Assignment - 5

CSE 537 Artificial Intelligence

By

Komal Gyanani(111491719)

Aditi Ghotikar (111353349)

Aditi Nayak(111482252)

Amogh Reddy(111493841)

Contributions

All members contributed equally.

Aditi Nayak - contributed to Spam Filter

Aditi Ghotikar - contributed to Spam Filter

Amogh Reddy - contributed to Spam Filter

Komal Gyanani - contributed to Spam Filter

Spam Filter

We have implemented the Spam Filter using Naive Bayes Algorithm.

Naive Bayes spam filtering is a baseline technique for dealing with spam that can tailor itself to the email needs of individual users and give low false positive spam detection rates

Here, we assume *naively* that the likelihoods of the samples can be estimated from the training data instead of evaluating all the possibilities. The conditional probability becomes:

$$P(x \mid \omega_j) = P(x_1 \mid \omega_j) \cdot P(x_2 \mid \omega_j) \cdot \ldots \cdot P(x_d \mid \omega_j) = \prod_{k=1}^{d} P(x_k \mid \omega_j)$$

Where,

- xi be the feature vector of sample i,i∈{1,2,...,n}i,i∈{1,2,...,n},
- ωj be the notation of class j,j∈{1,2,...,m}j,j∈{1,2,...,m},
- and P(xi | ωj) be the probability of observing sample xi given that is belongs to class ωj(i.e spam or ham)

The posterior probability is found using the formula:

Posterior probability = (Conditional Probability * prior probability) ÷ Evidence

In our code, we have taken the logarithm (base 10) for our probabilities because the probability values are sometimes very low. Hence, there cannot be any clear distinction as float returns 0.0 for such small values(underflow).

Our program gives an accuracy of 90.5% on the given training and test sets.

We tried on various smoothing parameters using Laplace/Additive smoothing :

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d),$$

Smoothing parameter	Accuracy
0.1	88.7%
1	90.5%
10	90.8%

REFERENCES

1. Lecture slides
2. http://sebastianraschka.com/Articles/2014_naive_bayes_1.html
3. https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering