Universidade do Minho
Departamento de Informática

FE08

Curso: Mestrado Integrado em Informática – Engenharia do Conhecimento
U.C.: Descoberta do Conhecimento

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

1. Parte I

[1] Que tipo de dados a regressão linear espera para todos os atributos? Qual o tipo de dados do atributo previsto quando este for calculado?

[2] Porque é que os intervalos de 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 resultados da regressão linear são interpretados?

2. Parte II

[1] Selecione uma organização desportiva profissional de que goste ou que conheça. Localize o site da organização e pesquise estatísticas, factos e números sobre os atletas dessa organização. Crie um dataset (usando o Excel por exemplo) e defina 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, anos_como_pro, altura, peso, idade etc. A lista é potencialmente ilimitada, variará de acordo com o tipo de desporto que escolher e dependerá dos dados disponíveis. O objetivo deste exercício será prever o salário dos atletas, portanto este deve ser um atributo obrigatório. PS: Lembre-se que a regressão linear só trabalha com dados numéricos.

[2] Pesquise as estatísticas de cada um dos atributos que selecionou e insira-as como observações na sua folha. Tente encontrar o maior número possível – pelo menos 40, a fim de atingir pelo menos um nível básico de validade estatística. Quanto mais melhor. Divida as observações do seu *dataset* em duas partes: uma parte de treino e uma parte de *scoring*. Certifique-se que tem pelo menos 20 observações no dataset de treino e pelo menos 20 no *dataset* de *scoring*. Como vamos tentar prever o salário dos atletas do dataset de *scoring*, não precisa de procurar nem preencher a coluna do salário para estes atletas. Guarde

dois ficheiros CSV (treino e *scoring*), como nomes distintos, carregue-os no RapidMiner e arreste-os para um novo 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 *scoring*.

[4] Relate seus resultados:

(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, quais e por que você acha que eles não eram eficazes na previsão?

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

(d) Que outros atributos acha que ajudariam o seu modelo a prever melhor os salários dos atletas profissionais?