Term Assignment

Vincent de Vos (0741795)
v.j.h.d.vos@student.tue.nl

Thom Hurks (0828691)
t.p.hurks@student.tue.nl

April 6, 2015

Questions:

1. The data set contain some missing values. Explain and motivate how you will deal with this issue.
2. The data set is imbalanced (only roughly 11% of customers in this dataset subscribed to this product). Explain and motivate how you will deal with this issue.
3. Some variables are categorical. Explain and motivate how you will deal with this issue.
4. Explain and motivate what you will do with the date of the last contact (2 variables: date and month).
5. Which variables you think are relevant for the classification task? Motivate your answer.
6. Is the data normalization required for this data set? Motivate your answer.
7. Prepare the data set, so that it is in the proper form for the neural network and fuzzy inference modeling. [Hint: Consider preparing 2 data sets, one for FIS and one for NN]
8. Explain and motivate how you will use the available data to create the model.
9. Explain and motivate how you will evaluate the performance of the models you will create.
10. Build the fuzzy inference system model. Which variables are you using? What parameters are you using? Explain, why you decided to choose those values. What is the quality of your model? [Hint: to keep the model smaller you use only one of the categorical variables that has 3 or more values (indicate which variable did you choose)]
11. Build a neural network model. Which variables are you using? What parameters are you using? Explain, why you decided to choose those values. What is the quality of your model?
12. Provide a comparison of the fuzzy set and neural network models. Which model would you recommend to the B-BANK? Motivate your answer.

Solutions

Question 1

We inspected the unique values present in each column of the input dataset to verify if there are missing values and if the present values are sane. We found that each column can contain values that indicate missing data, as shown in the following table:

Value	**Description**
NaN	For numerical columns, no number is present
<undefined>	For categorical columns, no entry is present
“unknown”	Actual value in the data indicating that the value is unknown
-1	In the case of “nr of days since..”, -1 must indicate a missing value

The variables in the dataset have the following number of missing values:

Variable	**NaN**	**“unknown”**	**<undefined>**	**-1**	**Total**
age	87	0	0	0	87
job	0	288	61	0	349
marital	0	0	72	0	72
education	36	0	0	0	36
credit_card	95	0	0	0	95
balance	89	0	0		89
mortgage	80	0	0	0	80
loan	132	0	0	0	132
contact	0	13020	0	0	13020
day	149	0	0	0	149
month	0	0	31	0	31
duration	68	0	0	0	68
campaign	56	0	0	0	56

pdays	76	0	0	36887	36963
previous	62	0	0	0	62
poutcome	0	36933	34	0	36967
target	0	0	0	0	0

We will deal with this issue by...how? We can remove entries (probably too harsh), replacing by a single “unknown” value or just leaving them as-is. They all have trade-offs, so we are probably expected to just pick one we like…

One “hybrid” method would be to delete entries if the number of missing values in the column is low (eg. as with job and loan). If a column has a large number of missing values (e.g. pdays, contact), replace by a single “missing” indicator value.

Question 2

Notice that we only have two classes: subscribed (11%) and unsubscribed (89%). A good approach would be to balance our dataset with undersampling. We will sample our balanced data set in such a way that we fully include all of our subscribed observations and let this make up for 30% of the dataset.

The remaining 70% will be containing unsubscribed data. However this means we won't be using the full 89% of unsubscribed data but only a subset. Therefore it is useful to ensure that this subset only contains usable observations. Therefore we will filter out some incomplete observations beforehand. (Depends on answer of question 1)

The 70/30 ratio was chosen in this case, because it is a tradeoff between two extremes; if we keep the 11/89 ratio then our data is too much skewed towards the unsubscribed outcome, and since we want to predict the subscribers that data shouldn’t be too scarce. If we would balance the dataset to a 50/50 ratio, then we would have to throw lots of data away which can cause us to miss certain patterns. So the 70/30 ratio is a good tradeoff between the two scenarios.

Once we have our balanced data set we can split up this data into a training set, on which we will train our models, and a test set in order to determine the performance. When creating training and validation sets we will always use stratification to ensure that all those sets are representative of the balanced dataset. For the final test set we will not use stratification, to make sure it has the same subscribed/unsubscribed ratio as the real data, because otherwise the final test performance would not be representative in reality.

Question 3

A variable that has categorical values can be converted to vectors of variables where each variable has the value 0 or 1. This approach should only be used with (rule of thumb) < 20 different possible categorical values. For example, the variable Colour {“Red”, “Green”, “Blue”} can be converted to the variables Red_0_1, Green_0_1 and Blue_0_1, where Red_0_1 has the value 1 and Green_0_1 and Blue_0_1 are 0 if Colour=”Red”.

With our number of different categorical values, this method is possible. No column has more than 20 different values, as can be seen in the following table:

Variable name	**Number of categorical values (excluding “unknown”)**
contact	2
job	11
marital	3
poutcome	3
target	2

Question 4

Date of last contact is problematic because it is stored in two variables; day and month. Since these two variables together form a date, they are obviously strongly tied together and should not be fed into a classifier in a separated way. Unfortunately, we cannot convert the two to a date variable, since classifiers cannot just handle dates. Besides that, we do not have a year variable. A solution would be to remove the month variable, and increase the day variable by the amount of days in the months that we removed. As a result we have one variable that tells us the number of days ago since last contact.

Optioneel (eerst de decision tree bekijken)

We can then discretize this variable by dividing it by 7 and rounding it up, which gives us a number then tells us roughly how many weeks have passed since last contact. This will reduce the number of possible values of this feature (from 365 initially, to 52) and probably yields better classification results because it is far more discriminative.

(Todo: look up more in slides)

Question 5
All variables are relevant for the classification task. If a variable is not, then the classification algorithm together with proper validation and testing will make sure that irrelevant variables get a very low or even 0 weight in the final classifier. Since we are not marketing or banking experts, it is not up to us to remove variables from the classification task beforehand based in our intuition; an intuition that could very well be wrong and hurt the classifier quality. *(Todo: slides mention something about this and using experts, I think. Update: misschien moeten we toch variabelen niet meenemen, maar ik heb geen idee hoe we dat zouden moeten kiezen. Misschien gewoon alle combinaties testen en kijken welke de beste classifier oplevert?)*

Question 6
Yes, data normalization is necessary, especially for neural networks. Not all variables have the same magnitude and not all classification algorithms can handle that. All variables should be normalized such that the magnitude of the variables is roughly the same. For fuzzy systems normalizations is not necessary per se. We can use two versions of the dataset: one normalized and another not normalized. *(Todo: slides have more info on this)*

Question 7
Work-in-progress: Do the data preparation for the NN and FIS model in MatLab.

Question 8
Todo: Explain and motivate how you will use the available data to create the model. (How will we feed it into the NN/FIX algorithms?)

For FIS we've chosen to use only specific columns that are suitable for modeling into a fuzzy system.

Question 9
Todo: Explain and motivate how you will evaluate the performance of the models you will create. (What performance measures, perhaps the same ones as Assignment 1? That is the most easy anyway.)

Questions 10 through 12 are about making the models in MatLab and comparing them.