Universidade do Minho
Departamento de Informática

FE08

Curso: Mestrado Integrado em Engenharia Biomédica – Informática Médica
U.C.: Sistemas de Aprendizagem e Extração do Conhecimento

Folha de Exercícios FE08	
Docente	Diana Ferreira
Tema	RapidMiner – Regressão Linear
Turma	PL
Ano Letivo	2019-20 – 2º Semestre
Duração da aula	2 horas

1. Parte I

[1] A regressão linear exige que todos os atributos sejam de um determinado tipo de dados. Qual é este tipo de dados? Qual é o tipo de dados do atributo previsto quando este for calculado?

[2] Porque é que os intervalos dos atributos são tão importantes ao realizar *data mining* através de regressão linear?

[3] O que são coeficientes de regressão linear? O que significa 'peso' neste contexto?

[4] Qual é a fórmula matemática de regressão linear e como é organizada?

[5] Como é que se interpretam os resultados da regressão linear?

2. Parte II

[1] Faça o *download* do *dataset* "NBA-dataset" e seleccione alguns atributos (pelo menos três ou quatro) para armazenar dados sobre cada atleta. Alguns atributos possíveis que pode considerar podem ser o salário anual, pontos_por_jogo, altura, peso, idade etc. O objetivo deste exercício será prever o salário dos atletas, portanto este deve ser um atributo obrigatório. [Nota: Lembre-se que a regressão linear só trabalha com dados numéricos.]

[2] Divida as observações do seu *dataset* em duas partes: uma parte de treino e uma parte de teste. Certifique-se que tem pelo menos 20 observações no *dataset* de treino e pelo menos 20 no *dataset* de teste. Como vamos tentar prever o salário dos atletas do *dataset* de teste, não precisa de preencher a coluna do salário para estes atletas. Guarde dois ficheiros CSV (treino e teste) com nomes distintos, carregue-os no RapidMiner e arreste-os para uma nova janela de processo.

[3] Repita os passos no RapidMiner tal como descritos nos slides da aula e após executar o seu modelo, na secção dos resultados, examine os coeficientes dos atributos e as previsões para os salários dos atletas no conjunto de teste.

[4] Relate os resultados que obteve, respondendo às seguintes questões.

(a) Que atributos têm maior peso?

(b) Algum atributo foi removido do conjunto de dados por não ter uma boa capacidade de previsão? Em caso afirmativo, qual(ais) e por que motivo acha que ele(s) não era(m) eficaz(es) na previsão?

(c) Procure os salários de alguns dos atletas nos dados de teste e compare o salário real com o previsto. Está perto?

(d) Que outros atributos poderiam ajudar o modelo a prever melhor os salários dos atletas profissionais?