Naive Bayes
EXERCISE

Krikun Gosha

October 30, 2016

Exercise Repository the ugliest implementation (github) - contains source code, data sets in recourse directory, sbt configuration, and about everything in `Main` class.

Based on Naive Bayes algorithm (multinomial model) [1, Chapter 13, figure 13.2]

Conditional probabilities are computed based on train set (350 nonspam messages and 350 spam messages), which give as prior probabilities:

$$P(\texttt{ns}) = \frac{350}{700} = 0.5 \text{ and } P(\texttt{s}) = \frac{350}{700} = 0.5.$$

Conditional probabilities for every token computes by dictionaries of doc, with multiplication on token frequency (actually it's equals computations by tokens, *but I already conduct dictionaries for each doc*).

After that evaluates on test sets with 130 nonspam messages and 130 spam messages. As soon as we use logarithm normalization on probability values $\in (0..1)$, thus all scores are negative.

Applying on `ns` test set Within 130 nonspam messages 7 was wrongly classified as spam:

docID	spam	nonspam
6-380msg1.txt	-1204.08	-1241.75
6-453msg1.txt	-108.18	-109.31
6-437msg3.txt	-1486.73	-1546.93
6-890msg3.txt	-1745.23	-1753.74
6-790msg1.txt	-2711.84	-2797.29
6-809msg3.txt	-2226.56	-2365.50
6-781msg5.txt	-455.77	-457.09

ACCURACY - 0.9461538

Applying on `s` test set And almost all spam messages was properly classified. Except one:

docID	spam	nonspam
spmsga125.txt	-1268.22	-1258.18

ACCURACY - 0.99230766

References

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, 2009.