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° 7
Abril, 2022

Tema

Seleção de atributos e Validação de Modelos

Enunciado

Uma exploração aprofundada dos dados permite retirar ilações, muitas vezes escondidas, que poderão ser importantes para compreender o domínio e o problema em mãos.

Com este enunciado prático os alunos deverão aplicar um conjunto de técnicas que permitam explorar e tratar *datasets*, por forma a permitir explorar estratégias de seleção de características (*feature selection*) e usar diversas abordagens de validação de modelos (*cross validation*, etc.).

Tarefa 1

Numa primeira fase devem descarregar o *dataset users_sentiment* disponível na plataforma de *e-learning* da UMinho, secção [Conteúdo]. Este *dataset* contém dados de um conjunto de utilizadores de uma determinada plataforma *web* assim como o seu sentimento em relação à mesma.

T1. Carregar, no *Knime*, o *dataset users_sentiment* e aplicar nodos de exploração de dados como forma de permitir a análise dos dados em relação à sua:

a. Tendência central;
b. Dispersão estatística;
c. Correlação entre *features*;

tendo sempre em consideração o tipo e respetiva qualidade de dados e quantidade de conhecimento "acondicionadas" no *dataset*.

T2. Obter perspetivas gráficas de análise dos dados criando *plots* para visualização dos dados.

T3. Aplicar nodos para tratamento de dados de forma a, por exemplo:

a. Excluir todas as colunas do tipo *Double*;
b. Tratar valores em falta (*missing values*);
c. Remover registos duplicados;
d. Criar 3 *bins* de igual frequência para a *feature age*;
e. Para cada registo, extrair o ano, mês e dia da semana da *feature birthday*;
f. Excluir utilizadores da plataforma que tenham uma atividade na plataforma (*WebActivity*) inferior a 1 hora e que tenham mais de 70 anos;
g. Excluir todos os registos que contenham a *sub-string* "co" no produto.

Tarefa 2

Na sequência da **Tarefa 1** e utilizando o mesmo *dataset users_sentiment*, aplique um conjunto de estratégias de seleção de características (*feature selection*) e validação de modelos de aprendizagem automática.

Atendendo aos passos aplicados na **Tarefa 1 - T1**, **T2** e **T3**, resolva as seguintes tarefas:

T1. Aplicar nodos de tratamento de dados de forma a permitir o treino e validação de modelos de aprendizagem, incluindo, mas não se limitando a:

a. Transformar os valores categóricos em representações de valores numéricos;
b. Remover da análise de dados colunas potencialmente indesejadas;
c. Aplicar estratégias de normalização de dados a colunas numéricas que entenda relevante;
d. Tratar casos que sejam considerados *outliers*.

T2. Construir modelos de decisão, como, por exemplo:

a. Árvore de decisão para prever o valor de *Sentiment Analysis*;
b. Aplicar técnicas de validação de modelos (*k-fold cross-validation*), obtendo a tabela com os resultados de validação do modelo para cada *fold*;
c. Explorar o treino e validação de outras técnicas de modelos de aprendizagem supervisionadas.

T3. Usar os resultados obtidos, em particular, com as técnicas de *feature selection*, para desenvolver novos modelos de aprendizagem e comparar os diferentes modelos.