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° 5
Março, 2022

Tema

Modelos de Aprendizagem baseados em Árvores de Decisão

Enunciado

Com este enunciado prático pretende-se que sejam desenvolvidos modelos de aprendizagem baseados em árvores de decisão, 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 lojas em diferentes regiões do país, sendo que cada loja contém vários departamentos (desporto, cozinha, produtos alimentícios, 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 aprendizagem que, com base num conjunto relevante de *features*, permita estimar as vendas mensais de cada loja. A empresa possui dois *datasets*: o primeiro (*dataset* [lojas] disponível na plataforma *e-learning* da UMinho, secção [Conteúdo]) contém informação sobre cada uma das lojas, incluindo o seu tipo e tamanho; o segundo (*dataset* [sales_training_set] disponibilizado de modo idêntico) 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 (*dataset* [sales_test_set] disponível nos mesmos termos) deve ser utilizado, única e exclusivamente, como conjunto de teste por ocasião do desenvolvimento dos modelos de aprendizagem, de modo a garantir que estes são avaliados com dados que desconhecem.

Deve ser desenvolvido um *workflow* para:

T1. Carregar no *Knime* os dois primeiros *datasets*, juntá-los e explorar os dados utilizando nodos de visualização gráfica que permitam interpretar a análise efetuada;

T2. Proceder ao tratamento e limpeza dos dados:

a) Fazer *label encoding* da *feature isHoliday* (o valor 1 deverá 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 de 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 (ativar 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 capaz de prever o valor de vendas de cada mês para cada uma das 17 lojas;
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) Guardar e analisar todos os resultados obtidos para cada combinação de hiper-parâmetros averiguada. Qual a combinação que oferece melhor desempenho? Existem grandes discrepâncias?

T5. Treinar um modelo de aprendizagem de floresta aleatória (*random forest*). Guardar e analisar todos os resultados obtidos para cada combinação de hiperparâmetros averiguada;

T6. Analisar e comparar os desempenhos dos modelos treinados anteriormente. Que conclusões se podem tirar?