1 Topic

Feature selection for categorical predictor variables in logistic regression.

The aim of the logistic regression is to build a model for predicting a binary target attribute from a set of explanatory variables (predictors, independent variables), which are numeric or categorical. They are treated as such when they are numeric. We must recode them when they are categorical. The dummy coding is undeniably the most popular approach in this context.

We take a simple example to clarify ideas. We have a variable X with 3 levels \{A, B, C\}. We create 2 dichotomous variables: \(X_A\) which takes the value 1 when \(X = A\), 0 otherwise; \(X_B\) which takes the value 1 when \(X = B\), 0 otherwise. We identify the case \(X = C\) when both \(X_A\) and \(X_B\) take the value 0. “C” is called the reference group. The choice of the reference group influences the interpretation of the regression coefficients. But it has no influence of the performance of the model i.e. regardless of the chosen reference group, the error rate of the model - for instance - is not modified.

The situation becomes more complicated when we perform a feature selection. The idea is to determine the predictors that contribute significantly to the explanation of the target attribute. There is no problem when we consider a numeric variable. It is either excluded or either kept in the model. But how to proceed when we handle a categorical explanatory variable? Should we treat the dichotomous variables associated to a categorical predictor as a whole that we must exclude or include into the model? Or should we treat the each dichotomous variable independently? How to interpret the coefficients of the selected dichotomous variables in this case?

In this tutorial, we study the approaches proposed by various tools: R 3.1.2, SAS 9.3, Tanagra 1.4.50 and SPAD 8.0. We will see that feature selection algorithms rely on specific criteria according to the software. We will see also that they use different approaches when we are in the presence of the categorical predictor variables.

2 Dataset

We use the “heart-disease” dataset from the UCI Repository\(^1\). We transform the target attribute NUM into a binary variable with two levels {absence, presence}.

The predictors are numeric or categorical. Here are their descriptions:

\(^1\) [https://archive.ics.uci.edu/ml/datasets/Heart+Disease](https://archive.ics.uci.edu/ml/datasets/Heart+Disease)
We can observe the levels of each categorical variable.

3 Logistic regression and variable selection in R

We import the « heart-c.xlsx » data file using the “xlsx” package.

```r
#clear the memory
rmlist=ls())
#modify the default directory
setwd("...")
#load the data file
library(xlsx)
heart<-read.xlsx(file="heart-c.xlsx",sheetIndex=1,header=T)
print(summary(heart))
```

We obtain the following summary. “Disease” is the target attribute.
### 3.1 Regression with all the predictors

We perform the logistic regression with the `glm()` procedure (“stats” package). The tool can handle both numeric and categorical (R data type `factor`) predictors.

```r
# regression disease vs. all the other variables of the data frame
lr.all <- glm(disease ~ ., data = heart, family = binomial)
# display the results
print(summary(lr.all))
```

We get the following classifier, the Akaike criterion equals AIC = 230.42

![Regression Results](image)

**Figure 2 - Classifier with all the candidate variables - glm() under R**

We observe that the categorical predictors are automatically coded. R chooses the first level (in the alphabetic order) as the reference group. Thus, for the categorical variables of our study, the reference groups are: `SEX = FEMALE`, `CP = ASYMPT`, `FBS = F`, `RESTECG = LEFT_VENT_HYPER`, `EXANG = NO`, `SLOPE = DOWN`, `THAL = FIXED_DEFECT`.

This choice is important for the interpretation of the results. For instance, we see that the coefficient for the dummy variable (`SEX = MALE`) is significant at the 1% level i.e. the coefficient is significantly different from 0. We observe also that its sign is positive. It means that, compared to the women (reference group), the men have more chance to be sick (disease = presence). To the contrary, if we have defined `SEX = MALE` as the reference group,
we would get the same coefficient in absolute value, but with the opposite sign. It means that, compared to the men, the women have less chance to be sick.

The choice of mode of reference has no influence on the prediction of the classifier. Whatever the reference chosen for each categorical predictor variable, we will get exactly the same prediction when we classify a new instance.

### 3.2 Backward selection - AIC criterion

In the backward elimination process, we start with all candidate variables. We test the deletion of each variable and measure the AIC criterion. We remove the variable which minimizes the AIC. We continue as long as the AIC decreases i.e. we stop when the deletion of the variable leads to an increase of the AIC criterion.

```r
#backward elimination - AIC criterion
library(MASS)
lr.back <- stepAIC(lr.all, direction = "backward")
print(summary(lr.back))
```

We obtain the following classifier:

```
Call:  

dlm(formula = disease ~ sex + cp + trestbps + thalach + exang + 
oldpeak + slope + ca + thal, family = binomial, data = heart)

Deviance Residuals:  

Min 1Q Median 3Q Max  
-2.7652 -0.4907 -0.1557 0.3275 2.7477

Coefficients:  

Estimate Std. Error z value Pr(>|z|)  
(Intercept) -3.02306 2.22560 -1.359 0.17408  
sexmale 1.54105 0.50238 3.068 0.00216 **  
cpangina -0.93653 0.55707 -1.681 0.09273 .  
cpangina -1.95688 0.48481 -4.039 5.36e-05 ***  
ctng -2.10448 0.65130 -3.231 0.00123 **  
trestbps 0.01857 0.01034 1.796 0.07253 .  
thalach -0.01471 0.01006 -1.465 0.14295 
extah -0.79254 0.43179 1.835 0.06843 .  
onpeak 0.43673 0.22331 1.957 0.05032 .  
slope 0.77596 0.83198 0.933 0.35099  
slope 0.34071 0.90105 -0.478 0.63265 
ca 1.24240 0.25821 4.812 1.50e-06 ***  
ca 0.28579 0.76239 0.375 0.70776  
ca 1.59118 0.74694 2.130 0.03135 *  
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 417.64 on 302 degrees of freedom  
Residual deviance: 195.90 on 289 degrees of freedom  
AIC: 223.9

Number of Fisher Scoring iterations: 6
```

**Figure 3 - Classifier after the backward selection - stepAIC() under R**

Compared with the regression incorporating all the variables (AIC = 230.42, Figure 2), the new classifier is better with a lowest AIC (AIC = 223.9, Figure 3). We note especially, and this
is what interests us in this tutorial, that the dichotomous variables coming from the
categorical predictors with more than 2 levels are treated as a whole: either they are all
excluded (CP, RESTEG), or retained (CP, SLOPE, THAL).

![Figure 4 - First step of the backward elimination process - StepAIC() under R](image)

This behavior appears clearly when we inspect the details of the calculations. In the first step
(Figure 4), stepAIC seeks to exclude the worst variable: RESTECG minimizes the AIC criterion
(227.94). The two corresponding dichotomous variables are excluded at the same time as
evidenced by the degree of freedom (DF - "degree of freedom" column, DF = 2 to RESTECG)
in the table showing the intermediate results.

**Conclusion:** This behavior seems consistent with the initial specifications. We attempt to
determine the best subset of variables for the regression. When a variable is expressed
through a set of indicators, we cannot ungroup them.

We will see below that this reasoning can be called into question because, beyond the
selection of variables, we are looking to build the most performing model. The possibility to
deal individually with the indicators of categorical variables provides an additional freedom
in searching for solutions, and leads to perhaps more efficient models. The challenge is to
correctly interpret the classifier obtained.

## 4 Regression and variable selection in SAS

We analyze the variable selection in SAS. We go the heart of the matter. For a deeper
description of the PROC LOGISTIC, please refer to a previous article ("Introduction to SAS
proc logistic", July 2012).
4.1 Regression with all the variables

The HEART dataset is stored in the DATAREG bank that we created. We set the following commands to perform a logistic regression.

```sas
proc logistic data = datareg.heart;
class sex cp fbs restecg exang slope thal / param = reference ref = first;
model disease (event = last) = age sex cp trestbps chol fbs restecg thalach
exang oldpeak slope ca thal;
run;
```

The parameter `CLASS` indicates the categorical predictors. `PARAM` specifies the type of encoding to use and `REF` the reference group.

<table>
<thead>
<tr>
<th>Statistiques d’ajustement du modèle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critère</td>
</tr>
<tr>
<td>AIC</td>
</tr>
<tr>
<td>SC</td>
</tr>
<tr>
<td>Z Log</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimations par l’analyse du maximum de vraisemblance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paramètre</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>age</td>
</tr>
<tr>
<td>sex male</td>
</tr>
<tr>
<td>cp atyp_engina</td>
</tr>
<tr>
<td>cp non_angular</td>
</tr>
<tr>
<td>cp typ_engineia</td>
</tr>
<tr>
<td>trestbps</td>
</tr>
<tr>
<td>chol</td>
</tr>
<tr>
<td>fbs t</td>
</tr>
<tr>
<td>restecg normal</td>
</tr>
<tr>
<td>restecg st_t_wave_abnor</td>
</tr>
<tr>
<td>thalach</td>
</tr>
<tr>
<td>exang yes</td>
</tr>
<tr>
<td>oldpeak</td>
</tr>
<tr>
<td>slope flat</td>
</tr>
<tr>
<td>slope up</td>
</tr>
<tr>
<td>ca</td>
</tr>
<tr>
<td>thal normal</td>
</tr>
<tr>
<td>thal reversible_defect</td>
</tr>
</tbody>
</table>

**Figure 5 - Logistic regression in SAS**

In our example, `PARAM = REFERENCE` recommends the dummy coding; with `REF = FIRST`, we specify that the first value (in alphabetical order) will correspond to the reference group.
Thus we have the default settings of R (section 3.1). The obtained classifiers are identical (Figure 5, AIC = 230.418). But SAS does not forget that some variables are indicators coming from a categorical predictor. It proposes a table where it checks the relevance of the variables (Figure 6).

![Figure 6 - Type 3 - Analysis of effects - SAS](image)

When the predictor variable is numeric or binary, the degrees of freedom, the test statistic and p-value are identical to those of the previous table (Figure 5). By contrast, when the variable is categorical with more than 2 levels, SAS performs a test for the simultaneous nullity of all the coefficients associated to its dichotomous variables.

Let us study the variable THAL at the 1% level. Apparently, in comparison with the reference group THAL = FIXED DEFECT, neither THAL = NORMAL, nor THAL = REVERSIBLE DEFECT, individually seems not to lead a significant increase or decrease of the chance to be sick (Figure 5). But when we perform the global test for the variable i.e. we test if the coefficients are simultaneously zero, we reject the null hypothesis (Figure 6, p-value = 0.0036).

In fact, we observe that treating the indicators individually or as a whole related to categorical variable correspond to different behaviors.

### 4.2 Variable selection in SAS

The backward selection is based on Wald test in SAS. Like R, the indicators coming from a categorical variable are processed as a group. The calculations are based on the type 3 analysis of effects (Figure 6) table. At each step SAS removes the less relevant variable.
The variable selection process works as follows: (1) SAS detects the least relevant variable, one that provides the highest p-value; (2) the variable is removed from the model if the p-value is greater than a threshold specified by the SLSTAY parameter. (3) SAS fits again the model on the dataset with the remaining variables, and then performs the steps (1) and (2) until all of the variables are relevant.

For our dataset, according to the table above (Figure 6), it would be AGE that would be eliminated at the first iteration with a p-value equal to 0.5809, superior to the parameter SLSTAY = 0.01. Here is the SAS code for the backward selection using the "proc logistic".

```
proc logistic data = datareg.heart;
class sex cp fbs restecg exang slope thal / param = reference ref = first;
model disease (event = last) = age sex cp trestbps chol fbs restecg thalach
exang oldpeak slope ca thal / selection = backward slstay = 0.01;
run;
```

SAS describes in a table the variables removed at each step (Figure 7).

Finally, we obtain a classifier with 5 predictor variables: SEX, CP, SLOPE, CA and THAL. All the variables are statistically relevant at the 1% level (Figure 8). It cannot be otherwise. This is the consequence of this type of selection process.

```
class sex cp fbs restecg exang slope thal / param = reference ref = first;
```

```
model disease (event = last) = age sex cp trestbps chol fbs restecg thalach
exang oldpeak slope ca thal / selection = backward slstay = 0.01;
```

```
run;
```
But some indicators may be non-significant (at the same significance level) into the table of coefficients (e.g. the indicators of SLOPE and THAL) (Figure 9). This is not inconsistent. SAS uses the Type 3 Analysis of effects to detect the variables to remove. We have seen above that an indicator may appear non-significant while the corresponding variable is significant (Figure 5 and Figure 6).

Note: About the AIC criterion, that of SAS is higher than that of R (323.167 vs. 223.9; the higher is the AIC, the worst is the model). This is not surprising. SAS does not use explicitly this criterion in the detection of the best model.

5 Regression in Tanagra

With Tanagra, we must recode explicitly the categorical predictors before performing a logistic regression\(^2\). There are several reasons for this: It seemed to me pedagogically desirable that students perform explicitly the coding for understanding the necessity and the influence of this step in regression; the user can introduce further analysis by using other types of coding (see the CLASS parameter into SAS); Tanagra relies on a specific strategy in variable selection process, it deals with the indicators as independent variables, the resulting classifier is of a different nature compared to previous approaches.

5.1 Regression with all the variables

5.1.1 Import the dataset

We load the file “heart-c.xlsx” into the Excel. We select the data range that we send to Tanagra using the `tanagra.xla` add-in.

---

5.1.2 Dummy coding of categorical predictor variables

Tanagra is automatically launched. We have 14 variables and 303 instances. In order to code the categorical predictor variables into indicators, we must first specify the variables to recode using the DEFINE STATUS component available in the toolbar.

We set as INPUT all the categorical variables except DISEASE which is the target attribute. Then we insert the 0_1_BINARIZE (tab FEATURE CONSTRUCTION) component into the diagram. It performs the coding process. We click on the VIEW pop-up menu to obtain the

---

results. We can visualize the set of indicators for each variable. The omitted value corresponds to the reference group.

5.1.3 Logistic regression
We insert again the DEFINE STATUS component to define the target attribute (TARGET = DISEASE) and the input ones (INPUT = all the numeric variables + the indicators from the categorical predictors). Then we add the BINARY LOGISTIC REGRESSION (SPV LEARNING tab) component. We click on the VIEW pop-up menu to get the results of the regression.

![Binary Logistic Regression in Tanagra](image)

We have exactly the same model that R and SAS as attested by the deviance (-2LL) i.e. applied on the same individuals, these classifiers will produce the same estimated class membership posterior probabilities and, a fortiori, the same predicted values. But, because the reference groups are not the same for the categorical predictors, the set of estimated coefficients are different.

Let us compare the results of R (Figure 2) and Tanagra (Figure 11) for some categorical predictors:

- For the variable SEX, FEMALE is the reference group for the two tools R and Tanagra, the regression coefficient is the same 1.658410.
- For FS, R has selected the "F" modality as reference, Tanagra used "T", the coefficients are not the same, respectively -0.348979 and 0.348979. Because FS is a binary variable, the coefficients have the same absolute value, but they are of opposite sign.
- For SLOPE which is not binary (SLOPE has 3 levels), R chooses "DOWN" while Tanagra used "UP" as reference group. The coefficients of R are \( a_{\text{FLAT/DOWN}} = 0.705005 \) and \( a_{\text{UP/DOWN}} = -0.465609 \); the coefficients for Tanagra are \( a_{\text{FLAT/UP}} = 1.170614 \) and \( a_{\text{DOWN/UP}} = 0.465609 \). We note that it is very easy to find the coefficients of R using those of Tanagra, indeed:

\[
\begin{align*}
    a_{\text{UP/DOWN}} &= -a_{\text{DOWN/UP}} \\
    a_{\text{FLAT/DOWN}} &= a_{\text{FLAT/UP}} - a_{\text{DOWN/UP}}
\end{align*}
\]
Consequently, regardless of the reference group used in the coding of the dummy variables, the behavior of the classifier and the interpretations of the coefficients are preserved.

![Model Fit Statistics](image)

**Attributes in the equation**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Coef.</th>
<th>Std-dev</th>
<th>Wald</th>
<th>Signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>-4.918515</td>
<td>3.9539</td>
<td>1.5474</td>
<td>0.2135</td>
</tr>
<tr>
<td>age</td>
<td>-0.013652</td>
<td>0.0247</td>
<td>0.3048</td>
<td>0.5809</td>
</tr>
<tr>
<td>trestbps</td>
<td>0.018961</td>
<td>0.0112</td>
<td>2.8591</td>
<td>0.0909</td>
</tr>
<tr>
<td>chol</td>
<td>0.004058</td>
<td>0.0041</td>
<td>1.0006</td>
<td>0.3171</td>
</tr>
<tr>
<td>thalach</td>
<td>-0.017445</td>
<td>0.0110</td>
<td>2.5244</td>
<td>0.1121</td>
</tr>
<tr>
<td>oldpeak</td>
<td>0.408770</td>
<td>0.2309</td>
<td>1.3414</td>
<td>0.0767</td>
</tr>
<tr>
<td>ca</td>
<td>1.288356</td>
<td>0.2769</td>
<td>21.6498</td>
<td>0.0000</td>
</tr>
<tr>
<td>sex_male_1</td>
<td>1.658410</td>
<td>0.6398</td>
<td>9.4404</td>
<td>0.0021</td>
</tr>
<tr>
<td>cp_asympt_1</td>
<td>2.085757</td>
<td>0.6603</td>
<td>9.6934</td>
<td>0.0018</td>
</tr>
<tr>
<td>cp_non_angular_1</td>
<td>0.169367</td>
<td>0.6563</td>
<td>0.0668</td>
<td>0.7961</td>
</tr>
<tr>
<td>cp_atype_angular_1</td>
<td>1.132273</td>
<td>0.7443</td>
<td>2.1986</td>
<td>0.1381</td>
</tr>
<tr>
<td>fbs_f_1</td>
<td>0.248979</td>
<td>0.5818</td>
<td>0.3598</td>
<td>0.5486</td>
</tr>
<tr>
<td>restecg_left_vent_hyper_1</td>
<td>-0.521230</td>
<td>2.5072</td>
<td>0.0432</td>
<td>0.8353</td>
</tr>
<tr>
<td>restecg_normal_1</td>
<td>-0.977185</td>
<td>2.5054</td>
<td>0.1521</td>
<td>0.6965</td>
</tr>
<tr>
<td>exng_Yes_1</td>
<td>0.821227</td>
<td>0.4396</td>
<td>3.4894</td>
<td>0.0618</td>
</tr>
<tr>
<td>slope_fat_1</td>
<td>1.706164</td>
<td>0.4711</td>
<td>6.1742</td>
<td>0.0130</td>
</tr>
<tr>
<td>slope_down_1</td>
<td>0.465609</td>
<td>0.9257</td>
<td>0.2530</td>
<td>0.6150</td>
</tr>
<tr>
<td>the_normal_1</td>
<td>0.135676</td>
<td>0.7817</td>
<td>0.0301</td>
<td>0.8622</td>
</tr>
<tr>
<td>the_reversible_defect_1</td>
<td>1.453132</td>
<td>0.7704</td>
<td>3.5581</td>
<td>0.0593</td>
</tr>
</tbody>
</table>

*Figure 11 - Regression with all the variables (numeric + indicators) - Tanagra*

### 5.2 Backward selection in Tanagra

Tanagra has really a different behavior in the variable selection process. Indeed, the tool for the variable selection deals with the indicators without taking account of their relationship with the original variables. We add the BACKWARD-LOGIT component (FEATURE SELECTION) into the diagram. We insert the tool after the DEFINE STATUS 2 where we specified the candidate explanatory variables for the regression. The default level of significance is 1%.
We observe that 6 numeric or indicator variables are selected. We observe also that some indicators coming from the same variable are treated separately (e.g. CP, SLOPE, THAL).

We perform the regression on the selected variables using the BINARY LOGISTIC REGRESSION component.
The AIC of the classifier equals 22.978 (not visible into the screenshot). Unlike SAS (Figure 9), all numeric and indicator variables are significant at the 1% level (Figure 13). It does not mean that this solution is better. This is simply the consequence of the selection strategy used by the tool that processes the indicator variables individually.

**Note:** Does this model make sense? The answer is yes. The missing levels become the reference group. We consider the variable SLOPE to support our argument. We have seen in the regression involving all indicators that coefficient $a_{\text{DOWN/UP}} = 0.465609$ was not significant with a $p$-value equals 0.6150 (Figure 11). By removing the DOWN indicator variable, the system tells us that it makes sense to merge the terms DOWN and UP, as decision trees algorithms that perform merges of leaves during splitting process (e.g. CART or CHAID). Thus, the coefficient $a_{\text{FLAT/\{UP,DOWN\}}} = 1.246519$ indicates the increase in chance to be sick associated to FLAT compared to the union of \{UP, DOWN\} levels.

### 6 Regression in SPAD - Comparison of the approaches

SPAD is a software package which provides tools for the whole data mining process\(^4\). We have study it in some circumstances previously (e.g. *interactive decision tree learning*). It works just like the other popular data mining / data science tools such as IBM SPSS Modeler or SAS Enterprise Miner.

SPAD incorporates the two approaches highlighted above. It can achieve the backward selection based on the Wald test by processing as a group the indicators associated to categorical predictors, like SAS; it can also treat them individually, like Tanagra. In this section, we will see the settings to be specified to get the behavior that we want.

#### 6.1 Logistic regression in SPAD

We must create a new project to start an analysis process. We use the FEUILLE EXCEL to import the data file. The operation is simple. We must nevertheless ensure that the CA column is encoded as real type. We set this in the METADONNEES tab of the dialog settings of the tool.

Then, we insert the REGRESSION LOGISTIQUE component (branch SCORING and MODELISATION), to which we link the Excel source. We click on the PARAMETRES menu to access the settings dialog box.

Into the tab “MODELE”, we define the target attribute (DISEASE) and the predictors.

Into the tab “PARAMETRES”, we specify that we include all the candidate predictors into the model (METHODE DE SELECTION = PAS DE SELECTION DE VARIABLES) (Figure 16).
We click on the OK button. The calculations are automatically launched. We can visualize the results by clicking on the Excel icon in the "Resultats" section (Figure 17).

The outputs are splitted in several sheets. Into "REG_MODEL", we can visualize the overall evaluation of the model, and the table with the estimated coefficients of the classifier (Figure 18).
Figure 18 - Description of the classifier - SPAD

SPAD mixes the table of coefficients and the type 3 analysis of effects. Thus, we can identify at a glance the relevant indicators and the relevant variables.

For instance, at the 1% level, the two indicators from THAL are not relevant (THAL = NORMAL, p-value = 0.862; THAL = REVERSIBLE_DEFECT, p-value = 0.059). But we cannot remove them from the model because we cannot consider that they are simultaneously equal to zero (THAL, p-value = 0.004). It may seem confusing as I said above. But we must not forget that these two indicators are not independent.
6.2 Backward selection I

First, we perform the backward elimination (at 1% level) where we treat as a whole the indicators coming from the same categorical predictors.

We validate the settings by clicking on the button OK. Here are the details of the results (Figure 19):

- (A) The list of the variables removed at each step. This table is the same as the summary table provided by SAS (Figure 7).
- (B) The overall evaluation of the model. We observe among others that the AIC = 232.167.
- (C) The likelihood ratio test which assesses the global significance of the model.
- (D) The coefficients of the model, including the Type 3 Analysis of Effects.
The results are identical in all respects to those of SAS.

6.3 Backward selection II

In this section, we want to get the same behavior as Tanagra. We open again the settings dialog box. We specify that we treat individually the indicator variables i.e. in French: "Autoriser la sélection individuelle des modalités d’une variable" (red arrow).
The selection table traces the elimination of the variables (numeric variables or indicator variables) (Figure 20).

**Sélection des variables du modèle**

<table>
<thead>
<tr>
<th>Ordre</th>
<th>Variable</th>
<th>Modalité</th>
<th>Direction</th>
<th>Khi2</th>
<th>P-valeur</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>thal</td>
<td>normal</td>
<td>sortie</td>
<td>0.030</td>
<td>0.862</td>
</tr>
<tr>
<td>1</td>
<td>restecg</td>
<td>st_t_wave_abnormality</td>
<td>sortie</td>
<td>0.045</td>
<td>0.832</td>
</tr>
<tr>
<td>2</td>
<td>slope</td>
<td>up</td>
<td>sortie</td>
<td>0.229</td>
<td>0.633</td>
</tr>
<tr>
<td>3</td>
<td>age</td>
<td></td>
<td>sortie</td>
<td>0.302</td>
<td>0.583</td>
</tr>
<tr>
<td>4</td>
<td>fbs</td>
<td>t</td>
<td>sortie</td>
<td>0.345</td>
<td>0.557</td>
</tr>
<tr>
<td>5</td>
<td>chol</td>
<td></td>
<td>sortie</td>
<td>0.810</td>
<td>0.368</td>
</tr>
<tr>
<td>6</td>
<td>restecg</td>
<td>normal</td>
<td>sortie</td>
<td>2.049</td>
<td>0.152</td>
</tr>
<tr>
<td>7</td>
<td>thalach</td>
<td></td>
<td>sortie</td>
<td>2.195</td>
<td>0.138</td>
</tr>
<tr>
<td>8</td>
<td>trestbps</td>
<td></td>
<td>sortie</td>
<td>2.550</td>
<td>0.110</td>
</tr>
<tr>
<td>9</td>
<td>cp</td>
<td>atyp_angina</td>
<td>sortie</td>
<td>3.131</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Compared with Tanagra, as the references groups are not the same for categorical variables, the removed indicators may be different. Thus, at the end, we can get - and that is true for our dataset - different models despite identical settings (significance level for the Wald test).

**Conclusion:** In the case of variable selection context, the choice of the reference group has an influence on the characteristics of the final model when we process individually the indicator variables.

In SPAD, we get the following final classifier (Figure 21):
What is the best approach?

"What is the best approach?" is an inevitable question in the present study. Actually, there is not really a consensus about the best method in variable selection. The various approaches provide scenarios for solutions. All depends on what we do with the classifiers thereafter.

The only objective evaluation would be to measure the performance of the resulting model (after the feature selection process) on a test set. But we must have a test set with sufficient number of instances to get a reliable evaluation of the error rate.

Drawing a parallel with decision trees, we could say that: treating the indicators from a categorical variable as a whole is similar to generate branches for each possible value of the variable during the splitting process of a node; treating individually the indicators enables to merge certain values as with CHAID induction tree algorithm (or others).

Conclusion

In this tutorial, we have studied the approaches implemented in various data mining tools for dealing with categorical predictors in variable selection context. All are based on the dummy coding of the categorical variables, but the indicators coming from one variable may be treated as whole or individually in the selection process. The resulting classifier can be different following the approaches.