# Subject

In this tutorial, we use the stepwise discriminant analysis (STEPDISC) in order to determine useful variables for a classification task.

## Feature selection for supervised learning

**Feature selection.** The determination of relevant variables is an important step in a classification process. The model complexity is reduced; it is easier to interpret. Moreover, the diminution of the variables to collect is an advantage during the deployment of the model. In some cases, the variable selection enables to improve the model accuracy.

Manual selection by an expert domain is certainly the best approach. But because the number of candidate descriptors is often large, it is not always possible in practice. So, we must select automatically the best variables. We can also use the automatic process as a preliminary approach in order to filter out the really irrelevant attributes.

**STEPDISC approach.** STEPDISC (Stepwise Discriminant Analysis) is always associated to discriminant analysis because it relies on the same criterion i.e. the WILKS' LAMBDA. So it is often presented such as a method especially intended for the discriminant analysis. In effect, it could be useful for various linear models because they are based upon the same representation bias (e.g. logistic regression, linear SVM, etc.). However, it is not really adapted to non-linear model such as nearest neighbor or multi layer perceptron.

We implement the FORAWRD and the BACKWARD strategies in TANAGRA. In the FORWARD approach, at each step, we determine which is the variable that really contributes to the discrimination between the groups. We add this variable if its contribution is significant. The process stops when there is no attribute to add in the model. In the BACKWARD approach, we begin with the complete model with all descriptors. We search which is the less relevant variable. We remove this variable if the removing does not significantly damage the discrimination between groups. The process stops when there is no variable to remove.

**Stopping rule.** The stopping rule is a key point of the algorithm. Using the standard statistical framework of hypothesis testing is not adapted here. The interpretation of the computed p-value as a probability of rejecting the null hypothesis – if it is true – is erroneous. Indeed, the variable to test is selected among several variables, it maximizes the statistic of the test, and the variables used at one step are the resulting selected (or unselected) variables at the previous step. So the comparison of the p-value with a predefined significance level must be done with caution. It must be viewed mainly such as an adjustment tool that enables to direct the process to the desired solutions (a learning bias). In TANAGRA, we can compare the p-value with a predefined "significance level", or compare the statistic F with a predefined threshold value. We can also directly set the number of selected variables.

# Dataset

We use the SONAR dataset (SONAR\_FOR\_STEPDISC.XLS). The goal is to determine the kind of object (MINE or ROCK) starting from parameters gathered with a sensor. Variable selection process seems necessary here. The ratio between the number of candidate attributes (60) and the number of observations (208) seems unfavorable. The overfitting problem -- the model is too specific to the learning set -- often occurs in this situation.

# Linear discriminant analysis

In a first step, we use all the descriptors for the construction of the prediction model. It is our reference situation; it will enable us to evaluate the efficiency of the STEPDISC method.

### Data importation

In order to initialize a new diagram and import the dataset, we activate the FILE/NEW menu. We use the EXCEL file format in this tutorial; the dataset must be in the first sheet of the workbook.

TANAGRA 1.4.13							
Ek Diagram Window Help							
Data mini	Choose your dataset an Diagram title : Default title Data mining diagram D/TemplExeldefau Dataset (*.s.t.*arti*	nd start download m file name : it.bdm xde) :	× • • • • • • • • • • • • • • • • • • •	, , , , , , , , , , , , , , , , , , ,		-	
			0		<u> </u>		
		OK	Tanagra	(m. 1)			2 🛛
<u></u>			Regarder dans :	discriminant_analy	vsis 💌	G 🕫 🖻 🖽-	
Data visualization	Statistics	Nonparametric statistics	A	sonar_for_stepdisc	ds		
Feature selection	Regression	Factorial analysis	Mes documents				
Spv learning	Meta-spv learning	Spv learning assessment			1		
Correlation scatterplot	Scatterplot	🔛 View multiple scatte	Bureau		1		
Export dataset	🔣 View dataset		Darcuu		1		
			1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		1		
			Mes documents		1		
					i		
			Poste de travail		i		
			~		¥		
				Nom du fichier :	sonar_for_stepdisc.xls	×	Ouvrir
			Favoris réseau	Fichiers de type :	Excel File (97 & 2000)	~	Annuler

We check that there are really 61 variables (60 descriptors + the class attribute) and 208 instances in our file.

🌋 TANAGRA 1.4.13 - [Da	taset (sonar_for_stepdis	sc.xls)]				
Tile Diagram Component	t Window Help					_ 8 ×
🗅 📽 🔚 🙀						
	Default title		Workbook into	rmation		~
Dataset (sonar_for_	Dataset (sonar for stepdisc.xls)			2		
			Selected sheet	sonar		
			Sheet size	209 × 61		
			Dataset size	209 × 61		
			Datasource pro	ocessing		
			Computation time	e 63 ms		
			Allocated memory	117 KB		
			Attribute Catego	escript	nations	
			<			>
		Com	ponents			
Data visualization	Statistics	Nonpar	ametric statistics	Instar	nce selection	
Feature construction	Feature selection	F	Regression	Fact	orial analysis	
PLS	Clustering	S	pv learning	Meta	-spv learning	
Spv learning assessment	Scoring	Association				
Correlation scatterplot	🚔 Export dataset	🛃 Si	catterplot	1940 V	ïew dataset	🚉 View multir
<						>
						.::

We use the DEFINE STATUS component in order to set the INPUT attributes (predictive attributes, V1...V60) and the TARGET attribute (class attribute, CLASS).



Then we add the LDA component into the diagram.

💇 TANAGRA 1.4.13 - [Suj	pervised Learning 1 (Line	ar discrimin	ant analysis	)]					(	_ 🗆 🔀
T File Diagram Component	Window Help									- 8 ×
	Default title						Results			~
🖃 🎹 Dataset (sonar_for_s	tepdisc.xls)		Class	. <b>c</b> :		·		-		
🖃 🎇 Define status 1	🖮 🎦 Define status 1			ILLI	er peri	orn	nance	5		
Supervised L	earning 1 (Linear discrimina)	nt analysis)	Fi	TOP	rate		1-	1, 1010	•	
l ^			∀atue	s pre	ediction		Confu	sion matrix		
			Value Re	call	1-Precision		Rock	Mine	Sum	
			Rock 0.8	969	0.1122	Rock	87	10	97	
1			Mine 0.9	9009	0 0909	Mine	11	100	111	
1					0.0707	Sum	98	110	208	
۱										
	1 1 1		Clace	fi	er cha	ract	oricti	~		~
	1		Components							
Data visualization	Statistics	Nonparame	tric statistics		Instance	selectio	on F	eature const	ruction	
Feature selection	Regression	Factori	al analysis		PL	.s		Clusteri	ng	
Spv learning	Mota spy loarning	Spv learnin	g assessment		Sco	ring		Associati	ion	
<sup>1</sup> / <sub>1+2</sub> *Binary logistic regression	C-SVC		🛏 🕍 Linear	disc	riminant anal	ysis	😡 Naive I	oayes		🔣 SVM
24.5	asC4.5 🚼 Encision List			g TR	IRLS		🔁 Protot	ype-NN		
🕅 C-PLS 🗛 ID3			Multilayer perceptron 🔅 Radial basis function				n			
A; C-RI		1+E* Multine	omia	Logistic Reg	ression	🐴 Rnd Tr	ee			
<										>

We see below the confusion matrix. The resubstitution error rate is **0.1010**. The detailed result of the learning process points out that several descriptors seem not relevant in the classification model.

Several questions can arise: "can we remove directly these variables?"; "all these variables or some of them?"; "the remaining variables are all relevant in the resulting model?"... The feature selection process must provide responses to these questions.



We know that the resubstitution error rate is often optimistic. We use a resampling method -- bootstrap -- for obtaining an honest error estimate.



We observe that the true error rate is about **0.2544**, broadly higher than the indicated value by the resubstitution error rate.

## STEPDISC feature selection method

Now we insert the STEPDISC approach in the learning framework. The key questions are: "how many variables are sufficient for the classification?"; "what is the consequence of the selection on the subsequent model?"

### STEPDISC

We add the STEPDISC component (FEATURE SELECTION tab) under the DEFINE STATUS 1 into the diagram. We activate the PARAMETERS menu.



We ask the FORWARD strategy and set the comparison of the computed statistic F to 3.84<sup>1</sup> as the stopping rule. We activate the VIEW menu in order to obtain the results. We observe that **7 descriptors** are selected. They are listed in a table.



<sup>&</sup>lt;sup>1</sup> This is the default threshold value in SPSS.

The detailed results are displayed in the following table. Only the results for the 5 (can be modified) best attributes are displayed. At each step, we observe the best attributes. We can check especially if the best one has really a significant better F-value than the following attributes.

De	taile	d resu	lts				
N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(1,206)	<b>v11</b> L : 0.813 F : 47.23 p : 0.0000	<b>v11</b> L : 0.813 F : 47.23 p : 0.0000	<b>v12</b> L : 0.846 F : 37.36 p : 0.0000	<b>v49</b> L:0.882 F:27.57 p:0.0000	<b>v45</b> L : 0.884 F : 27.11 p : 0.0000	<b>v10</b> L : 0.884 F : 26.94 p : 0.0000
2	(1, 205)	<b>v47</b> L : 0.731 F : 23.23 p : 0.0000	<b>v47</b> L:0.731 F:23.23 p:0.0000	<b>v46</b> L:0.738 F:21.11 p:0.0000	<b>v49</b> L : 0.743 F : 19.37 p : 0.0000	<b>v48</b> L:0.749 F:17.69 p:0.0000	<b>v45</b> L:0.750 F:17.46 p:0.0000
3	(1,204)	<b>v36</b> L:0.679 F:15.67 p:0.0001	<b>v36</b> L : 0.679 F : 15.67 p : 0.0001	<b>v37</b> L : 0.689 F : 12.45 p : 0.0005	<b>v21</b> L : 0.693 F : 11.01 p : 0.0011	<b>v35</b> L : 0.696 F : 10.09 p : 0.0017	<b>v22</b> L:0.697 F:9.78 p:0.0020
4	(1, 203)	<b>v45</b> L : 0.653 F : 8.09 p : 0.0049	<b>v45</b> L:0.653 F:8.09 p:0.0049	<b>v44</b> L:0.655 F:7.31 p:0.0074	<b>v4</b> L:0.657 F:6.53 p:0.0113	<b>v21</b> L : 0.658 F : 6.35 p : 0.0125	<b>v43</b> L:0.660 F:5.81 p:0.0168
5	(1, 202)	<b>v4</b> L : 0.630 F : 7.19 p : 0.0079	<b>v4</b> L:0.630 F:7.19 p:0.0079	<b>v21</b> L : 0.635 F : 5.58 p : 0.0191	<b>v31</b> L : 0.637 F : 4.91 p : 0.0279	<b>v54</b> L : 0.638 F : 4.55 p : 0.0342	<b>v23</b> L:0.638 F:4.46 p:0.0359
6	(1,201)	<b>v15</b> L : 0.611 F : 6.15 p : 0.0140	<b>v15</b> L:0.611 F:6.15 p:0.0140	<b>v16</b> L:0.612 F:6.07 p:0.0146	<b>v23</b> L : 0.616 F : 4.60 p : 0.0331	<b>v21</b> L : 0.616 F : 4.53 p : 0.0346	<b>v22</b> L:0.617 F:4.27 p:0.0401
7	(1,200)	<b>v21</b> L:0.585 F:9.03 p:0.0030	<b>v21</b> L:0.585 F:9.03 p:0.0030	<b>v20</b> L:0.589 F:7.57 p:0.0065	<b>v22</b> L:0.592 F:6.58 p:0.0111	<b>v31</b> L : 0.594 F : 5.86 p : 0.0164	<b>v23</b> L:0.597 F:4.85 p:0.0288
8	(1, 199)	-	<b>v52</b> L:0.577 F:2.89 p:0.0908	<b>v54</b> L:0.578 F:2.30 p:0.1312	<b>v49</b> L:0.579 F:2.15 p:0.1440	<b>v43</b> L:0.579 F:2.03 p:0.1556	<b>v31</b> L:0.580 F:1.65 p:0.2008

We can make several comments from this table. At the first step, V11 is the best attribute; the difference with the following variable (V12) seems large (+27%). At the second step, V47 is the best attribute, but the difference with the following variable is not obvious.... At the  $6^{th}$  step, the choice between V15 and V16 is not conclusive. Adding or removing a small number of instances in the dataset can invert the result.

N.B. STATISTICA, with the same parameters, gives exactly the same succession of results.

ANALYS DISCRI	hese Anal E M	yse Increment Etape	F	stepdisc.sta)	dl 2	niveau p	Nb. de var inc	Lambda	Valeur F	dl 1	dl 2	-U	×
V11	-(E)	1	47.23461	1	206	.000000	1.000000	.813475	47.23462	1	206	.000000	
V47	-(E)	2	23.23266	1	205	.000003	2.000000	.730668	37.78254	2	205	. 000000	
V36	-(E)	3	15.67116	1	204	.000104	3.000000	.678543	32.21473	3	204	. 000000	
V45	-(E)	4	8.09142	1	203	.004903	4.000000	.652533	27.02379	4	203	. 000000	
V4	-(E)	5	7.19146	1	202	.007931	5.000000	.630101	23.71670	5	202	. 000000	
V15	-(E)	6	6.14801	1	201	.013979	6.000000	.611400	21.29227	6	201	.000000	
V21	-(E)	7	9.03302	1	200	.002991	7.000000	.584979	20.27033	7	200	. 000000	-
•												•	

### STEPDISC + Discriminant Analysis + Bootstrap

According to STEPDISC, 7 variables seem sufficient for the classification process. Now, we want to check the efficiency of the resulting model. We add the Linear Discriminant Analysis and the Bootstrap components into the diagram. The simplest way is to drag and drop the corresponding branch of the diagram under STEPDISC 1.



We observe that all the variables are relevant in the classification model (at the 5% level). The resubstitution error rate is **0.1875**.

## **Classifier performances**

	Error	rate		0.1875				
Values prediction				Confu	ision matrix			
Value	Recall	1-Precision		Rock	Mine	Sum		
Rock	0.8041	0.2041	Rock	78	19	97		
Mine	0.8198	0.1727	Mine	20	91	111		
			Sum	98	110	208		

#### LDA Summary

	Classification	n functions	Statistical Evaluation				
Attribute	Rock	Mine	Wilks L.	Partial L.	F(1,200)	p-value	
v11	8.637494	16.695728	0.656707	0.890776	24,52325	0.000002	
v47	29,753123	36,964342	0.600735	0.973772	5,38679	0.021298	
v36	10.181178	6,494436	0.650366	0.899462	22,35513	0.000004	
v45	-7.377507	-2.882663	0.601739	0.972147	5,73010	0.017601	
v4	2.268385	15,468078	0.613508	0.953499	9.75378	0.002055	
v15	3,120895	-0.313122	0.616228	0.949291	10.68357	0.001272	
v21	11.106784	13.585777	0.611400	0.956787	9.03302	0.002991	
constant	-8.278780	-11.421379		-			

And the bootstrap error rate estimate is **0.2584**. We note that this model with 7 variables is as accurate as the model with 60 variables.

💯 TANAGRA 1.4.13 - [Bo	otstrap 2]					
Tile Diagram Component	: Window Help			_ 7 ×		
🗅 📽 🖪   🎎						
	Default title		Bootstrap 2			
🖃 🥅 Dataset (sonar_for_	_stepdisc.xls)		Parameters			
😑 🚰 Define status 1		Replications : 25	Replications : 25			
Supervised L	.earning 1 (Linear discriminar	nt analy:				
Bull Stendisc 1	рі		Docute			
in topuse : in topuse : in topuse : Supervis:	ed Learning 2 (Linear discrim	inant ar				
E. P. Boot	strap 2 <mark>&gt; </mark>	Boostrap e	rror estimation			
		Er	ror rate	<b>_</b>		
		.632+ bootstrap	0.2584			
		.632 bootstrap	0.2501			
		Resubstitution	0.1875			
		Avg test set	0.2865	_		
		Computation time Created at 26/01/	Computation time : 1766 ms. Created at 26/01/2007 09:14:33			
<		>				
		Components				
Data visualization	Statistics	Nonparametric statistics	Instance selection			
Feature construction	Feature selection	Regression	Factorial analysis			
PLS	Clustering	Spv learning	Meta-spv learning			
Spv learning assessment	Scoring	Association				
🛃 CFS filtering 🛛 🖪 F	CBF filtering 🛛 💾 Fisher	filtering 🛛 🔂 MODTree filt	ering 👜 Runs filtering			
👪 Define status 🛛 🖪 F	eature ranking 💦 🔣 MIFS fi	iltering 🛛 🔢 Remove cons	stant 🛛 🛃 Stepdisc			

# **BACKWARD Selection**

In this section, we modify the parameter of STEPDISC. We select a backward strategy<sup>2</sup>.

Stepdisc Parar	neter	
Parameters	Report	
Seard Seard For Stopp Stopp Stopp Stopp Stopp Stopp Stopp Stopp Stopp Stopp Stopp Stopp Stopp Stopp	ch strategy ckward search ward search ing rule o enter or remove nificance level oset size	10 0.05 5
	04	Cancel Help

<sup>&</sup>lt;sup>2</sup> The threshold value 10 is suggested by STATISTICA for this problem.

The process selects also 7 variables, but the selected subset is noticeably different of the subset selected with the FORWARD approach.

#### Selection results

[7] selected attributes on [60]

#### Selected attributes' subset

N°	Selected atts
1	v4
2	v12
3	v30
4	v31
5	v32
6	v36
7	v49

#### **Detailed results**

N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	(1, 147)	<b>v26</b> L:0.380 F:0.00 p:0.9929	<b>v26</b> L:0.380 F:0.00 p:0.9929	<b>v45</b> L:0.380 F:0.00 p:0.9846	<b>v33</b> L:0.380 F:0.01 p:0.9210	<b>v46</b> L:0.380 F:0.01 p:0.9048	<b>v38</b> L:0.380 F:0.03 p:0.8683
2	(1, 148)	<b>v45</b> L : 0.380 F : 0.00 p : 0.9854	<b>v45</b> L:0.380 F:0.00 p:0.9854	<b>v33</b> L:0.380 F:0.01 p:0.9208	<b>v46</b> L : 0.380 F : 0.02 p : 0.9021	<b>v38</b> L:0.380 F:0.03 p:0.8680	<b>v58</b> L : 0.380 F : 0.04 p : 0.8407
3	(1, 149)	<b>v33</b> L : 0.380 F : 0.01 p : 0.9212	<b>v33</b> L:0.380 F:0.01 p:0.9212	<b>v46</b> L : 0.380 F : 0.02 p : 0.8784	<b>v38</b> L:0.380 F:0.03 p:0.8682	<b>v37</b> L:0.380 F:0.04 p:0.8403	<b>v58</b> L:0.380 F:0.04 p:0.8394
4	(1, 150)	<b>v46</b> L : 0.380 F : 0.03 p : 0.8713	<b>v46</b> L:0.380 F:0.03 p:0.8713	<b>v38</b> L:0.380 F:0.03 p:0.8687	<b>v58</b> L:0.380 F:0.04 p:0.8438	<b>v37</b> L : 0.380 F : 0.04 p : 0.8431	<b>v43</b> L:0.380 F:0.06 p:0.8118

# FORWARD – BACKWARD comparison

When we compare the results of the two strategies, we observe that two variables only are shared by the resulting subsets.

N٥	Forward	Backward
14	TOIWalu	Dackwalu
1	v4	v4
2	v11	v12
3	v15	v30
4	v21	v31
5	v36	v32
6	v45	v36
7	v47	v49

This kind of result leaves often puzzled the users. What is the "true" best subset?

I think there is no numerical – statistical – satisfying response to this question. These tools give us some indications on the possible relevant attributes. Only the domain knowledge and the specification of the study enable us to select an appropriate result.