Subject

Coding categorical predictive attributes for logistic regression.

When we want to use predictive categorical attributes in a logistic regression or a linear discriminant analysis, we must recode them. The most used strategy is certainly dummy variables. The coding scheme is the following: we create a dummy variable for each category of the original attribute. If there are K categories, we build (K-1) dummy variables; the last category is deduced from the other variables.

In this tutorial, we display how to use the 0_1 _BINARIZE component in order to transform categorical predictive attributes for logistic regression.

Dataset

We use CATEGORICAL_HEART.XLS in this tutorial. We want to predict the values of COEUR based on several predictive attributes (SEXE,...,VAISSEAU). The dataset contains 270 examples.

Dummy variables for categorical predictive attributes

Diagram creation

The simplest way to create a diagram is to open the dataset in the EXCEL spreadsheet. If you have installed the TANAGRA.XLA add-in, a new menu is now available (this add-in is come from 1.4.11 version). We select the range of cells and activate the TANAGRA / EXECUTE TANAGRA menu.

۵	🗃 🖬 🛛	😂 🖪 🚏 🕺	• • •	😻 10 × 1	a 🖉 🍓	Σ <i>f</i> * 🙀	Execute Tan	agra 6	• 🤉 .	<u>ð</u> -		2
	G271	• =	presence				¥					_
	Α	B	С	D	E	F	G	Н	E	J	K	
1	sexe	type_douleu	r sucre	electro	angine	vaisseau	coeur					
2	masculin	D	A	C	non	D	presence					2005
3	reminin	C	A	с	non	A	absence					
4	masculin	В	A	A	non	A	presence					
5	faminin	P		A C	oui	D	absence					
7	manculin	D	~		non	A	absence					
8	maeculin	C	R	C	oui	R	bresence					
9	masculin	D	4	c	oui	B	presence					
10	masculin	D	4	c	non	c	presence			-		
11	feminin	D	2	c	non	D	presence					
12	masculin	D	A	A	non	A	absence					
13	masculin	D	A	C	oui	A	absence					
14	masculin	c	A	c	non	A	absence					
15	masculin	A	A	A	non	C	presence					
16	feminin	D	A	с	non	в	absence					
17	feminin	D	A	A	non	A	absence					
18	masculin	D	A	A	oui	с	presence					
19	masculin	D	в	С	oui	A	presence					
20	masculin	A	A	С	oui	A	absence					
21	masculin	A	A	A	oui	A	absence					
22	masculin	D	A	С	oui	С	presence					
23	masculin	В	A	C	non	A	absence					
21	maeculin	D. /	٨	. ^	non	٨	abganca					2

A dialog box appears. We check that the range selection is right and we validate.

Micr	osoft E			t.xls								
Eich	ier <u>E</u> dit	ion <u>A</u> ffichage <u>I</u> r	nsertion Fo	rma <u>t</u> <u>O</u> utils	Données I	Fe <u>n</u> être <u>?</u> T	anagra Sipina	3			-16	a ×
		5 D. 🖤 🐰	h (2)	S 10 -	a - 18	Σ f* 💽		100%	- 2.	<u>ð</u> , -		*
A	1	× =	presence		1.00	100				4		_
	A	В	C	D	E	F	G	Н	I.	J	K	-
1 sex	xe	type_douleur	sucre	electro	angine	vaisseau	coeur					
2 ma	sculin	D	A	С	non	D	presence					
3 fem	ninin	с	A	C	non	A	absence					
4 ma	sculin	В	A	A	non	A	presence					
5 ma	sculin	D	A	A	oui	в	absence					
6 fem	ninin	В	A	С	oui	в	absence					
7 ma	sculin	D	^	•	non		shoonco I			2		
8 ma	sculin	C Ex							×	-		
9 ma	sculin	D										
10 ma	sculin	D	Datase	t range (inclu	iding the nam	e of the attrib	utes first ro	w):				
11 fem	ninin	D	\$A\$1	L:\$G\$271								
12 ma	sculin	D	1				4		3			
13 ma	sculin	D				ок		Cancel		-		
14 ma	sculin	с							_			
15 ma	sculin	Α	17.00	11.505								
16 fem	ninin	D	A	C	non	В	absence					
17 fem	ninin	D	A	A	non	A	absence					
18 ma	sculin	D	A	A	oui	С	presence					
19 ma	sculin	D	в	C	oui	A	presence					
20 ma	sculin	A	A	С	oui	A	absence				-	
21 ma	sculin	A	A	A	oui	A	absence					
22 ma	sculin	D	A	C	oui	C	presence				-	
23 ma	sculin	В	A	С	non	A	absence					
24 ima 4 ↓ ▶) H\he	art /	•	•	.non	•	absence '					
De <u>s</u> sin •	- 🗟 (Formes autor	matiques 🕶	× × □	0 🖉 🗸	1 😰 🔌	· _ · A	• = = ;	🗄 🗖 🍘	•		
Pointer										NUM		-

TANAGRA is automatically executed. We check that 270 observations and 7 variables are available.

🌋 TANAGRA 1.4.16 - [Da	itaset (tan85.txt)]			
🏆 File Diagram Componen	t Window Help			- 8 ×
D 📽 🖬 🔛				
	Analysis			~
			Dataset (tan85.txt)	
		Database : CODS	Parameters	
		Database . C. Do	Come - rimaison (Cockes- ritemp (Carlos, Coc	
			Results	=
		Download	l information	
		Datasource pro	ocessing	
		Computation tin	ne Orns	_
		Allocated memor	v 10 KB	
		Dataset d	escription	
		7 attribute(s)		
		270 example(s)	<u> </u>	~
]		<u> </u>
Data visualization	Statistics	Components Nonnarametric statistics	Instance selection	
Feature construction	Feature selection	Regression	Factorial analysis	
	Chatadaa	Cerriterenter		
PL5	clustering	spy tearning	weta-spy tearning	
5pv learning assessment	scoring	Association		
Correlation scatterplot	Export dataset	🌌 Scatterplot	🖉 Scatterplot with label 🛛 😫	View datase
<				>

Dividing the dataset into learning and test set

In order to obtain an honest error rate estimate, we must evaluate the classifier on a test set i.e. a dataset that is not used in the learning phase. So we divide the dataset in two parts using a random sampling: 200 examples for the training phase, and 70 examples for the test phase.

We insert the SAMPLING component (INSTANCE SELECTION tab) into the diagram. We activate the PARAMETERS menu and we select 200 examples.

TANAGRA 1.4.16 - [Dataset (tan85.t	xt)]	_ 🗆 🗙
Tile Diagram Component Window Help		_ 8 ×
🗅 🛩 🖬 🗱		
Analysis		~
🖃 🥅 Dataset (tan85.txt)	Sampling parameter	
Sampling 1		
	Parameters	
\hat{i}	Sample size definition	8
1		
	O proportion size	
1	Dabashuta siza	
i		
	OK Cancel Help	
N. State Sta	7 attribute(s)	
	270 example(s)	~
	Components	
Data visualization Statis	tics Nonparametric statistics Instance selection	
Feature construction - Feature se	election Regression Factorial analysis -	
PLS Ctuste	ring Spv learning Meta-spv learning	
Spv learning assessment Score	ing Association	
🖉 Continuous select examples 🖄 Recov	er examples	
📝 Discrete select examples 🛛 🥂 Rule-b	ased selection 🛛 📝 Stratified sampling	

Coding categorical attributes

The 0_1_BINARIZE component enables to transform categorical attributes into dummy (binary) variables. By default, (K-1) attributes are created from a categorical variable with K values. But we can modify the parameters settings for some statistical methods.

We insert DEFINE STATUS component into the diagram, the simplest way is to use the shortcut in the toolbar. We set as INPUT the 6 categorical predictive variables.



We add the 0_1 BINARIZE component into the diagram; we activate the VIEW menu in order to visualize the results.

TANAGRA 1.4.16 - [0_1_Binarize 1]		
Analusis	0 1 Binarize 1	
Dataset (tap85 tyt)	Parameters	
□ → Sampling 1	Used values : K-1 (ignore last value)	
🖻 🎦 Define status 1	Results	
	Attribute bioprination	
N N	seve (seve_mascumr)	deux P. 43
N N	cype_dodiedr (cype_dodiedr_0_1,cype_dodiedr_0_1,cype_dod	lieur_b_1)
N N	sucre (sucre_A_1)	
	electro (electro_C_1,electro_A_1)	
	angine (angine_non_1)	
i i	vaisseau (vaisseau_D_1,vaisseau_A_1,vaisseau_B_1)	
I		¥
N N	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	
[🕂 🗓 🛄 🛄 U_1_Binarize) 📶 EqFreq Disc 🛛 📓 Formula	🙊 RBF 🛛 🖏 Rnd Proj 🔶 Trend	
ABinary binning ABeqWidth Disc :: MDLPC	🚓 Residual Scores 🛛 🎏 Standardize	

For each categorical attribute, a set of dummy variables is created.

Logistic regression

These dummy variables can be used in a logistic regression now. We add again a DEFINE STATUS component into the diagram. We set as INPUT these binary variables, and we set as TARGET the class attribute (COEUR).



Then we add the BINARY LOGISTIC REGRESSION (SPV LEARNING tab) into the diagram. We activate the contextual VIEW menu.



The resubstitution error rate is 19.0%. For a 0.05 significance level, only 4 binary attributes are significant. We use a very rough variable selection strategy. We simply remove all the variables that are not significant. Perhaps, more sophisticated approaches give better results but the goal is to display the utilization of the variable transformation component in this tutorial.

Adjustement quality

Predicted attribute	coeur
Number of examples	200
-2 Log Likelihood	159.4172
Chi-2	115,4167
P(>Chi-2)	0.0000

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	1.216147	-	-	-
sexe_masculin_1	1.812569	0.4704	14.8506	0.0001
type_douleur_D_1	2,452703	0.8692	7.9628	0.0048
type_douleur_C_1	1.057939	0.8852	1.4283	0.2320
type_douleur_B_1	0.663332	0.9715	0.4662	0.4947
sucre_A_1	0.384675	0.5761	0.4459	0.5043
electro_C_1	-1.776383	1.8446	0.9274	0.3355
electro_A_1	-2.006760	1.8413	1.1878	0.2758
angine_non_1	-1.389320	0.4613	9.0719	0.0026
vaisseau_D_1	-0.174881	1.1084	0.0249	0.8746
vaisseau_A_1	-2.661483	0.7256	13.4557	0.0002
vaisseau_B_1	-0.902098	0.7982	1.2771	0.2584

To do that, we select the DEFINE STATUS 2 component and activate the parameters menu. We clear the current selection and select only the 4 significant predictive attributes.

TANAGRA 1.4.16 - [0_1_Binarize 1]			(_ 🗆 🛛
Tile Diagram Component Window Help				- 8 ×
Analysis	A Distance A			
🖃 🥅 Dataset (tan85.txt)		Parameters		
🖮 🌶 Sampling 1	Use Define attribute st	atuses		
🖮 🚰 Define status 1				
⊡ <mark>}</mark> 0_1_Binarize 1	Parameters			
Define status ?	Attributes :	Target	Input Illustrative	
Supervised ogi sti		sexe_mascul	in_1	
Execute View	sex D sucre	angine non		
	tyr D electro	alsseau_A_1		
	Suc D coeur	•		
	ele C sexe_maso	ulin_1 ur_D_1		
	and C type_doule	ur_C_1 ur B 1		
	C sucre_A_1	1		
	E E	Cle	ar selection	
	Con			
	Cre	OK	Cancel Help	
-	Components			
Data visualization Statistics Nonpa	arametric statistics	Instance selection	Feature construction	
Feature selection Regression Fa	ctorial analysis	PLS	Clustering	
Spv learning Meta-spv learning Spv le	arning assessment	Scoring	Association	
1/10% Binary logistic regression	🞇 C-PLS	A .c-	RT	<mark>i</mark> ≰c-s
				>
				.:

We activate the VIEW menu of the logistic regression. We obtain the following results.

Classifier performances

Error rate				0.185			
Values prediction				Confusion matrix			
Value	Recall	1-Precision	on presence absence Su				
presence	0.7416	0.175	presence	66	23	89	
absence	0.8739	0.1917	absence	14	97	111	
			Sum	80	120	200	

Classifier characteristics

Data description

Target attribute	values)
# descriptors	4

Adjustement quality

Predicted attribute	coeur
Number of examples	200
-2 Log Likelihood	164.1061
Chi-2	110.7279
P(>Chi-2)	0

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-0.127911	-	-	-
sexe_masculin_1	1.659938	0.4354	14.5359	0.0001
type_douleur_D_1	1.772698	0.4165	18.1156	0
angine_non_1	-1.317853	0.4465	8.7096	0.0032
vaisseau_A_1	-2.024704	0.4146	23.8492	0

The resubstitution error rate is now 18.5%. All the predictive attributes are significant for a 5% significance level.

We want now to compute the error rate on the test set. We insert the DEFINE STATUS component. We set as TARGET variable the class attribute (COEUR); the INPUT variable is PRED_SPV_INSTANCE_1. It is automatically generated by the supervised learning component.



We use the TEST component in order to compute the error rate on the test set. We note that the "true" error rate is 20%.

1.4.16 - [Test 1]									
🕎 File Diagram Component Window Help 💶 É									Ξ×
Analysis	Test 1							_	
🖃 🎹 Dataset (tan85.txt)	Parameters								
Evaluation set : unselected examples									
🖢 🎇 Define status 1									
e lo 1 D Results									
🖮 🚰 Define status 2	pred SpyInstance 1								
Supervised Learning 1 (Binary logistic regressive)									
🖻 🚰 Define status 3	Error rate				0.2000				_
Test 1		values prediction			Confusion matrix				=
1	Val	ue	Recall	1-Precision		presence	absence	Sum	
;	pres	ence	0.8065	0.2424	presence	25	6	31	
1	abse	nce).7949	0.1622	absence	8	31	39	
į.					Sum	33	37	70	
	Computation time : 0 ms.								
Created at 04/04/2007 10:19:30									~
Components									
Data visualization Statistics Nonparametric	statistics Instance selection				- Fe	Feature construction			
Feature selection Regression Factorial an	nalvsis PLS					Clustering			
Sou learning					Association				
Beller in the second se						HSSOCIACI			
Pr Blas-variance decomposition Pr Bootstrap					→ ¹ / ¹ / ¹ / ₂ ¹ / ¹ / ₂ ¹ /2				

Conclusion

Predictive dummy variables can be used for logistic regression or linear discriminant analysis. In the literature, one advises to use logistic regression instead of discriminant analysis because the Gaussian assumption of this last approach seems not suited for binary variables. In practice, on the majority of the situations, we do not really observe a different behavior of these two supervised learning methods on predictive binary variables.

In our example, the linear discriminant analysis applied to the same variables gives the same performances: the test error rate is 20%.