1 Objectif

Mettre en œuvre le composant STEPDISC de **TANAGRA** pour sélectionner automatiquement les variables pertinentes pour l'analyse discriminante linéaire.

2 La sélection de variables

Nécessité de la sélection de variable. La sélection de variables est un processus très important en apprentissage supervisé. Nous disposons d'une série de variables candidates, nous cherchons les variables les plus pertinentes pour expliquer et prédire les valeurs prises par la variable à prédire. Les objectifs sont bien souvent multiples : nous réduisons le nombre de variables à recueillir pour le déploiement du système ; nous améliorons notre connaissance du phénomène de causalité entre les descripteurs et la variable à prédire, ce qui est fondamental si nous voulons interpréter les résultats pour en assurer la reproductibilité ; enfin, mais pas toujours, nous améliorons la qualité de la prédiction, le ratio nombre d'observations et dimension de représentation étant plus favorable.

La meilleure approche pour sélectionner les variables pertinentes est certainement la sélection experte. Seule la connaissance du domaine permet de bien comprendre les causalités sous-jacentes, discerner les vrais liens des simples artefacts, mettre en évidence les interactions, etc. Malheureusement, elle n'est pas toujours possible, notamment parce que le nombre de variables candidates est élevé, une sélection manuelle devient vite inextricable. De toute manière, dans une démarche exploratoire, il paraîtrait bien étrange finalement que tout soit connu à l'avance, on se demande alors à quoi servirait la fouille de données dans ce contexte.

Il nous faut nous tourner vers les méthodes automatiques. Ne serait-ce que pour défricher le terrain et proposer aux experts les groupes de variables les plus intéressantes, quitte à procéder manuellement dans une deuxième phase.

La méthode STEPDISC. Dans ce tutoriel, nous présentons la méthode STEPDISC (Stepwise Discriminant Analysis). Elle repose le critère du LAMBDA de WILKS. Géométriquement, il s'agit de trouver le sous-espace de représentation qui permet un écartement maximal entre les centres de gravité des nuages de points conditionnels c.-à-d. les nuages de points associés à chaque valeur de la variable à prédire. Elle est donc particulièrement bien adaptée à l'analyse discriminante linéaire qui

utilise également le même critère, d'où son appellation. Elles sont systématiquement associées dans les logiciels.

En réalité, rien ne nous empêche de l'utiliser comme préalable à telle ou telle méthode d'apprentissage. Il faut simplement garder à l'esprit qu'elle est adaptée pour les méthodes induisant un séparateur linéaire (régression logistique, SVM linéaire, etc.). En revanche, il est totalement illusoire d'en attendre un quelconque bénéfice si nous souhaitons l'utiliser dans le cadre de la méthode des plus proches voisins par exemple¹. Les idées qui les sous-tendent – le biais de représentation – ne sont pas du tout en adéquation.

Le LAMBDA de WILKS est la transposition multidimensionnelle de la décomposition de la variance, il représente le rapport entre l'inertie intra classes et l'inertie totale. Si les nuages sont totalement confondus, LAMBDA = 1 ; plus LAMBDA se rapproche de 0, plus les nuages conditionnels sont distincts. TANAGRA implémente deux stratégies. L'approche FORWARD consiste à partir de l'ensemble vide, choisir la variable induisant la meilleure amélioration du LAMBDA, et la sélectionner si amélioration est statistiquement significative ; nous procédons itérativement en ajoutant unes à unes les variables jusqu'à ce que l'adjonction d'une variable n'apporte plus d'amélioration. A l'inverse, l'approche BACKWARD, part de l'ensemble des variables candidates, recherche la variable dont le retrait entraînerait la dégradation la plus faible du LAMBDA, et la retire effectivement si cette dégradation n'est pas statistiquement significative.

Il existe une variante dite STEPWISE qui mixe ces deux approches : après avoir ajouté une variable, nous regardons s'il n'est pas nécessaire de retirer certaines variables parmi celles qui ont déjà été introduites. Puis nous regardons à nouveau s'il n'est pas possible d'en ajouter de nouvelles, etc. Cette option n'est pas implémentée dans TANAGRA.

Définition de la règle d'arrêt. Pour évaluer le rôle significatif d'une variable, nous utilisons la statistique F qui, a priori, suit une loi de Fisher. Il suffirait donc de comparer la p-value calculée pour la variable à évaluer et la comparer avec le niveau de signification qu'a choisi l'utilisateur.

Cette démarche n'est malheureusement pas exempte de reproches. En effet, la loi calculée dans le contexte d'un test d'hypothèse statistique est mise en œuvre dans un contexte exploratoire. Nous

¹ Mis à part peut-être qu'en réduisant la dimensionalité, nous améliorons la stabilité des estimations locales des probabilités. Il s'agit plus dans ce cas d'une conséquence de la réduction du nombre de variables plutôt que la définition d'un sous-espace adapté à la méthode.

avons d'abord choisi la variable maximisant (forward) ou minimisant (backward) F, puis nous avons testé cette variable. A l'étape suivante, nous procédons de même, sachant que les évaluations actuelles sont dépendantes des calculs à l'étape précédente, etc. L'interprétation de la p-value en termes de risque n'est donc plus possible. La loi n'est pas adaptée. Et les techniques de correction usuelles plus ou moins restrictives du calcul de la p-value (Bonferroni, Sidak, etc.) posent problème du fait que les calculs sont en cascade, les évaluations à chaque étape sont tributaires des sélections réalisées à l'étape précédente. Pour cette raison, certains logiciels -- STATISTICA par exemple -- se refusent à comparer la p-value calculée dans un contexte de tests d'hypothèses avec un niveau de signification défini par l'utilisateur. D'autres – SPSS – la proposent quand même mais en concurrence avec d'autres règles d'arrêt.

Il est vrai que la p-value a un autre défaut, elle a tendance à prendre des valeurs très basses lorsque l'échantillon est grand, laissant à croire que toutes les variables sont pertinentes si nous la comparons avec les seuils usuels utilisés en statistique (5%, 1%, etc.)

Si la p-value ne convient pas, quel critère utiliser ? La plupart des logiciels contournent le problème en proposant de comparer la statistique F calculée avec un seuil défini de manière ad hoc par l'utilisateur. Reste alors à définir la valeur seuil. Il n'y a pas vraiment de repères. SPSS propose par défaut la valeur 3.84 qui ressemble à s'y méprendre à la valeur critique d'un test à 5% lorsque nous travaillons sur un échantillon de quelques milliers d'individus. Ce qui nous ramène au problème précédent. STATISTICA propose des valeurs seuils par défaut sans que l'on sache vraiment ce qu'il faut en penser.

Enfin, un seuil est toujours arbitraire. Si nous fixons le seuil F à 3.84, une variable avec un F calculé de 3.839 ne passera pas dans la sélection FORWARD. Ce n'est pas très satisfaisant. Surtout si l'information nous a échappé.

Si du point de vue statistique, tout cela semble bien compliqué, du point de vue exploratoire, c'est moins gênant. Il faut surtout voir ces seuils comme autant d'outils mis à la disposition du praticien pour guider la recherche vers les solutions qui répondent à son cahier des charges. S'il désire sélectionner un ensemble de variables plus large pour pouvoir choisir lui-même dans ce sousensemble réduit, il choisira un seuil plus permissif. S'il veut en revanche réduire de manière drastique le nombre de variables parce qu'il en a un très grand nombre au départ, il fixera le seuil en conséquence. Pour cette raison, nous proposons dans **TANAGRA** les deux approches ci-dessus. Nous permettons également à l'utilisateur d'imposer directement le nombre de variables à sélectionner. Et surtout, nous fournissons le détail des calculs : nous avons la possibilité d'inspecter les variables qui ont été en compétition, celles qui ont été éliminées parce que redondantes (dans un processus FORWARD, elles sont en bonne position dans les premières étapes, puis n'apparaissent plus) ; celles qui sont complémentaires (toujours dans un FORWARD, elles ne sont pas présentes dans les premières étapes, puis sont très significatives par la suite). L'analyse fine des résultats permet comprendre les relations entre les variables.

3 Fichier de données

Nous utilisons le fichier SONAR (SONAR_FOR_STEPDISC.XLS). L'objectif est de prédire l'appartenance d'un objet (Roc ou Mine) à partir des relevés fournis par un détecteur.

La sélection de variables est particulièrement recommandée car nous disposons de relativement peu d'observations (208 individus) pour 60 variables candidates. Il y a risque de sur-apprentissage.

4 Analyse discriminante linéaire dans TANAGRA

Dans un premier temps, nous mettons en œuvre une analyse discriminante linéaire, sans chercher à réduire le nombre de variables. Le résultat servira de référence, elle permettra d'évaluer l'efficacité de la méthode de sélection de variables STEPDISC.

4.1 Importation des données et création d'un diagramme

Nous créons un nouveau diagramme et nous importons les données. Nous activons pour cela le menu FILE/NEW. Notre fichier est au format EXCEL, les données doivent être situées sur la première feuille de calcul.

TANAGRA 1.4.13							
File Diagram Window Help							
□ 글_ ↓ #							
Data min	ing diagram (empty)						
		· · · · · · · · · · · · · · · · · · ·					
	Choose your dataset ar	nd start download	1. J.				
			- N.				
	Diagram title :		- N.				
	Default title		<u>``</u>				
	Data mining diagra	m file name :		<u>``</u>			
	D:\Temp\Exe\defau	ilt.bdm		<u> </u>			
	Dataset (*.txt,*.arff,*.	xls) :		ト			
		ОК	Tanagra				? 🛛
]	Regarder dans :	😂 discriminant_anal	lysis 💌 🤇	3 💋 📂 🛄-	
		Components		sonar_for_stepdis	c.xls		
Data visualization	Statistics	Nonparametric statistics		ivine_quality.xls			
Feature selection	Regression	Factorial analysis	récents		i i		
Spv learning	Meta-spv learning	Spv learning assessment					
Correlation scatterplot	🛃 Scatterplot	🔛 View multiple scatter	Bureau		1		
Export dataset	🔛 View dataset				1		
					1		
[Mes documents		i		
			-				
					1		
			Poste de travail		<u> </u>		
			S	Nom du fichier :	sopar for stendisc vis	~	Ouwrit
				Fichiera de tupe :	Eural File (97 & 2000)		Aprular
		1	Pavons reseau	nonieis de type :	Excel File (97 & 2000)	×	Annuler

Vérifions que nous avons bien 61 variables et 208 observations dans notre fichier.

💯 TANAGRA 1.4.13 - [Da	taset (sonar_for_stepdis	c.xls)]				
🏆 File Diagram Component	: Window Help					- 8 ×
🗅 📽 🔚 👪						
	Default title	ĺ	Workbook into	rmation		
Dataset (sonar for	stendisc.xls)		Number of sheets	2		
			Selected sheet	sonar		
			Sheet size	209 × 61		
			Dataset size	209 × 61		
			Datasource pro	ocessing		
			Computation time	e 63 ms		
			Allocated memory	117 KB		
			61 attribute(s) 208 example(s) Attribute Catego	ory Inform	ations	
			<			>
		Corr	ponents			
Data visualization	Statistics	Nonpara	metric statistics	Instan	ce selection	
Feature construction	Feature selection	F	egression	Facto	orial analysis	
PLS	Clustering	S	ov learning	Meta-	spv learning	
Spv learning assessment	Scoring	A	ssociation			
Correlation scatterplot	Export dataset	🛃 So	atterplot	🔛 Vi	iew dataset	🔃 View multip
<						>
						.::

4.2 Analyse discriminante

Nous plaçons le composant DEFINE STATUS dans le diagramme à partir du raccourci dans la barre d'outils. Nous plaçons en INPUT les variables V1...V60 ; CLASS est la variable à prédire (TARGET).



Puis nous insérons le composant Analyse Discriminante Linéaire.



La matrice de confusion indique un taux de mauvais classement de 0.1010.

En consultant le détail des résultats, nous constatons qu'à un niveau de signification de 5%, plusieurs variables ne sont pas déterminantes dans la modélisation c.-à-d. leur retrait n'entraînerait pas une dégradation significative du Lambda de Wilks.

🕱 TANAGRA 1.4.13 - [Supervised Learning 1 (Linear discriminant analysis)]										
Tile Diagram Component Window Help								_ M -	Ξ×	
Default title	LDA Summary								~	
Dataset (sonar_for_stepdisc.xls)	Classification functions Statistical Evaluation						valuation		·	
🖃 🏠 Define status 1		Attribute	Rock	Mine	Wilks L.	Partial L.	F(1,147)	p-value		
Supervised Learning 1 (Linear discriminant ar	n: v1		38.682162	62.320660	0.383801	0.990494	1.41078	0.23684	3	
	V2	:	5.729074	31.719773	0.385168	0.986977	1.93963	0.16581	1	
	v3		40.363501	-22.797330	0.411132	0.924649	11.97930	0.00070	4 🗐	
	V4	I	18.065991	58.625878	0.400148	0.950029	7.73215	0.00613	5	
	VB	;	0.852600	-6.189806	0.381312	0.996959	0.44835	0.50417	1	
	vé		-15.588328	-4.725685	0.383034	0.992476	1,11447	0.29284	4	
	v7		2,495369	-11.402162	0.385101	0.987148	1.91378	0.16864	1	
	VB		4.468222	-2.370412	0.382169	0.994723	0.77990	0.37861	4	
	V9		-0.235151	7.206803	0.382380	0.994175	0.86132	0.35489	2	
<	2 v1	0	15.326616	9.421990	0.381846	0.995563	0.65510	0.41960	3 🔽	
		Corr	ponents							
Data visualization Statistics N	onparam	etric statistic:	s Insta	nce selection	Featu	ure constructi	on			
Feature selection Regression	Factor	rial analysis		PLS		Clustering				
Spv learning Meta-spv learning S	pv learni	ng assessment	t	Scoring		Association				
1 ter∗Binary logistic regression I KC-SVC		🕍 Linear	discriminant	analysis	🕅 Naive baye	s	<u>الا الا</u>	//A		
asC4.5		1+E* Log-Re	eg TRIRLS		🍡 Prototype-	NN				
🔀 C-PLS 🚓 ID3		📴 Multila	ayer perceptr	on	🔅 Radial basi:	s function				
🕰 C-RT 🔄 K-NN		<mark>₁.e</mark> *Multin	iomial Logistic	Regression	🖧 Rnd Tree					
<									>	

4.3 Evaluation Bootstrap de l'erreur

Le taux d'erreur de 0.1010 calculé sur les données ayant servi à l'apprentissage – dit erreur en resubstitution – est très souvent trop optimiste. Optimisme d'autant plus marqué qu'il y a surapprentissage : le modèle « colle » trop aux données. Afin d'obtenir une évaluation plus réaliste des performances, nous utilisons des méthodes de rééchantillonnage. Nous plaçons à cet effet le composant BOOTSTRAP dans le diagramme.

💯 TANAGRA 1.4.13 - [Bo	ootstrap 1]					
Tile Diagram Componen	t Window Help				-	a ×
🗅 📽 🖬 🔤						
D	efault title			Bootstrap 1		^
🖃 🏢 Dataset (sonar_for_	_stepdisc.xls)			Parameters		
🖃 🚰 Define status 1		Repli	cations : 25			
🖻 🕩 Supervised	Learning 1 (Linear discriminar	nt an:				
E Bootstra	эр 1			Docutto		
\uparrow				Results		=
/ · · · ·		Boo	ostrap err	or estimation		
			Error	rate		
		.632	+ bootstrap	0.2544		
		.632	bootstrap	0.2263		
		Resu	bstitution	0.1010		
		Avg	test set	0.2992		
	····)					<u>~</u>
		Com	ponents			
Data visualization /	Statistics	Nonparamet	ric statistics	Instance selection	Feature construction	
Feature selection	Regression	Factoria	l analysis	PLS	Clustering	
Spv learning	Meta-spv learning 🗡	Spv learning	assessment	Scoring	Association	
Bias-variance decompos	ition ? Train-test					
P Bootstrap						
Patross-validation						
						:

Les résultats montrent que lorsque nous déploierons le modèle de prédiction dans population, la probabilité de réaliser une mauvaise prédiction est de 0.2544, et non de 0.1010 comme semblait l'indiquer l'évaluation en resubstitution.

5 Sélection de variables STEPDISC

Dans un deuxième temps, nous procédons à une sélection de variables à l'aide de la méthode STEPDISC. Nous évaluerons alors les conséquences de cette sélection sur les performances de l'analyse discriminante.

5.1 STEPDISC

Nous plaçons le composant STEPDISC à la suite du DEFINE STATUS 1. Puis nous le paramétrons en activant le menu contextuel PARAMETERS.

💯 TANAGRA 1.4.13 - [Bo	ootstrap 1]						
Tile Diagram Component	t Window Help					-	a ×
D 📽 🔚 🙀							
D	efault title			Destates a			^
🖃 🏢 Dataset (sonar_for_	_stepdisc.xls)		Stepdisc Paramete	r			
🖕 🙀 Define status 1			Rep Parameters F	leport			
🦯 📄 💽 Supervised I	Learning 1 (Linear discrimina	nt an:					
/ P Bootstra	ар 1		Search str	ategy			
/ Stepdisc 1	Daviasa barra		Backwa	rd search			
	Parameters		Bo O Forward	Isearch			
N. Contraction of the second s	Execute						
L N	VIEW		Stopping r	ar or remove 3.84			
<u> </u>			.63				
			.63				
			Res	size			
Te II			Avg				
		<u>i</u>		OK Cancel	Help		×
		<u>``</u>	Cor nponents				
Data visualization	Statistics	Nonpar	ametric statistics	Instance selection	Feat	ure construction	
➡ Feature selection	Regression	Fac	torial analysis	PLS		Clustering	
Spv learning	Meta-spv learning	S <mark>p</mark> v lea	rning assessment	Scoring		Association	
🖪 CFS filtering 🛛 🕂 F	Fisher filtering 🛛 📺 Runs f	iltering					
🚰 Define status 🛛 🛃 A	AIFS filtering 👘 📝 Stepdi	sc ¹					
🛃 FCBF filtering 🔂 🗛	AODTree filtering						
💾 Feature ranking 🔳 F	Remove constant						
[

Nous souhaitons une recherche FORWARD, avec pour règle d'arrêt la comparaison de la valeur F calculée avec le seuil 3.84 (proposée par défaut dans SPSS par ex.). L'exécution du composant (Menu contextuel VIEW) nous indique que **7** variables ont été sélectionnées par la procédure.

💯 TANAGRA 1.4.13 - [S	tepdisc 1]							
🏆 File Diagram Componer	it Window Help			_ 8 ×				
🗅 📽 🔚 🎎								
Defau	ılt title			~				
🖃 🏢 Dataset (sonar_for	_stepdisc.xls)		Results					
🖃 🎎 Define status 1		Selection results						
🖃 🕨 Supervised	Learning 1 (Linear discrim							
P Bootstr	ap 1	[/] selected attributes on [60]						
Stepdisc 1		Selected attribut	es' subset					
		N [®] Selected atts						
		1 v11						
		2 v47						
		3 v36	3 v36					
		4 v45						
		5 v4						
		6 v15						
		7 v21						
				×				
		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 🛃	FCBF filtering 🛛 🛃 Fisher	filtering 🛛 🔣 MODTree filt	tering 👜 Runs filtering					
📬 Define status 🛛 🖪	Feature ranking 🛛 🔣 MIFS f	iltering 👖 Remove con:	stant 🛛 🛃 Stepdisc					
				.::				

Le détail des calculs est affiché dans un tableau. Pour ne pas alourdir la présentation, seules les 5 (paramétrable) meilleures solutions sont affichées à chaque étape.

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

Nous observons que la variable V11 a été introduite à la première étape, le F calculé est de 47.23. A la seconde étape, la variable V47 a été sélectionnée. Nous constatons le choix entre V47 et V46 s'est jouée à peu de choses (F = 23.23 contre F = 21.11). En travaillant sur un autre échantillon, il se peut très bien que V46 passe devant V47. La sélection incrémentielle continue ainsi jusqu'à la huitième étape où la meilleure variable V52 propose un F calculé (2.89) inférieur au seuil que l'on s'est choisi (3.84). Le processus est donc stoppé.

N'ayant aucune connaissance dans le domaine, nous n'allons pas nous risquer à interpréter les résultats. Tout juste ferons-nous remarquer que la variable V12 qui arrive en seconde position dans

la première étape, n'apparaît plus dans les meilleures solutions par la suite. Cela laisse à penser que cette variable est fortement corrélée avec V11, le fait d'avoir introduit cette dernière dans la sélection exclut V12. Si les variables V11 et V12 revêtent des significations particulières pour l'expert, il pourra s'appuyer sur ces informations pour guider la recherche dans la bonne direction.

N.B. A titre de comparaison, nous constatons que STATISTICA fournit exactement la même séquence de résultats.

🔚 Synth	🐨 Synthèse Analyse Incrément. (son _for_stepdisc.sta) _ 🗆 🗙											
ANALYS DISCRI	E M.	Etape	F inc./exc	dl 1	dl 2	niveau p	Nb. de var inc	Lambda	Valeur F	dl 1	dl 2	niveau p 📥
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 🚽
4												•

5.2 STEPDISC + Analyse discriminante + Evaluation bootstrap

Au-delà de l'analyse fine des résultats, un des objectifs de la sélection de variables est de produire un espace de représentation plus performant. Pour évaluer cette idée, nous plaçons à la suite du composant STEPDISC la séquence Analyse Discriminante – Bootstrap. Le plus simple est de procéder par glisser-déposer dans le diagramme lui-même.

TANAGRA 1.4.13 - [Stepdisc 1]									
Tile Diagram Component Window Help									- 8 ×
Default title									~
🖃 🏢 Dataset (sonar_for_stepdisc.xls)	lts								
E Carlos Status 1	N "	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5	
 □ Supervised Learning 1 (Linear discriminant analys □ 1 Bootstrap 1 □ 2 Stepdisc 1 □ 1 Supervised Learning 2 (Linear discriminant ar 	1	(1, 206)	v11 L : 0.813 F : 47.23 p : 0.0000	v11 L : 0.813 F : 47.23 p : 0.0000	v12 L:0.846 F:37.36 p:0.0000	v49 L:0.882 F:27.57 p:0.0000	v45 L:0.884 F:27.11 p:0.0000	v10 L : 0.884 F : 26.94 p : 0.0000	
Bootstrap 2	2	(1, 205)	v47 L : 0.731 F : 23.23 p : 0.0000	v47 L : 0.731 F : 23.23 p : 0.0000	v46 L:0.738 F:21.11 p:0.0000	v49 L:0.743 F:19.37 p:0.0000	v48 L:0.749 F:17.69 p:0.0000	v45 L:0.750 F:17.46 p:0.0000	
	3	(1, 204)	v36 L : 0.679 F : 15.67 p : 0.0001	v36 L:0.679 F:15.67 p:0.0001	v37 L:0.689 F:12.45 p:0.0005	v21 L:0.693 F:11.01 p:0.0011	v35 L:0.696 F:10.09 p:0.0017	v22 L:0.697 F:9.78 p:0.0020	
	4	(1, 203)	v45 L : 0.653 F : 8.09 p : 0.0049	v45 L:0.653 F:8.09 p:0.0049	v44 L:0.655 F:7.31 p:0.0074	ν4 L:0.657 F:6.53 p:0.0113	v21 L:0.658 F:6.35 p:0.0125	ν43 L:0.660 F:5.81 p:0.0168	
	5	(1, 202)	v4 L : 0.630 F : 7.19 p : 0.0079	v4 L:0.630 F:7.19 p:0.0079	v21 L:0.635 F:5.58 p:0.0191	v31 L:0.637 F:4.91 p:0.0279	v54 L:0.638 F:4.55 p:0.0342	v23 L:0.638 F:4.46 p:0.0359	•
	Co	mponent	ts						
Data visualization Statistics Nonpa	arametr	ic statist	ics	Instance :	selection	Fe	ature cons	struction	
Feature selection Regression Fa	ctorial	analysis		PL	S		Cluster	ing	
Spv learning Meta-spv learning Spv le	arning	assessme	nt	Scor	ing		Associa	tion	
CFS filtering Define status H FCBF filtering	H	Feature	ranking	🕂 Fishe	er filtering	R. M	IFS filterin	s H	MODTree >

Nous activons le menu VIEW de l'analyse discriminante. Nous constatons que l'erreur en resubstitution est maintenant de 0.1875, et que les 7 variables sont toutes pertinentes à 5%.

20

98

91

110

97

111

208

Classifier performances										
	Еггог	rate	0.1875							
٧a	lues pro	ediction		Confu	ision matrix					
Value	Recall	1-Precision	Rock Mine							
Rock	0.8041	0.2041	Rock	78	19					

Mine

Sum

0.1727

LDA Summary

Mine 0.8198

	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		-				

Reste alors à évaluer les performances de tout ce processus à l'aide du composant BOOTSTRAP. Il nous indique qu'en déploiement, notre modèle a une probabilité d'erreur de 0.2584, très proche finalement du premier modèle, sauf qu'ici nous n'utilisons plus que 7 variables.

🌋 TANAGRA 1.4.13 - [Bo	ootstrap 2]			
Tile Diagram Componen	t Window Help			_ 8 ×
D 📽 🔚 🎎				
	Default title		Bootstrap 2	
🖃 🏢 Dataset (sonar_for	_stepdisc.xls)		Parameters	
🖃 🎇 Define status 1		Replications : 25		
Supervised	Learning 1 (Linear discrimina)	nt analy:		
Bootstra	арі		Results	
ia in Supervis	ed Learning 2 (Linear discrim	ninant ar	nosaus	
Boot	tstrap 2 🚬	Boostrap e	error estimation	
		En	ror rate	
		.632+ bootstrap	0.2584	
		.632 bootstrap	0.2501	
		Resubstitution	0.1875	
		Avg test set	0.2865	_
<	m	Computation time Created at 26/01/	: : 1766 ms. 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	FCBF filtering 🛛 Fisher Feature ranking 🔛 MIFS fi	filtering I MODTree filt iltering II Remove con:	ering 👜 Runs filtering stant 🗹 Stepdisc	

6 Sélection BACKWARD

Voyons maintenant ce qu'il en est si l'on met en œuvre la sélection BACKWARD. Nous revenons sur le composant STEPDISC et nous activons le menu PARAMETERS. Nous paramétrons la méthode de la manière suivante (la valeur seuil 10 est fixée en référence à la valeur par défaut proposée par STATISTICA).

nepon	
Search strategy	
O Backward search	
O Forward search	
⊂ Stopping rule	
• E to enter or remove	10
 Significance level 	0.05
O Subset size	5
0.000	

Après calcul (menu VIEW), 7 variables ont été retenues pour la discrimination. Le détail des opérations nous indique que la variable V26 a été la première variable exclue, puis V45, etc.

Se	Selection results							
[7] :	[7] selected attributes on [60]							
Se	Selected attributes' subset							
N °	Selecte	d atts						
1	v4							
2	v12							
3	v30							
4	v31							
5	v32							
6	v36							
7	v49							
De	Detailed results							
N°	d.f	Best	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5	
1	(1, 147)	v26 L : 0.380 F : 0.00 p : 0.9929	v26 L:0.380 F:0.00 p:0.9929	v45 L:0.380 F:0.00 p:0.9846	v33 L:0.380 F:0.01 p:0.9210	v46 L:0.380 F:0.01 p:0.9048	v38 L : 0.380 F : 0.03 p : 0.8683	
2	(1, 148)	v45 L : 0.380 F : 0.00 p : 0.9854	v45 L:0.380 F:0.00 p:0.9854	v33 L:0.380 F:0.01 p:0.9208	v46 L:0.380 F:0.02 p:0.9021	v38 L:0.380 F:0.03 p:0.8680	v58 L : 0.380 F : 0.04 p : 0.8407	
3	(1, 149)	v33 L : 0.380 F : 0.01 p : 0.9212	v33 L:0.380 F:0.01 p:0.9212	v46 L:0.380 F:0.02 p:0.8784	v38 L:0.380 F:0.03 p:0.8682	v37 L : 0.380 F : 0.04 p : 0.8403	v58 L : 0.380 F : 0.04 p : 0.8394	
4	(1, 150)	v46 L : 0.380 F : 0.03 p : 0.8713	v46 L:0.380 F:0.03 p:0.8713	ν38 L:0.380 F:0.03 p:0.8687	ν58 L:0.380 F:0.04 p:0.8438	v37 L : 0.380 F : 0.04 p : 0.8431	v43 L:0.380 F:0.06 p:0.8118	

7 Conclusion -- Comparaison FORWARD - BACKWARD

En comparant les résultats fournis par les deux approches, nous observons que le sous-ensemble final diffère assez sensiblement. Nous recensons les variables dans le tableau suivant.

N°	Forward	Backward		
1	v4	v4		
2	v11	v12		
3	v15	v30		
4	v21	v31		
5	v36	v32		
6	v45	v36		
7	v47	v49		

Seules les variables V4 et V36 ont été sélectionnées dans les deux cas.

Bien souvent ce type de résultats laisse l'utilisateur perplexe. Mais quel est le bon sous-ensemble de variables alors ?

Il n'y a pas de réponse définitive en réalité. Il faut garder à l'esprit qu'il s'agit là de techniques numériques, avec ses avantages (rapidité entre autres) et ses inconvénients (calcul + tri : une différence de quelques millièmes est une différence entérinée). Nous disposons de scénarios de résultats. Il faut s'en servir avant tout pour mieux défricher les données, cerner les phénomènes sous-jacents. lci commence le vrai rôle du data scientist.