

文本分类的初体验

伯努利模型

小胖

目录

ONE 前情回顾

文本的特征提取

TWO 伯努利模型简介

模型假设与参数估计

THREE 对生僻字的工程处理

平滑项

前情回顾

文本的特征提取

$$P(\mathbf{X}|y) = P(x_1, x_2, \dots, x_n | y) = \prod_{i=1}^n P(x_i | y)$$

之前的建模场景

建模对象本身是数字化的

特征提取

使用模型

只能处理数字

文本分类的场景

文本并没有数字化

感性到理性

数字化

特征提取

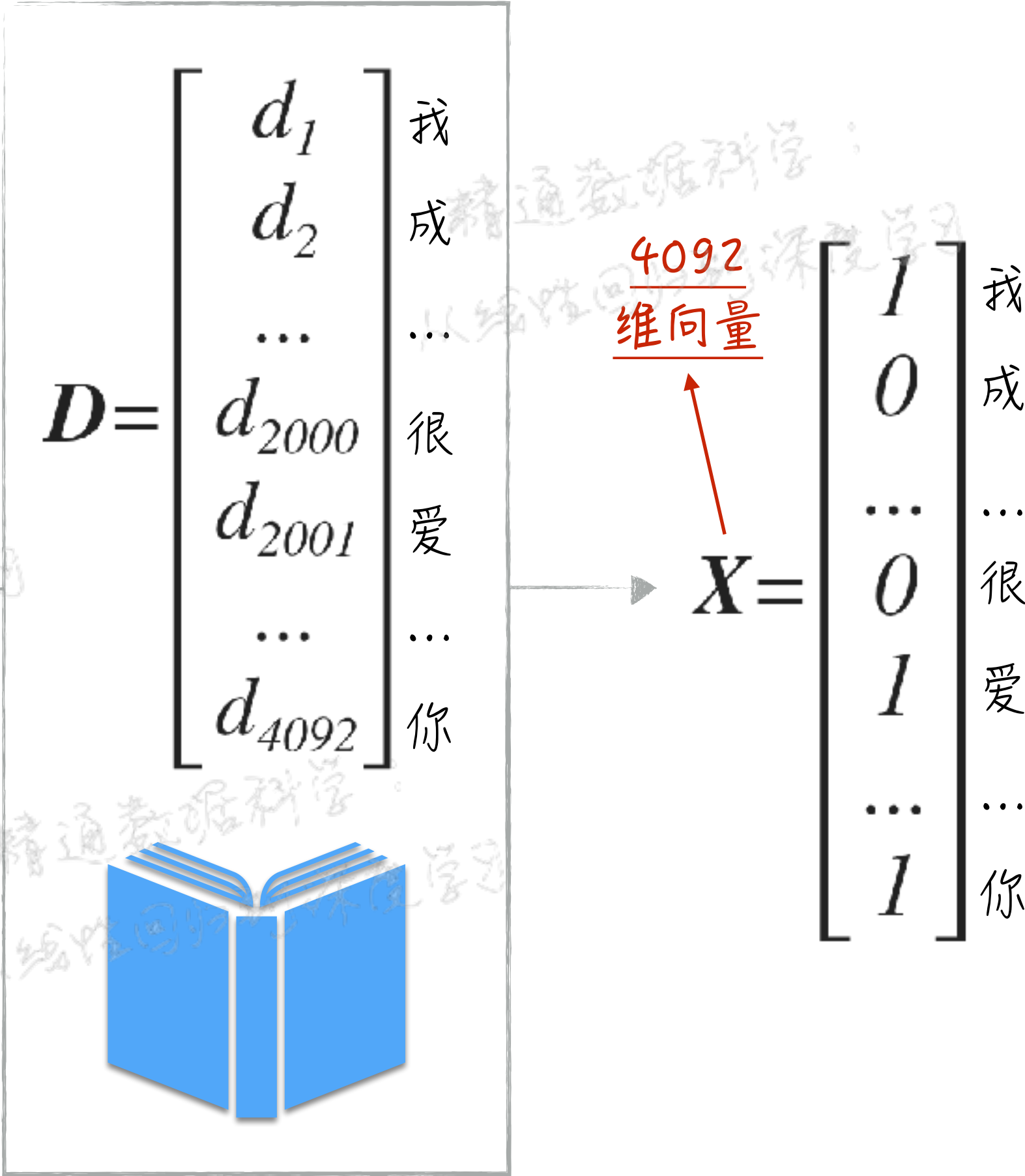
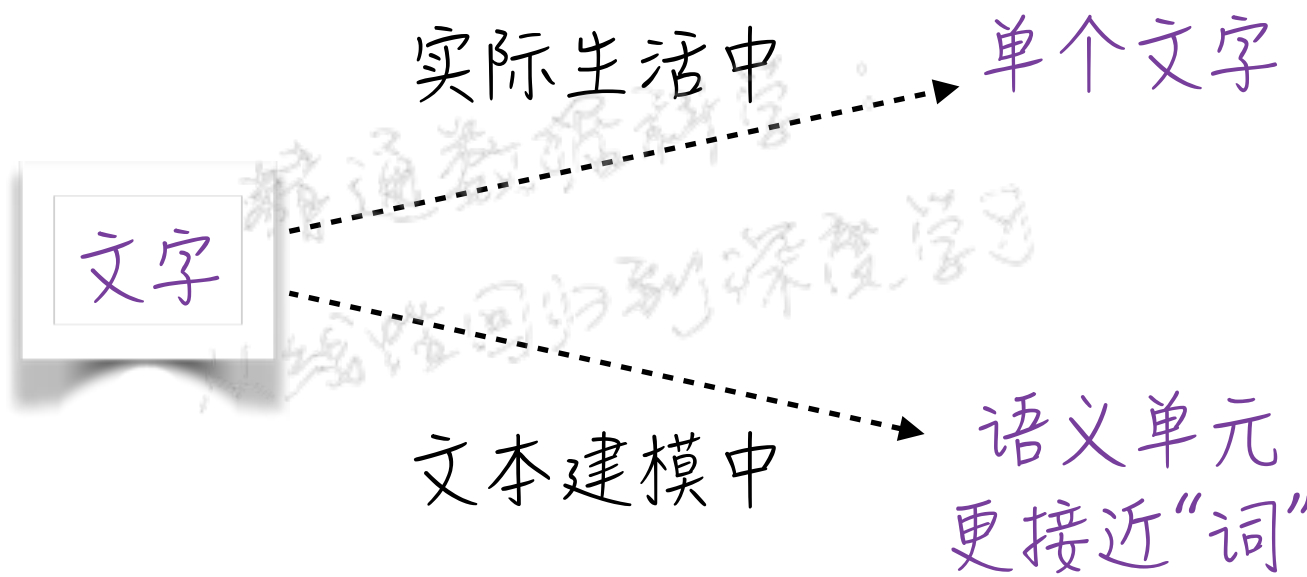
使用模型

前情回顾

文本的特征提取

将文字转换为数字最直接的方法是利用字典进行提取：

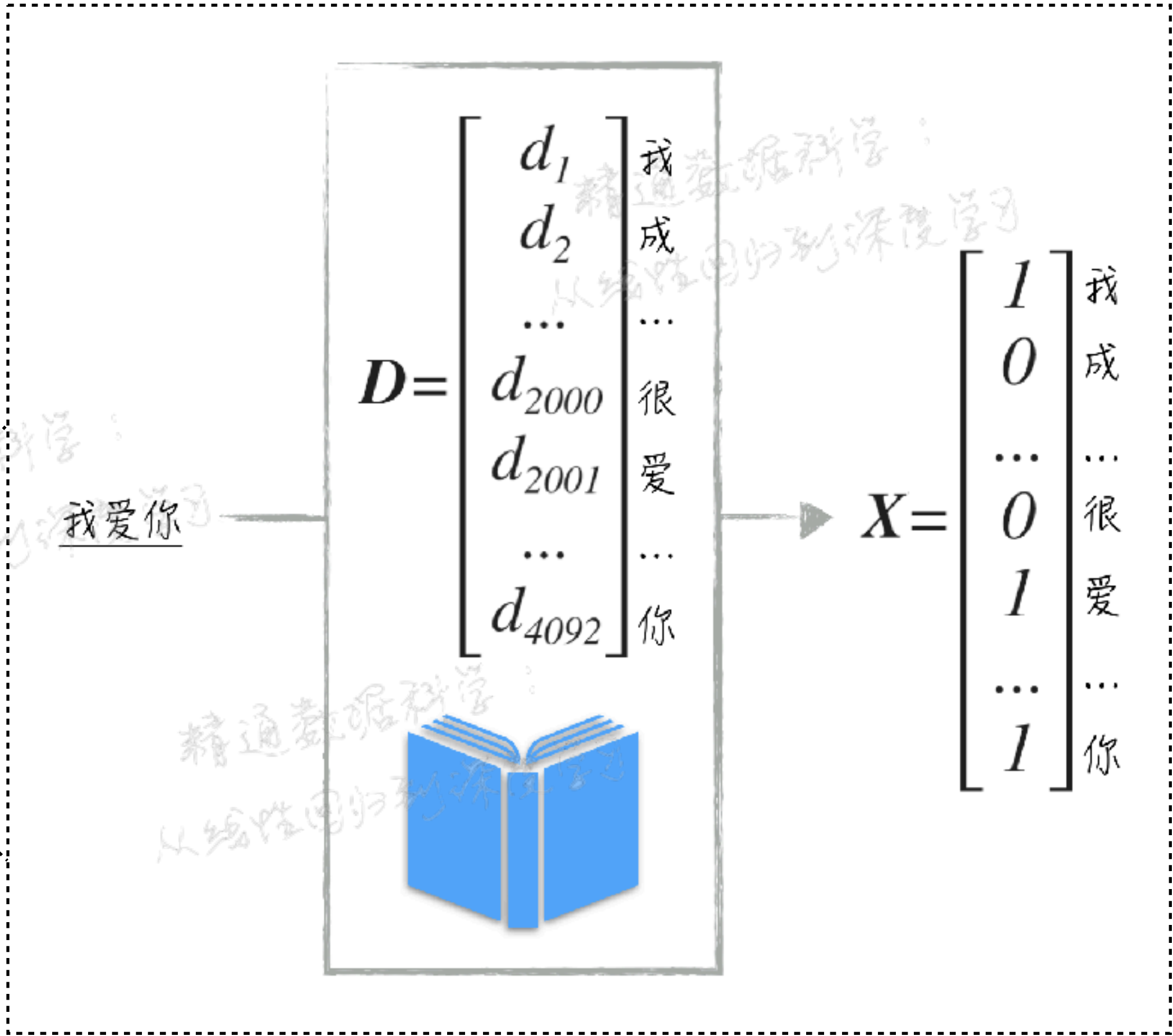
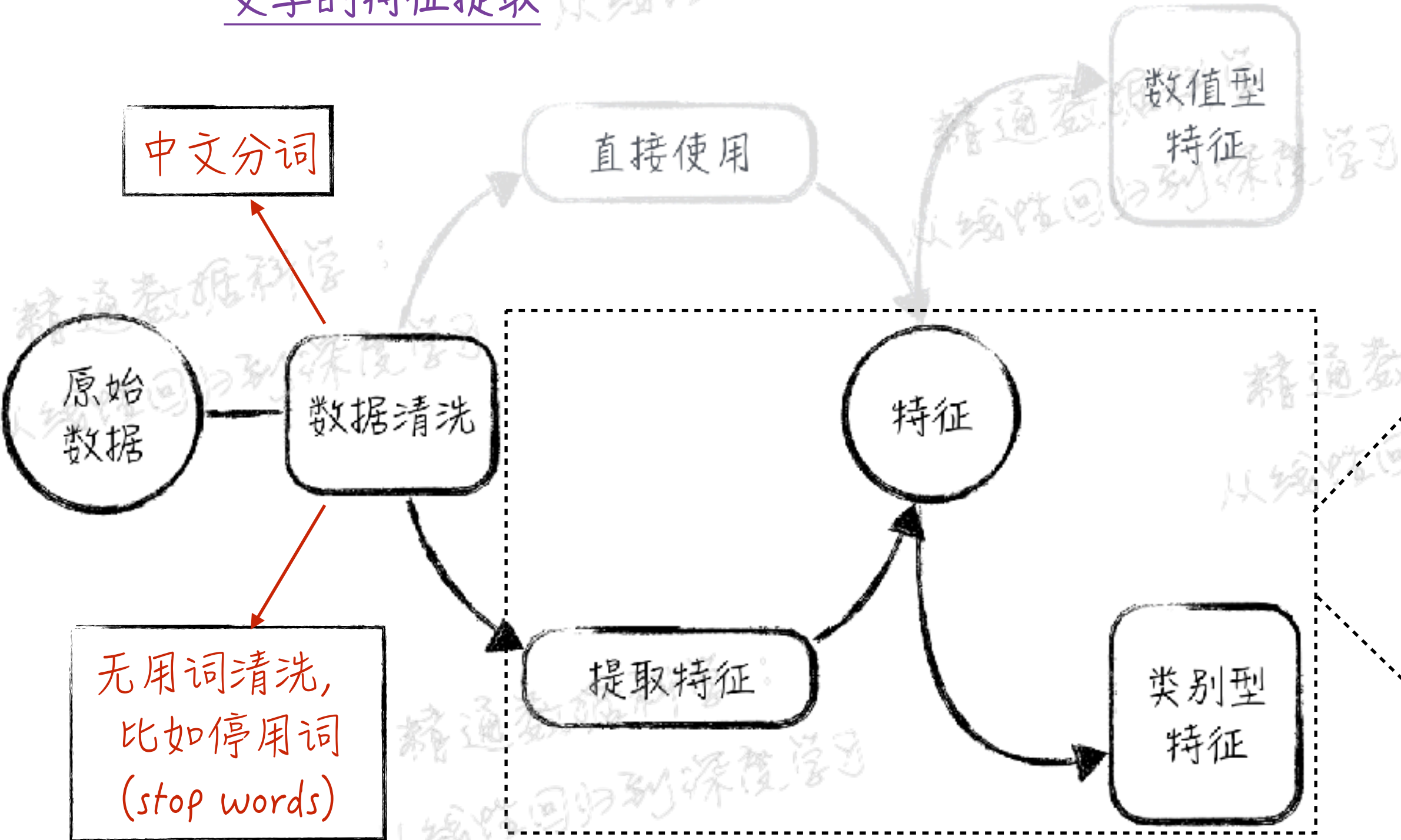
- 将所有文字进行排序，形成一个 **n维** 的字典
- 每一个文字对应一个变量 d_i
- 用一个 **n维** 的向量来表示文本：当文字出现在文本里，则相应的变量等于1，否则等于0



前情回顾

文本的特征提取

文字的特征提取



目录

ONE 前情回顾

文本的特征提取

TWO 伯努利模型简介

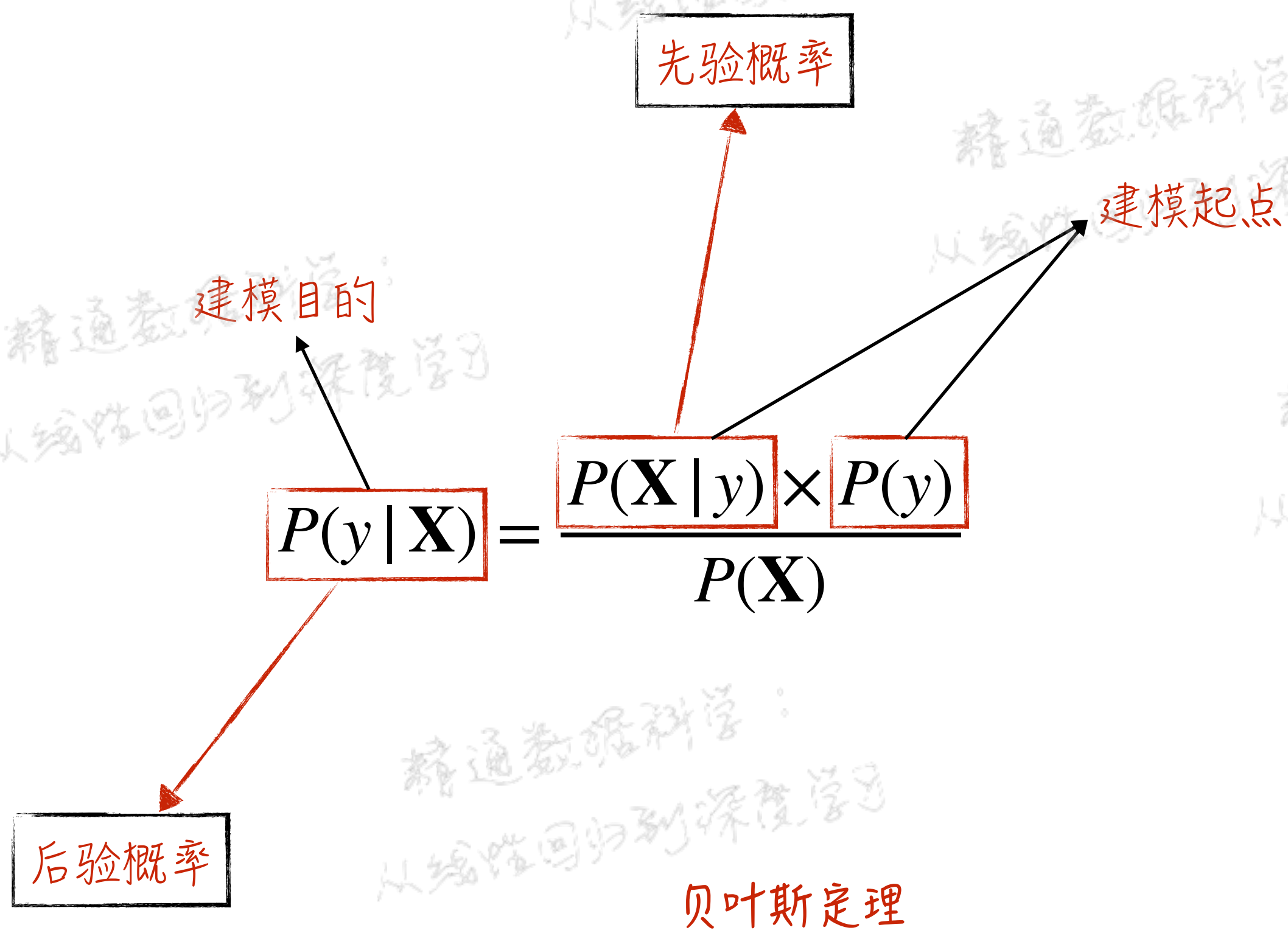
模型假设与参数估计

THREE 对生僻字的工程处理

平滑项

伯努利模型简介

模型假设



朴素贝叶斯模型的假设：

- 在给定类别下，各特征相互独立

$$P(\mathbf{X} | y) = P(x_1, x_2, \dots, x_n | y) = \prod_{i=1}^n P(x_i | y)$$

伯努利模型假设

× 只有两种取值 $x_i = 1 \quad or \quad x_i = 0$

“字”出现的概率
随类别的不同而不同

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

文本类别的分布 $P(y = l) = \theta_l$

伯努利模型简介

进一步推导

伯努利模型假设

×只有两种取值

$$x_i = 1 \quad \text{or} \quad x_i = 0$$

“字”出现的概率

$$P(x_i = 1 | y) = p_{i,y}$$

随类别的不同而不同

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

文本类别的分布

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

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

$$P(x_{i,j} | y_i) = x_{i,j}p_{j,y_i} + (1 - x_{i,j})(1 - p_{j,y_i})$$

$$L = \prod_i L_i \xrightarrow{\text{最大似然估计法}} \hat{\theta}_l, \hat{p}_{j,l} = \operatorname{argmax}_{\theta,p} L$$

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

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

文本类别出现的比例

在某类别下，“文字”出现的比例

最终预测公式

$$\hat{y}_i = \operatorname{argmax}_l \prod_j \hat{p}_{j,l} \hat{\theta}_l$$

伯努利模型简介

一个简单的例子

文本	类别
我爱你	正面
我喜欢你	正面
我不爱你	负面

Bernoulli 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

$$P(\text{我} \mid \text{正面}) = 2/2$$

$$P(\text{我} \mid \text{负面}) = 1/1$$

$$P(\text{爱} \mid \text{正面}) = 1/2$$

$$P(\text{爱} \mid \text{负面}) = 1/1$$

$$P(\text{你} \mid \text{正面}) = 2/2$$

$$P(\text{你} \mid \text{负面}) = 1/1$$

$$P(\text{喜} \mid \text{正面}) = 1/2$$

$$P(\text{喜} \mid \text{负面}) = 0/1$$

$$P(\text{欢} \mid \text{正面}) = 1/2$$

$$P(\text{欢} \mid \text{负面}) = 0/1$$

$$P(\text{不} \mid \text{正面}) = 0/2$$

$$P(\text{不} \mid \text{负面}) = 1/1$$

一个很简单的例子

(不考虑分词以及平滑项)

目录

ONE 前情回顾

文本的特征提取

TWO 伯努利模型简介

模型假设与参数估计

THREE 对生僻字的工程处理

平滑项

对生僻字的工程处理

平滑项

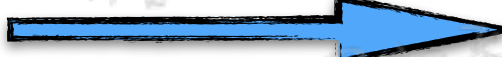
伯努利模型的参数估计

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

文本类别出现的比例

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

在某类别下，“文字”出现的比例



当有生僻字（没有出现在训练文本中）
出现在预测文本时

$$\hat{p}_{j,l} = 0$$

无法得到预测结果

加入平滑项

最终预测公式

$$\hat{y}_i = \operatorname{argmax}_l \prod_j \hat{p}_{j,l} \hat{\theta}_l$$

$$\hat{p}_{j,l} = \frac{\sum_{i=1}^m 1_{\{x_{i,j}=1, y_i=l\}} + \alpha}{\sum_{i=1}^m 1_{\{y_i=l\}} + 2\alpha}$$

$$0 < \alpha \leq 1$$

THANK YOU

精通数据挖掘科学：
从线性回归到深度学习