Universidade do Minho
Departamento de Informática

FE05

Course: Mestrado Integrado em Informática – Engenharia do Conhecimento
U.C.: Knowledge Discovery

Exercise Sheet FE05	
Teacher	Cristiana Neto
Theme	RapidMiner - Correlations
Class	PL
Year	2019-20 – 2nd Semester
Duration	2 hours

1. Part I

[1] What are some of the limitations of correlation models?

[2] What is a correlation coefficient? How is it interpreted?

[3] What is the difference between a positive and a negative correlation?

(a) If two attributes have values that decrease at essentially the same rate, is that a negative correlation? Why or why not?

[4] How is correlation strength measured? What are the ranges for strengths of correlation?

[5] The number of heating oil consuming devices was suggested as a possibly interesting attribute that could be added to the example data set for this chapter. Can you think of others? Why might they be interesting?

2. Part II

[1] Access to the *mpg_dataset.csv* file.

[2] Execute the *Data Understanding* operation taking into account that:

cylinders: number of cylinders in the engine
displacement: engine displacement
horsepower: engine power
weight: vehicle weight
acceleration: vehicle acceleration, i.e. time in seconds to accelerate from 0 to 60
model year: vehicle model year in the 1900s
origin: car origin (1: American, 2: European, 3: Japanese)
mpg: fuel consumption / efficiency (miles per gallon - mpg)

[3] Perform Data Preparation step on Weka. Don't forget to look for ‘outliers’ and ‘missing values’. Once properly processed, save the data in a .csv file that allows the correlation process to be executed in the rapidminer.
- Import csv into RapidMiner.
- Add csv to the design.

[4] Document which attributes can influence or explain fuel consumption/efficiency in a given vehicle (mpg).