### **Módulo 7 - Aprendizagem não supervisionada** Aula 51a - Introdução a Clustering

**Objetivos de aprendizagem**

1. Analisar formato e pré-processamento para aplicar algoritmos de clustering
2. Apresentar os algoritmos mais usados (k-means, hierarchical, etc.)
3. Fazer e interpretar uma análise de cluster com K-Means
4. Apresentar conceitos de avaliação para o ajuste de clustering

**Conteúdo**

1. Medidas de semelhança - distância
2. Algoritmo K-means
3. Avaliação de clusters

**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 da aprendizagem supervisionada * Pergunta motivadora: Qual é a característica fundamental dos problemas de aprendizagem supervisionada?   + Resposta: Existe uma variável resposta para modelar * O que acontece se essa variável não existe? |

|  |  |
| --- | --- |
| 1. **Introdução à aprendizagem não supervisionada** | **15 minutos** |

|  |  |
| --- | --- |
| hablar.png | Expor as diferenças entre aprendizagem não supervisionada e supervisionada   * Ausência de variável de resposta * Complexidade da avaliação dos modelos |

|  |  |
| --- | --- |
| 1. **Introdução a clustering - Apresentação do algoritmo K-Means** | **20 minutos** |

|  |  |
| --- | --- |
| hablar.png | * Visão geral sobre clustering (tipos de clustering: algoritmos, fuzzy, probabilísticos, etc.) * Revisão de métricas de distância (Minkowski, Manhattan e Euclidiana) * Descrição do algoritmo k-means   + Atribuição aleatória de pontos a clusters e cálculo de centroides   + Descrição da função objetivo: minimizar a distância até os centroides dos clusters   + Repetição até convergência   + Necessidade de predefinir o número de clusters (k) |

|  |  |
| --- | --- |
| 1. **Prática guiada: K-Means Clustering em Python** | **30 minutos** |

|  |  |
| --- | --- |
|  | * Executar um notebook com a implementação de um clustering baseado em k-means sobre o conjunto de dados “IRIS”. * Antes disso, fazer um breve pré-processamento dos dados. |

|  |  |
| --- | --- |
| 1. **Algumas limitações do algoritmo K-means** | **5 minutos** |

|  |  |
| --- | --- |
| hablar.png | * Como começa com uma atribuição aleatória de clusters, pode ser que os resultados variem em diferentes execuções * É necessário ajustar em parâmetro K (número de clusters) * Pelas características do algoritmo, o mínimo global pode não ser alcançado * K-means só consegue captar separações lineares entre os dados. |

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

|  |  |
| --- | --- |
| hablar.png | * O processo de clustering pode ser muito útil para tentar entender um set de dados não classificado (não supervisionado). * De acordo com o algoritmo utilizado para o processo, é possível obter uma classificação categórica da classe (K-Means) ou probabilística (Fuzzy C-means) * O resultado de K-Means depende das condições iniciais para os centroides e a quantidade de k clusters escolhidos (convergência local). * Representar os dados por pares de variáveis ajuda a escolher um número de k clusters. * Silhouette score é uma boa métrica para medir a qualidade dos clusters obtidos. |