文本分类的初体验

伯努利模型

小胖

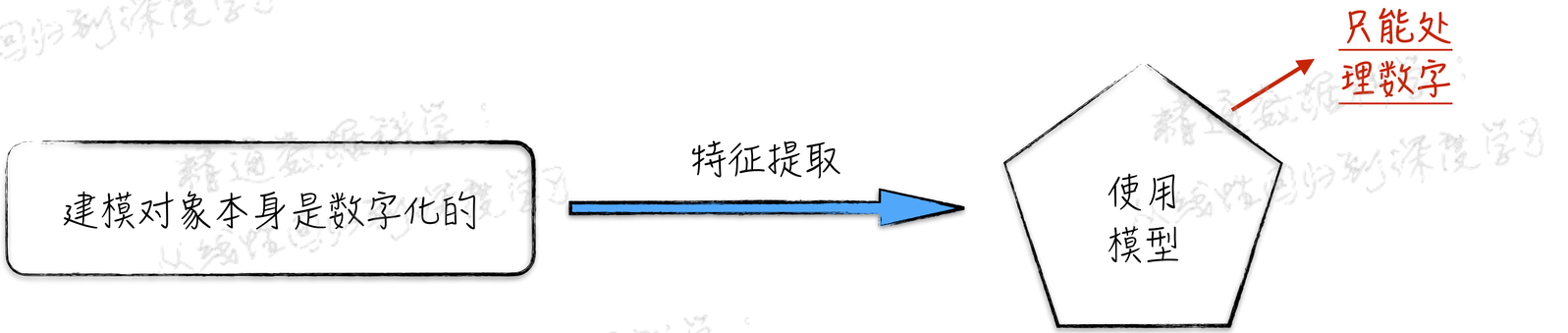
目录

前情回顾

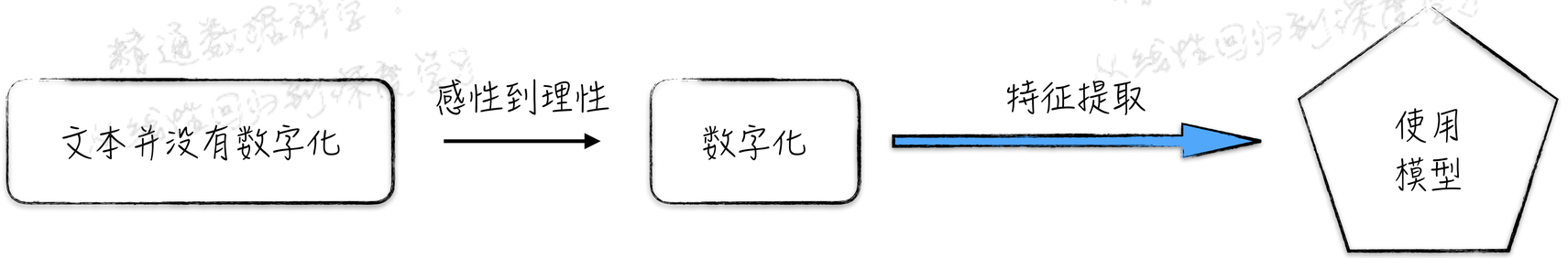
文本的特征提取

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

之前的建模场景



文本分类的场景

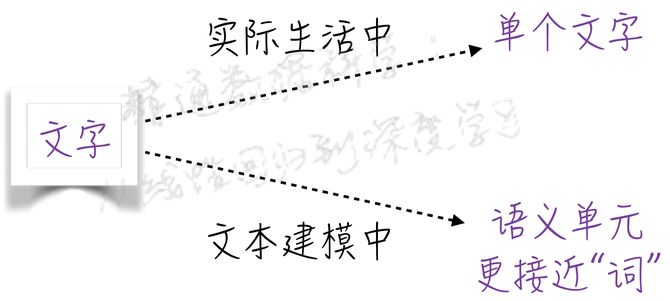


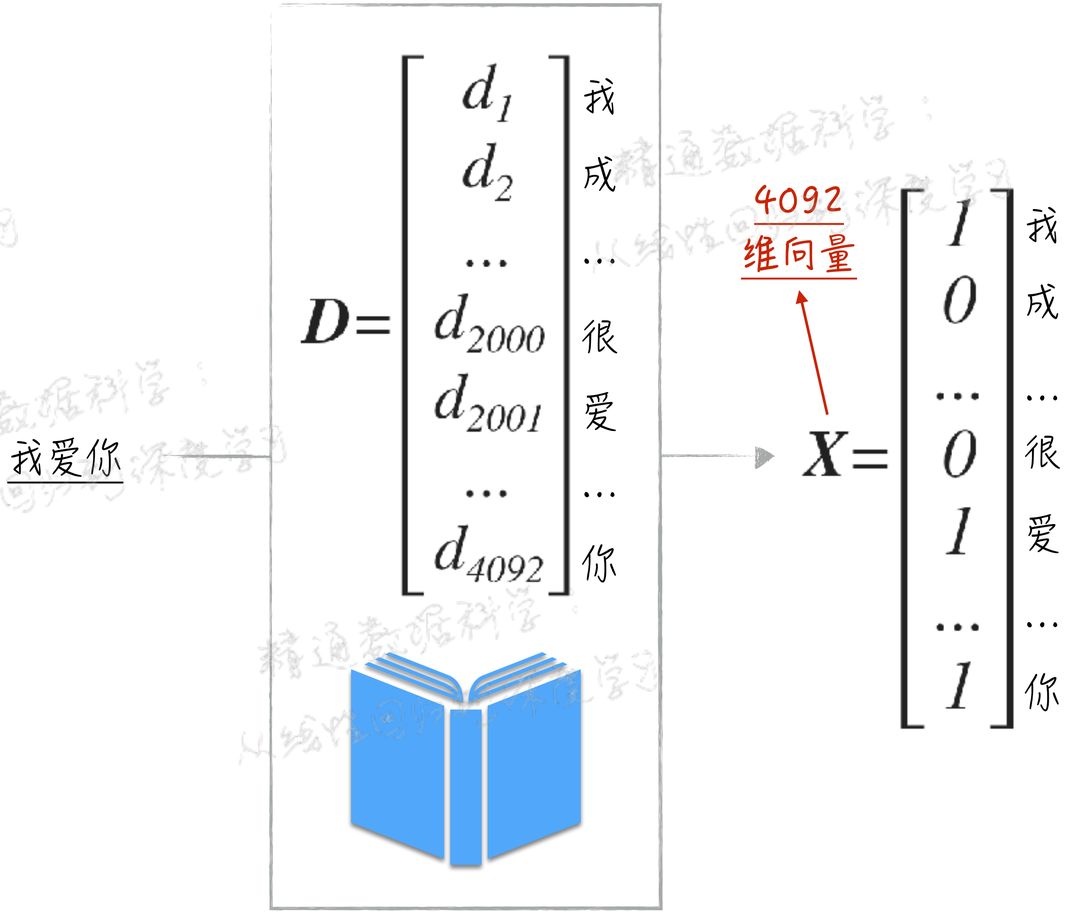
前情回顾

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：

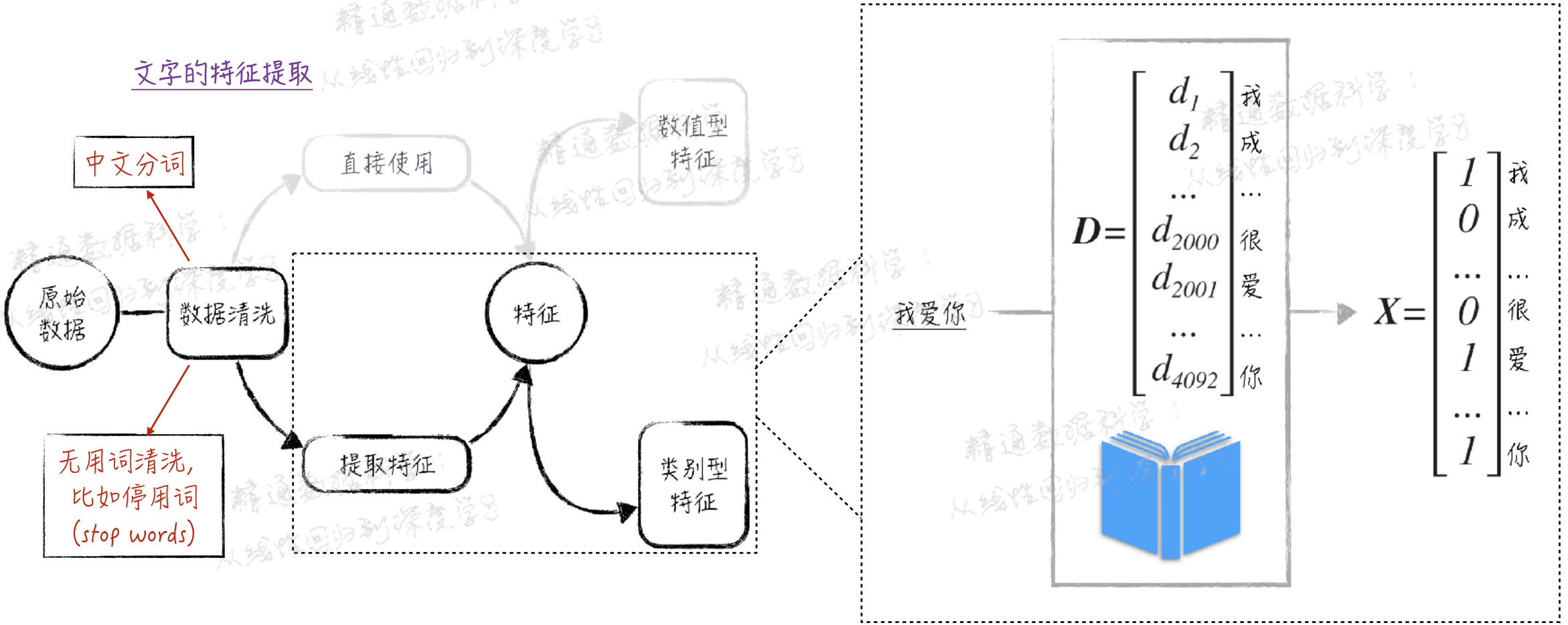
- 将所有的文字进行排序，形成一个**n维**的字典
- 每一个文字对应一个变量d_i
- 用一个**n维**的向量来表示文本：当文字出现在文本里，则相应的变量等于1，否则等于0





前情回顾

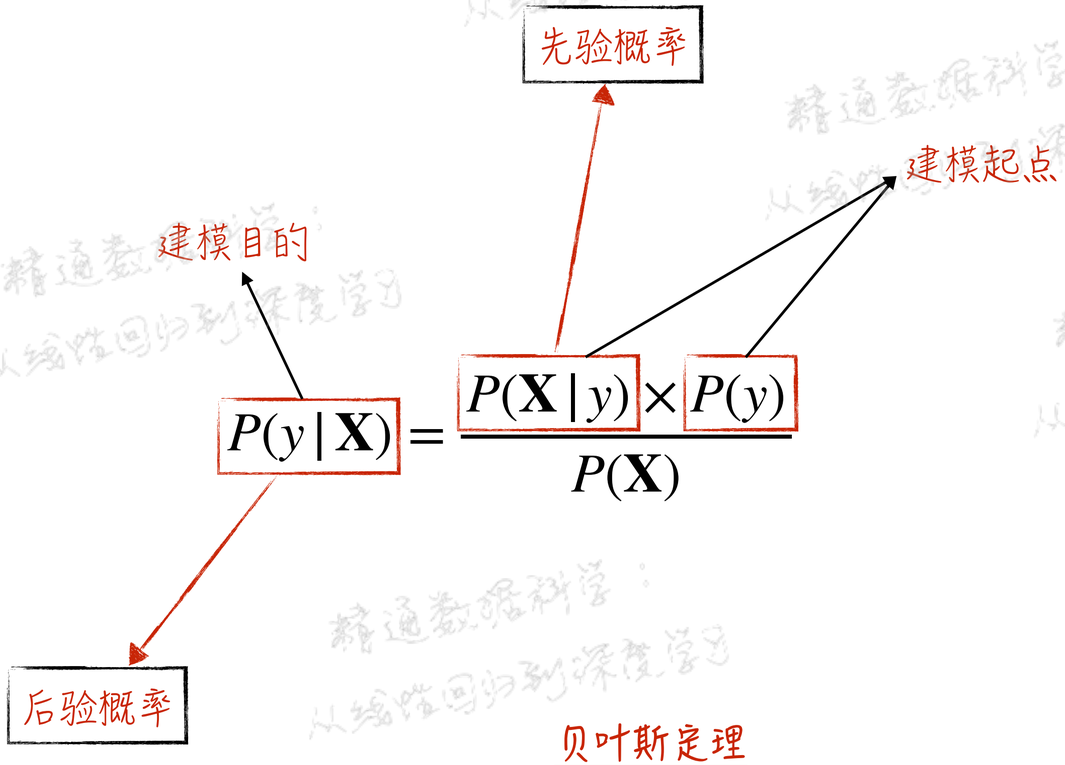
文本的特征提取



目录

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

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

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

x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

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

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

文本类别的分布 $P(y = l) = \theta_l$

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

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

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

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

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

最终预测公式 $\hat{y}_i = argmax_l \prod_j \hat{p}_{j,l}\hat{\theta}_l$

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(正面) = 2/3　　P(负面) = 1/3

特征条件概率

P(我 | 正面) = 2/2　　P(我 | 负面) = 1/1

P(爱 | 正面) = 1/2　　P(爱 | 负面) = 1/1

P(你 | 正面) = 2/2　　P(你 | 负面) = 1/1

P(喜 | 正面) = 1/2　　P(喜 | 负面) = 0/1

P(欢 | 正面) = 1/2　　P(欢 | 负面) = 0/1

P(不 | 正面) = 0/2　　P(不 | 负面) = 1/1

一个很简单的例子

(不考虑分词以及平滑项)

目录

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

最终预测公式

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

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

$$\hat{p}_{j,l} = \frac{\sum_{i=1}^{m} 1_{\{x_{i,j}=1, y_i=l\}} + \alpha}{\sum_{i=1}^{m} 1_{\{y_i=l\}} + 2\alpha} \qquad 0 < \alpha \leq 1$$

THANK YOU