Subject

ROC graphs (Receiver Operating Characteristics) with TANAGRA.

ROC graphs enable to compare two or more supervised learning algorithms, they have properties that make them especially useful for domains with skewed class distribution and unequal classification error costs.

An ROC graph depicts relative trade-offs between true positives rate and false positives rate. It needs continuous output of classifier, an estimate of an instance's class membership probabilities. In fact, a "score", a numeric value that represents the degree to which an instance is a member of a class is sufficient.

AUC (Area Under Curve) reduces ROC performances to a single scalar value, which enables to compare several classifiers: this area is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

In this tutorial, we compare linear discriminant analysis (LDA) and support vector machine (SVM) on a heart-diseases detection problem.

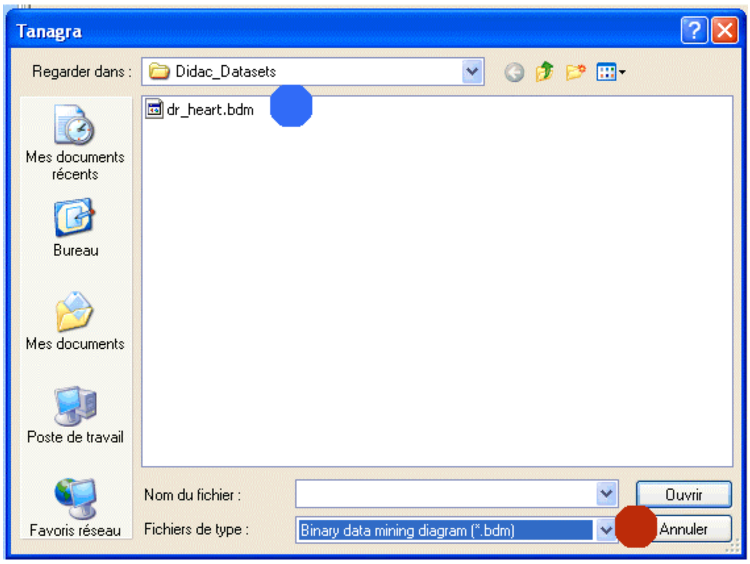
Dataset

HEART dataset; detect heart-disease from patient's characteristics.

ROC graphs

Download the dataset

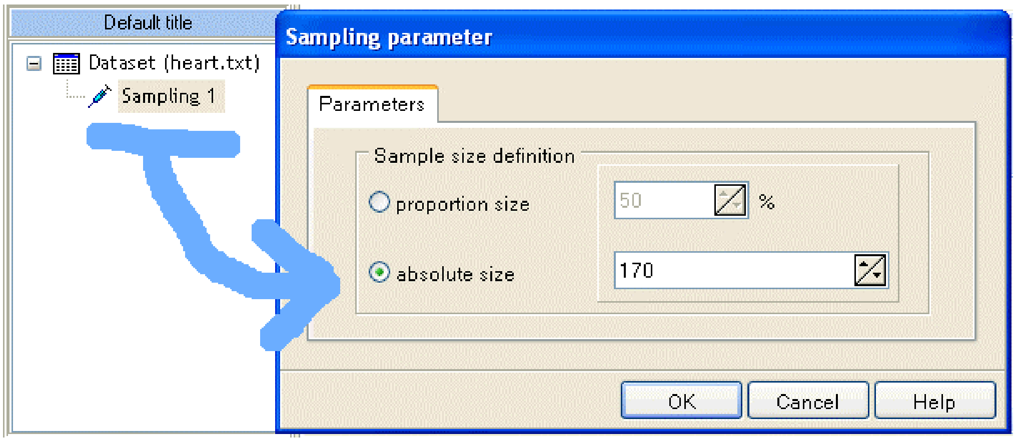
Open the DR_HEART.BDM dataset, click on File / Open menu.



Define training and test set

In order to obtain an unbiased estimate of the performances of the classifiers, we will divide the whole dataset into two parts: 170 instance as a training set; 100 instances for the evaluation of the classifiers.

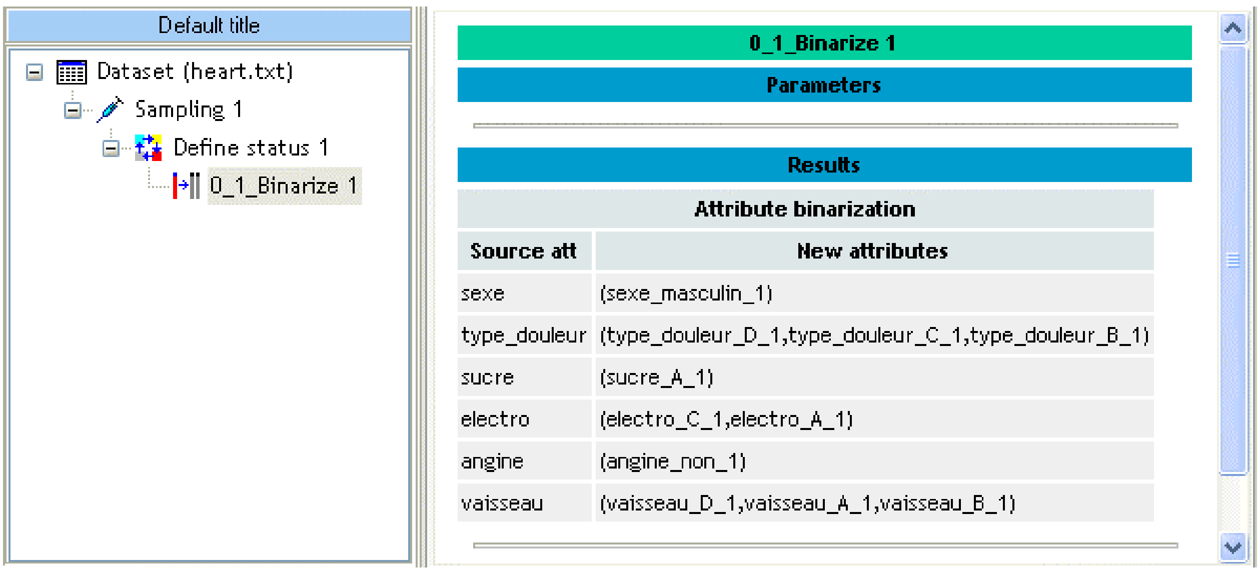
Add the SAMPLING component in the stream diagram, set the following parameters.



Feature construction

LDA and SVM cannot handle discrete attributes, we must transform them.

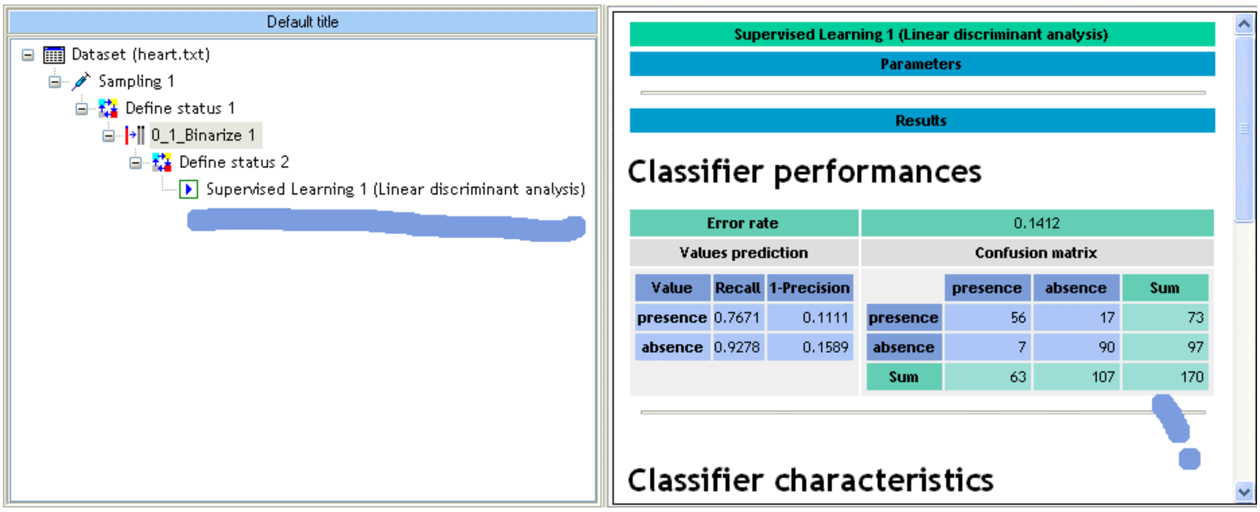
Add the 0_1_BINARIZE component after having selected all discrete attributes (except the class attribute COEUR).



LDA

We can run the LDA learning algorithm now. Set as INPUT all continuous attributes and as TARGET the COEUR attribute, add the SUPERVISED LEARNING component.

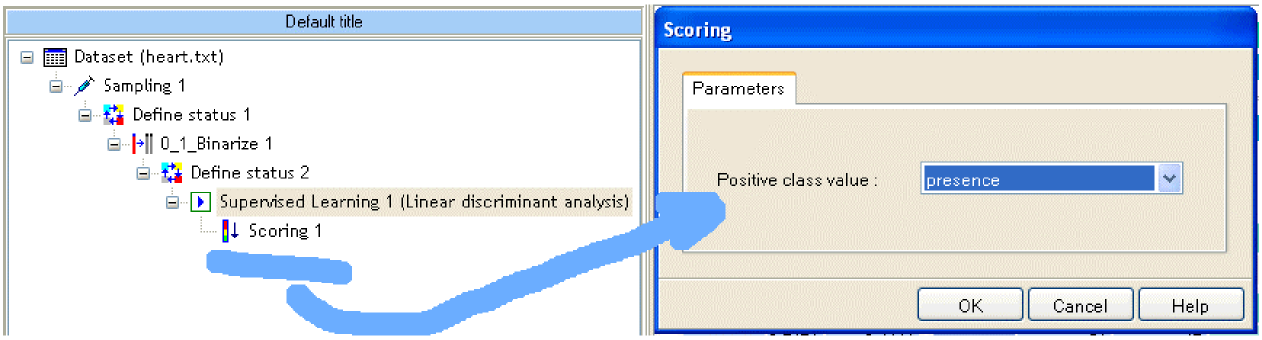
We obtain the following results; we see that we have built the prediction model on 170 instances.



SCORING for LDA

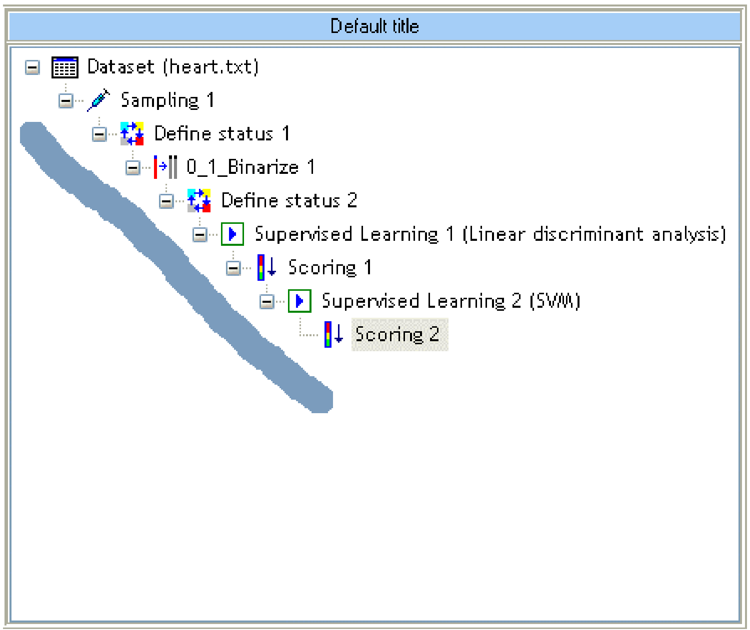
We must now compute the "score" of each instance for the positive class value on the whole dataset.

Add the SCORING component, and set as positive the "PRESENCE" value.



SCORING for SVM

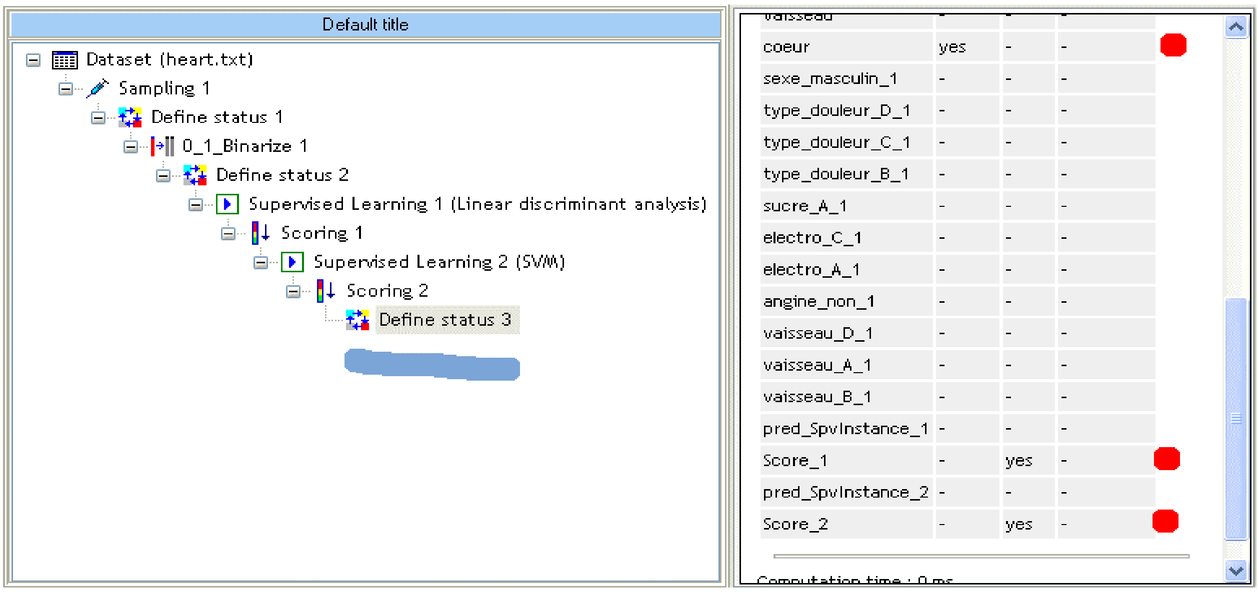
Add also the SVM component in the stream diagram and build the "score" value of each instance.



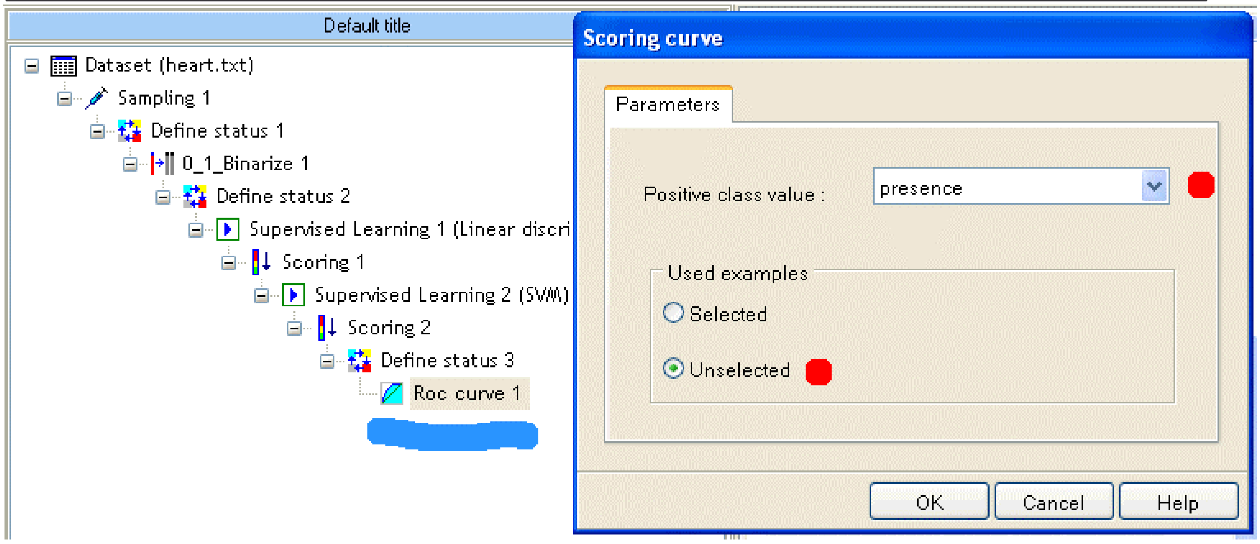
ROC graphs table

In order to build the ROC graphs, we must specify the TARGET attribute and the SCORE attributes.

Add the DEFINE STATUS component in the diagram, set COEUR as TARGET, SCORE_1 (from LDA) and SCORE_2 (from SVM) as INPUT. We note that we can add other score attributes, for instance a score which is provided by a domain expert.



Set the right parameter of the ROC component: the positive class value; the ROC graphs must be computed on the test set.



We obtain the following results:

- AUC.
- For each target size (True Positive + False Positive), we have the true positive rate and the false positive rate.
- Generally, score values are not comparable for different methods.

Roc curve 1

Parameters

Positive class value : presence
Used examples : Unselected

Results

ROC Curve

Sample size : 100
Positive examples : 47
Negative examples : 53

Score Attribute	Score_1			Score_2		
AUC	0.9065			0.8884		
Target size (%)	Score	FP-Rate	TP-Rate	Score	FP-Rate	TP-Rate
0	0.5069	0.0000	0.0000	0.9639	0.0000	0.0000
5	0.5059	0.0000	0.1064	0.9374	0.0000	0.1064
10	0.5054	0.0000	0.2128	0.9127	0.0000	0.2128
15	0.5040	0.0189	0.2979	0.8728	0.0189	0.2979
20	0.5027	0.0189	0.4043	0.7869	0.0189	0.4043
25	0.5023	0.0189	0.5106	0.7357	0.0189	0.5106
30	0.5018	0.0189	0.6170	0.6366	0.0377	0.5957
35	0.5004	0.0566	0.6809	0.5392	0.0566	0.6809
40	0.4997	0.1132	0.7234	0.4996	0.0943	0.7447
45	0.4990	0.1509	0.7872	0.4192	0.1887	0.7447
50	0.4987	0.2075	0.8298	0.3535	0.2453	0.7872
55	0.4984	0.2642	0.8723	0.3200	0.3019	0.8298
60	0.4979	0.3396	0.8936	0.2562	0.3396	0.8936
65	0.4971	0.3962	0.9362	0.2345	0.3962	0.9362
70	0.4966	0.4528	0.9787	0.1861	0.4717	0.9574
75	0.4963	0.5472	0.9787	0.1461	0.5660	0.9574
80	0.4960	0.6415	0.9787	0.1258	0.6415	0.9787
85	0.4954	0.7170	1.0000	0.1038	0.7358	0.9787
90	0.4944	0.8113	1.0000	0.0621	0.8113	1.0000
95	0.4937	0.9057	1.0000	0.0435	0.9057	1.0000
100	0.4930	1.0000	1.0000	0.0270	1.0000	1.0000

For the heart disease problem, we see that LDA and SVM have similar performances (because we have randomly selected the training and test set, you obtain a little different results).

ROC graphs with a spreadsheet

In order to draw the ROC graph, you can copy the results in a spreadsheet and build a scatter plot.

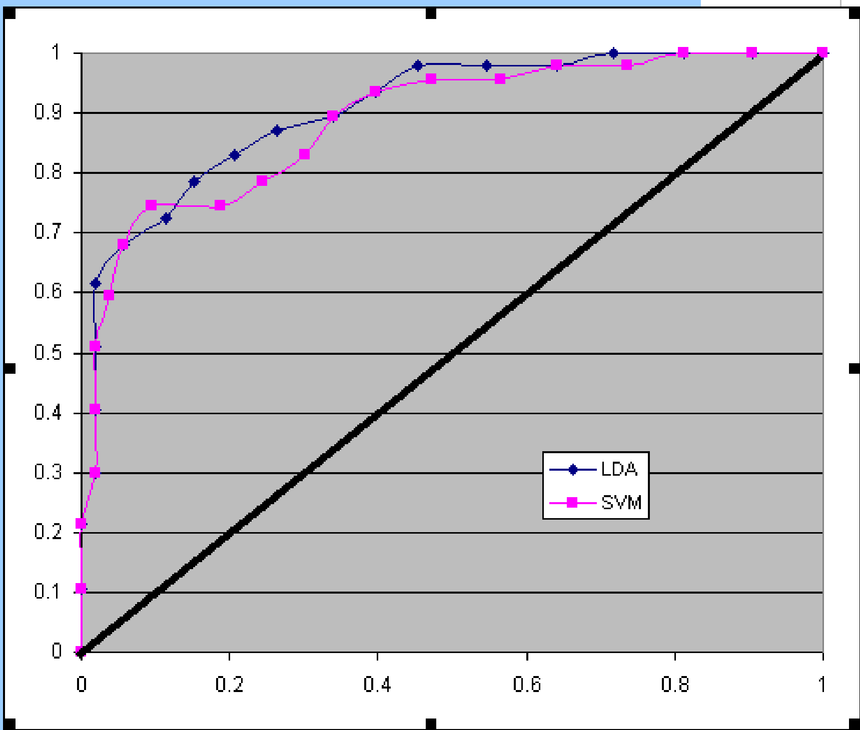
Click on the COMPONENT / COPY RESULTS menu and paste the grid in a spreadsheet, to build the graph is easy.

Results

ROC Curve

Sample size : 100
Positive examples : 47
Negative examples : 53

Score Attribute		Score_1			Score_2	
AUC		0.9065			0.8884	
Target size (%)	Score	FP-Rate	TP-Rate	Score	FP-Rate	TP-Rate
0	0.5069					
5	0.5059					
10	0.5054					
15	0.504					
20	0.5027					
25	0.5023					
30	0.5018					
35	0.5004					
40	0.4997					
45	0.499					
50	0.4987					
55	0.4984					
60	0.4979					
65	0.4971					
70	0.4966					
75	0.4963					
80	0.496					
85	0.4954					
90	0.4944					
95	0.4937					
100	0.493					



Computation time : 0 ms.
Created at 29/04/2005 10:49:42