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 2019/2020

Ficha Prática nº 6
07 de novembro de 2019

Tema

Segmentação/*Clustering*.

Enunciado

Criação e análise de *clusters* sobre dois *datasets*. O primeiro refere-se a um *dataset* de vinhos e contém um ficheiro para aprendizagem e um outro para teste (https://goo.gl/8jjW8t). O segundo *dataset* contém dados referentes a 3 fabricantes de automóveis (https://goo.gl/Eap319).

Tarefas

Para o primeiro *dataset* o objetivo deste exercício é o de criar clusters de acordo com as características químicas dos vinhos usando algoritmos de segmentação, como o *k-means*. Deve assim ser desenvolvido um *workflow* na plataforma *Knime* para:

- Tratar o atributo *quality* de forma a torná-lo num inteiro;
- Normalizar os atributos numéricos utilizando a transformação linear Min-max de forma a produzir um *input* normalizado sobre o qual deve ser aplicado o algoritmo de segmentação;
- Atribuir diferentes cores por qualidade do vinho e diferentes formas aos clusters;
- Criar *scatter plots* e *scatter matrixes* que permitam ter uma noção gráfica, em duas dimensões, dos atributos e dos clusters criados;
- 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;
- Guardar o modelo no formato PMML.

Para o segundo *dataset*, deve-se proceder à criação de clusters de forma similar à descrita em cima, i.e.:

- Tratar os atributos e normalizá-los;
- Aplicar diferentes algoritmos de segmentação sobre os dados normalizados e sobre os dados não normalizados;
- Atribuir diferentes cores e formas aos clusters. Criar *plots* que permitam ter uma noção gráfica dos atributos. Qual o impacto da normalização dos dados?
- Aplicar uma Análise de Componentes Principais (*Principal Component Analysis*) de forma a projetar os dados em apenas duas dimensões. Qual o impacto da normalização dos dados no PCA?
- Guardar o modelo no formato PMML.