodos de clustering aprendidos até agora

**Conteúdo**

1. Visão geral do algoritmo DBSCAN
2. Diferenças entre DBSCAN e os outros
   1. Clustering baseado em densidade versus clustering baseado em distâncias
   2. Definição de “ruído” em DBSCAN
   3. Capacidade de detectar bordas “não lineares” entre clusters
3. Desvantagens do DBSCAN

**Modalidades de aula incluídas**

|  |  |
| --- | --- |
| hablar.png | Exposição teórica oral e interativa, com material de apoio por escrito |

|  |  |
| --- | --- |
|  | Práticas interativas, grupais e individuais, usando notebooks |

### Roteiro da aula

|  |  |
| --- | --- |
| 1. **Boas-vindas** | **5 minutos** |

|  |  |
| --- | --- |
| hablar.png | * Verificar se os PCs estão funcionando corretamente * Fazer uma revisão das noções gerais de clustering k-means e sua avaliação |

|  |  |
| --- | --- |
| 1. **DBSCAN: algoritmo** | **25 minutos** |

|  |  |
| --- | --- |
| hablar.png | * Visão geral do algoritmo   + Explicação dos parâmetros a definir:     - **épsilon:** é a distância máxima entre dois pontos para que sejam considerados pertencentes ao mesmo cluster (ou seja, o raio máximo do cluster)     - **min\_points:** ou seja, o mínimo de pontos necessários para formar um cluster (a ideia desse parâmetro é evitar a formação de clusters pequenos demais)   + Explicação rápida do algoritmo * Visualização gráfica: <https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/> |

|  |  |
| --- | --- |
| 1. **Diferenças em relação a K-Means e Clustering hierárquico e desvantagens** | **15 minutos** |

|  |  |
| --- | --- |
| hablar.png | * Diferenças e vantagens:   + A diferença fundamental em relação a k-means é que se baseia na “densidade” em vez de começar pela distância até um ponto central (centroide).   + Não exige a definição antecipada de um número de clusters   + Também é muito útil quando a distribuição dos dados é muito densa:     - nesses casos, é provável que k-means apresente apenas um cluster. DBSCAN deveria poder dividir esse agrupamento em partes menores   + Pode encontrar clusters com limites de formas “irregulares”   + Incorpora no processo o conceito de “ruído”, que deixa os outliers mais robustos   + É determinístico (depois que os parâmetros iniciais são definidos) * Desvantagens:   + Conjuntos de dados com zonas com densidades muito diferentes no espaço de pontos: nesse caso, pode não funcionar corretamente   + Se a escala dos dados não está clara, pode ser difícil encontrar os parâmetros. |

|  |  |
| --- | --- |
| 1. **Prática guiada: DBSCAN no conjunto de dados IRIS** | **30 minutos** |

|  |  |
| --- | --- |
|  | * Fazer um clustering DBSCAN com o conjunto de dados * Visualizar, avaliar e conversar sobre os resultados |

|  |  |
| --- | --- |
| 1. **LAB: Clustering hierárquico e DBSCAN sobre o conjunto de dados de câncer de mama (Scikit-Learn)** | **30 minutos** |

|  |  |
| --- | --- |
|  | * O LAB tem três objetivos:   + conseguir estimar e avaliar um clustering hierárquico   + conseguir estimar e avaliar um clustering gerado por DBSCAN (testando diferentes valores de `eps` e `min\_samples` para DBSCAN)   + conseguir interpretar os resultados dos dois métodos:   + De forma geral, os alunos deveriam perceber que o DBSCAN não funciona bem e que, provavelmente, não exista uma estrutura clara nesse conjunto de dados. Dessa forma, é possível apresentar o problema da alta dimensionalidade (o conjunto de dados tem 30 features) e também mostrar o problema de redução de dimensionalidade (tema das próximas aulas) |

|  |  |
| --- | --- |
| 1. **Conclusão** | **5 minutos** |

|  |  |
| --- | --- |
| hablar.png | * Características de DBSCAN:   + Não exige a definição prévia de um número de clusters   + É determinístico - não apresenta um resultado diferente cada vez que é executado   + Pode encontrar “bordas” de clusters com formas irregulares   + É robusto para outliers |