北风网
IBEIFENG.COM

# 人工智能之机器学习

## 垃圾邮件过滤

上海育创网络科技股份有限公司

主讲人：刘老师(GerryLiu)

# 课程要求

- 课上课下“九字”真言
  - 认真听，**善摘录，勤思考**
  - **多温故，乐实践**，再发散
- 四不原则
  - 不懒散惰性，不迟到早退
  - 不请假旷课，不拖延作业
- 一点注意事项
  - 违反“四不原则”，不推荐就业

# 课程内容

- 垃圾邮件过滤
- 音乐系统文件分类/音频样本的标签化
- 金融反欺诈项目

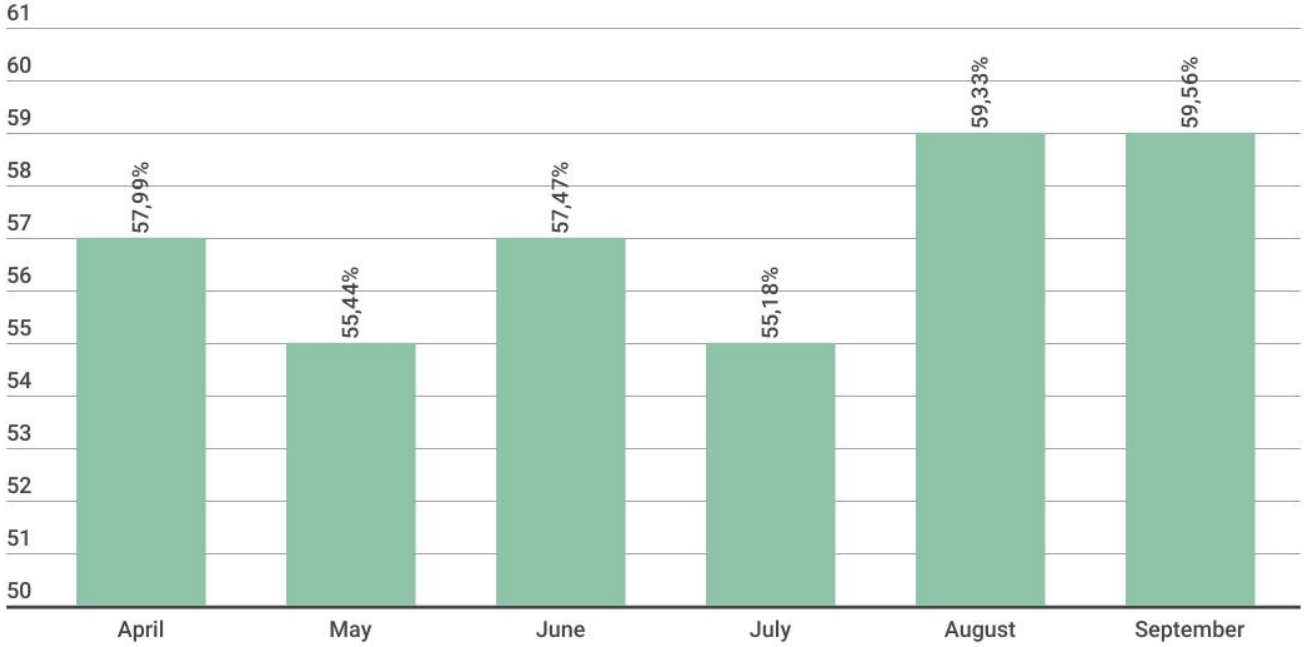
# ML常见考点



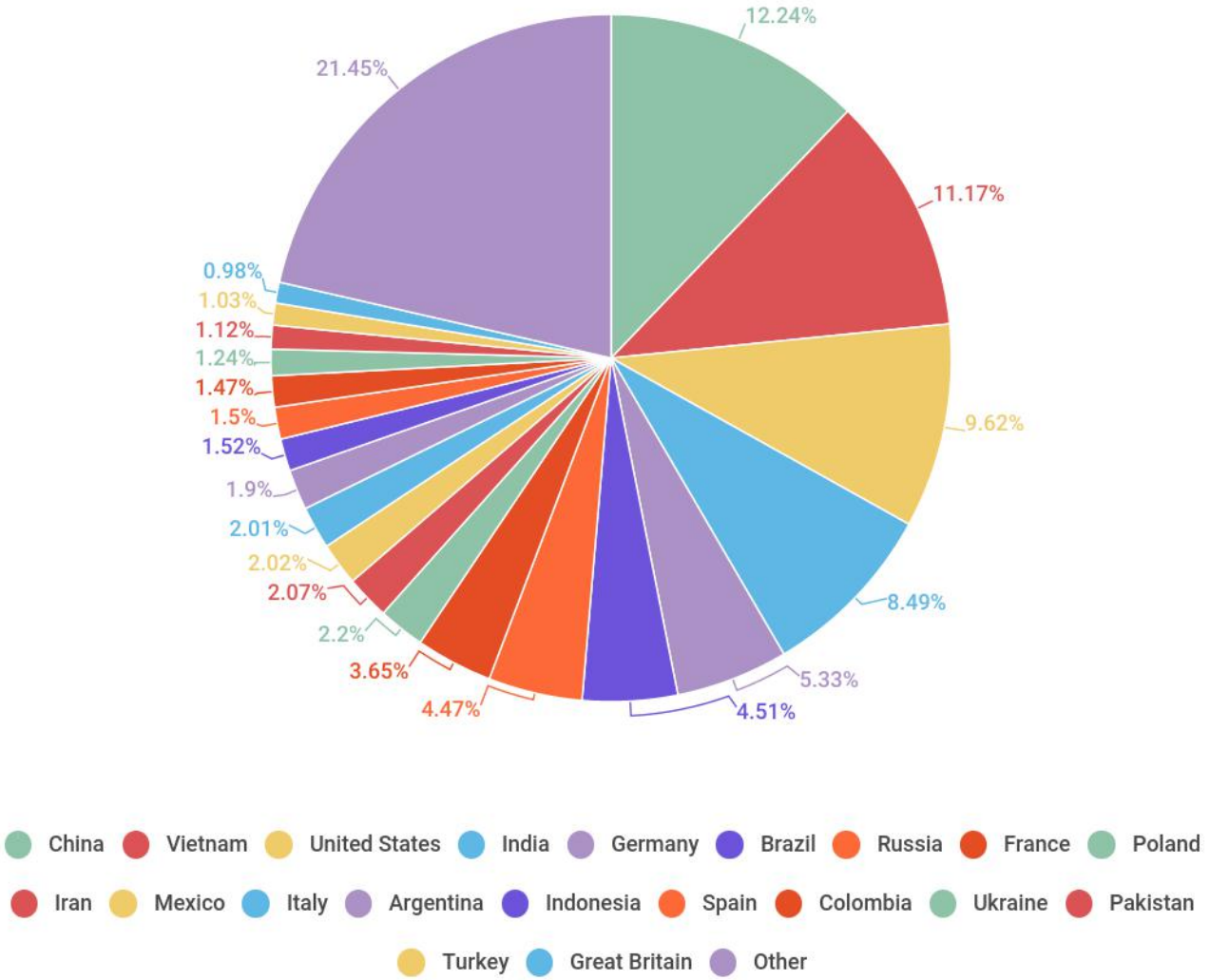
- 基础知识：什么是有监督/无监督算法， 什么是过拟合/欠拟合，如何避免过拟合，什么是交叉验证，什么是bagging/boosting?
- 常问算法：LR、SVM、Random Forest、GBDT、KMeans以及简历中写到的算法
- 算法细节：LR的优势、Random Forest里面的树是否修剪、GBDT算法中梯度下降的求解公式、KMeans算法的k如何选择
- 算法描述：给出KMeans/KNN算法的伪代码、LR/SVM的推导公式
- 其它：算法的优劣比较、L1/L2正则项比较、online learning

# 垃圾邮件过滤概述

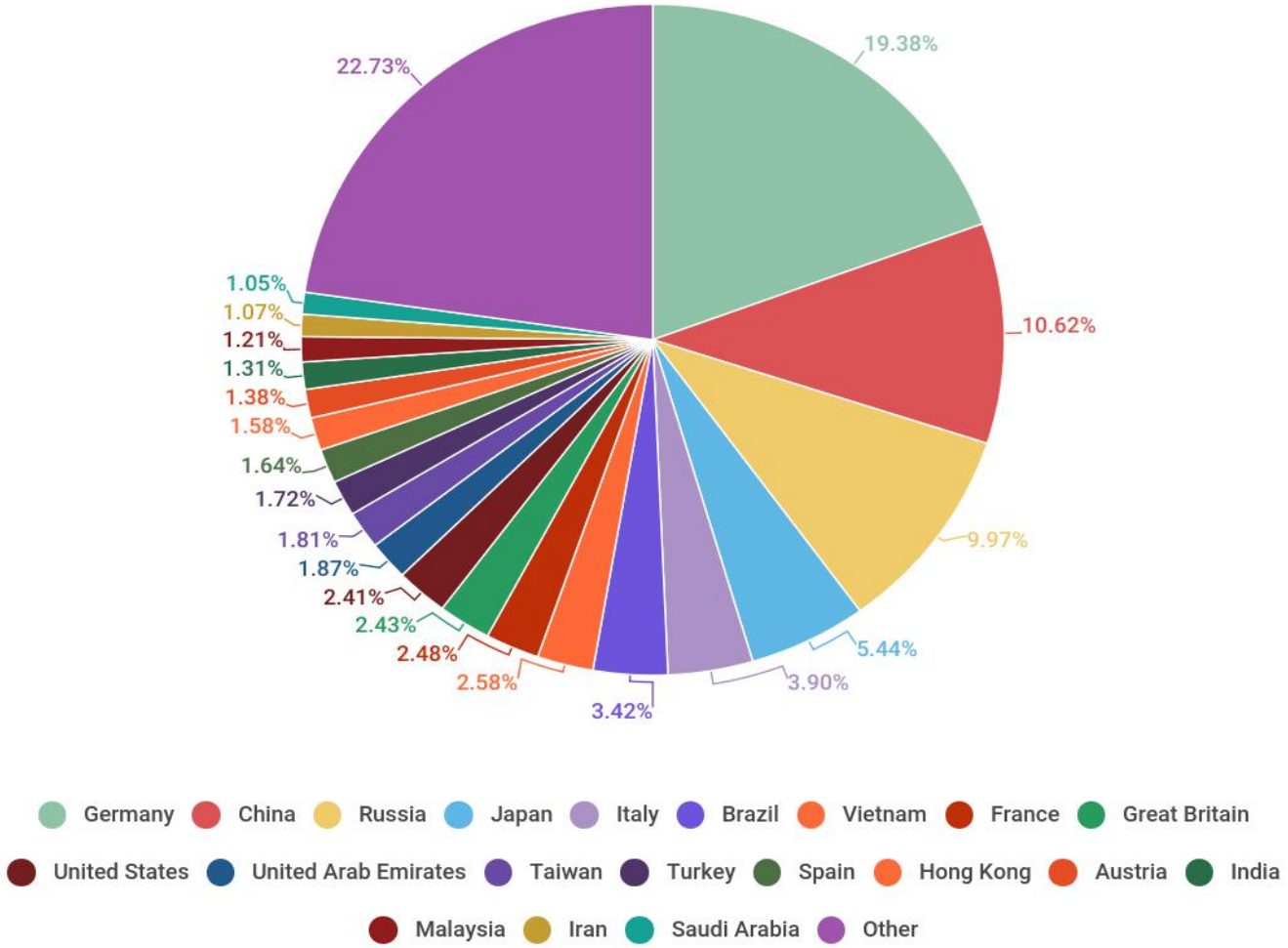
- 随着互联网的蓬勃发展，电子邮件已经成为互联网上最普遍的通讯方式之一；据最新调查显示，2017Q3季度中国是世界最大垃圾邮件产生国和第二大受恶意邮件袭击的国家。垃圾邮件的内容主要包括欺诈邮件、新闻议程、钓鱼攻击邮件、站点宣传邮件、病毒邮件等等



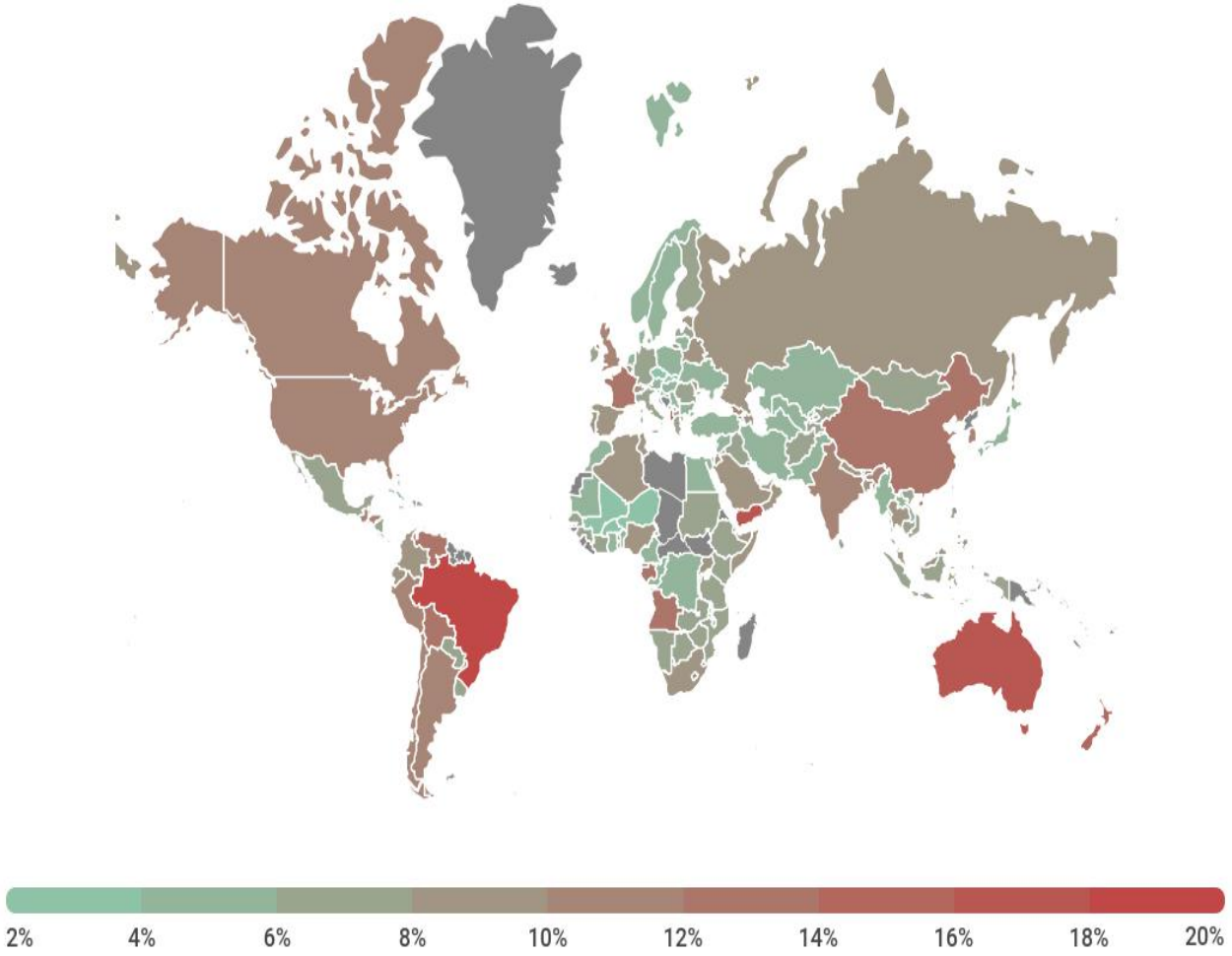
# 2017垃圾邮件_来源分布情况图



# 2017垃圾邮件_遭恶意邮件袭击国家情况图



# 2017垃圾邮件_网络钓鱼全球分布情况



# 垃圾邮件的影响

- 垃圾邮件主要影响的因素如下：
  - 占用网络带宽，造成邮件服务器拥塞，进而降低整个网络的运行效率。
  - 骗取钱财，传播色情内容等。
  - 携带病毒程序，可能导致接收邮件的机器/服务器感染病毒。

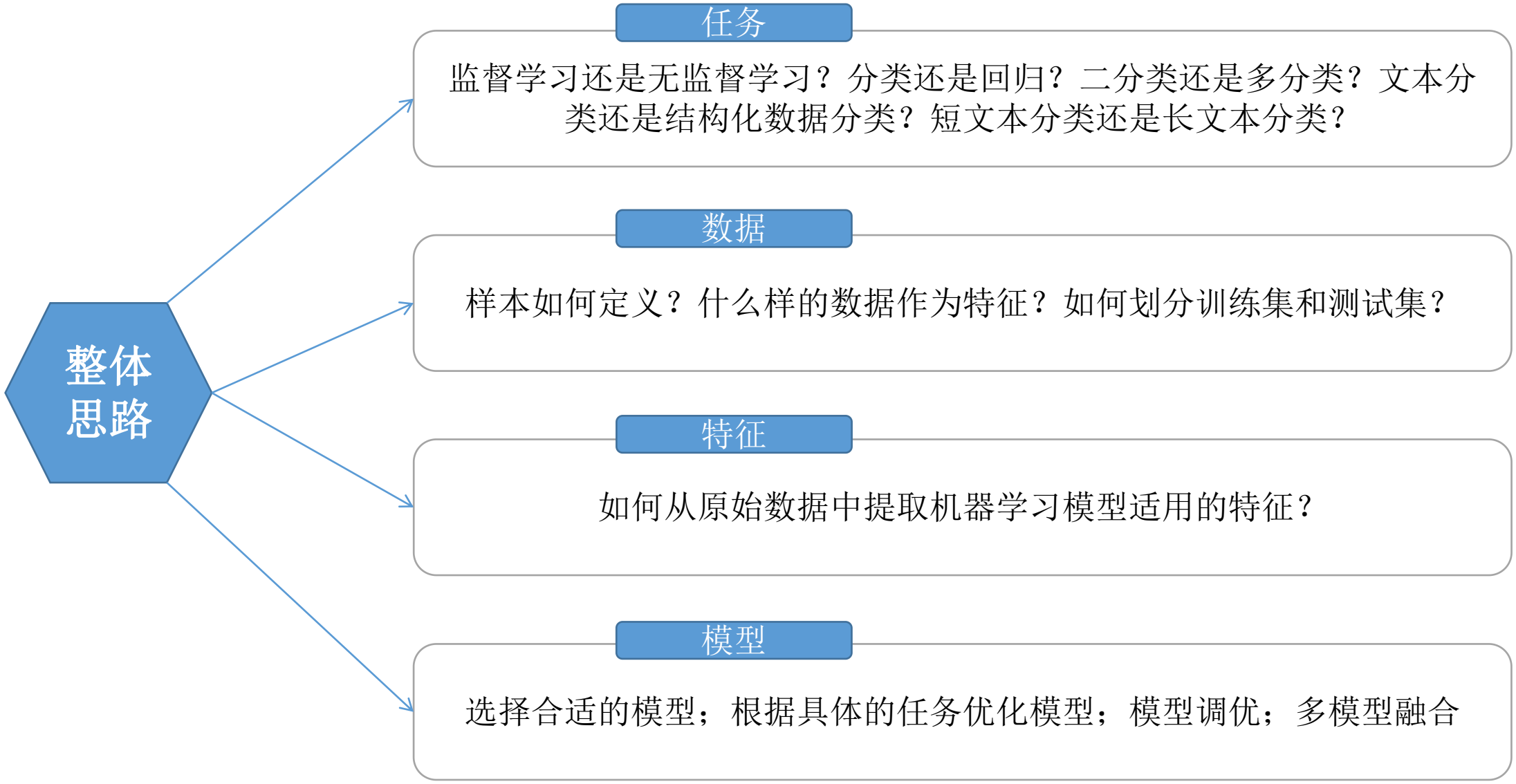
# 垃圾邮件过滤技术方案



- 正确的识别垃圾邮件的技术难度比较大，常用的垃圾邮件过滤方式有：关键词法、校验码法、主题/源Email地址/IP地址/附件审计、白名单/黑名单机制、贝叶斯算法过滤等；
- 其中贝叶斯算法过滤垃圾邮件是一种基于统计学的过滤器，是建立在已有的统计结果之上的，所以贝叶斯算法过滤垃圾邮件模型属于一种有监督的分类算法。基于贝叶斯算法的过滤垃圾邮件也属于一种比较常用的算法模型。

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)}$$

# 算法介绍


整体思路
任务
监督学习还是无监督学习？分类还是回归？二分类还是多分类？文本分类还是结构化数据分类？短文本分类还是长文本分类？
数据
样本如何定义？什么样的数据作为特征？如何划分训练集和测试集？
特征
如何从原始数据中提取机器学习模型适用的特征？
模型
选择合适的模型；根据具体的任务优化模型；模型调优；多模型融合

# 邮件数据格式

| 属性 | 描述 |
| --- | --- |
| 发件人 | 发送邮件的邮箱号码 |
| 收件人 | 接收邮件的邮箱号码 |
| 邮件发送时间 | 发送邮件的时间点 |
| 邮件内容 | 邮件具体内容 |

```
From: =?GB2312?B?1cW6o8TP?= <jian@163.con>
Subject: =?gb2312?B?uavLvtK1zvEutPq/qreixrGjoQ==?=
To: xing@ccert.edu.cn
Content-Type: text/plain;charset="GB2312"
Date: Sun, 14 Aug 2005 10:17:57 +0800
X-Priority: 2
X-Mailer: Microsoft Outlook Express 5.50.4133.2400

尊敬的贵公司(财务/经理)负责人您好!
        我是深圳金海实业有限公司（广州。东莞）等省市有分公司。
我司有良好的社会关系和实力，因每月进项多出项少现有一部分
发票可优惠对外代开税率较低，增值税发票为5%其它国税.地税.
运输.广告等普通发票为1.5%的税点，还可以根据数目大小来衡
量优惠的多少，希望贵公司.商家等来电商谈欢迎合作。

    本公司郑重承诺所用票据可到税务局验证或抵扣!
欢迎来电进一步商谈。
电话：13826556538（24小时服务）
信箱：szlianfen@163.com
联系人：张海南

    顺祝商祺

                深圳市金海实业有限公司
```

# 需求分析



任务分析

基于邮件内容、发件人、发件时间等相关内容，准确地、完整地识别出垃圾邮件和正常邮件

01

邮件数据有类别标签，属于监督学习类型

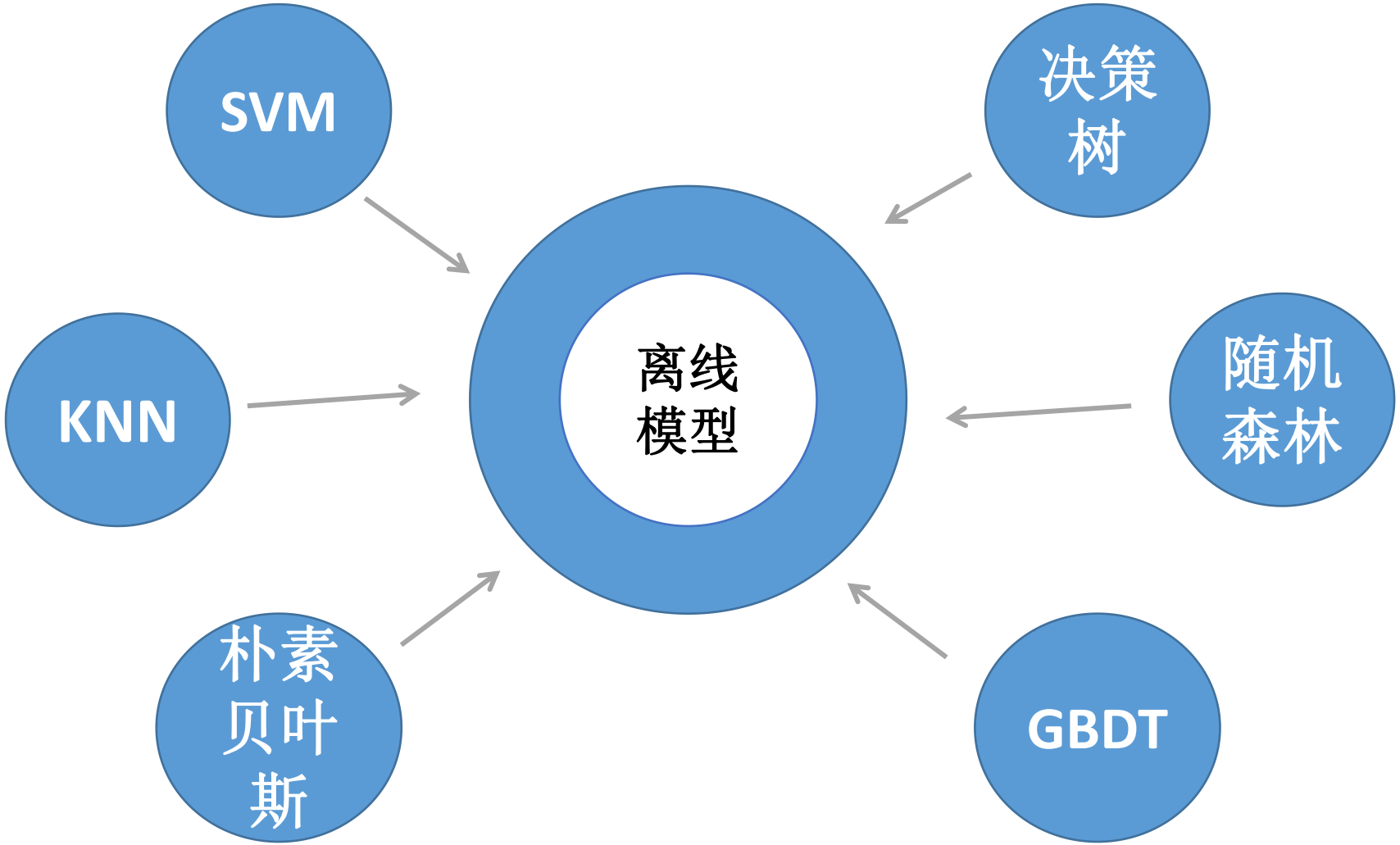
02

二分类：垃圾邮件(1)，正常邮件(0)

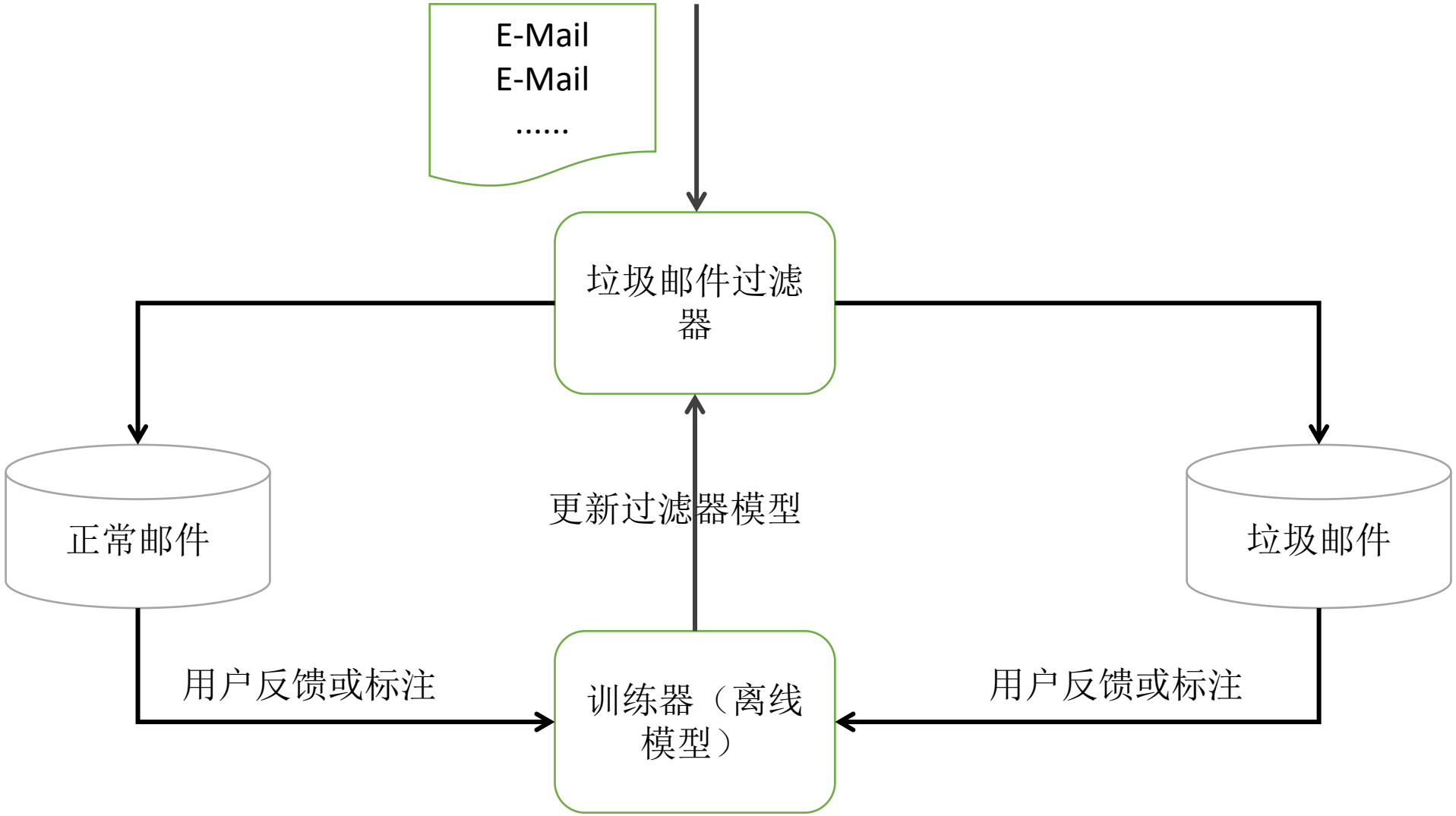
03

主要是非结构化的邮件文本数据，属于长文本分类问题

# 模型选择



# 垃圾邮件过滤技术方案


E-Mail
E-Mail
.....
垃圾邮件过滤器
正常邮件
更新过滤器模型
垃圾邮件
用户反馈或标注
训练器（离线模型）
用户反馈或标注

# 数据清洗



- 从原始数据中，将邮件数据转换称为结构化类型的数据，并且去掉其它不需要的字段信息，只需要保留发件人、收件人、发送时间、邮件内容这四部分的内容，对于这四个字段信息，如果这四个字段为空，那么将为空的属性设置为unknown。

# 特征工程

- 发件人和收件人邮箱服务器提取，如果没有发件人或者收件人的邮件地址的，直接将该字段的值设置为unknown。
- 通过对服务器地址字段的分析，可以得出在最终的算法模型中，该特征属性不需要使用的结论

```
=======to address=======
ccert.edu.cn    64407
unknown           193
yahoo.com.cn        8
163.net             3
quanso.com          2
Name: to_address, dtype: int64
总邮件接收服务器类别数量为:(12,)
```

```
=======from address=======
163.com                  7500
mail.tsinghua.edu.cn     6498
126.com                  5822
tom.com                  4075
mails.tsinghua.edu.cn    3205
Name: from_address, dtype: int64
总邮件发送服务器类别数量为:(3567,)
发送邮件数量小于10封的服务器数量为:(3202, 1)
```

# 特征工程

- 邮件发送时间提取，主要提取出来星期、小时、时间段(上午&下午&晚上&凌晨)等时间的表示字段信息。
- 通过对时间提取字段信息的分析，可以得到时间对于垃圾邮件的分类，作用不大，在后续的模型训练中可以不考虑该字段特征属性。同时从数据上我们也可以看出如果一个邮件没有发送时间，那么一定属于垃圾邮件，所以可以在最终模型中加入这个特征属性。

```
=========星期属性字段的描述=========
fri    10859
sat    10316
thu     9780
Name: date_week, dtype: int64
date_week  label
fri        0.0      3884
           1.0      6975
mon        0.0      2568
           1.0      5491
sat        0.0      3681
```

```
=========小时属性字段的描述=========
19    2835
16    2772
15    2750
Name: date_hour, dtype: int64
date_hour  label
00         0.0       904
           1.0      1716
01         0.0       925
           1.0      1791
02         0.0       868
           1.0      1736
```

```
=========时间段属性字段的描述=========
19    2835
16    2772
15    2750
Name: date_hour, dtype: int64
date_time_quantum  label
0                  0.0      4396
                   1.0      8893
1                  0.0      5756
                   1.0     10570
```

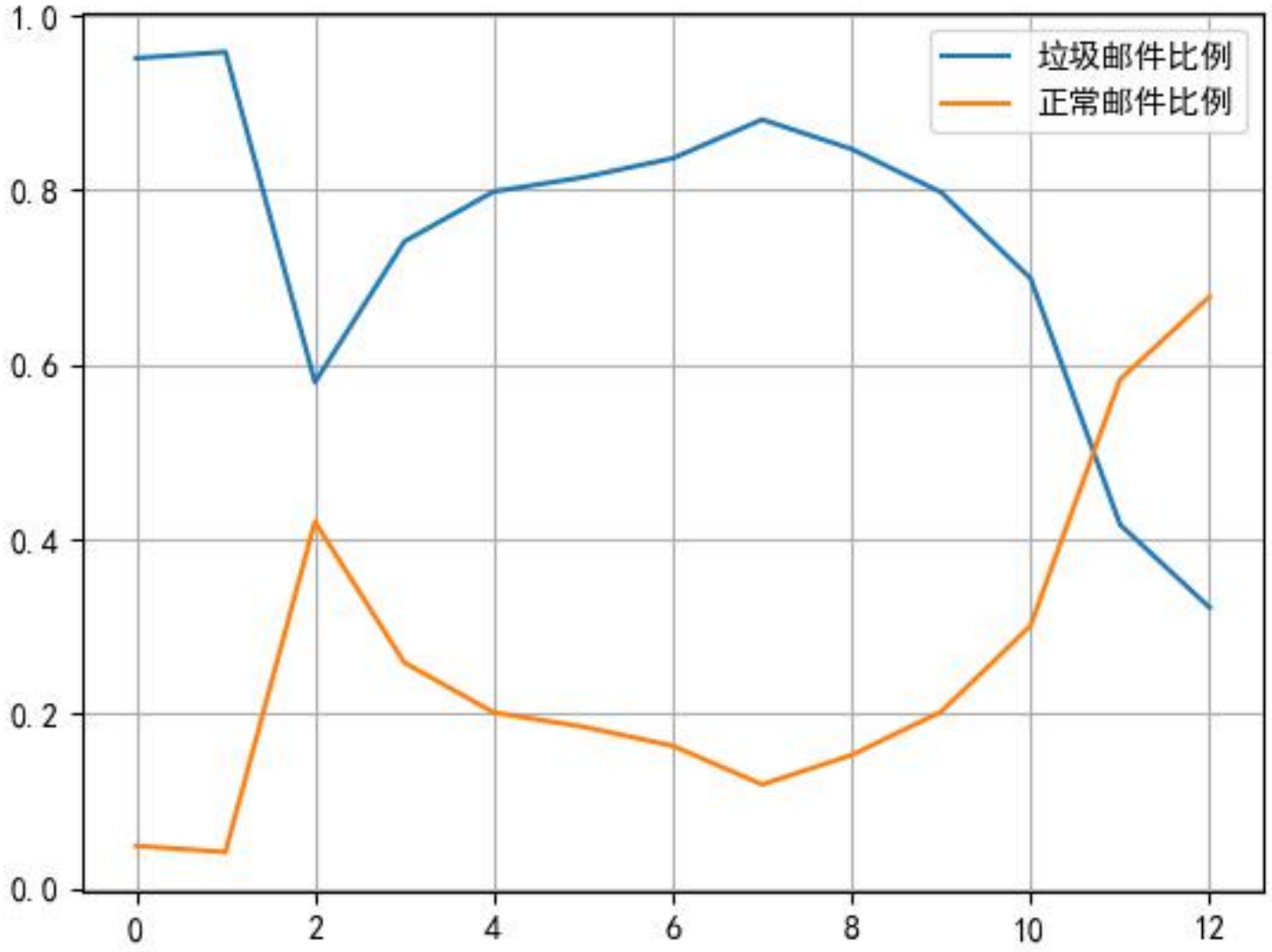
# 特征工程

- 中文分词
  - 利用开源的分词工具jeba分词处理

```
>>> import jieba
>>> l = jieba.cut('我来自北风网')
>>> ' '.join(l)
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\ibf\AppData\Local\Temp\jieba.cache
Loading model cost 1.061 seconds.
Prefix dict has been built succesfully.
'我 来自 北风 网'
```

# 特征工程

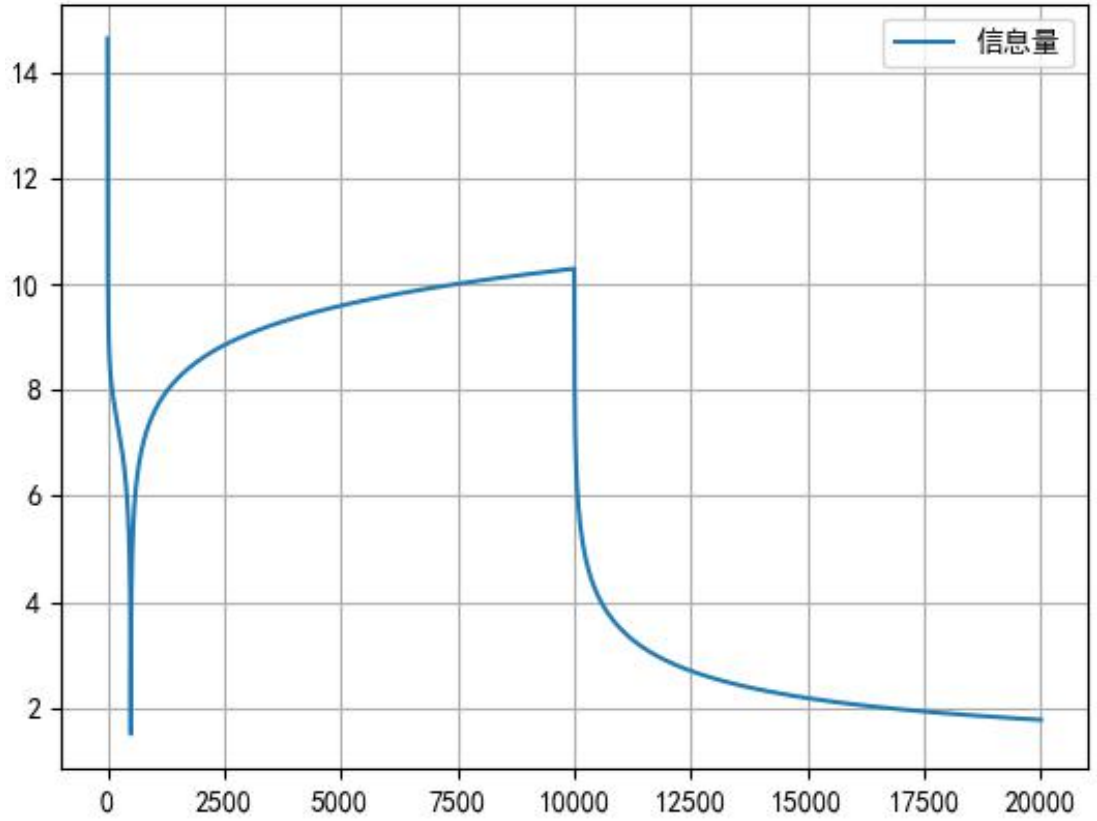
- 信息量特征
  - 正常邮件的内容长度一般都在一定范围内，即不会太长也不会太短；但是一般情况下，邮件的内容越短，那么该邮件就越有可能是垃圾邮件。
  - 信号量：值越大，就越有可能是属于垃圾邮件。



# 特征工程

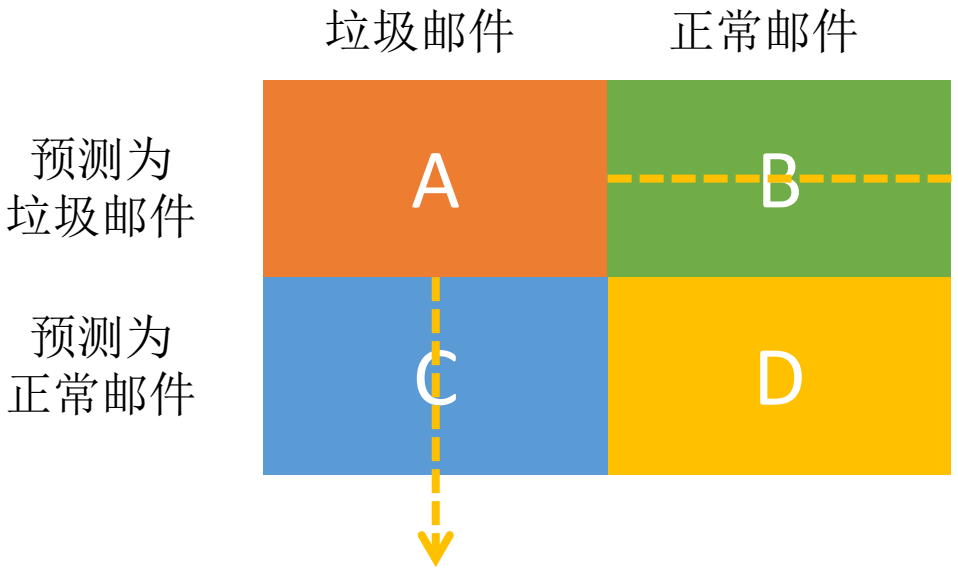
$$信号量=\begin{cases}\dfrac{1}{\exp(\log_{10}(x)-\log_{10}(L_1))*2}+\log_e(|x-L_1|+B_1)-\log_e(x-L_2)+B_2, x>L_2\\ \dfrac{1}{\exp(\log_{10}(x)-\log_{10}(L_1))*2}+\log_e(|x-L_1|+B_1)+B_2, x\le L_2\end{cases}$$

- x表示文本长度
- $L_1$和$L_2$为调节因子，在该项目中，分别设置为500和10000
- $B_1$和$B_2$为信息量平滑因子，在该项目中，全部设置为1



# 模型效果评估

- 在进行垃圾邮件过滤的时候，即需要注意垃圾邮件的拦截率(召回率)，也需要注意正常邮件被当成垃圾邮件的错判率(精确率)，在当前项目中，我们主要考虑召回率这个指标。




$$P(\text{精确率}) = \frac{A}{A+B}$$

$$R(\text{召回率}) = \frac{A}{A+C}$$

$$F1 = \frac{2PR}{P+R}$$

# 模型选择



- 分别选择KNN、SVM、Bayes、DecisionTree、RandomForest、GBDT这几种算法，并比较各种不同算法的效果。

| 算法 | 精确率 | 召回率 | F1 | 模型训练耗时 |
|---|---|---|---|---|
| KNN | **0.97996** | 0.98725 | **0.98359** | 988ms |
| SVM(SVC-rbf) | 0.87646 | 0.98678 | 0.92835 | 3.3min |
| DecisionTree | 0.96911 | 0.98182 | 0.97542 | 635ms |
| RandomForest | 0.94326 | 0.98925 | 0.96571 | 6.26s |
| GBDT | 0.94986 | 0.98866 | 0.96887 | 15.6s |
| Bayes | 0.94667 | **0.98937** | 0.96755 | 58.2ms |

# 总结



- 1. 垃圾邮件过滤一般常用的基础算法有Bayes、KNN、LR等。一般最常用的算法选择Bayes算法。
- 2. 垃圾邮件过滤系统中一般采用算法过滤+其它过滤统计结合的方式来进行垃圾邮件过滤。
- 3. 在垃圾邮件过滤中主要是需要进行分词操作，中文邮件一般可以选择使用jieba(python)、ANSJ(java)等工具进行分词处理。
- 4. 在垃圾邮件过滤中一般注意召回率，也就是说一般情况下，需要尽可能的提高垃圾邮件过滤的成功率。

# 作业

- 1. 修改jieba分词部分的实现逻辑，添加自定义分词词典。
- 2. 所有代码整理(选择一个最优的最终算法模型)，封装成为class或者API的形式。
- 3. 基于上课所将的特征工程提取出来的全部特征信息，使用SVM、GBDT、随机森林、KNN等分类算法，查看一下效果；最终使用GridSearchCV对任意一个模型进行模型参数优化的过程。
- 4. 使用保存好的模型对完整的原始邮件数据做一个判断/预测。(代码)


北风网
IBEIFENG.COM