概率论与数理统计

扩展任务 8 试验报告

GHe

GitHub：https://github.com/GHe0000/SpamEmailBayesClassifier

Build: 2025-04-04

Typst Version: 0.13.1

目录

1 程序设计和原理阐述

1.1 词频-逆文档频率（TF-IDF）

TF-IDF 是一种统计方法，旨在评估某个词对于一个文档的重要程度，其核心思想是：一个词在文档中出现的频率越高（TF），同时在语料库中的常见程度越低（IDF），则认为这个词越能代表文档，其重要性越高.

Definition 1.1.1 (词频(TF)): 词频衡量了某个词在文档中出现的频率，词频越大，词对当前文档的重要性可能越大.

$$\mathrm{TF}(t,d)=\frac{\text{词 t 在文档 d 中出现的次数}}{\text{文档 d 的总词数}}$$

Definition 1.1.2 (逆文档频率(IDF)): 逆文档频率衡量了某个词在整个语料库中的普遍性或稀缺性，逆文档频率越高，词在整个语料库中越稀有，则认为这个词越能代表文档，其重要性越高.

$$\mathrm{IDF}(t,D)=\log\left(\frac{\text{语料库的文档总数 N}+1}{\text{包含词 t 的文档数}+1}\right)$$

注意在上述 IDF 的计算中，并没有使用完全标准的 IDF 计算公式，而是经过了一个所谓的“+1”平滑，这是为了避免文档频率为 0 的“除零”问题，从而增强了数值的稳定性. 同时这种处理本身也类似于 Laplace 平滑，对 IDF 值的影响不大，同时也提升了模型的稳健性.

Remark: 所谓的 *“+1”* 平滑，可以看成是 *Bayes* 学派中的**先验分布**. 当我们通过数据得到频率后，相当于“更新”了这个先验，得到后验. 当数据量足够大时，先验的取值并不会影响最终的结果.

所谓的 TF-IDF，就是将 TF 和 IDF 两个指标综合起来，作为衡量词对于文档的重要性的最终指标. 其计算公式如下：

$$\text{TF-IDF}(t,d,D)=\mathrm{TF}(t,d)\cdot\mathrm{IDF}(t,D)$$

Mark: 这里有一个有意思的小问题：为何 TF 直接是频率和 IDF 经过了 log 变换？

TF 不需要 log 变换很好理解，因为其反映的是词在文档中的**局部信息**，其信息量关于词频是**线性**的. 若某个词出现的次数是另一个词的两倍，则其包含的信息也认为是另一个词的两倍.

但 IDF 不同，一个词出现在文档中越少，则其携带的信息越多，且这种信息的增长量是**非线性**的. 例如一个词在 1 篇文档中出现与在 10 篇文档中出现，其重要性差异远大于在 100 篇文档中出现与在 1000 篇文档中出现的差异. 而 log 正好能将这种非线性的增长量转化为线性的增长量.

从信息论的角度看，一个事件 x 的**自信息**定义为：

$$I(x) = -\log(P(x))$$

其核心思想是：事件发生的概率越低，其发生时携带的信息量越大. 这里 log 将概率的乘法关系转化为信息量的加法关系（例如，独立事件联合概率的信息量为各事件信息量之和）.

因此实际上，IDF 就是“某个词在文档中出现”这一个事件带来的**信息量**.[1]

这里我们将 TF 和 IDF 两个指标相乘，既通过 IDF 抑制了常见词，也通过 TF 强调了文档中的重要词. 因此 TF-IDF 常常作为衡量一个词对于一个文档的重要性的最终指标.

1.2 Bayes 分类器具体算法

现在假设我们有某个邮件 $\boldsymbol{d}$，其包含了 n 个词：$\boldsymbol{d} = [t_1, t_2, ..., t_n]$. 然后我们有训练集 $D = \{\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_m\}$，其中 $\boldsymbol{d}_m$ 是训练集中的邮件. 每一个邮件 $\boldsymbol{d}_m$ 都有一个标签 l_m，表示该邮件是否是垃圾邮件（Spam）或正常邮件（Ham）.

则 Bayes 分类器的训练过程如下：

- **生成词汇表**：从训练数据集 D 中生成词汇表 $V = [v_1, v_2, ..., v_n]$，其中包含了最高频出现的前 n 个词（提前去除停用词），词汇表 V 即 Bayes 分类器的特征空间.
- **计算 IDF**：对于词汇表中的每个词 $v_m \in V$，计算其在训练集 D 中的 IDF 值 $i_m = \mathrm{IDF}(v_i, D)$，从而得到词汇表 V 对应的 IDF 值 $I = [i_1, i_2, ..., i_n]$. 其中 $\mathrm{IDF}(v_i, D)$ 的计算公式如下：

$$\mathrm{IDF}(v_i, D) = \log\left(\frac{m+1}{m_{v_i}+1}\right)$$

其中 m 是训练集中邮件的总数，m_{v_i} 是词 v_i 在训练集中被包含的邮件数.

[1] 这里在 Shannon 的论文 The Mathematical Theory of Communication 中对于信息均默认取以 2 为底的对数，但实际上 log 的底数是可以任取的，这里取 10 为底.

- **计算先验**：对于每个标签 $l_m \in L$，计算先验概率 $P(l_m)$. 这里我们用频率来估计先验概率，即 $P(l_m) = m_l/m$，其中 m_l 是训练集中标签为 l_m 的邮件的数量，m 是训练集中邮件的总数.
- **计算条件概率**：对于每个标签 $l_m \in L$，提取出训练集中所有为 l_m 的邮件 D_{l_m}

2 程序实现和结果分析

2.1 源码代码及其实现原理