

文本分类的改进

更有效的特征提取

小胖

目录

ONE 前情回顾

伯努利模型

TWO 多项式模型

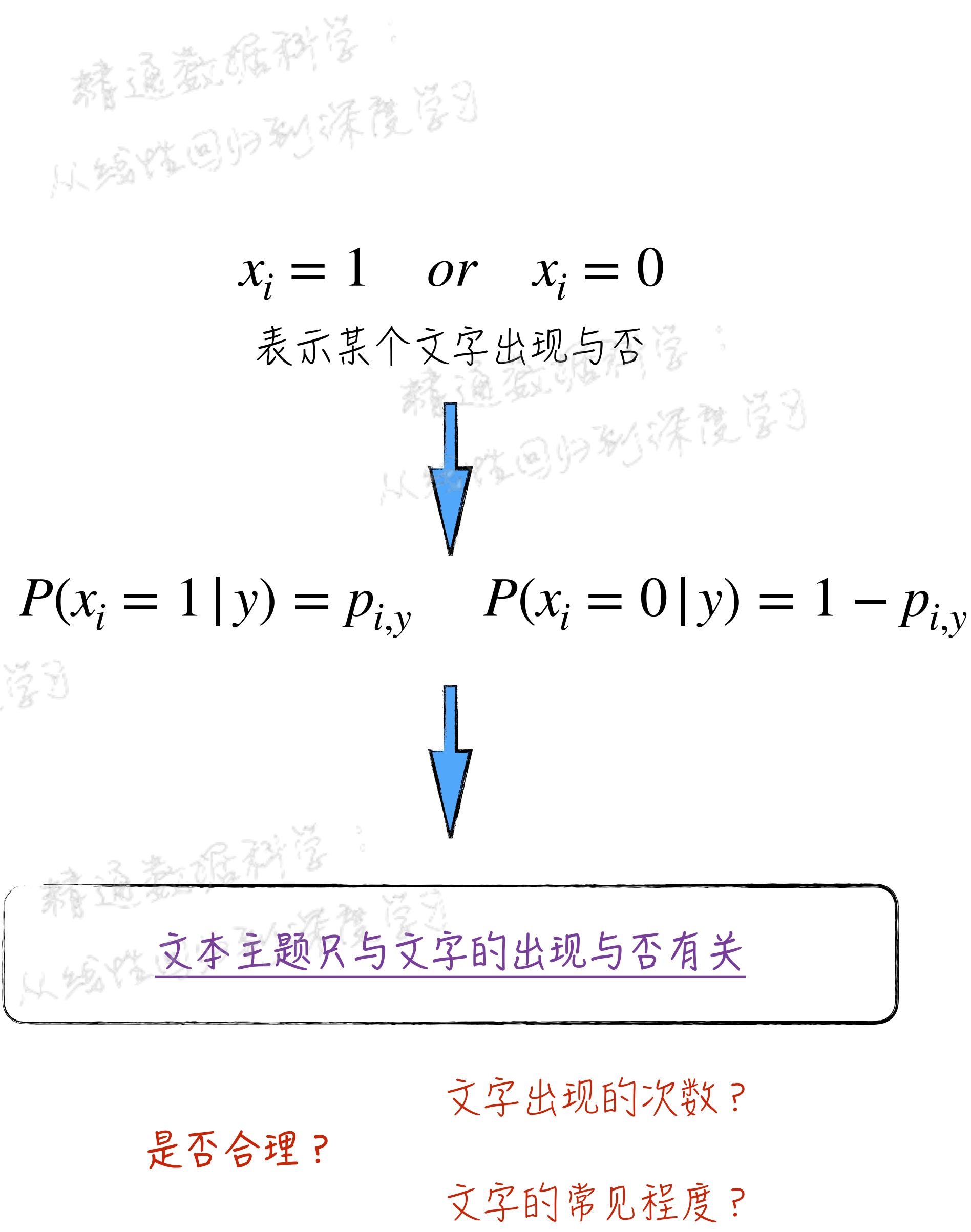
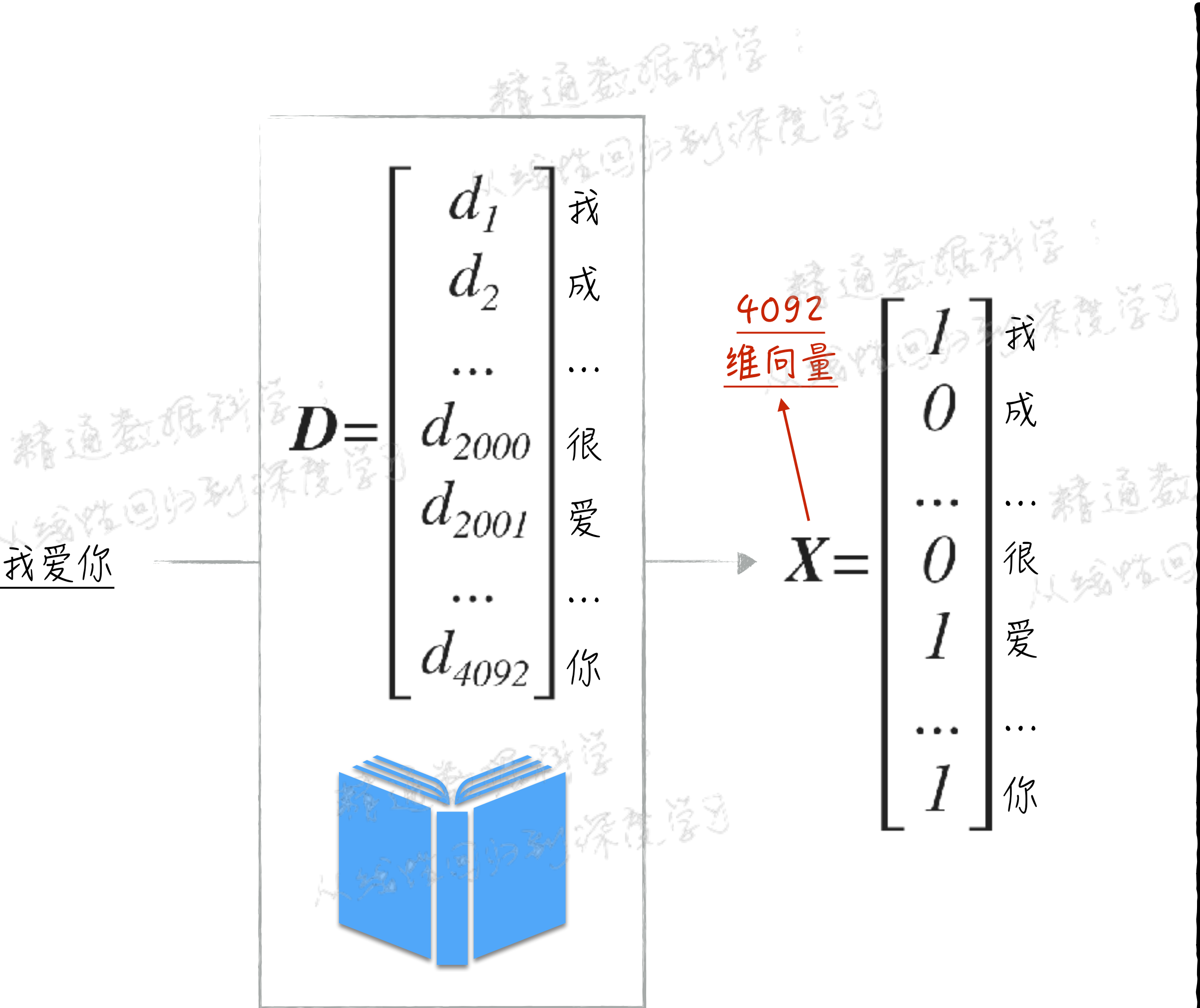
用出现次数来衡量权重

THREE TF-IDF

更有效地衡量文字权重

前情回顾

伯努利模型



前情回顾

伯努利模型

$$x_i = 1 \quad or \quad x_i = 0$$

表示某个文字出现与否



$$P(x_i = 1 | y) = p_{i,y} \quad P(x_i = 0 | y) = 1 - p_{i,y}$$



文本主题只与文字的出现与否有关

是否合理？

多项式模型



文字出现的次数？

+

文字的常见程度？

TF-IDF
+
多项式模型

目录

ONE 前情回顾

伯努利模型

TWO 多项式模型

用出现次数来衡量权重

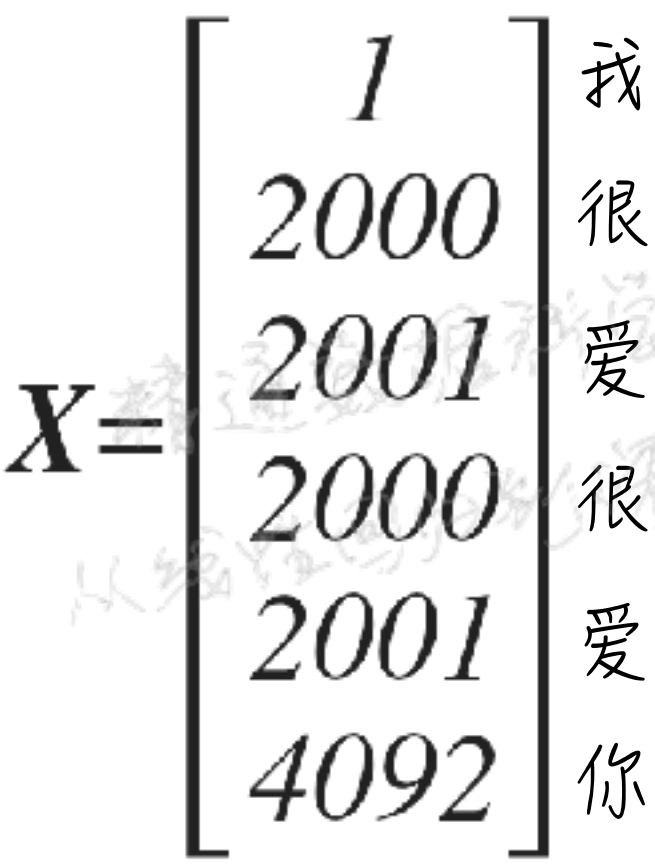
THREE TF-IDF

更有效地衡量文字权重

多项式模型

理论推导

我很爱很爱你



第*i*位置出现第*k*个字的概率

$$P(x_i = k | y) = p_{k,y}$$

文本类别的分布

$$P(y = l) = \theta_l$$

$$L_i = P(\mathbf{X}_i, y_i) = P(\mathbf{X}_i | y_i) P(y_i) = \prod_{j=1} P(x_{i,j} | y_i) P(y_i)$$

在某类别下，“文字”出现次数的比例

$$\hat{p}_{k,l} = \frac{\sum_{i,j} 1_{\{x_{i,j}=k, y_i=l\}}}{\sum_k \sum_{i,j} 1_{\{x_{i,j}=k, y_i=l\}}}$$


$$\hat{\theta}_l = \frac{\sum_{i=1}^m 1_{\{y_i=l\}}}{m}$$

文本类别出现的比例

多项式模型

出现次数代替是否出现

我很爱
很爱你

$$D = \begin{bmatrix} d_1 \\ d_2 \\ \dots \\ d_{2000} \\ d_{2001} \\ \dots \\ d_{4092} \end{bmatrix}$$


我
成
...
很
爱
...
你

4092
维向量

$$X = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 2 \\ 2 \\ \dots \\ 1 \end{bmatrix}$$

我
成
...
很
爱
...
你

$$\mathbf{X}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n})$$

$$\hat{\theta}_l = \frac{\sum_{i=1}^m 1_{\{y_i=l\}}}{m}$$

$$\hat{p}_{k,l} = \frac{\sum_{i,j} 1_{\{x_{i,j}=k, y_i=l\}}}{\sum_k \sum_{i,j} 1_{\{x_{i,j}=k, y_i=l\}}} \rightarrow \hat{p}_{k,l} = \frac{\sum_{i=1}^m x_{i,k} 1_{\{y_i=l\}}}{\sum_k \sum_{i=1}^m x_{i,k} 1_{\{y_i=l\}}}$$

$$P(\mathbf{X}_i | y) = \prod_{j=1}^n P(x_{i,j} | y) \qquad \hat{P}(\mathbf{X}_i | y = l) = \prod_k \hat{p}_{k,l}^{x_{i,k}}$$

如果理论上使用这
样的特征提取

$$P(x_{i,j} | y_i) = ?$$

无法定义

多项式模型

一个简单的例子

| 文本 | 类别 |
|--------|----|
| 我很爱很爱你 | 正面 |
| 我喜欢你 | 正面 |
| 我不爱你 | 负面 |

multinomial
naive Bayes

类别概率分布

$$P(\text{正面}) = 2/3$$

$$P(\text{负面}) = 1/3$$

特征条件概率

$$P(\text{我} \mid \text{正面}) = 2/10$$

$$P(\text{我} \mid \text{负面}) = 1/4$$

$$P(\text{很} \mid \text{正面}) = 2/10$$

$$P(\text{很} \mid \text{负面}) = 0/4$$

$$P(\text{爱} \mid \text{正面}) = 2/10$$

$$P(\text{爱} \mid \text{负面}) = 1/4$$

$$P(\text{你} \mid \text{正面}) = 2/10$$

$$P(\text{你} \mid \text{负面}) = 1/4$$

$$P(\text{喜} \mid \text{正面}) = 1/10$$

$$P(\text{喜} \mid \text{负面}) = 0/4$$

$$P(\text{欢} \mid \text{正面}) = 1/10$$

$$P(\text{欢} \mid \text{负面}) = 0/4$$

$$P(\text{不} \mid \text{正面}) = 0/10$$

$$P(\text{不} \mid \text{负面}) = 1/4$$

一个很简单的例子

(不考虑分词以及平滑项)

加入平滑项:

$$\hat{p}_{k,l} = \frac{\sum_{i=1}^m x_{i,k} 1_{\{y_i=l\}} + \alpha}{\sum_k \sum_{i=1}^m x_{i,k} 1_{\{y_i=l\}} + n\alpha}$$

n为字典大小

目录

ONE 前情回顾

伯努利模型

TWO 多项式模型

用出现次数来衡量权重

THREE TF-IDF

更有效地衡量文字权重

TF-IDF

更有效地衡量文字权重

文字对文本类别的影响

文字出现次数越多，与
文本类别强相关

TF

$$TF_{i,k} = \frac{x_{i,k}}{\sum_k x_{i,k}}$$

IDF

$$IDF_k = \ln \frac{m}{\sum_i 1_{\{x_{i,k} > 0\}}}$$

$$TFIDF_{i,k} = TF_{i,k} IDF_k$$

文字在其他文本也经常出
现，则与文本类别弱相关

TF-IDF

在多项式模型里使用TF-IDF

在多项式模型里使用TF-IDF:

- 理解参数估计公式的含义
- 直接修改参数估计公式

虽然损失了理论的优雅性，但提升了模型效果

更合理的权重定义

$$TFIDF_{i,k} = TF_{i,k} IDF_k$$

出现次数作为文字权重

$$\hat{p}_{k,l} = \frac{\sum_{i=1}^m x_{i,k} 1_{\{y_i=l\}}}{\sum_k \sum_{i=1}^m x_{i,k} 1_{\{y_i=l\}}}$$

多项式模型 + TF-IDF

THANK YOU

精通数据科学：
从线性回归到深度学习