朴素贝叶斯

最适合简单的文本分析的算法

朴素贝叶斯 — 核心思想

请在此分隔线以上回复内容

Email

您提交的 # 3152号工单：来自于李先生的留言 有更新。

请点击以下链接查看工单处理进度：
https://tingyun.kf5.com/hc/request/view/3152/
要添加另外的工单评论，请回复此邮件。

正常 垃圾

在垃圾邮件里经常出现**“链接”，“点击”**这种单词。

Intuition：

假如一个邮件里包含了这些单词，这个邮件很可能是垃圾邮件。

一个简单的例子： 垃圾邮件分类

Part 1: 朴素贝叶斯核心

垃圾邮件分类 (监) supervised

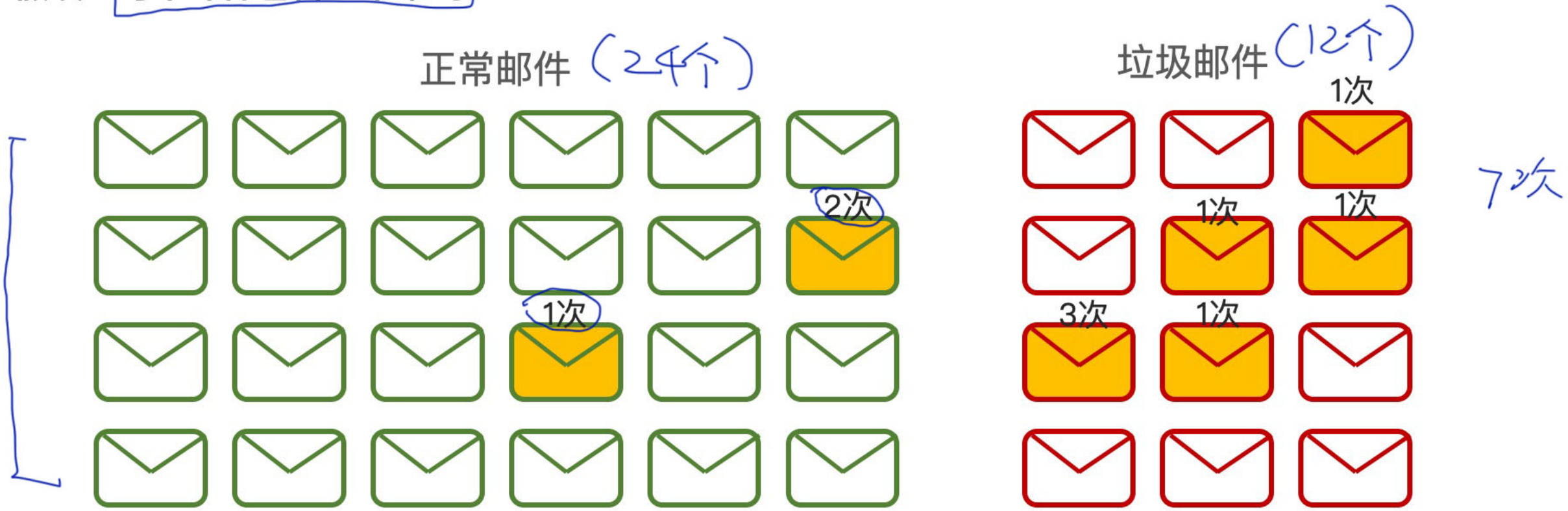
正常邮件

垃圾邮件

Q: 新的邮件，属于哪一个分类?

对于“购买”单词

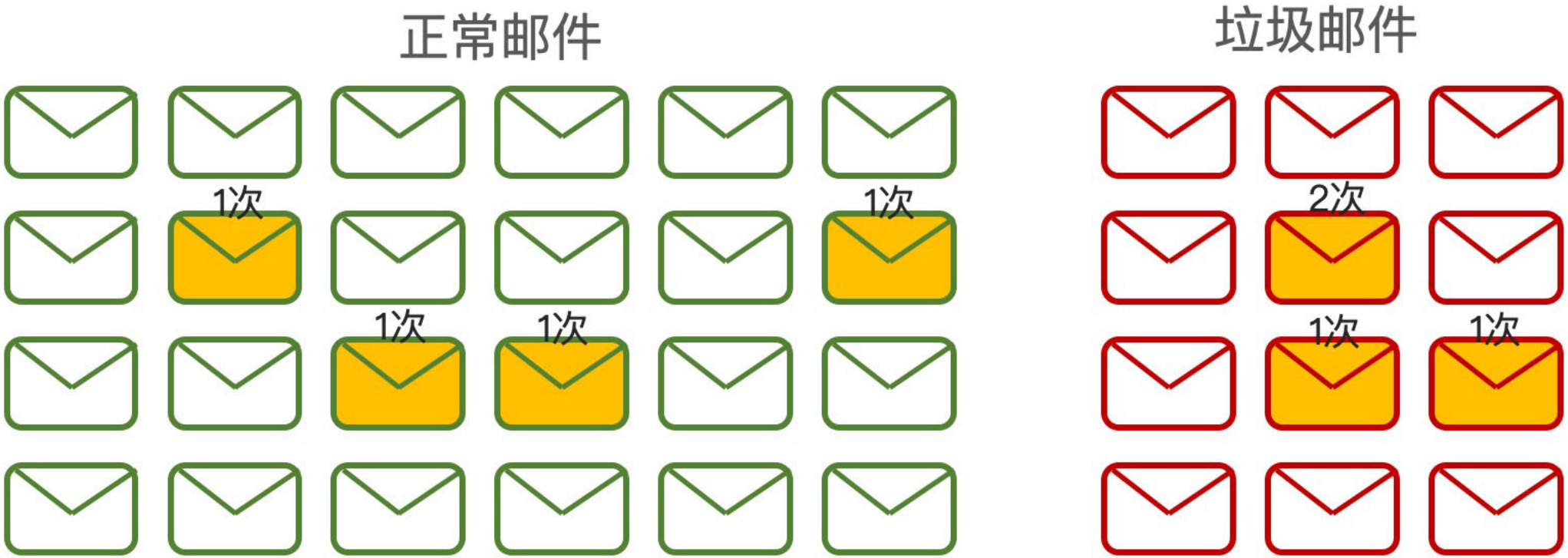
假设：每个邮件包含10个单词



正常邮件含有“购买”词的概率多少？ p(“购买”|正常) = 3/(24*10) = 1/80

垃圾邮件含有“购买”词的概率多少？ p(“购买”|垃圾) = 7/(12*10) = 7/120

对于“物品”单词



正常邮件含有“物品”词的概率多少? $P(\text{物品} \mid \text{正常}) = \frac{4}{24\times10} = \frac{1}{60}$

垃圾邮件含有“物品”词的概率多少? $P(\text{物品} \mid \text{垃圾}) = \frac{4}{12\times10} = \frac{1}{30}$

对于“不是”单词

正常邮件含有“不是”词的概率多少? $P(\text{不是}|\text{正常}) = \frac{4}{240} = \frac{1}{60}$

垃圾邮件含有“不是”词的概率多少? $P(\text{不是}|\text{垃圾}) = \frac{3}{120} = \frac{1}{40}$

对于“广告”单词

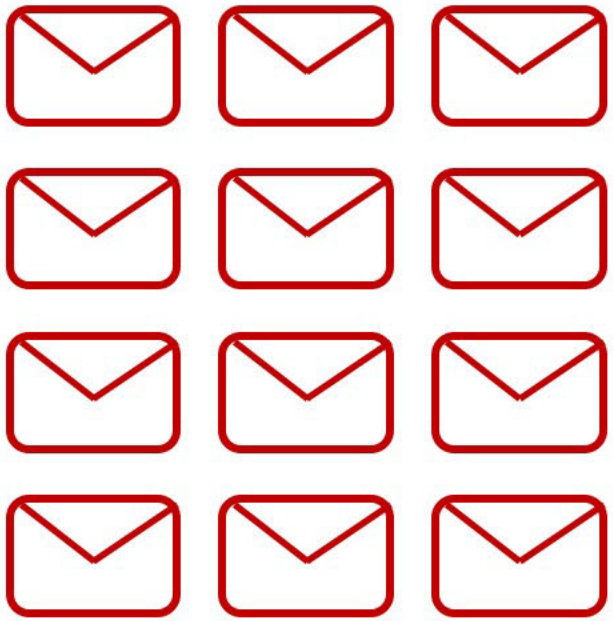
正常邮件含有“广告”词的概率多少? $P(广告|正常) = \frac{5}{240} = \frac{1}{48}$

垃圾邮件含有“广告”词的概率多少? $P(广告|垃圾) = \frac{4}{120} = \frac{1}{30}$

还有一个概率需要统计 (prior) 先验

正常邮件 (24)

垃圾邮件 (12)



有多少邮件是正常邮件（百分比）？ $\frac{24}{24+12} = \frac{2}{3}$

有多少邮件是垃圾邮件（百分比）？ $\frac{12}{24+12} = \frac{1}{3}$

利用朴素贝叶斯识别垃圾邮件

前面计算出了很多零零散散的概率，怎么整合这些信息来完成识别任务？

P(单词i | 正常)
P(单词i | 垃圾)
P(正常)、P(垃圾)

已知（计算过）

从概率统计的角度

P(垃圾|邮件内容)：一个邮件内容为垃圾邮件的概率

P(正常|邮件内容)：一个邮件内容为正常邮件的概率

邮件

如何做判断？

如果 P(垃圾|邮件内容) > P(正常|邮件内容)，则可以认为是垃圾邮件

如果 P(垃圾|邮件内容) <= P(正常|邮件内容)，则可以认为是正常邮件

P(垃圾 | 内容)
vs
P(正常 | 内容)

新的问题：P(垃圾|邮件内容)， P(正常|邮件内容)怎么计算？？？

回顾：贝叶斯定理

$$P(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)}$$

$\Leftrightarrow$ Important!

$$p(x,y) = p(x|y) \cdot p(y)$$

$$= p(y|x) \cdot p(x)$$

正常邮件　　垃圾邮件

新邮件

是垃圾邮件还是正常邮件?

Bayes

$$P(\text{正常} \mid \text{邮件内容}) = \frac{P(\text{邮件内容}|\text{正常}) * p(\text{正常})}{p(\text{邮件内容})}$$

近似

Conditional Independence

$$= \frac{P(\text{“购买”}, \text{“物品”}, \text{“不是”}, \text{“广告”}|\text{正常}) * p(\text{正常})}{p(\text{邮件内容})}$$

$$= \frac{P(\text{“购买”}|\text{正常}) * P(\text{“物品”}|\text{正常}) * P(\text{“不是”}|\text{正常}) * P(\text{“广告”}|\text{正常}) p(\text{正常})}{p(\text{邮件内容})}$$

$$= \frac{\frac{1}{80} \cdot \frac{1}{60} \cdot \frac{1}{60} \cdot \frac{1}{48} \cdot \frac{2}{3}}{\cancel{P(\text{邮件内容})}} \quad ①$$

$$\Rightarrow P(\text{垃圾} \mid \text{邮件内容}) = \frac{P(\text{邮件内容}|\text{垃圾}) * p(\text{垃圾})}{p(\text{邮件内容})}$$

$$= \frac{\frac{7}{120} \cdot \frac{1}{30} \cdot \frac{1}{40} \cdot \frac{1}{30} \cdot \frac{1}{3}}{\cancel{P(\text{邮件内容})}} \quad ②$$

① < ②

∴ 垃圾邮件

P(“购买”|正常) = $\frac{1}{80}$　　P(“购买”|垃圾)= $\frac{7}{120}$

P(“物品”|正常) = $\frac{1}{60}$　　P(“物品”|垃圾)= $\frac{1}{30}$

P(“不是”|正常) = $\frac{1}{60}$　　P(“不是”|垃圾)= $\frac{1}{40}$

P(“广告”|正常) = $\frac{1}{48}$　　P(“广告”|垃圾)= $\frac{1}{30}$

P(正常) = $\frac{2}{3}$　P(垃圾) = $\frac{1}{3}$

200-300 怎么处理概率为0的情况?

4个

$$P(\text{正常} \mid \text{邮件内容}) = \frac{P(\text{“购买”}\mid\text{正常}) * P(\text{“物品”}\mid\text{正常}) * P(\text{“不是”}\mid\text{正常}) * P(\text{“广告”}\mid\text{正常})\, p(\text{正常})}{p(\text{邮件内容})}$$

=0

Add-one smoothing

某些单词没有出现在正常邮件(训练) ⟹ 平滑(smoothing)

$P(\text{购买}\mid\text{正常}) = 0$

$$P(\text{购买}\mid\text{正常}) = \frac{\#\text{ of “购买”} + 1}{\#\text{ of words (total)} + V}$$

词库的大小

手推一个完整的例子

$P(正常|新邮件) = \frac{\frac{2}{15}\cdot\frac{1}{15}\cdot\frac{1}{15}\cdot\frac{1}{15}\cdot\frac{1}{2}}{P(新邮件)}$

垃圾邮件 (8)

1. 点击 更多 信息
2. 最新 产品
3. 信息 点击 链接

正常邮件 (6)

1. 开会
2. 信息 详见 邮件
3. 最新 信息

训练数据

新邮件

(最新 产品 ~~实惠~~ 点击 链接)

属于正常邮件还是垃圾邮件?

词库 = {点击, 更多, 信息, 最新, 产品, 链接, 开会, 详见, 邮件} $\boxed{V=9}$

① 条件概率

$P(点击|垃圾) = \frac{2+1}{8+9} = \frac{3}{17}$

$P(更多|垃) = \frac{1+1}{8+9} = \frac{2}{17}$

$P(信息|垃) = \frac{2+1}{8+9} = \frac{3}{17}$

$P(最新|垃) = \frac{1+1}{8+9} = \frac{2}{17}$

$P(产品|垃) = \frac{1+1}{8+9} = \frac{2}{17}$

$P(链接|垃) = \frac{1+1}{8+9} = \frac{2}{17}$

$P(开会|垃) = \frac{0+1}{8+9} = \frac{1}{17}$

$P(详见|垃) = \frac{0+1}{8+9} = \frac{1}{17}$

$P(邮件|垃) = \frac{0+1}{8+9} = \frac{1}{17}$

$P(点击|正常) = \frac{0+1}{6+9} = \frac{1}{15}$

$P(更多|正常) = \frac{0+1}{6+9} = \frac{1}{15}$

$P(信息|正常) = \frac{2+1}{6+9} = \frac{3}{15}$

$P(最新|正常) = \frac{1+1}{6+9} = \frac{2}{15}$

$P(产品|正常) = \frac{0+1}{6+9} = \frac{1}{15}$

$P(链接|正常) = \frac{0+1}{6+9} = \frac{1}{15}$

$P(开会|正常) = \frac{1+1}{6+9} = \frac{2}{15}$

$P(详见|正) = \frac{2}{15}$

$P(邮件|正) = \frac{2}{15}$

② 先验

$P(正常) = \frac{1}{2}$

$P(垃圾) = \frac{1}{2}$

A. 训练过程

手推一个完整的例子

利用严格的数学来表示朴素贝叶斯过程

给定一个文本向量 $x = (x_1, \dots, x_N)$, 以及两种类型标签 $y = \{0,1\}$,

其中N为词典库的大小，θ为模型的参数，其中θ_{i0} 和θ_{i1}

分别表示第i个参数出现在分类0和分类1的概率。x_i表示词典中第i个单词出现的次数。

Coding Time–1

Part 2: 文本表示

单词的表示 对

词典：[我们，去，爬山，今天，你们，昨天，跑步]

7个单词

每个单词的表示：

我们：(1, 0, 0, 0, 0, 0, 0) $\in R^7$

爬山：(0, 0, 1, 0, 0, 0, 0) $\in R^7$

跑步：(0, 0, 0, 0, 0, 0, 1) $\in R^7$

昨天：(0, 0, 0, 0, 0, 1, 0) $\in R^7$

句子的表示（boolean） vector

词典：［我们，又，去，爬山，今天，你们，昨天，跑步］

每个句子的表示

我们 今天 去 爬山：（1, 0, 1, 1, 1, 0, 0, 0）

你们 昨天 跑步：（0, 0, 0, 0, 0, 1, 1, 1）

你们 又 去 爬山 又 去 跑步：（0, 1, 1, 1, 0, 1, 0, 1）

句子的表示（count） vector

词典：［我们，又，去，爬山，今天，你们，昨天，跑步］

每个句子的表示

我们 今天 去 爬山：(1, 0, 1, 1, 1, 0, 0, 0)

你们 昨天 跑步：(0, 0, 0, 0, 0, 1, 1, 1)

你们 又 去 爬山 又 去 跑步：(0, 2, 2, 1, 0, 1, 0, 1)

句子的表示

句子1：He is going from Beijing to Shanghai

句子2：He denied my request, but he actually lied.

句子3：Mike lost the phone, and phone was in the car

句子1：(0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0)

句子2：(1, 0, 0, 1, 0, $\boxed{1}$, 0, 0, $\boxed{2}$, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0)

句子3：(0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 2, 0, 0, 2, 0, 1)

denied he

这种表示有什么缺点？

句子的表示

词典

句子1：He is going from Beijing to Shanghai

句子2：He denied my request, but he actually lied.

句子3：Mike lost the phone, and phone was in the car

count

句子1：(0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0)

句子2：(1, 0, 0, 1, 0, 1, 0, 0, 2, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0)

句子3：(0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 2, 0, 0, 2, 0, 1)

denied he

a, an, the

①并不是出现的越多就越重要！

②并不是出现的越少就越不重要！

$w_d = (0, 1, 2, 1, \ldots)$

次数

Tf-idf表示

次数

$$tfidf(w) = tf(d, w) * idf(w)$$

重要性

文档d中w的词频

$$\log \frac{N}{N(w)}$$

文档的总数 $N=100$

w 在多少个文档

$N(w)=10$

N: 语料库中的文档总数

N(w): 词语w出现在多少个文档?

log

$$tfidf(w) = tf(d,w) * idf(w)$$

词典：{今天，上，机器学习，课程，的，有，意思，数据，也}

$\boxed{V=9}$

1 今天 上 机器学习 课程

2 今天 的 课程 有 意思

3 数据 课程 也 有 意思

$N=3$

句子1 $\Rightarrow (1\cdot\log\frac{3}{2}, 1\cdot\log\frac{3}{1}, 1\cdot\log\frac{3}{1}, 1\cdot\log\frac{3}{3}, 0, 0, 0, 0, 0)$

$\Rightarrow (\log 1.5, \log 3, \log 3, \log 1, 0, 0, 0, 0, 0)$

句子2 $\Rightarrow (1\cdot\log\frac{3}{2}, 0, 0, 1\cdot\log\frac{3}{3}, 1\cdot\log\frac{3}{1}, 1\cdot\log\frac{3}{2}, 1\cdot\log\frac{3}{2}, 0, 0)$

$\Rightarrow (\log 1.5, 0, 0, \log 1, \log 3, \log 1.5, \log 1, 0, 0)$

$idf = \log\frac{N+V}{N(w)+1}$

句子3 $\Rightarrow$ (? Easy)

Why? log：

$N=10000$

$N(\text{的}) = 2$, $N(\text{运动}) = 1000$, $idf(\text{的}) = 5000$, $idf(\text{运动}) = 10$

Coding Time–1

Part 3: Extensions（可选）

1. 当特征为实数型的时候

$$P(\text{购买} \mid \text{正常}) = \frac{\#\text{ of 购}}{\#\text{total words}}$$

坏

词子 ⇒ (0, 1, 2, 3, 2, 0, 0)

word2Vec

词子 ⇒ (0.1, 0.2, 0.3, 0.1)

Gaussian NB

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

2. 为什么叫“朴素”？— 条件独立

“朴素” + 贝叶斯

$$P(\text{正常} \mid \text{邮件内容}) = \frac{P(\text{邮件内容} \mid \text{正常}) * p(\text{正常})}{p(\text{邮件内容})}$$

条件独立

$$= \frac{P(\text{“购买”}, \text{“物品”}, \text{“不是”}, \text{“广告”} \mid \text{正常}) * p(\text{正常})}{p(\text{邮件内容})}$$

$$= \frac{P(\text{“购买”} \mid \text{正常}) * P(\text{“物品”} \mid \text{正常}) * P(\text{“不是”} \mid \text{正常}) * P(\text{“广告”} \mid \text{正常}) p(\text{正常})}{p(\text{邮件内容})}$$

3. 朴素贝叶斯的最大似然

$p(x|y)=$

$$p(D) = \prod_{i=1}^{N} p(x^i, y^i) = \prod_{i=1}^{N} p(x^i|y^i)p(y^i)$$

$$= \prod_{i=1}^{N} p(x_1^i, x_2^i, ..., x_{m_i}^i|y^i)p(y^i)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{m_i} p(x_j^i|y^i)p(y^i)$$

https://zhuanlan.zhihu.com/p/71960086

4. 生成模型与判别模型

目标函数

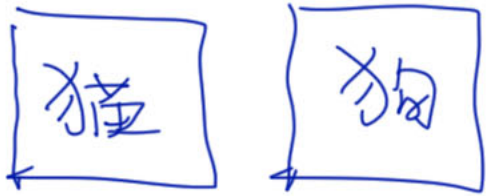
逻辑回归

$$\boxed{\prod_{i=1}^{m} p(y_i|x_i)}$$

条件

⇓

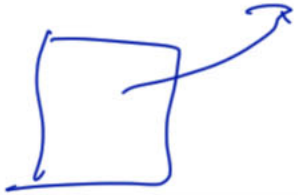
判别



VS

① 猫 VS 狗

区别



朴素贝叶斯

$$\boxed{\prod_{i=1}^{m} p(x_i, y_i)}$$

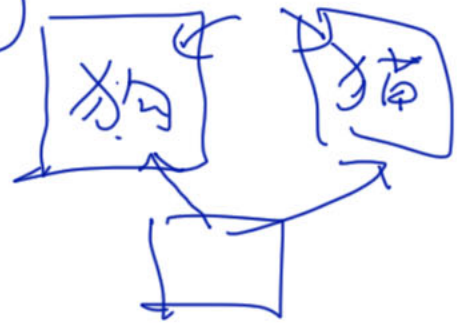
+ 联合

⇓

生成 ⇒ 数据

GAN (图片, 代码, 文章, 音乐)

②



$$P(x_1 x_2 x_3 x_4 \mid y=1) = P(x_1 \mid y=1) \cdot P(x_2 \mid y=1) \cdot P(x_3 \mid y=1)\, P(x_4 \mid y=1)$$

$$P(x_1 x_2 x_3 x_4 \mid y=0) = P(x_1 \mid y=0)\, P(x_2 \mid y=0)\, P(x_3 \mid y=0)\, P(x_4 \mid y=0)$$

$$\text{underflow} \leftarrow \overset{0.01}{(0.01)^{20}} \Rightarrow 0 \qquad \text{inf}$$

$$\log\left(P(x_1 \mid y=1) \cdot P(x_2 \mid y=1)\, P(x_3 \mid y=1)\, P(x_4 \mid y=1) \right)$$

$$= \underline{\log P(x_1 \mid y=1)} + \underline{\log P(x_2 \mid y=1)} + \underline{\log P(x_3 \mid y=1)} + \cdots$$

似然：likelihood

样本的 | Model (θ)

厂家

$\boxed{\text{coin} \quad p(\text{正面}) = \theta}$

未知世界

样本

正 反

$\Rightarrow$ H H T H H T $\quad \theta = \frac{2}{3}$

$p(\overbrace{HHTHHT} \mid \theta)$

最大化（最大似然概率）

likelihood (似然)

$$P(x|y) = \frac{P(y|x)\cdot P(x)}{P(y)}$$

$P(x)$ → prior (先验)

$P(x|y)$ → 后验 (posterior)

$P(y)$ → Normalization (很难计算)

条件独立: $P(x_1, x_2, x_3 \cdots x_n | y)$

$$= P(x_1|y)\cdot P(x_2|y) \cdots P(x_n|y)$$

$$= \prod_{i=1}^{n} P(x_i|y)$$

$$P(x_1, x_2) = P(x_1)\cdot P(x_2)$$

x_1, x_2 独立的