Toy problem: Email Classification

General Dictionary

1. This
2. Hello
3. How
4. Offer
5. Buy
6. Thanks
7. Shipping
8. Meet
9. Tomorrow
10. Lunch
11. Spicy
12. Morning
13. Evening
14. Function
15. Party

**At the beginning of a classification problem, we will be given with a dataset. In case of email classification, we will have two sets, 1) Ham 2) Spam; having 100 emails in each set.**

Ham Set

100 emails

Spam Set

100 emails

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Spam set analysis**   |  |  |  | | --- | --- | --- | | **Word Index** | **Freq** | **Likelihood**  **P(word | Spam)** | | **1** | **500** | **=500/4623** | | **2** | **100** | **=100/4623** | | **3** | **700** | **=700/4623** | | **4** | **900** | **=900/4623** | | **5** | **850** | **=850/4623** | | **6** | **600** | **=600/4623** | | **7** | **950** | **=950/4623** | | **8** | **20** | **=20/4623** | | **9** | **3** | **=3/4623** | | **10** | **0** | **=0** | | **11** | **0** | **=0** | | **12** | **0** | **=0** | | **13** | **0** | **=0** | | **14** | **0** | **=0** | | **15** | **0** | **=0** | | **Total** | **4623** | **=1** | | **Ham set analysis**   |  |  |  | | --- | --- | --- | | **Word Index** | **Freq** | **Likelihood**  **P(word | Ham)** | | **1** | **700** | **=700/2821** | | **2** | **40** | **=40/2821** | | **3** | **750** | **=750/2821** | | **4** | **10** | **=10/2821** | | **5** | **10** | **=10/2821** | | **6** | **600** | **=600/2821** | | **7** | **1** | **=1/2821** | | **8** | **300** | **=300/2821** | | **9** | **100** | **=100/2821** | | **10** | **70** | **=70/2821** | | **11** | **50** | **=50/2821** | | **12** | **40** | **=40/2821** | | **13** | **30** | **=30/2821** | | **14** | **70** | **=70/2821** | | **15** | **50** | **=50/2821** | | **Total** | **2821** | **=1** | |

**Prior probability: Let us assume that we have a data that a group of people received 50000 emails during last one month and out of the 15000 emails were Spam emails.**

**P(Spam) = 15000/50000 = 0.3**

**P(Ham) = 1 – 0.3 = 0.7**

**Marginal Probability:**

**P(B) = P(B | A) P(A) + P(B | not A) P(not A)**

**P(A | B) = P(B | A) P(A) / P(B)**

**Posterior probability = likelihood x prior probability / marginal probability**

**Assume that we have a new email: This Offer, Buy**

**Spamity Calculation:**

**P( Spam | Word = This) = P(Word = This | Spam) P(Spam)/P(Word = This) = 500/4623 x 0.3 /0.2 = 0.162**

**P(Word = This) = P(Word = This | Spam) P(Spam) + P(Word = This | Not Spam) P(not Spam)**

**We know Not spam is “Ham”**

**= 500/4623 X 0.3 + 700/2821 x 0.7 = 0.2**

**Hamity calculation:**

**P(Ham | Word = This) = P(Word = This | Ham) P(Ham)/P(Word = This)**

**= 700/2821 x 0.7/ (500/4623 X 0.3 + 700/2821 x 0.7) = 0.8426**