Universidade do Minho
Licenciatura em Engenharia Informática
Mestrado integrado em Engenharia Informática

Aprendizagem e Decisão Inteligentes
3°/4° Ano, 2° Semestre
Ano letivo 2021/2022

Enunciado Prático n° 6
Março, 2022

Tema

K-means Clustering

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. Os modelos de aprendizagem não supervisionada apresentam como objetivo o agrupamento de um conjunto de casos de estudo não "rotulados" (i.e., sem *label*), atendendo à semelhança das suas características.

Tarefas

Parte I

Para esta tarefa, deverão agrupar um conjunto de flores em diferentes *clusters* (i.e., de acordo com a sua espécie: "iris-setosa", "iris-versicolor", "iris-virginica"), atendendo às suas características. Devem, para isso, descarregar o *dataset* [iris_data] disponível na plataforma de *e-learning* da UMinho, secção [Conteúdo]. Atendendo às características apresentadas, foi decidido aplicar um conjunto de modelos não-supervisionados, especificamente a segmentação *K-means*.

T1. Carregar, no *Knime*, o *dataset* [iris_data] e aplicar nodos de exploração de dados de modo a permitir a análise dos dados;

T2. Particionar os dados, de forma aleatória, utilizando 80% para aprendizagem e 20% para teste;

T3. Remover dos dados de aprendizagem as colunas "Id" e "Species";

T4. Aplicar o nodo **k-Means** para treinar o respetivo modelo de aprendizagem não supervisionada, como forma de classificar cada caso de estudo como "iris-setosa", "iris-versicolor" ou "iris-virginica" (*number of clusters* = 3):

T5. Aplicar os nodos de visualização *Color Manager*, *Shape Manager* e *Scatter Plot* para representar os diferentes casos de estudo e respetivos *clusters* associados;

T6. Aplicar o nodo *Cluster Assigner* para inferir sobre os dados de teste utilizando o modelo treinado no nodo **k-Means**. Após esta tarefa, aplique o nodo *Rule Engine* para adequar o nome dos *clusters* atribuídos para cada caso de estudo ao respetivo nome da espécie da planta (coluna "*Species*");

T7. Avalie o desempenho dos modelos de aprendizagem **k-Means** treinados na T4 através do uso de matrizes de confusão e métricas de avaliação (use o nodo *Scorer (JavaScript)*). Quais os resultados obtidos? Em que situações o modelo acerta/falha? Como melhorar o modelo de aprendizagem proposto?

Tarefas

Parte II

Com esta tarefa é pretendido agrupar um conjunto de universidades em dois grupos: instituições **privadas** ou instituições **públicas**. Devem descarregar o *dataset* [college_data] disponível na plataforma de *e-learning* da UMinho, secção [Conteúdo]. Atendendo às características apresentadas, foi decidido aplicar um conjunto de modelos não-supervisionados, especificamente o agrupamento **k-Means**, como forma de resolver este problema de classificação binária.

T1. Carregar, no *Knime*, o *dataset* [college_data_train] e aplicar nodos de exploração de dados como forma de permitir a análise dos dados;

T2. Proceder ao tratamento e limpeza dos dados;

T3. Aplicar o nodo **k-Means** para treinar o respetivo modelo de aprendizagem não supervisionado, de modo a classificar cada caso de estudo como "instituto privado" ou "instituto público" (*number of clusters* = 2):

T4. Aplicar os nodos de visualização *Color Manager*, *Shape Manager* e *Scatter Plot* para representar os diferentes casos de estudo e respetivos *clusters* associados;

T5. Carregar, no *Knime*, o *dataset* [college_data_test] que apresenta o conjunto de dados de teste (com a adição do atributo "*Private"*, representando se a universidade é um instituto privado ou público). Proceder ao seu tratamento e limpeza dos dados;

T6. Aplicar o nodo *Cluster Assigner* para inferir sobre os dados de teste utilizando o modelo treinado no nodo **k-Means**. Após esta tarefa, aplique o nodo *Rule Engine* para adequar o nome dos *clusters* atribuídos para cada caso de estudo à respetiva classificação do instituto (coluna "*Private"*);

T7. Avalie o desempenho dos modelos de aprendizagem **k-Means** treinados na T3 através do uso de matrizes de confusão e métricas de avaliação (use o nodo *Scorer (JavaScript)*). Quais os resultados obtidos? Em que situações o modelo acerta/falha? Como melhorar o modelo de aprendizagem proposto?