### **Módulo 7 - Aprendizagem não supervisionada** Classe 51b - Ajuste e avaliação de clustering

**Objetivos de aprendizagem**

1. Entender a composição dos clusters e como foram construídos
2. Compreender e incorporar conceitos relativos à avaliação do resultado da análise realizada por K-Means
3. Introduzir o conceito de “avaliação interna” e “avaliação externa” de um processo de clustering
4. Apresentar algumas medidas de avaliação

**Conteúdo**

1. Medidas de avaliação interna:
   1. Análise visual
   2. Silhouette Score
   3. Calinski-Harabaz Index
2. Medidas de avaliação externa:
   1. Accuracy
   2. F1-Score
   3. Confusion Matrix

**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 * Revisão de noções gerais de clustering (homogeneidade interna/heterogeneidade externa) |

|  |  |
| --- | --- |
| 1. **Técnicas de avaliação de clusters: Medidas externas** | **10 minutos** |

|  |  |
| --- | --- |
| hablar.png | * Utilizam alguma variável “correlacionada” ao clustering para avaliar o desempenho dos clusters. * As medidas de avaliação externa são iguais às utilizadas em problemas de classificação, mas a lógica é diferente: clustering é uma questão de aprendizagem não supervisionada, portanto não modela uma variável resposta, mas sim, usa uma variável “externa” para avaliar os resultados. Vamos ver as seguintes medidas (destacar que existem muitas outras):   + Accuracy   + Medida F   + Confusion Matrix |

|  |  |
| --- | --- |
| 1. **Técnicas de avaliação de clusters: Medidas internas** | **10 minutos** |

|  |  |
| --- | --- |
| hablar.png | * Buscam avaliar a semelhança entre os membros de um cluster (homogeneidade) e a diferença entre eles e o restante dos clusters. Ou seja, se baseiam nas informações intrínsecas do conjunto de dados disponível. Vamos ver as seguintes medidas (destacar que existem muitas outras):   + Técnicas visuais (úteis, porém limitadas a dados com baixa dimensionalidade)   + Silhouette Score   + Calinski-Harabaz Index |

|  |  |
| --- | --- |
| 1. **Prática guiada: Ajuste de K-Means Clustering em Python** | **20 minutos** |

|  |  |
| --- | --- |
|  | * Executar um notebook com a implementação de um clustering baseado em k-means sobre o conjunto de dados “IRIS”. * Calcular as diversas medidas de avaliação sobre o conjunto de dados   + Para as de avaliação externa, utilizar como variável extra a espécie da flor * A partir das medidas aprendidas, buscar o k ideal para o conjunto de dados. |

|  |  |
| --- | --- |
| 1. **LAB: Clustering e avaliação sobre o conjunto de dados de câncer de mama (Scikit-Learn)** | **40 minutos** |

|  |  |
| --- | --- |
|  | * Para esta seção de laboratório, vamos concluir um processo completo de K-Means. * Trabalhar com o conjunto de dados de câncer de mama elaborado pela Universidade de Wisconsin. As features foram calculadas a partir de imagens digitalizadas de tecido mamário. Elas descrevem características dos núcleos celulares nas imagens. * Objetivo: encontrar grupos de tecidos similares em função dessas características. * Começar com um k=4 e avaliar esse processo de clustering * Por último, encontrar o k ideal para o conjunto de dados e conversar sobre as diferenças entre as medidas de avaliação. |

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

|  |  |
| --- | --- |
| hablar.png | * Existem diferentes métodos para avaliar a qualidade de uma análise (incluindo visualização, silhouette scores, F-metrics e matrizes de confusão). * Depois de analisar os clusters, talvez seja necessário redefinir a quantidade “k” de clusters procurados. * Sempre é uma boa ideia examinar várias métricas para entender a qualidade da análise realizada. |