JAMES COOK UNIVERSITY AUSTRALIA

Practical 4
Data Pre-Processing Review Questions

What are we doing?

Using what you learned from lectures, you will answer some review questions. These questions are for your self-review on topics covered: Data exploration and pre-processing. You will need to review the lecture to answer questions.

Submission:
You are required to submit one document containing your answers via the weekly-practical submission box (available on CP1407 LearnJCU). For Laboratory questions, screen capture your computer screen after the completion of each task and include the captured image in your document to submit.

Self-Review Questions

1. The table below shows a sample of a larger dataset containing details of policy holders at an insurance company. The attributes included in the table describe each policy holders' ID, occupation, gender, age, the value of their car, the type of insurance policy they hold, and their preferred contact channel.

ID	Occupation	Gender	Age	Motor Value	Policy Type	Pref Channel
1	lab tech	female	43	42,632	planC	sms
2	farmhand	female	57	22,096	planA	phone
3	biophysicist	male	21	27,221	planA	phone
4	sheriff	female	47	21,460	planB	phone
5	painter	male	55	13,976	planC	phone
6	manager	male	19	4,866	planA	email
7	geologist	male	51	12,759	planC	phone
8	messenger	male	49	15,672	planB	phone
9	nurse	female	18	16,399	planC	sms
10	fire inspector	male	47	14,767	planC	email

   a) State whether each descriptive attribute contains numeric, ordinal, nominal, or textual data. (For example, the 'Gender' feature is nominal data.)

   b) How many pre-defined data values does each nominal and ordinal attribute have? (For example, 'Gender' attribute has 2 pre-defined data values (male, female))

2. The table below contains sample data about the employees of an IT company.

Emp ID	Name	Year of Birth	Gender	Status	Salary
100	Smith	1954	M	Director	$200.000
125	Jones	1967	F	Technician	$36,000
167	Highley	1975	F	Senior Technician	$70,000
200	Millins	1987	M	Technician	$32,000
205	Dujevic	1985	M	Technician	$34,000
216	Isovic	1985	F	Technician	$34,000
220	Sun	1986	F	Senior Technician	$66,000
301	Bean	1955	M	Deputy Director	$160,000

Answer the following questions.

a) An input data set consists of individual data objects, also known as data records, instances or samples. All data sets have the common properties: Type, Size, Dimensionality, Sparsity. Describe the properties of the above dataset in relation to its type, size, dimensionality and sparsity.

b) If the 'Salary' attribute needs to be discretised into three pay bands (3 groups), suggest a simple yet sensible solution for the discretisation.

c) If Mr Dujevic's salary was unknown and the unknown value needed to be imputed, what is a sensible replacement value and why?

d) Among the employee records, which record can be considered as an outlier? What harm can an outlier sample cause to the understanding of the data set?

Laboratory Questions

1. The table below presents a data set about student homework and examination results.

Student ID	Home-work 1	Home-work 2	Home-work 3	Exam
1		94	34	42
2	35	94	85	45
3	31	46	22	48
4	46	90	60	50
5	52	94	49	50
6	58	94	30	51
7	47	90		52
8	37	94	25	52
9	35	94	45	54
10	57	94	100	54
11	51	94	5	54
12	45	94	33	55
13	44	0	35	55
14	52	95	36	56
15	35	94		57
16	57	97	57	57
17	45	90	71	57
18	39	94	54	57
19	31	94	63	57
20	45	94		59
21	35	90	84	59
22	37	90	40	61
23	83	97	26	61
24	68	97	55	62
25	50	95	56	62
26	77	93		63
27	84	48	18	63

Student ID	Home-work 1	Home-work 2	Home-work 3	Exam
28	45	90	21	63
29	62	95	38	63
30	38	94	40	64
31	50	90		64
32	32	90	38	64
33	44	90	43	65
34	57	94	52	68
35	50	94	39	70
36	55	90	62	71
37	43	94	54	72
38	50	90	30	74
39	54	90	82	77
40	64	95	5	78
41	85	95		79
42	63	90	62	82
43	75	90	35	83
44	86	97	39	84
45	77	95	79	84
46	79	94	35	86
47	86	98	57	87
48	71	90	9	89
49	45	94	72	90
50	90	94	68	92
51	89	94	53	93
52	90	98	79	98
53	57	92	40	
54	36	94	54	22

Use MS Excel and WEKA to create an ARFF file for the data set in the table above. (Store this example data in MS Excel → save as .csv file → open the csv file in WEKA → save as .arff file in WEKA)

Open the ARFF file in Weka, and perform the following tasks .

a) Observe the summary data for the data set and the histograms for all attributes on the 'Preprocess' tap page. Use the Visualize tab page to view the scatter plots between the variables of the data set.

b) Apply the unsupervised Discretize filter to the exam marks.

c) Practise filling in missing values in Weka both manually in the Viewer window and by using filters. We introduced the 'ReplaceMissingValues' filter as one of other useful filters in Weka (Practical 2).

Numeric values are replaced with the sample mean and nominal values are replaced with the sample mode. The user can also fill in missing values manually in the viewer window (using "Edit" menu). For numeric

attributes, the user may enter any value. For nominal attributes, the user can only select one of the nominal labels that already exists in the attribute domain. If the label does not exist (for instance, it is a special code indicating unknown), the label can be added into the attribute domain by using “AddValues” filter.

2. [**Extension**: this task is optional and are not required to complete or submit]

   Principal component analysis (PCA) is a useful tool to reduce dimensionality of a given data set. WEKA is equipped with a PCA filter, which can be used on the Select Attribute tab of the Explorer. Open one of the data sets provided in WEKA, such as *cpu.arff*. On the *Select Attribute* tab, press the *Choose* button and select the PricipalComponents filter. Press the Start button and observe the output in the Attribute Selection Output window. The window should show the new attributes in eigenvectors, each of which is a linear combination of the original attributes and the ranking among them according to their significance. Discuss your findings.