

**2021年《机器学习》**

**工程报告**



**课 程：** 机器学习

**姓 名：**

**完成时间：** 2021.6.15

一． 工程摘要与每人贡献

摘要：总体介绍本工程的工作。

每人分工：

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 角色  （组长，组员，独自完成） | 工作量  比例 | 负责内容 |
|  | 组长 | 40% | 实验代码实现；报告的第五部分撰写 |
|  | 组员 | 20% | 数据集搜集；报告的第三部分撰写 |
|  | 组员 | 20% | 实验可视化设计；报告的第四部分撰写 |
|  | 组员 | 20% | 实验模型的训练；报告的第二部分撰写 |

二、研究背景与意义

说明本工程所涉及的研究背景研究的意义，包括主要涉及领域，主要研究方法，主要存在问题，现有解决方案等等，可以自行补充其他内容。

**垃圾邮件**(未经许可的商业邮件或不受欢迎的非法邮件)成为1990 年以来一直困扰着互联网的问题。那时，互联网越来越商业化，越来越接近消费者。一开始只是感觉到少数人制造一些电子广告，而现在垃圾邮件占用了目前大部分的邮件空间。垃圾邮件造成恶性商业形象，损害了邮件使用者的利益，占用网络资源，而且它对公司的网络系统和网络可信度,效率和安全性造成严重的威胁。大多数关于垃圾邮件的讨论围绕在技术层面。然而，驱使垃圾邮件不断增加的却是经济原因,要求发送者的商业公司为发送垃圾邮件支付的巨额费用。反垃圾邮件是一项长期而艰巨的工作。  
**第一代反垃圾邮件技术:** 1、基础MTA控制:MTA协议应该能控制基于每个域名的通讯连接，比如，防止“开放转发”并不能当作一种独立的反垃圾邮件技术。因为协议是在早期还没有产生垃圾邮件问题时被提出的，并没有预见到会有这些基本的安全隐患。  
 2、白名单和黑名单:黑名单(B1ack List)和白名单(White List)。 分别是己知的垃圾邮件发送者或可信任的发送者IP地址或者邮件地址。现在有很多组织都在做b1(blocklist) ,将那些经常发送垃圾邮件的IP地址(甚至IP地址范围)收集在一起,做成blocklist。目前很多邮件接收端都采用了黑白名单的方式来处理垃圾邮件,包括MUA和MTA,当然在MTA中使用得更广泛，这样可以有效地减少服务器的负担。  
 3、简单的关键字搜索:简单的关键字搜索一直是对抗垃圾邮件的基本方法。这一功能存在于垃圾邮件成为互联网的主要问题之前，那时作为内容过滤的一部分和基于反病毒产品的解决方案和服务。  
**第二代反垃圾邮件技术:**  
 1、实时黑名单:尽管在基于网址和域名上它是一个DNS测试，RBLs 是真实的反垃圾邮件技术，在RBLs之后的概念是简单的维护-一个发送垃圾邮件的网址，以阻止垃圾邮件的继续发送。这种技术会有一-定的效果，但容易被绕过。比如，改变IP地址，或者利用第三方的服务器来发送垃圾邮件。同样地，域名很容易被获得，并被垃圾邮件发送者利用，因而不能完全依赖它来判别垃圾邮件。

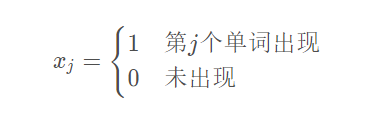
2、电子签名:这是对于垃圾邮件防御有重大意义的一项技术。电子签名技术就是,如果垃圾邮件以大量的相同信息发送，可以用电子签名技术产生一个唯一的电子签名来收集和辨別垃圾邮件。如果能够获得充足的垃圾邮件样本，对于降低垃圾邮件的比率有重要意义。但是这种技术需要及时操作才能达成效果。

**第三代反垃圾邮件技术:**  
 贝叶斯过滤是利用统计学的方法检测垃圾邮件,基于垃圾邮件中单个词语的出现概率来判定，这是反垃圾邮件技术上的第一个突破，贝叶斯过滤技术的发展从根本上把反垃圾邮件的重点从网络和协议改变为邮件内容。梭子鱼的贝叶斯规则库在出厂之前都经过近万封邮件的培训，到达用户之后，用户继续对其进行培训，被“有效培训”以后，过滤垃圾邮件的准确率达到99%。  
**第四代反垃圾邮件技术:**  
 垃圾邮件的存在原因还有一部分是因为，在SMTP创造之初，只是用于学校，政府和军队，因为是一个封闭的系统，所以不存在非法使用和电子邮件的滥用。1990 年起，互联网广:泛的应用于商业，但是之前的技术隐患仍然存在。垃圾邮件和反垃圾邮件活动必将长期存在，经过了十几年的发展,新兴的反垃圾邮件技术也层出不穷。但无论哪一种技术，都无法完全应对多变的垃圾邮件。

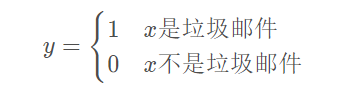
三、模型方法

详细说明本工程所使用的模型方法和理论等

我们令向量 x 表示垃圾邮件的特征向量，该向量包含了 100 个按字母序排序的单词特征，这些单词通常为垃圾邮件常出现的词汇：discount，deal，now 等等：



令 y 标签表示该邮件是否是垃圾邮件：



那么垃圾邮件分类就是一个 0/1 分类问题，可以用逻辑回归完成，这里不再重复介绍逻辑回归的过程了，我们考虑如何降低分类错误率：

尽可能的扩大数据样本：Honypot 做了这样一件事，把自己包装成一个对黑客极具吸引力的机器，来诱使黑客进行攻击，就像蜜罐（honey pot）吸引密封那样，从而记录攻击行为和手段。

添加更多特征：例如我们可以增加邮件的发送者邮箱作为特征，可以增加标点符号作为特征（垃圾邮件总会充斥了？，！等吸引眼球的标点）。

预处理样本：正如我们在垃圾邮件看到的，道高一尺，魔高一丈，垃圾邮件的制造者也会升级自己的攻击手段，如在单词拼写上做手脚来防止邮件内容被看出问题，例如把 medicine 拼写为 med1cinie 等。因此，我们就要有手段来识别这些错误拼写，从而优化我们输入到逻辑回归中的样本。

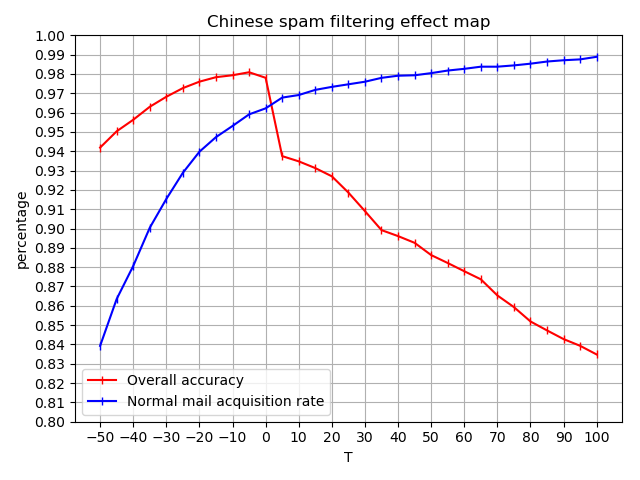
四、系统设计

系统的**详细**设计，系统流程，系统的每一步的具体流程，例如，如何处理语料，如何训练模型，如何测试模型，最后对模型进行评估等。

1. 数据准备：收集数据与读取  
2. 数据预处理：处理数据  
3. 训练集与测试集：将先验数据按一定比例进行拆分  
4. 提取数据特征，将文本解析为词向量   
5. 训练模型：建立模型，用训练数据训练模型。即根据训练样本集，计算词项出现的概率P(xi|y)，后得到各类下词汇出现概率的向量   
6. 测试模型：用测试数据集评估模型预测的正确率  
7. 预测一封新邮件的类别  
8. 考虑如何进行中文的文本分类

五．实验结果分析、对比和讨论

对实验结果进行分析说明，测试一些课程中演示的样例，根据结果说明为什么对或者为什么错等具体分析；对不同模型和参数的对比进行分析；最后对系统提出改进方案等。



图

实验的数据来源于http://plg.uwaterloo.ca/~gvcormac/treccorpus06/，我们项目组下载了大量的中文邮件进行训练，这里面包括大量的宣传信息、招聘广告、生活邮件、正常的通信等等生活中常见的邮件类型，包括有价值的邮件和很多广告垃圾邮件。我们的模型通过训练，在大量采用数据集进行训练后，我们可以从上图——中文垃圾邮件过滤效果图中看出，随着训练次数的提升，训练的结果也越发清晰。总体的准确率逐渐下降，而正常邮件的获取率则在逐渐上升。

这很有力的证实了训练的效果。我们在训练过程中，不断的进行参数的调整，可以看出，实验效果在逐渐变好，正常邮件获取的程度在逐渐上升。但是也普遍存在一些问题。包括随着训练的进行，总体准确率在不断下降。这一点无法避免。我们使用的朴素贝叶斯算法属于第三代过滤算法，具有较好的过滤效果，以及较快的学习速度。我们可以看出实验后期的训练结果稳定维持在98%以上，是较为稳定的过滤结果。

在改进方面，我们可以调整在模型训练中的学习率，使得模型在训练过程汇总不会出现实验的准确率快速下跌的情况，让实验结果更具可信性，可行性。另一个方面，我们也会对词频进行进一步整理，让训练获得的词频模型在测试集下取得更好的成果。

六．对本门课的感想、意见和建议

机器学习是一门多领域交叉学科。他通常用于研究计算机怎样模拟或实现人类的[学习行为](https://baike.baidu.com/item/%E5%AD%A6%E4%B9%A0%E8%A1%8C%E4%B8%BA/5482132" \t "https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/_blank)，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。它也是[人工智能](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD/9180" \t "https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/_blank)核心，是使计算机具有智能的根本途径。随着[大数据](https://baike.baidu.com/item/%E5%A4%A7%E6%95%B0%E6%8D%AE/1356941" \t "https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/_blank)时代各行业对数据分析需求的持续增加，通过机器学习高效地获取知识，已逐渐成为当今机器学习技术发展的主要推动力。

我们作为计算机的学生，学习这门课也充满了动力。这门专业课很好的体现了本专业的特性，这也激发了我的兴趣。机器学习，让我们了解如何构筑基本的网络，让机器，或者说算法了解如何处理问题，这是一个十分迷人的手段。

最后我们真切的希望这门课加一些实际的案例，由老师带领学生做一遍，这样会让我们对模型的了解更加深刻，这是我们认为更有必要的事情。