

WEKA ΠΑΡΑΔΕΙΓΜΑΤΑ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμηση Στατιστικής και Αναλογιστικών –
Χρηματοοικονομικών Μαθηματικών
Πανεπιστήμιο Αιγαίου
Σαμος

2021

Προ-επεξεργασία

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** [Apply] [Stop]

Current relation
Relation: breast-cancer
Instances: 286
Attributes: 10
Sum of weights: 286

Attributes
All | None | Invert | Pattern

No.	Name
1	<input type="checkbox"/> age
2	<input type="checkbox"/> menopause
3	<input checked="" type="checkbox"/> tumor-size
4	<input type="checkbox"/> inv-nodes
5	<input type="checkbox"/> node-caps
6	<input type="checkbox"/> deg-malg
7	<input type="checkbox"/> breast
8	<input type="checkbox"/> breast-quad
9	<input type="checkbox"/> irradiat
10	<input type="checkbox"/> Class

Remove

Selected attribute
Name: tumor-size
Missing: 0 (0%)
Distinct: 11
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	0-4	8	8.0
2	5-9	4	4.0
3	10-14	28	28.0
4	15-19	30	30.0
5	20-24	50	50.0
6	25-29	54	54.0
7	30-34	60	60.0
8	35-39	19	19.0
9	40-44	22	22.0
10	45-49	3	3.0
11	50-54	8	8.0
12	55-59	0	0.0

Class: Class (Nom) [Visualize All]

Label	Count
0-4	8
5-9	4
10-14	28
15-19	30
20-24	50
25-29	54
30-34	60
35-39	19
40-44	22
45-49	3
50-54	8
55-59	0

Status: OK [Log] x 0

EN 6:41 μμ 25/5/2021

Προ-επεξεργασία

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None**

Current relation: Relation: breast-cancer, Instances: 286

Attributes: All

No.	Name
<input type="checkbox"/>	age
<input type="checkbox"/>	menopause
<input type="checkbox"/>	tumor-size
<input type="checkbox"/>	inv-nodes
<input type="checkbox"/>	node-caps
<input type="checkbox"/>	deg-malig
<input checked="" type="checkbox"/>	breast
<input type="checkbox"/>	breast-quad
<input type="checkbox"/>	irradiat
<input type="checkbox"/>	Class

Type: Nominal
Unique: 0 (0%)
Weight: 152.0, 134.0

Visualize All

Status: OK | Log | x 0

6:45 μμ
25/5/2021

Category	Blue Class	Red Class
0	8	4
1	28	30
2	50	54
3	60	19
4	22	3
5	8	0

Category	Blue Class	Red Class
0	68	150
1	201	85

Κατηγοριοποίηση

The screenshot displays the Weka Explorer application window. The main interface is titled "Classifier" and shows the "ZeroR" classifier selected. The "Test options" section includes radio buttons for "Use training set", "Supplied test set", "Cross-validation" (selected), and "Percentage split". The "Cross-validation" option is set to "Folds: 10". Below this, there are "Start" and "Stop" buttons. The "Result list" area is currently empty.

Two dialog boxes are open over the main interface:

- Classifier evaluation options:** This dialog has several checked options: "Output model", "Output per-class stats", "Output confusion matrix", "Store test data and predictions for visualization", and "Collect predictions for evaluation based on AUROC, etc.". It also includes a field for "Output predictions" set to "Null", a "Random seed for XVal / % Split" set to "1", and an "Evaluation metrics..." button.
- Manage evaluation metrics:** This dialog shows a list of metrics with checkboxes. The "All" button is selected. The list includes:

No.	Name	Checked
1	Correct	Yes
2	Incorrect	Yes
3	Kappa	Yes
4	Total cost	Yes
5	Average cost	Yes
6	KB relative	Yes
7	KB information	Yes
8	Correlation	Yes
9	Complexity 0	Yes

The Windows taskbar at the bottom shows the system tray with the date "25/5/2021" and time "6:58 μμ".

Κατηγοριοποίηση

- Στην καρτέλα Classify παρέχονται πολλοί αλγόριθμοι αλλά και μέθοδοι κατηγοριοποίησης των δεδομένων.
- Στο πεδίο classifier υπάρχει η δυνατότητα επιλογής πολλών αλγορίθμων. Ο έλεγχος της απόδοσης τους μπορεί να πραγματοποιηθεί με τέσσερις τρόπους:

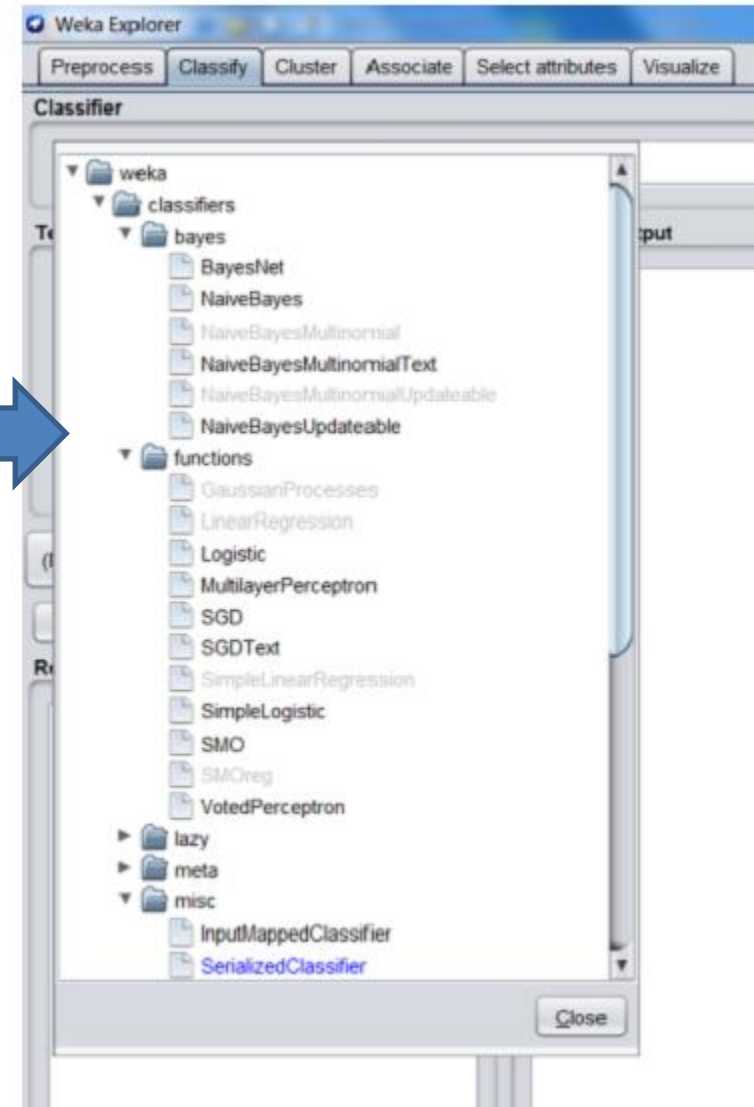
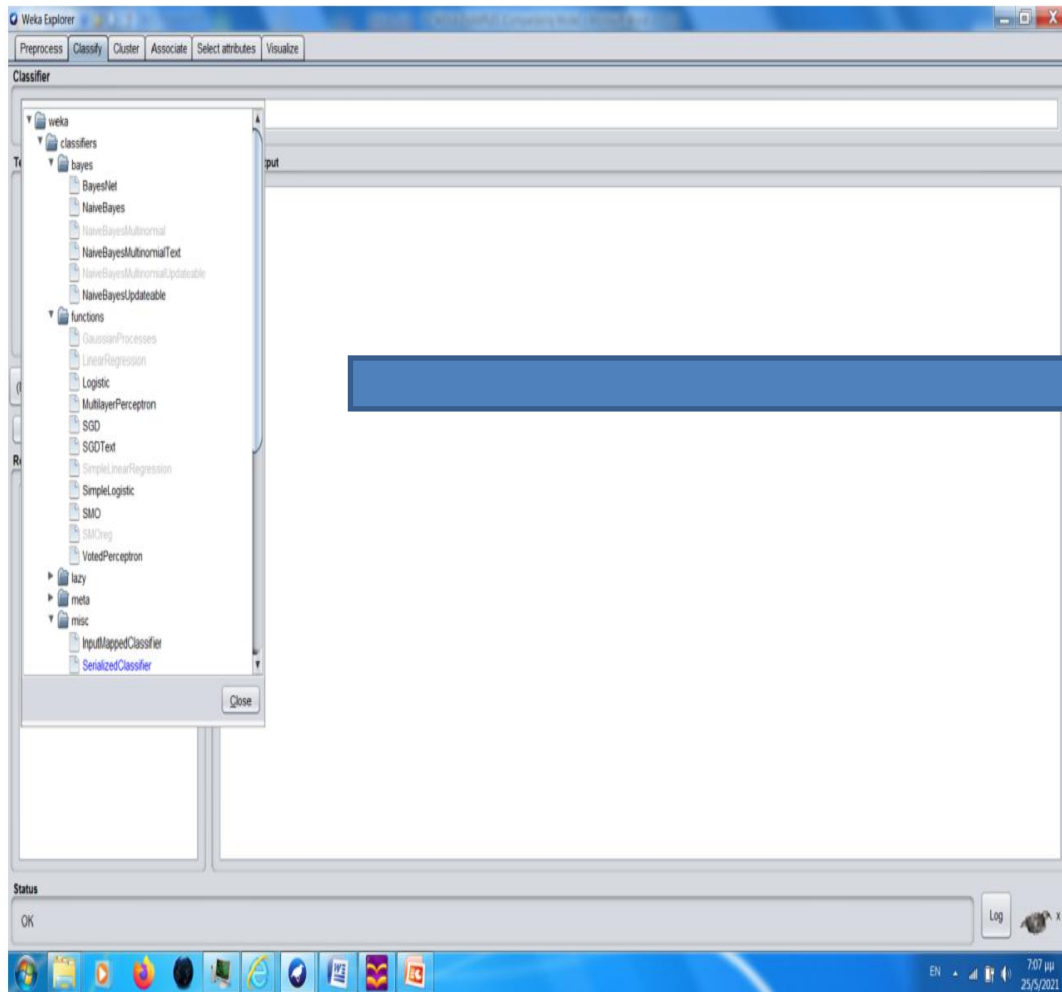
Κατηγοριοποίηση

1. στο σύνολο των δεδομένων εκπαίδευσης (use training set) :
εδώ ο classifier αποτιμάται στο πόσο καλά μπορεί να προβλέψει την κλάση (class) των παραδειγμάτων (instances) που εκπαιδεύτηκε,
2. με supplied test set : ο classifier αποτιμάται στο πόσο καλά προβλέπει την class από το set των instances που φορτώθηκαν από το αρχείο,

Κατηγοριοποίηση

3. με την μέθοδο cross validation: όπου ο classifier αποτιμάται από cross validation μια διαδικασία διασταυρωμένης επικύρωσης, χρησιμοποιώντας τον αριθμό των folds που εισάγονται στο ανάλογο πεδίο
4. percentage split: δηλαδή σε ένα συγκεκριμένο ποσοστό από τα αρχικά δεδομένα. Ο classifier αποτιμάται στο πόσο καλά προβλέπει ένα συγκεκριμένο ποσοστό των δεδομένων (certain percentage) που προσφέρονται για έλεγχο. Τα οποία δεδομένα εξαρτώνται από την τιμή που εισάγεται στο πεδίο.

Κατηγοριοποίηση



Κατηγοριοποίηση

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

Use training set
 Supplied test set Set...
 Cross-validation Folds **10**
 Percentage split % 66
More options...

(Nom) breast

Start Stop

Result list (right-click for options)

19:09:53 - bayes.NaiveBayes

Classifier output

```
right_row 10.0 11.0
central 12.0 11.0
[total] 156.0 139.0
```

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	189	66.0839 %
Incorrectly Classified Instances	97	33.9161 %
Kappa statistic	0.3187	
Mean absolute error	0.4433	
Root mean squared error	0.473	
Relative absolute error	88.9973 %	
Root relative squared error	94.7889 %	
Total Number of Instances	286	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,684	0,366	0,680	0,684	0,682	0,319	0,642	0,619	left
	0,634	0,316	0,639	0,634	0,637	0,319	0,642	0,593	right
Weighted Avg.	0,661	0,342	0,661	0,661	0,661	0,319	0,642	0,607	

=== Confusion Matrix ===

```
 a  b  <-- classified as
104 48 | a = left
 49 85 | b = right
```

Status

OK

Log x0

EN 7:10 μμ 25/5/2021

Naïve Bayes

Ο Naïve Bayes αλγόριθμος ταξινόμησης είναι χρήσιμος για να χαρακτηρίσει ακόμα και σύνολα δεδομένων με υψηλό όγκο πληροφοριών, καθώς εκτελείται αποτελεσματικά και είναι εύκολο να εφαρμοστεί.

$$P(c|X) = [P(X|c) * P(c)] / P(X)$$

- $P(c|x)$ = εκ των υστέρων πιθανότητα
- $P(x|c)$ = Δεσμευμένη πιθανότητα
- $P(c)$ = εκ των προτέρων πιθανότητα κλάσης
- $P(X)$ = εκ των προτέρων πιθανότητα ταξινομητή

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Naïve Bayes

Η εκ των υστέρων πιθανότητα υπολογίζει την πιθανότητα του αποτελέσματος που προκύπτει από μια νέα πληροφορία. Στο $P(c|x)$, το c αναπαριστά την κλάση που ταξινομούνται τα δεδομένα και το x τον ταξινομητή. Η δεσμευμένη πιθανότητα είναι η πιθανότητα να βρίσκεται ο ταξινομητής(χαρακτηριστικό) μέσα στην κλάση.

Naïve Bayes

Βάσει λοιπόν των πιο πάνω, μπορούμε να υπολογίσουμε τη μέγιστη a Posteriori πιθανότητα $P(c|X)$ των δεδομένων.

$$c_{MAP} = \arg \max_{c \in C} P(c|X)$$

$$= \arg \max_{c \in C} \frac{P(X|c)*P(c)}{P(X)}$$

$P(X)$ σταθερό => Μπορεί να

παραλειφθεί

$$= \arg \max_{c \in C} P(X|c)*P(c)$$

Naïve Bayes

Υποθέτοντας ότι όλες οι πιθανότητες c είναι το ίδιο πιθανές, δηλαδή $P(c_i) = P(c_j)$, μπορούμε να παραλείψουμε το και το $P(c)$. Άρα:

$$c_{MAP} = \arg \max_{c \in C} P(X | c)$$

Maximum Likelihood Hypothesis

Κατηγοριοποίηση

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	189	66.0839 %
Incorrectly Classified Instances	97	33.9161 %
Kappa statistic	0.3187	
Mean absolute error	0.4433	
Root mean squared error	0.473	
Relative absolute error	88.9973 %	
Root relative squared error	94.7889 %	
Total Number of Instances	286	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,684	0,366	0,680	0,684	0,682	0,319	0,642	0,619	left
	0,634	0,316	0,639	0,634	0,637	0,319	0,642	0,593	right
Weighted Avg.	0,661	0,342	0,661	0,661	0,661	0,319	0,642	0,607	

=== Confusion Matrix ===

a	b	<-- classified as
104	48	a = left
49	85	b = right

Μετρικές

$$acc = \frac{a + d}{a + b + c + d}$$

$$prec = \frac{a}{a + c}$$

$$sen = \frac{a}{a + b}$$

$$spec = \frac{d}{c + d}$$

- a (ή TP) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 1 και ταξινομήθηκαν στην 1
- b (ή FN) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 1, αλλά ταξινομήθηκαν στην 2
- c (ή FP) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 2, αλλά ταξινομήθηκαν στην 1
- d (ή TN) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 2 και ταξινομήθηκαν στην 2

$$Fmeasure = \frac{2 * sen * prec}{sen + prec}$$

Συσταδοποίηση

The screenshot displays the Weka Explorer software interface. The 'Clusterer' tab is active, showing the 'SimpleKMeans' algorithm selected. The 'Cluster mode' section on the left includes options for training set usage, percentage split (66%), and visualization settings. The 'Clusterer output' pane on the right shows the execution results, including the number of iterations, within-cluster sum of squared errors, initial starting points, final cluster centroids, and clustered instances.

Clusterer

Choose **SimpleKMeans** -k 2 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

- Use training set
- Supplied test set (Set...)
- Percentage split % 66
- Classes to clusters evaluation (Nom) breast
- Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

- 19:58:08 - SimpleKMeans

Clusterer output

```
=====  
Number of iterations: 2  
Within cluster sum of squared errors: 146.0  
  
Initial starting points (random):  
  
Cluster 0: left_up,right  
Cluster 1: left_low,left  
  
Missing values globally replaced with mean/mode  
  
Final cluster centroids:  
Attribute      Full Data      Cluster#  
                (286.0)      (170.0)      (116.0)  
=====  
breast-quad    left_low    left_up    left_low  
breast         left       right     left  
  
Time taken to build model (full training data) : 0.01 seconds  
  
=== Model and evaluation on training set ===  
  
Clustered Instances  
  
0      170 ( 59%)  
1      116 ( 41%)
```

Status

OK Log x 0

7:58 μμ
25/5/2021

αλγόριθμος K-means

ROC curve

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Classifier output' pane displays the following text:

```
central 12.0 11.0  
[total] 156.0 139.0  
  
Time taken to build model: 0.01 seconds  
  
=== Stratified cross-validation ===  
=== Summary ===  
  
Correctly Classified Instances 189 66.0839 %  
Incorrectly Classified Instances 97 33.9161 %  
Kappa statistic 0.3187  
Mean absolute error 0.4433  
Root mean squared error 0.473  
Relative absolute error 88.9973 %  
Red error rate 94.7889 %  
Instances 286  
  
Accuracy By Class ===  
  
P Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class  
,684 0,366 0,680 0,684 0,682 0,319 0,642 0,619 left  
,634 0,316 0,639 0,634 0,637 0,319 0,642 0,593 right  
,661 0,342 0,661 0,661 0,661 0,319 0,642 0,607
```

A context menu is open over the 'ROC Area' column, with an orange arrow pointing to the 'Visualize ROC curve' option. The menu items are:

- View in main window
- View in separate window
- Save result buffer
- Delete result buffer(s)
- Load model
- Save model
- Re-evaluate model on current test set
- Re-apply this model's configuration
- Visualize classifier errors
- Visualize tree
- Visualize margin curve
- Visualize threshold curve
- Cost/Benefit analysis
- Visualize cost curve

The status bar at the bottom shows 'OK' and a 'Log' button. The system tray at the bottom right displays '8:09 μμ' and '25/5/2021'.

ROC curve

