Weka Practical II

Note:

- Briefly, OneR (one-attribute-rule) algorithm generates rules that only include one attribute plus the class. ID3 is an algorithm that generates a decision tree, which can be converted to several rules. It only accepts nominal values. J48 (C4.5) is an improved version of ID3. It accepts both numerical and nominal values and generates a pruned decision tree. For example, see the output by J48 on the weather data on page 5 in How to Use Weka. From the J48 pruned tree, we can generate a rule like this

  IF outlook is sunny and humidity is greater than 75 **THEN** don't play.

- Before you start this exercise please read the notes: *How to Use Weka* **first**.
- Don't change any parameters until you understand them.
- Study the data before you apply any algorithm on it.
- Scheme J48 in this exercise means **weak.classifiers.j48.J48**
- In general, you are expected to find patterns (rules) from models generated by weka. These patterns are normally not easily found by just studying the datasets manually.
- For all exercises, use different test options: Use training set, Cross Validation and Percentage split. What are the differences? You are expected to check the model evaluation as well.
- Datasets for this exercise can be found either under the weka's data directory or on the module's Moodle.
- If the dataset provided in .xlsx format, you need to convert it to .arff first.

1. Complete practical 1

2. Convert the Fruit data (see Lecture notes) into ARFF format and use the following learning schemes to analyse it: OneR, ID3and J48

   Compare the results you obtained using these three classifiers individually.

3. Use OneR, ID3 and J48 to analyse the life style data set, lifestyle.x.sx, which can be downloaded from moodle. There are 34 attributes. See the Appendix for the metadata for this dataset. You are expected to find interesting patterns from the dataset. You might need to do preparation before you start to analyse it. You might also need to select different group of attributes at each play.

References

R. Kirkby and E. Frank: WEKA Explorer User Guide for Version 3.5.8, 2008

Appendix

Life Style Data Set

Attribute and Meta Data

No.	Attribute Name	Data Type	Meta Data
1	URN	String	Primary Key
2	Title	String	Mr, Mrs, MS etc
3	Gender	Double	1: Male
			2:Female
4	Marital Status	Double	0: Unknown
			1: Married
			2: Single/Never Married
			3: Divorced/Separated
			4: Widowed
			5:Living with partner
5	Aged (Band)	Double	1: 18~24
			2: 25~29
			3: 30~34
			4: 35~39
			5: 40~44
			6: 45~49
			7: 50~54
			8: 55~59
			9: 60~64
			10: 65~69
			11: 70~74
			12: 75+
6	Income	Double	0: Unknown
			1: Under £5,000
			2: £5000 - £9,999
			3: £10,000 - £14,999
			4: £15,000 - £19,999
			5: £20,000 - £24,999
			6: £25,000 - £29,999
			7: £30,000 - £34,999
			8: £35,000 - £39,999
			9: £40,000 +
7	Home Ownership	Double	0: Unknown
			1: Home Owners
			2: Private Renters
			3: Council Renters
			4: Living with Parents
8	Bingo	Double	1: yes, 0: No
9	Reading Books	Double	1: yes, 0: No
10	Current Affairs	Double	1: yes, 0: No
11	Collectibles	Double	1: yes, 0: No
12	Computing	Double	1: yes, 0: No

13	DIY	Double	1: yes, 0: No
14	Eating Out	Double	1: yes, 0: No
15	Fine Art /Antiques	Double	1: yes, 0: No
16	Fashion	Double	1: yes, 0: No
17	Further Education	Double	1: yes, 0: No
18	Foreign Travel	Double	1: yes, 0: No
19	Gardening	Double	1: yes, 0: No
20	Good Food & Wine	Double	1: yes, 0: No
21	Golf	Double	1: yes, 0: No
22	Health Foods	Double	1: yes, 0: No
23	Hiking	Double	1: yes, 0: No
24	Jogging / Exercise	Double	1: yes, 0: No
25	CDs & Music	Double	1: yes, 0: No
26	Pets	Double	1: yes, 0: No
27	Photography	Double	1: yes, 0: No
28	Does The Pools	Double	1: yes, 0: No
29	Going Down The Pub	Double	1: yes, 0: No
30	Puzzles / Crosswords	Double	1: yes, 0: No
31	Coins & Stamps	Double	1: yes, 0: No
32	Theatre / Cultural	Double	1: yes, 0: No
33	National Trust / Voluntary	Double	1: yes, 0: No
34	Charities & Voluntary	Double	1: yes, 0: No
35	UK Holidays	Double	1: yes, 0: No
36	European Holidays	Double	1: yes, 0: No