Making predictions on new data using Weka

Daniel Rodríguez daniel.rodriguezg@uah.es University of Alcala

One we have learned a model, it can be used to classify new unseen data. These notes describe the process of doing some both graphically and from the command line.

First, the file with cases to predict needs to have the same structure that the file used to learn the model. The difference is that the value of the class attribute is "?" for all instances (question marks represent missing values in Weka). For example assuming that we have learnt a decision tree using the diabetes datasets included weka, the following file will be used to predict the 5 cases included in the arff file:

<pre>@relation pima_diabetes</pre>
@attribute 'preg' real
@attribute 'plas' real
@attribute 'pres' real
@attribute 'skin' real
@attribute 'insu' real
@attribute 'mass' real
@attribute 'pedi' real
@attribute 'age' real
<pre>@attribute 'class' { tested_negative, tested_positive}</pre>
@data
6,148,72,35,0,33.6,0.627,50,?
1,85,66,29,0,26.6,0.351,31,?
8,183,64,0,0,23.3,0.672,32,?
1,89,66,23,94,28.1,0.167,21,?
0,137,40,35,168,43.1,2.288,33,?

Using Weka's Explorer

First, we load the saved model with the right click menu on the "Result list" panel:

🗿 Weka Explorer			
Preprocess Classify Cluster Classifier Choose ZeroR	Associate Select attribu	utes Visualize Forecast	
Test options © Use training set © Supplied test set	Set	Classifier output	
 Cross-validation Percentage split % 	10 66		
More options			
(Nom) class Start Result list (right-click for option)	Stop		
	View in main w View in separat Save result buff Delete result bu	indow e window er ffer	
	Load model Save model Re-evaluate mo	idel on current test set	
	Visualize classif Visualize tree Visualize margi Visualize thresh	ier errors n curve old curve	þ.
Status OK	Cost/Benefit an Visualize cost c	alysis urve	Log x0

In the "Test Options", we have to select "Supplied test set", and once the file is loaded we select "No class" from the list of attributes.

process Classify Cluster Assoc	iate Select attributes Visualize Forecast		
assifier			
Choose ZeroR			
st options	Classifier output		
) Use training set	S Test Instances		1
Supplied test set Set	Relation: pima_diabetes	Attributes: 9	
Cross-validation Folds 10	Instances: ?	Sum of weights: ? all F-Measure	Ē
Percentage split % 66	Open file Open URL		
More options		NaN	
om) dass 🗸 🗸		Close	
Start Stop			Í
sult list (right-click for options)	0 0 a = tested_negative		
:35:31 - trees.J48 from file 'j48.mod	0 0 b = tested_positive		
	Rector		
4 m			ř

Then, clicking "More Options", a new window opens and we choose **PlainText** from 'Output predictions'

Classifier evaluation options	
V Output model	
☑ Output per-dass stats	
Output entropy evaluation measures	
V Output confusion matrix	
Store predictions for visualization	
Output predictions Choose PlainText	
Cost-sensitive evaluation Set	
Random seed for XVal / % Split 1	
Preserve order for % Split	
Output source code WekaClassifier	
ОК]

Finally, we need to right click in the model and run "Re-evaluate model on current test set".

	·							No. of	
reprocess Class	TY Cluster	Associate	Select attributes	Visualize Fo	recast				
Classifier									
Choose Ze	roR								
lest options			Classifier outpu	t					
🔘 Use training s	et		inst#	actual p	redicted erro	r prediction	ĩ		
Supplied test	set S	et	1	1:? 2:	tested_p	0.717			
El Crass uslidati	ian Coldo	10	2	1:? 1:	tested_n	0.818			
Cross-Validat	ion rolas	10	3	1:? 2:	tested_p	0.917			
Percentage s	plit %	66	4	1:? 1:	tested_n	0.878			
More	e options		5	1:? 2:	tested_p	0.933			
iom) contact-len:	ses	-	=== Summar	y === er of Inst	ances	0			
Start	S	top	Ignored Cl	ass Unknow	n Instances	0	5		
esult list (right-cl	ick for options	1					1750		
1:35:31 - trees. 1	48 from file	, 	Dened		Class ===	=			
		View in	n main window						
		View in	n <mark>s</mark> eparate windo	w	FP Rate	Precision	Recall	F-Measure	ROC
		Save re	esult buffer		0	0	0	0	2
		D.L.	b L CC		0	0	0	0	2
		Delete	result buffer		N = N	NaN	NaN	NaN	
					INGIN	TACTA		- Hart	NaN
		Load r	nodel		Walk	ivait			NaN
		Load r Save n	nodel nodel		INGIN	HUIT		11011	NaN
		Load r Save n Re-eva	nodel nodel iluate model on	current test s	et	ITELY			NaN
	E	Load r Save n Re-eva Visuali	nodel nodel Iluate model on ze classifier erro	current test s	et ve ve				NaN
	E	Load r Save n Re-eva Visuali Visuali	nodel nodel Iluate model on ze classifier erro ze tree	current test s	et ve ve				NaN
	E	Load r Save n Re-eva Visuali Visuali	nodel nodel Iluate model on ze classifier erro ze tree	current test s	et ve ve				NaN
	C	Load r Save n Re-eva Visuali Visuali Visuali	nodel nodel iluate model on ze classifier erro ze tree ze margin curve	current test s	et ve ve		1		NaN
		Load r Save n Re-eva Visuali Visuali Visuali Visuali	nodel nodel iluate model on ze classifier erro ze tree ze margin curve ze threshold cur	current test s rs ve	et ve ve				NaN
tatus		Load r Save n Re-eva Visuali Visuali Visuali Visuali Cost/E	nodel nodel iluate model on ze classifier erro ze tree ze margin curve ze threshold cur Benefit analysis	current test s rs ve	et ve ve				NaN

The results are shown in the "Classifier output" panel, under "Predictions on test data". The "predicted" column contains tested_p or tested_n for each of the lines in the test file.

Using the command line

It is explained in the following link:

http://weka.wikispaces.com/Making+predictions

An example using our data:

```
java weka.classifiers.trees.J48 -T diabetes2.arff -1 j48.model -p 0
```

You need to add the weka.jar file into the CLASSPATH environment variable (or using -cp) and the 'bin' directory of your java installation in the PATH variable)

And the output should look like this:

```
C:\Windows\system32\cmd.exe
 C:\Users\drg\Desktop>dir *.arff
El volumen de la unidad C no tiene etiqueta.
El número de serie del volumen es: 80DC-6F5F
                                                                                                                     Ε
 Directorio de C:\Users\drg\Desktop
16/02/2012 14:15
                                              421 diabetes2.arff
                     1 archivos 421 bytes
0 dirs 250,942,443,520 bytes libres
C:\Users\drg\Desktop>java weka.classifiers.trees.J48 -T diabetes2.arff -l j48.mo
del -p 0
 === Predictions on test data ===
               actual predicted error
1:? 2:tested_p
1:? 1:tested_n
1:? 2:tested_p
1:? 1:tested_n
                                                  prediction
0.717
0.818
 inst#
       123
                                                   0.917
       4
                                                   0.878
       5
                    1:?
                          2:tested_p
                                                   0.933
C:\Users\drg\Desktop>_
```