Universidade do Minho
Departamento de Informática
Mestrado [Integrado] em Engenharia Informática

Dados e Aprendizagem Automática
1º Ano, 1º Semestre
Ano letivo 2021/2022

Enunciado Prático nº 6
18 de novembro de 2021

Tema *Clustering K-means* e *K-medoids*

Enunciado

A aprendizagem não-supervisionada é essencialmente utilizada para obter inferências de conjuntos de dados sem intervenção humana, em contraste com a aprendizagem supervisionada, em que os *labels* são fornecidos em conjunto com os dados. Duas técnicas aplicadas neste contexto são os modelos *K-means* e *K-medoids*. Ambos os algoritmos apresentam como objetivo o agrupamento de um conjunto de casos de estudo não "rotulados" (i.e., sem *label*), atendendo à semelhança das suas características. No entanto, enquanto que o *K-means* tenta minimizar as distâncias dentro do cluster, o *K-medoids* tenta minimizar a soma das distâncias entre cada ponto e o "*medoid*" do respetivo cluster.

Tarefas

Com este enunciado é pretendido agrupar um conjunto de universidades em dois grupos: institutos privados ou institutos públicos. O respetivo *dataset* encontra-se disponível em https://rb.gy/fwxvoq. Atendendo às características apresentadas, foi decidido aplicar um conjunto de modelos não-supervisionados, especificamente o agrupamento *K-means* e *K-medoids*, como forma de resolver este problema de classificação binária.

Após descarregarem o *dataset*, deverão:

T1. Carregar o *dataset*, utilizando a função *pandas.read_csv(…)*;

T2. Aplicar métodos para exploração e visualização de dados;

T3. Treinar um modelo de aprendizagem não-supervisionado de agrupamento *K-means* (*sklearn.cluster.KMeans*) e *K-medoids* (*sklearn_extra.cluster.KMedoids*), classificando cada caso de estudo como "instituto privado" ou "instituto público" (*n_clusters = 2*);

Nota: O atributo *"Private"* indica o rotulo de cada universidade, apresentando se a universidade é um instituto privado. Para efeitos de treino, deverá ser removido este atributo do *dataset*.

T4. Atendendo ao valor do atributo "*Private*", avaliar a performance de agrupamento de cada modelo, através da criação de uma matriz de confusão *sklearn.metrics.confusion_matrix(…)* e de um relatório de classificação *sklearn.metrics.classification_report(…);*

T5. Atendendo aos resultados obtidos em **T4**, quais as conclusões adquiridas? Em que situações o modelo acerta/falha? Como melhorar o modelo de aprendizagem proposto?