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 5
12 de novembro de 2020

Tema

Tuning de Modelos Baseados em Árvores

Enunciado

Pretende-se, com este enunciado prático, que seja feito o *tuning* de modelos baseados em árvore, abordando parâmetros nominais e numéricos como a medida de qualidade, o método de *pruning* e o número mínimo de registos por nodo, entre outros.

Tarefas

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. 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. Um terceiro *dataset* (http://bit.ly/2MoReLz) deve ser utilizado, única e exclusivamente, como conjunto de teste aquando do desenvolvimento dos modelos de *machine learning* de forma a garantir que estes são avaliados com dados que desconhecem.

Assim, deve agora ser desenvolvido um *workflow* para:

T1. Carregar, no *Knime*, os dois primeiros *datasets*, juntá-los e explorar os dados utilizando vistas gráficas que permitam perceber a análise efetuada;

T2. Tratar os dados, i.e.:

a. Fazer label encoding à feature *isHoliday* (1 deve corresponder ao valor *True*);
b. Adicionar, a cada registo, as *features* ano e mês;
c. 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;
d. Normalizar o somatório das vendas semanais utilizando a transformação linear *Min-Max* entre 0 e 1;
e. Criar 4 *bins* de igual frequência sobre o valor normalizado no passo anterior (ligando a opção *replace target column(s)*);
f. 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*.

T3. Treinar:

a. Uma árvore de decisão;
b. Carregar o *dataset* de teste e prever o valor de vendas de cada mês para cada uma das 17 lojas;
c. Mostrar, graficamente, uma tabela com a matriz de confusão do modelo.

T4. Fazer o *tuning* do modelo criado no passo anterior, experimentando:

a. Todos os valores, entre 2 e 10, para o número mínimo de registos por nodo;
b. Todas as possibilidades para a medida de qualidade;
c. Todas as possibilidades para o método de *pruning*;
d. Fazer o *tuning* dos parâmetros anteriores num único *workflow*. Guardar e analisar todos os resultados obtidos para cada combinação de hiper-parâmetros. Qual a combinação que oferece melhor performance? Existem grandes discrepâncias?

T5. Treinar e fazer o *tuning* de uma *Random Forest*. Guardar e analisar todos os resultados obtidos para cada combinação de hiper-parâmetros;

T6. Analisar e comparar as performances dos modelos treinados em *T4* e *T5*. Que conclusões se podem tirar?