Lab Exercise 2

Introduction to data classification - Decision Trees

A decision tree is a graphical method of supporting the decision-making process, used in decision theory. The decision tree algorithm is also used in machine learning to generate knowledge based on given examples.

The aim of this laboratory is to use the WEKA package to generate a decision tree (decision table).

I. Data loading and analysis

- 1. Download data pack: data.tar.gz (available on the course web site)
- 2. Open in a notebook (or other text editor) a file named **swimming.arff** and learn the structure of this learning file with its symbolic data vectors.
- 3. Start Weka, click the Explorer button and load the swimming.arff data file.
- 4. Analyse the first '*Preprocess*' tab and answer the questions below:

0		We	lorer					- + ×		
Preprocess Clas	sify Clu	ociate	attribute	s Visual	ize					
Open fil	erat Undo E			Edit	Edit Save					
Filter										
Choose None									Apply	
Current relation Relation: swimm Instances: 6		7	Selected attribute Name: sky Missing: 0 (0%) Distinct: 3					Type: Nominal Unique: 1 (17%)		
Attributes					No.		abel		Count	
All	2	sunny rainy cloudy		3 1 2						
No.		Name			3	ciouay			2	
1 sky 2 airTemp										
3 humidity										
4 wind 5 water		L				_				
6 forecast	Class: er	njoy (Nom)				Visualize All				
7 enjoy	Remo		2		1					
Status OK										Log 🛷 x O

- a) What is the size of the training set?
- b) How many attributes exist in the training set?
- c) How many instances are positive (Enjoy = yes) and how many negative?
- d) Which attribute best separates the data?
- e) How many elements from the data set have the humidity attribute set as high?

II. Load and analyse data

- 1. Open the Classify tab.
- 2. Select the J48 classifier using the Choose button.
- 3. Make sure that 'Use training' set is checked in the 'Test options' window. Attention! In the future, we will **not** use this form of testing we are forced here because of the small training set.

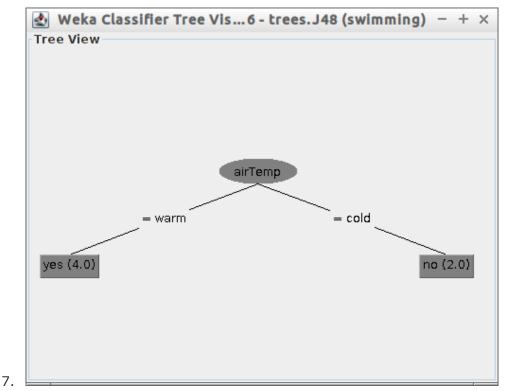
4. Click on the Start button. Look at the result. What do the results mean?

0				Weka Explorer							- +	. >	
Preprocess	Classify	Cluster	Associate	Select	attribute	s Visua	alize						
Classifier													
Choose	ld3												
Test options			Classif	ier outp	ut								
🖲 Use traini	ng set		TTHE C	aken tu	Durta mode	L. O Seco	nus					1	
O Supplied test set Set				Evaluation on training set									
O Cross-vali	dation F	olds 10											
O Percentad	ne solit	% 66			sified Ins		6		100	% 4			
				ectly Cl statisti	assified In	nstances	0		0	16			
M	ore option	s		bsolute			i i						
					red error		ŏ						
(Nom) enjoy			- Relati	ve absol	ute error		0	5					
trong orgoj					squared er		0	\$					
Start		Stop	Total	Number o	f Instance	5	6						
Result list (ri	ght-click fo	r options)	De'	tailed A	ccuracy By	Class	-						
12:24:57 - tre	es.Id3				TP Rate	FP Rate	Precision	Recall	E-Measure	RDC Area	Class		
					1	0	1	1	1	1	ves		
					ī	õ	ĩ	ī	ī	ī	no		
			Weight	ed Avg.	1	0	1	1	1	1			
			Co	nfusion	Matrix							1	
			ab	c., cla	ssified as								
				a = yes									
				b = no									
Status												-	
ок										Log	1000	×	

5. Select the J48 classifier using the Choose button and click Start, then visualize the tree as shown below:

0						W	eka Expl	orer					- +	Η×
Pre	process	Classify	Cluster	Assoc	ciate	Selec	t attribute	s Visua	alize					
Clas	ssifier													
	hoose	J48 -C 0.25 -	M 2											
Tes	t options			c	lassifi	er out	put							_
Use training set					Evaluation on training set Summary									
○ Supplied test set Set														
0	O Cross-validation Folds 10				Correctly Classified Instances 6 100									
01	O Percentage split % 66						lassified 1		0		0	% %		
	More options				Kappa s Mean ab				1					
		iore option	a				ared error		0					
(No	(Nom) enjoy			F	Relative absolute error Root relative squared error				0	\$				
					Root re Total N	lative umber i	squared er of Instance	ror	0	\$				
	Start		Stop											
Res	ult list (r	ight-click fo	r options)	-	Det	ailed	Accuracy By	Class	-					
12:2	24:57 - tro	es.Id3					TP Rate	FP Rate	Precision	Recall	F-Measure	RDC Area	Class	
1 Vi	iew in ma	in window					1	0	1	1	1	1	yes	
		parate wine	wob		ighte	d Avg.	1	0	1	1	1	1	no	
	ave resul				1									_
	Delete result buffer				= Confusion Matrix ===									
	oad mod						assified as	3						
	ave mode					a = ye b = no								
		te model on		st set	- 2	b = no								
		lassifier er	rors											-
	isualizet	ree hargin curv												
		hreshold curve										Log	100	x O
		fit analysis	nve.										-	

6. Does the tree look like this?



III. Classification accuracy

- Load the file credit-g.arff to Weki. It contains learning data for the system, which on the basis of the attributes contained in the file, should determine whether a given set of attribute values indicates a credibility of bank customers – i.e. whether the bank should grant him a loan or if it is too risky to do so.
- 2. Open Classify tab and choose J48 algorithm.
- 3. In the '*Test options*' area, select Percentage split and type in 66%. IT means that 66% of the data will be used for learning and 34% of this data set will be used for validation.
- 4. Run the algorithm. How many percent of cases were correctly classified? Is this a good result?
- 5. Change the classifier to ZeroR from the rules branch. What is the obtained result? Better or worse than J48?
- 6. Try other classifiers. What are their results?
- 7. Go to the '*Preprocess*' tab and see how the distribution of the attribute defines whether the set is good or bad. What would be the effectiveness of an algorithm that regardless of the value of attributes would "shoot" that the user is reliable or not?
- 8. Why is it worth taking a look at the data before attempting a classification task?