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 2018/2019

Trabalho Prático – 2ª Parte
Outubro, 2018

Tema

Sistemas Baseados em Similaridade - Árvores de Decisão

Objetivos de Aprendizagem

Com a realização deste enunciado prático pretende-se que os grupos de trabalho:

- Realizem um projeto utilizando modelos de *Machine Learning*, em particular Árvores de Decisão, através da plataforma KNIME.

Enunciado

Para a 2ª parte da componente prática de avaliação, os grupos de trabalho deverão realizar as seguintes tarefas:

- Consultar, analisar e selecionar um conjunto de dados (*dataset*) de entre os que estão acessíveis a partir do *Google Dataset Search* (ou outras fontes);
- Para além do *dataset* selecionado no ponto anterior:
  - Os grupos de **número par** deverão analisar, também, o *dataset* em https://www.kaggle.com/c/sbs2p2018;
  - Os grupos de **número ímpar** deverão analisar, também, o *dataset* em https://www.kaggle.com/c/sbs2i2018;
    - Os links anteriores redirecionam para a plataforma *Kaggle* onde foram criadas duas competições. Uma para os grupos de número par e outra para os grupos de número ímpar. Os *datasets* estão disponíveis nos links acima referidos assim como todos os detalhes do funcionamento da competição. Em suma, deverão criar um modelo, aplicá-lo e submeter os resultados na plataforma, a qual dará a *accuracy* do modelo desenvolvido.
- Utilizar a plataforma KNIME para desenvolver um, ou vários, *workflows* para:
  - Análise e tratamento dos dados dos dois *datasets* (seleção de atributos, tratamento de valores duplicados e em falta, *feature engineering*, etc.);
  - Extração de conhecimento dos dados;
  - Utilizar Árvores de Decisão como algoritmo de *Machine Learning* para aprendizagem supervisionada.
- Obtenção de resultados, incluindo o *tuning* (optimização) do algoritmo;
- Interpretar os resultados e a sua utilidade no contexto dos problemas subjacentes aos *datasets*. Determinar quais os resultados mais relevantes;

- Criação de objetos visuais que permitam ter uma noção gráfica dos modelos e dos resultados obtidos;
- Submeter os resultados obtidos no *Kaggle* de forma a obter a *accuracy* do modelo. Este passo refere-se, obviamente, apenas ao *dataset* de competição.

Os resultados obtidos deverão ser objeto de um relatório que contenha, entre outros:

- Introdução e Objetivos: quais os domínios a tratar, quais os objetivos e que benefícios se espera obter;
- Descrição do dataset e do tratamento dos dados: qual o *dataset* escolhido, o que o caracteriza e que *features* o compõe; que tratamentos foram aplicados aos dados dos dois *datasets*, como e porquê;
- Descrição dos *Workflows*: que *workflows* foram criados e com que objetivo; quais os principais nodos e como foram configurados; entre outros detalhes que seja oportuno fornecer;
- Modelos desenvolvidos e resultados obtidos: quais foram os modelos desenvolvidos e quais as suas características; como foi feito o *tuning* do modelo e sobre que parâmetros; sumário dos resultados obtidos e respetiva análise crítica;
- Recomendações: apresentação de sugestões após análise dos resultados e dos modelos.

Todo o processo deverá ser acompanhado de exemplos e indicações que permitam reproduzir todos os passos realizados assim com os resultados obtidos.

A data para a entrega do relatório final, que deve ser enviado por email aos docentes da cadeira, é o dia 18 de novembro de 2018 (23h59min).

A data para a sessão de apresentação do trabalho é o dia 22 de novembro de 2018, na sala DI-0.05, com início às 14h00min. Cada grupo disporá de 10 minutos para realizar a apresentação, utilizando os que meios que considerar mais adequados.

Avaliação

A avaliação deste trabalho de grupo contará com os seguintes elementos:

- Pelo documento produzido (75%);
- Pela apresentação realizada do trabalho desenvolvido (25%).