Support Vector Machines and kernel methods Big Data Management and Analytics

Exercise: Bank Marketing dataset models

Data cooking

Bank marketing dataset is being used to generate SVM models, data is loaded from file *bank-full.csv*, then cooked as in previous practice, with just *log10* most of times, scale and center, obtaining the following:

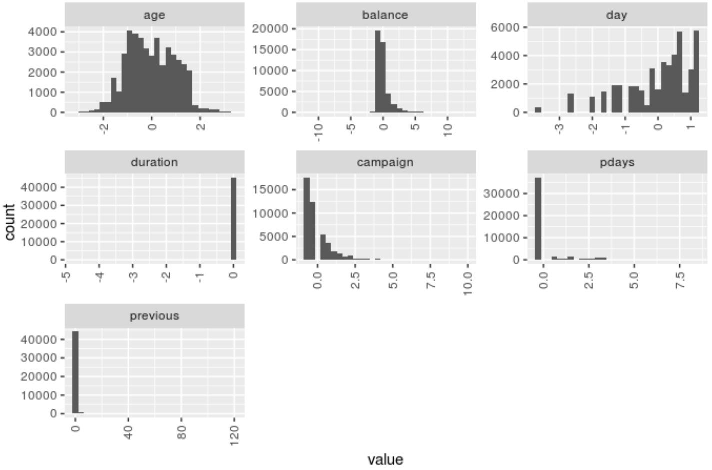

Figure 1 - continuous variables inspection, after log10 transform

R will create dummy variables from categorical automatically when fitting models:

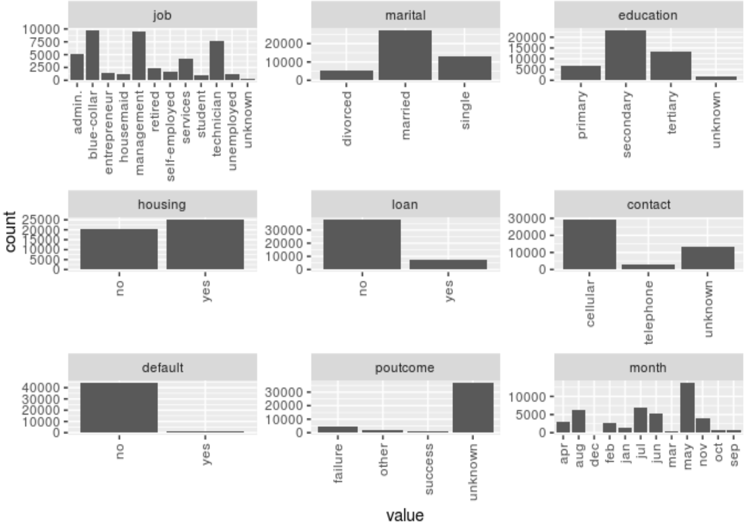

Figure 2 - categorical variables inspection

learn.data is set to 2N/3
test.data is set to N/3
with *N <- nrow(deposit)*

At this point we are ready to build our models.

SVM model1 (cost=1, gamma=1, epsilon=0.1)

A *linear* model is built, obtaining the following confusion matrix:

```
          Reference
Prediction   no   yes
   no     13125  1430
   yes      185   330          Accuracy : 0.8928
```

Obtaining better results than other models:

Algorithm	*Accuracy*
LDA	86.34%
LDA_CV	86.68%
RDA	85.90%
RDA_CV	85.75%
QDA	85.89%
Logistic regression	*Learning data: 9.55% ERROR* *Test data:* 9.78% ERROR
SVM *linear* (cost=1, gamma=1, epsilon=0.1)	***89.28%***

SVM model2 (cost=2, gamma=1, epsilon=0.1)

Let's try another model with cost=2, results are not better:

```
          Reference
Prediction   no   yes
   no     11870  1474
   yes     1440   286                     Accuracy : 0.8066
```

SVM model3 (cost=3, gamma=1, epsilon=0.1)

With *cost=3* results become fine:

```
           Reference
Prediction   no   yes
   no     13310  1760
   yes        0     0                Accuracy : 0.8832
```

However table does not look very healthy.

SVM model4 (cost=1, gamma=10, epsilon=25)

With the set parameters, model looks fine again:

```
          Reference
Prediction   no   yes
   no     13125  1430
   yes      185   330                Accuracy : 0.8928
```

SVM model5 (cost=1, gamma=10, epsilon=30)

A small variation in *epsilon* has no effects.

Tuning

From now on, the number of samples of *deposit* is reduced to *N=10000* for *svm* model creation not being much time consuming.

Linear

No matter what parameters are set, an accuracy of *0.9684968* is obtained with *kernel="linear"*.

Sigmoid

When *sigmoid* is used, changes in cost are reflected, the following tables are obtained:

cost	0.1	0.2	0.3	1	1.5	2
accuracy	0.9417942	0.9384938	0.9369937	0.9351935	0.9348935	0.9348935

Variations in *epsilon* or *gamma* do not affect accuracy.

Polynomial

Different models are created with different parameters, obtaining assorted accuracies, for example varying *gamma*:

gamma	0.1			
epsilon	0.1			
cost	0.1	1	3	5
degree				
1	**0.9684968**	**0.9684968**	**0.9684968**	**0.9684968**
2	**0.9684968**	**0.9684968**	0.9672967	0.9666967
3	0.9621962	0.9546955	0.949895	0.9450945
4	0.9291929	0.9168917	0.9135914	0.9135914
5	0.9048905	0.89979	0.8937894	0.8874887

gamma	1			
epsilon	0.1			
cost	0.1	1	3	5
degree				
1	0.969697	**0.9684968**	**0.9684968**	**0.9684968**
2	0.969697	**0.9684968**	0.9672967	0.9666967
3	0.9615962	0.9546955	0.949895	0.9450945
4	0.9372937	0.9168917	0.9135914	0.9135914
5	0.9159916	0.89979	0.8937894	0.8874887

gamma	1.5			
epsilon	0.1			
cost	0.1	1	3	5
degree				
1	**0.9684968**	**0.9684968**	**0.9684968**	**0.9684968**
2	**0.9684968**	**0.9684968**	0.9672967	0.9666967
3	0.9621962	0.9546955	0.949895	0.9450945
4	0.9291929	0.9168917	0.9135914	0.9135914
5	0.9048905	0.89979	0.8937894	0.8874887

gamma	2			
epsilon	0.1			
cost	0.1	1	3	5
degree				
1	**0.9684968**	**0.9684968**	**0.9684968**	**0.9684968**
2	**0.9684968**	**0.9684968**	0.9672967	0.9666967
3	0.9621962	0.9546955	0.949895	0.9450945
4	0.9291929	0.9168917	0.9135914	0.9135914
5	0.9048905	0.89979	0.8937894	0.8874887

Varying *epsilon* produces no significant changes:

gamma	1			
epsilon	**1**			
cost	0.1	1	3	5
degree				
1	0.9684968	0.9684968	0.9684968	0.9684968
2	0.9684968	0.9684968	0.9672967	0.9666967
3	0.9621962	0.9546955	0.949895	0.9450945
4	0.9291929	0.9168917	0.9135914	0.9135914
5	0.9048905	0.89979	0.8937894	0.8874887

gamma	1			
epsilon	**1.5**			
cost	0.1	1	3	5
degree				
1	0.9684968	0.9684968	0.9684968	0.9684968
2	0.9684968	0.9684968	0.9672967	0.9666967
3	0.9621962	0.9546955	0.949895	0.9450945
4	0.9291929	0.9168917	0.9135914	0.9135914
5	0.9048905	0.89979	0.8937894	0.8874887

gamma	2			
epsilon	**1.5**			
cost	0.1	1	3	5
degree				
1	0.9684968	0.9684968	0.9684968	0.9684968
2	0.9684968	0.9684968	0.9672967	0.9666967
3	0.9621962	0.9546955	0.949895	0.9450945
4	0.9291929	0.9168917	0.9135914	0.9135914
5	0.9048905	0.89979	0.8937894	0.8874887

Best found accuracy is *0.9684968*.

Radial

No big differences are observed but with *cost*.

gamma	1			
epsilon	1			
cost	0.1	1	3	5
accuracy	0.9684968	0.9684968	0.9678968	0.9618962

Conclusions

SVM *“linear”* model with *cost=1*, *gamma=1*, *epsilon=0.1* as our initial model, would be good enough, with *89.28%* accuracy; where model is built for full dataset.