

Classification of web-based discussions using Naive Bayes

Ekaterina Soukhikh
Group 8



Problem description

● is to investigate whether the Naive Bayes algorithm is applicable for classifying of mobile related web-based discussions.

3 research categories:

● Forum subjects

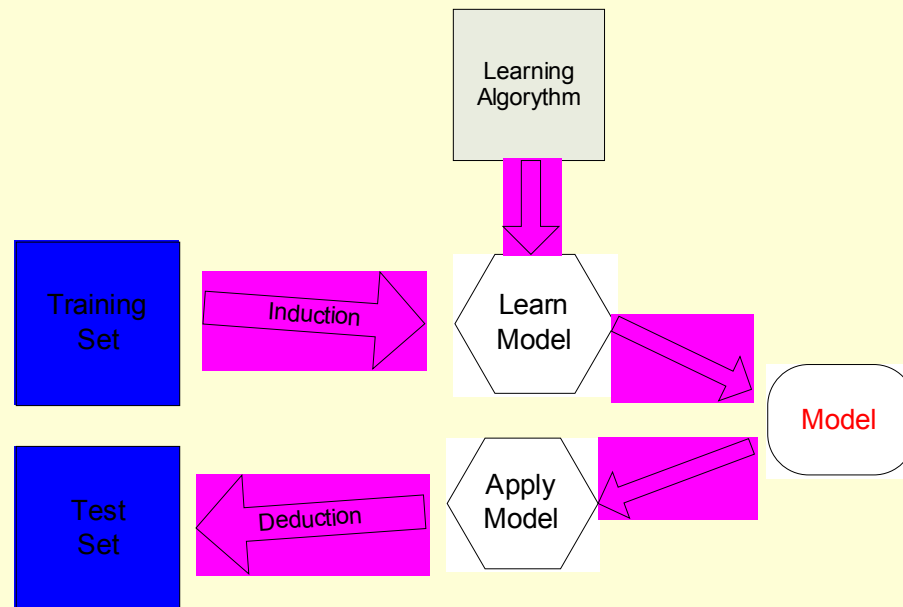
● Discussions' language recognition

● Positive/negative expressions



Learning in Text Classification

- Next model illustrates the task of text classification using learning algorithms.






Naive Bayes

- Looks at words in the text as independent attributes.

$$p(\text{Class}|\text{Document}) = p(\text{Class}) \prod_i p(\text{Word}_i|\text{Class})$$

- Then the probability that document belongs to a specific class is a product of the conditional probabilities for each attribute value.
- 

Third Party Application

- During the training, the application finds words that are presented more often and creates “vocabulary” (set of words with the appearance probability) for each category.
- When finding probability of a test document, only those words that presented in the vocabulary will be taken into consideration.

```
Read 101 words from sprinttest2.txt
Classifying...


=====
=      Table of Probability      =
=====
Nokia      4.5016843619955833E-4
Ericsson   1.200341248257459E-5
DoCoMo     0.13940457760449565
Motorola   0.024159166370298232
Sprint     0.8359740841765241

DONE!
```



Results

The general impression :


- Language recognition works best,
 - Forum subjects are a bit harder to classify,
 - Positive/negative expressions are even harder.
- 



Discussion

- Languages are relatively different, and easier to separate
- Topics / subjects differs less and are not always properly defined in each message.
- Positive/negative expressions are not always based on words that express happiness (like, fine, nice, good etc).


One example: big screen, small weight, 3 mgp camera!!! (positive expression) or a big weight, small screen, 1,2 megapixel Camera☹ (negative expression)





Conclusion

The conclusion is that Naive Bayes showed to work relatively fine, and can be used for classification of web discussions, especially in cases when:

- It is possible to collect the accurate training data, which will be similar to the data that are going to be tested.
 - It is possible to pre-define a limited number of classes / categories that tested data have to be sorted into.
- 

● Questions?

