1、对象：

**Data**

{type ,message}

**Classification**

{hamMap, spamMap, hamProbability, spamProbability,

hamWordsCount, spamWordsCount, count}

**短信**{类型，内容}

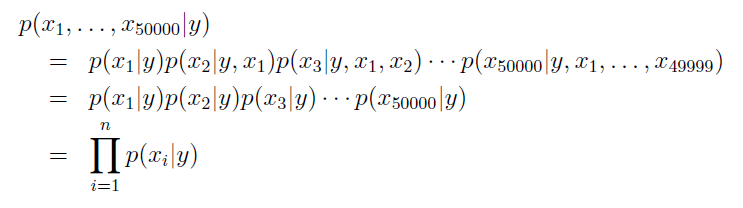
**分类器**{非垃圾短信词典，垃圾短信词典；非垃圾短信比例，垃圾短信比例；

非垃圾短信单词数，垃圾短信单词数；训练集不重复单词数}

2、公式：

朴素贝叶斯：

A表示短信，B表示类别



拉普拉斯平滑：



3、训练过程：

（1）for 训练集每一条邮件

建立（非）垃圾邮件单词集合

计算（非）垃圾邮件数量p(B)

计算（非）垃圾邮件单词数量count(y)

计算训练集不重复单词数量N

（2）for （非）垃圾邮件单词集合

建立（非）垃圾邮件映射表{单词，数量}

（3）计算（非）垃圾短信比例

建立分类器

4、测试过程：

（1）for 测试集每一条邮件

计算该邮件在（非）垃圾邮件的概率（拉普拉斯平滑）

比较该邮件在哪类的概率更大

（2）统计正确率