Universidade do Minho
Departamento de Informática
Mestrado [integrado] em Engenharia Informática
Mestrado em Engenharia de Sistemas

Perfil de Machine Learning: Fundamentos e Aplicações
Sistemas Baseados em Similaridade
4º/2º Ano, 1º Semestre
Ano letivo 2020/2021

Enunciado Prático nº 4
05 de novembro de 2020

Tema

Clustering

Enunciado

Pretende-se, com este enunciado prático, que sejam aplicados métodos de *clustering* sobre um *dataset* de vinhos, o qual contém um ficheiro para aprendizagem e outro para teste. Deverão também ser aplicadas técnicas para exploração e tratamento de dados, assim como para parametrização do *workflow* a desenvolver.

Tarefas

Numa primeira fase devem descarregar o *dataset* disponível em https://goo.gl/8jjW8t. Devem, de seguida:

T1. Carregar, no *Knime*, o *dataset* descarregado e explorar os dados;

T2. Tratar os dados, i.e.:

a. Fazer cast do atributo "*quality*" para inteiro;
b. Normalizar todos os atributos numéricos utilizando a transformação linear Min-max de forma a produzir um input normalizado entre 0 e 1;
c. Criar 4 *bins* de igual frequência para a *feature* "*citric acid*", substituindo a *feature* original;
d. Renomear cada *bin* de forma a que o primeiro corresponda a *Low*, o segundo a *Medium*, o terceiro a *High* e o quarto a *Very High*.
    - Dica: no passo anterior usar *Numbered* como *Bin Naming* – podem depois usar os nodos *Table Creator* e *Cell Replacer*.

T3. Aplicar:

a. Uma Análise de Componentes Principais (PCA) de forma a projetar os dados em apenas duas dimensões;
b. Utilizar um *scatter plot* para visualização dos resultados obtidos pelo PCA.

T4. Segmentar o *dataset*:

a. Aplicando o método *k-means*;
b. Atribuir diferentes cores por qualidade do vinho e diferentes formas aos clusters;
c. Criar *scatter plots* e *scatter matrixes* que permitam ter uma noção gráfica, em duas dimensões, dos atributos e dos clusters criados;
d. Ler e tratar os dados de teste de forma a que, com base no modelo desenvolvido nos passos anteriores, seja atribuído um cluster a cada registo deste ficheiro;
e. Guardar o resultado da atribuição num ficheiro csv.

T5. Parametrizar o *workflow*, utilizando variáveis de fluxo para definir o número de *bins*, o número de *clusters* e os títulos dos gráficos criados;

T6. Produzir o *workflow* de maneira a que seja possível visualizar, numa única página, todos os componentes visuais implementados;

T7. Experimentar, avaliar e comparar outros métodos de segmentação.