# 貝葉斯垃圾信息過濾器使用說明

## 基本原理

貝葉斯垃圾信息過濾器 (bayesian\_spam\_filter\cppjieba\test\demo\_test.cpp) 用於判斷一條message是否屬於垃圾信息（spam）。基本工作原理如下：

第一步，將接收到的message用結巴分詞拆成若干個詞組，例如 {你好嗎} 將會被拆成 {你好} 和 {嗎} 兩個詞組；

第二步，搜索已準備好的詞典庫，算出單個詞組為垃圾信息的概率

第三步，根據貝葉斯推斷，計算出message為垃圾信息的概率

第四步，把算出來的結果與預先設定的閾值（threshold）比較，若 , 則判斷該message為垃圾信息。

為了應用本代碼，需要預先在以下目錄準備好字典文件：

* bayesian\_spam\_filter\cppjieba\data\save\wnormal.txt
* bayesian\_spam\_filter\cppjieba\data\save\wspam.txt

## 使用方法

### 依賴軟件

* g++ (version >= 4.1 is recommended) or clang++;
* cmake (version >= 2.6 is recommended);

### 下載和編譯

>>cd cppjieba

>>mkdir build

>>cd build

>>cmake ..

>>make

### API 範例

#### predict\_spam

輸入信息（string）和閾值（double），判斷該信息是否垃圾信息。

**API:** bool predict\_spam(string msg, double threshold){

//參數 msg: 被檢測的輸入信息

//參數 threshold: 閾值，用於與預測的spam概率比較

//返回 : 該信息是垃圾信息（true），或不是垃圾信息（false）

}

根據測試結果（測試sample為11-Oct-17至31-Oct17的記錄），閾值為0.07時誤判率（正常信息被判斷為垃圾信息）為5.58%，而正確率（垃圾信息被判斷為垃圾信息）達到60.24%，這個結果在所有測試的閾值中效果最好。因此，推薦的閾值設置值為threshold=0.07.

|  |  |  |
| --- | --- | --- |
| Threshold | Normal As Spam (%) | Spam Detected (%) |
| 0.01 | 32.18% | 68.63% |
| 0.03 | 12.14% | 63.31% |
| 0.05 | 6.84% | 61.37% |
| 0.07 | 5.58% | 60.24% |
| 0.09 | 4.44% | 59.34% |

## 結果分析

由於訓練數據集是由人工分類得到的，因此數據本身的誤差影響了模型的表現。以下例子全部來自於測試數據中被人工歸類為垃圾信息的記錄。可以看到，有些本應該是正常信息的記錄，被人工歸類為垃圾信息了，對於這些被錯誤歸類的“垃圾信息”，模型預測為垃圾信息概率是非常低的，與正確歸類的垃圾信息有較明顯的差別。如果訓練數據集能夠更精確一些，相信模型的表現會有進一步改善。

|  |  |  |  |
| --- | --- | --- | --- |
|  | Spam Filter (th=0.07) | | Manual |
| Message | Spam Prob | Is Spam | Is Spam |
| 没看清的看好 | 0.000112 | No | Yes |
| 还是庄再来 | 0.000009 | No | Yes |
| 再来个庄 | 0.000012 | No | Yes |
| 怎么学啊 | 0.015442 | No | Yes |
| 继续闲闲闲。 | 0.000508 | No | Yes |
| 漂亮闲收对子收完美 | 0.000000 | No | Yes |
| 稳。打。稳。赢丅XX 丅WW | 0.276596 | Yes | Yes |
| 丅62丅35 认准钰 丅62丅35 | 0.914922 | Yes | Yes |
| 想赢，搜公纵號 〔龙庄姐〕 | 0.999903 | Yes | Yes |
| 搜工重号〖屠龙队〗解析赢钱秘方 | 1.000000 | Yes | Yes |