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 4
17 de outubro de 2019

Tema

Sistemas Baseados em Similaridade - Concepção de modelos de *machine learning*.

Enunciado

Uma multinacional na área do retalho possui o histórico de vendas semanais de 17 das suas lojas em diferentes regiões do país, sendo que cada loja contém vários departamentos (desporto, cozinha, produtos alimentícios e higiene pessoal, entre outros). A empresa realiza também vários eventos promocionais ao longo do ano, normalmente precedendo feriados importantes. A empresa pretende agora extrair informação relevante dos *datasets* e desenvolver um modelo de *machine learning* que, com base num conjunto relevante de *features*, permita estimar as vendas mensais de cada uma das suas lojas.

Tarefas

A empresa possui dois *datasets*: o primeiro (https://goo.gl/wxdAk4) contém informação sobre cada uma das lojas, incluindo o seu tipo e tamanho, enquanto que o segundo (http://bit.ly/2oMYLdZ) contém dados referentes às vendas semanais de cada departamento de cada loja, a data e um *boolean* indicando se houve um feriado durante essa semana. Deve agora ser desenvolvido um único *workflow* na plataforma Knime para:

- Carregar, juntar, explorar e analisar os referidos *datasets*, utilizando vistas gráficas que permitam ao utilizador perceber a análise efectuada;
- Num *Line Plot* mostrar a média semanal de vendas de cada uma das 17 lojas de forma descendente;
- Num *Bar Chart* mostrar a média semanal de vendas dos 10 departamentos que mais venderam.

Mais tarde a empresa forneceu um terceiro *dataset* (http://bit.ly/2MoReLz) que contém também dados de vendas de cada uma das lojas, mas referente a meses posteriores ao fornecido nos *datasets* anteriores. O objectivo passa por utilizar este terceiro *dataset* única e exclusivamente como conjunto de teste aquando do desenvolvimento de modelos de *machine learning* de forma a garantir que o modelo é avaliado com dados que desconhece. Para treino dos referidos modelos deverão utilizar os dois *datasets* fornecidos inicialmente. Devem assim desenvolver um modelo de *machine learning* que, utilizando árvores de decisão, seja capaz de prever as vendas mensais de cada loja. Para isso, devem aplicar os seguintes tratamentos aos dados de treino:

- Fazer binary encoding à feature *isHoliday* (1 deve corresponder ao valor *True*);
- Adicionar, a cada registo, as *features* ano e mês;

- Agrupar os registos por loja, tipo, tamanho, ano e mês, agregando de forma a obter o somatório das vendas semanais de cada loja e a indicação da existência de feriados nesse mês;
- Normalizar o somatório das vendas semanais utilizando a transformação linear *Min-Max* entre 0 e 1;
- Criar 4 *bins* de igual frequência sobre o valor normalizado no passo anterior;
- 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*;
- Utilizar estes dados tratados para treinar uma árvore de decisão. Analisar e mostrar graficamente a árvore criada pelo *Learner*;
- Carregar o *dataset* de teste e prever o valor (i.e., a classe) de vendas de cada mês de cada uma das 17 lojas;
- Mostrar, graficamente, uma tabela com a matriz de confusão do modelo.

Como extra:

- Produzir o *workflow* de maneira a que seja possível visualizar, numa só página, todos os componentes visuais implementados;
- Criar uma variável de fluxo com o valor “Sales Report and Modelling”, que deverá ser utilizada em cada gráfico para redefinir o título do mesmo.