Text Classification using Naive Bayes Classifier

Francesca Sallicati

21 febbraio 2018

Contents


```
## Loading required package: NLP
```

1 Introduction

The data used to apply the Naive Bayes Classifier come from a text dataset consisting of two variables:

• **TYPE**: the personality type result from the Meier-Brigg's test of the person. It is a factor between 16 levels.

TYPE	Description
ENFJ	Extroversion iNtuion Feeling Judgment
ENFP	Extroversion iNtuition Feeling Perception
ENTJ	Extroversion iNtuition Thinking Judgment
ENTP	Extroversion iNtuition Thinking Perception
ESFJ	Extroversion Sensing Feeling Judgment
ESFP	Extroversion Sensing Feeling Perception
ESTJ	Extroversion Sensing Thinking Judgment
ESTP	Extroversion Sensing Thinking Perception
INFJ	Introversion iNtuition Feeling Judgment
INFP	Introversion iNtuition Feeling Perception
INTJ	Introversion iNtuition Thinking Judgment
INTP	Introversion iNtuition Thinking Perception
ISFJ	Introversion Sensing Feeling Judgment
ISFP	Introversion Sensing Feeling Perception
ISTJ	Introversion Sensing Thinking Judgment
ISTP	Introversion Sensing Thinking Perception

• **POSTS**: a record of the last 50 posts online of the person

2 Naive Bayes Classification

2.1 16 Personality Types

The frequencies for the 16 types are:

```
##
## ENFJ ENFP ENTJ ENTP ESFJ ESFP ESTJ ESTP INFJ INFP INTJ INTP ISFJ ISFP ISTJ
##  190  675  231  685   42   48   39   89 1470 1832 1091 1304  166  271  205
## ISTP
##  337
```

Let's divide the data randomly into train and test sets; the new partitions contain the following number of observations, which will remain the same in the whole analysis:

```
length(indicesTrain)
```

```
## [1] 6072
```

```
length(indicesTest)
```

```
## [1] 2603
```

After having created the document term matrix in which the presence/absence of a word is record, let's search for the words which apperas more than 300 times but less than 4000. I choose this value after many attemps, trying to select the most representative words, avoiding the usual words that appers many times but are not contained in the stopwords function.

```
## [1] 1494
```

After building the two matrices for the train and test set referred to these words, we can apply the Naive Bayes classifier:

```
table(predictions,mbt_test$type)
```

```
##
## predictions ENFJ ENFP ENTJ ENTP ESFJ ESFP ESTJ ESTP INFJ INFP INTJ INTP
##        ENFJ   10    4    5    1    0    0    0    0    5    5    2    0
##        ENFP    8   61    4   12    2    2    0    1   20   33   12    9
##        ENTJ    1    2   17   10    0    0    2    0    4   11    7    6
##        ENTP    2   11    9   71    3    3    1    7   24   17   17   32
##        ESFJ    0    0    0    0    0    0    0    0    0    0    0    0
##        ESFP    0    4    3    3    0    1    0    1    6    5    9    3
##        ESTJ    0    0    0    0    0    0    0    0    0    0    0    0
##        ESTP    0    1    0    2    1    0    0    2    0    2    0    0
##        INFJ   17   35    3   10    3    0    1    1  183   78   29   28
##        INFP   14   46    2   10    1    3    2    2  107  292   28   40
##        INTJ    2   10   13   21    2    1    0    0   31   26  139   45
##        INTP    0    5    8   42    0    0    1    2   21   18   37  192
##        ISFJ    0    0    1    3    1    0    0    0    6    3    1    0
##        ISFP    2   24    4   11    1    1    0    3   28   48   23   24
##        ISTJ    0    2    1    2    0    0    1    0    3    3    5    2
##        ISTP    2    3    3    8    0    0    1    5    4    3    8   18
##
## predictions ISFJ ISFP ISTJ ISTP
##        ENFJ    0    1    1    0
##        ENFP    3    3    2    5
##        ENTJ    2    3    1    3
```

```
##         ENTP    0    4    5    8
##         ESFJ    0    0    0    0
##         ESFP    2    2    5    1
##         ESTJ    0    0    0    0
##         ESTP    0    0    0    0
##         INFJ    7   11    2    7
##         INFP    8   16    9    4
##         INTJ    2    4   11   10
##         INTP    1    1    4   12
##         ISFJ   18    2    3    0
##         ISFP    8   28   10    7
##         ISTJ    1    0    9    0
##         ISTP    1    2    6   43
```

```
accuracy<-1-(mean(predictions!=mbt_test$type));accuracy
```

```
## [1] 0.4095275
```

2.2 4 Main Types

Let's first group the original 16 types into the 4 main categories which are:

TYPE	Description
diplomat	"INFJ","INFP","ENFJ","ENFP"
analyst	"INTJ","INTP","ENTJ","ENTP"
sentinel	"ISTJ","ISFJ","ESTJ","ESFJ"
explorer	"ISTP","ISFP","ESTP","ESFP"

Their frequencies are:

Analysts	Diplomats	Explorers	Sentinels
3331	4167	745	452

Then after having built the train and test sets, let's consider the same range of frequencies for the word appereance. In this case the number of words considered is:

```
five_times_words <- findFreqTerms(mbt_dtm_train, 300,4000)
length(five_times_words)
```

```
## [1] 1503
```

After building the two matrices for the train and test set referred to these words, we can apply the Naive Bayes classifier:

	True Analyst	True Diplomat	True Explorer	True Sentinel
Predicted Analyst	666	214	56	33
Predicted Diplomat	170	841	45	40
Predicted Explorer	142	130	117	23
Predicted Sentinel	27	43	16	40

```r
accuracy<-1-(mean(predictions!=mbt_test$type));accuracy
```

```
## [1] 0.622743
```

We can deduce that the algorith improves to an accuracy of 0.641, which is better than the one obtained with 16 types, even though it is not a satisfiying result.

2.3 Introverts / Extroverts

Let's consider now Introverted VS Extroverted people. Grouping our data into this variables gives the following frequency table:

Extrovert	Introvert
1999	6676

Again let's consider the following number representing the words that appears within our range over the training sets:

```
## [1] 1503
```

and apply the classifier.

	True Extrovert	True Introvert
Predicted Extrovert	333	385
Predicted Introvert	270	1615

```
accuracy<-1-(mean(predictions!=mbt_test$type));accuracy
```

```
## [1] 0.7625816
```

Introverts and Extroverts are classified by the algorithm with the an accuracy of roughly 75%. In particular the partial correct classifications are the following:

Type	Correct Classification
Introvert	0.5522388
Extrovert	0.8075

which are quite satisfying for both classes.

2.4 Thinking / Feeling

By considering Thinking VS Feeling kind of people we obtain the following table:

Feeling
4694

The words within the range of frequiencies for the training set are:‘

```
## [1] 1532
```

After building the two matrices for the train and test set referred to these words, we can apply the Naive Bayes classifier:

	True Feeling	True Thinking
Predicted Feeling	1805	256
Predicted Thinking	324	938

```
accuracy<-1-(mean(predictions!=mbt_test$type));accuracy
```

```
## [1] 0.7602766
```

For Intuition and Sensing the algorithm classify quite good with the an accuracy of about 78%. In particular the partial correct classifications are the following:

Type	Correct Classification
Feeling	0.8757885
Thinking	0.7432647

which are quite satisfying for both classes.

2.5 Intuition / Sensing

Let's consider Intuition VS Sensing with the following frequencies:

Intuition
7478

Here the number of words is:

```
## [1] 1496
```

Let's apply again the classifier:

	True Intuition	True Sensing
Predicted Intuition	1890	162
Predicted Sensing	356	195

```
accuracy<-1-(mean(predictions!=mbt_test$type));accuracy
```

```
## [1] 0.7871687
```

For Intuition and Sensing the algorithm classify better with the highest accuracy found until here. In particular the partial correct classifications are the following:

Type	Correct Classification
Intuition	0.841496
Sensing	0.5462185

2.6 Judgment / Perception

Finally for the division Judgment/Perception the frequency table is:

Judgment
2343

while the number of words of the traininig set in the range (300,4000) is:

```
## [1] 1495
```

Here the Naive Bayes Classifier gives the following results:

```
tJP<-table(predictions,mbt_test$type)
```

	True Judgment	True Perception
Predicted Judgment	402	331
Predicted Perception	326	1221

```
a
```

```
## [1] 0.7118421
```

Type	Correct Classification
Judgment	0.5521978
Perception	0.7867268

3 Conclusion

With the initial dataset differentiating between the 16 types of personalities the Naive Bayes Classifier misclassifies observations between many similar classes reaching an accuracy of only 0.4; this may happen because the dataset is quite unbalanced. However by grouping people in bigger groups according to the test structure we could get some nice insights. For instances, by considering the 4 main groups (mediators, analysts,sentinels and explorers) the accuracy improves to 0.6 but the classes are still quite unbalanced. By further grouping according to one of 4 components of the type the accuracy grows between 0.7 and 0.8. The most successful application of the Naive Bayes Classifier is the one obtained from the division between Intuion and Sensing people which has globally an accuracy of about 0.8, even though the success rate in the two classes in quite unbalanced. On the contrary between Thinking and Feeling people the algorithm reaches a slightly low accuracy (0.75) but it has no significative difference in the successful rate between the two classes.