

现代信息检索

Modern Information Retrieval

第13讲 文本分类及朴素贝叶斯分类器
Text Classification & Naïve Bayes

授课人：王斌

<http://ir.ict.ac.cn/~wangbin>

提纲

- ① 上一讲回顾
- ② 文本分类
- ③ 朴素贝叶斯
- ④ 朴素贝叶斯理论
- ⑤ 文本分类评价

提纲

- ① 上一讲回顾
- ② 文本分类
- ③ 朴素贝叶斯
- ④ 朴素贝叶斯理论
- ⑤ 文本分类评价

统计语言建模IR模型(SLMIR)

- 马萨诸塞大学(University of Massachusetts, UMass)大学Ponte、Croft等人于1998年提出。随后又发展了出了一系列基于SLM的模型。代表系统Lemur。
 - **查询似然模型**：把相关度看成是每篇文档对应的语言下生成该查询的可能性
 - **翻译模型**：假设查询经过某个噪声信道变形成某篇文章，则由文档还原成该查询的概率(翻译模型)可以视为相关度
 - **KL距离模型**：查询对应某种语言，每篇文档对应某种语言，查询语言和文档语言的KL距离作为相关度量
- 本讲义主要介绍查询似然模型

查询似然模型QLM

- QLM计算公式

$$\begin{aligned} RSV(Q, D) &= P(Q | D) = P(Q | M_D) \\ &= P(q_1 q_2 \dots q_m | M_D) \\ &= P(q_1 | M_D) P(q_2 | M_D) \dots P(q_m | M_D) \\ &= \prod_{w \in Q} P(w | M_D)^{c(w, Q)} \end{aligned}$$

- 于是检索问题转化为估计文档 D 的一元语言模型 M_D ，也即求所有词项 w 的概率 $P(w|M_D)$

QLM求解步骤

- 第一步：根据文档 D (样本)，估计文档模型 M_D (总体)，在一元模型下，即计算所有词项 w 的概率 $P(w|M_D)$
- 第二步：计算在模型 M_D 下生成查询 Q 的似然(即概率)
- 第三步：按照得分对所有文档排序

几种QLM中常用的平滑方法

- Jelinek-Mercer(JM), $0 \leq \lambda \leq 1$, 文档模型和文档集模型的混合

$$p(w|D) = \lambda p_{ML}(w|D) + (1 - \lambda) p(w|C)$$

- 课堂提问, 对于 $w \in D$, 折扣后的 $P_{DML}(w|D)$ 是不是一定小于 $P_{ML}(w|D)$?

- Dirichlet Priors(Dir), $\mu \geq 0$, *DIR*和*JM*可以互相转化

$$p(w|D) = \frac{c(w, D) + \mu p(w|C)}{|D| + \mu}$$


- Absolute Discounting(Abs), $0 \leq \delta \leq 1$, $|D|_u$ 表示 D 中不相同的词个数(u =unique)

$$p(w|D) = \frac{\max(c(w, D) - \delta, 0)}{|D|} + \frac{\delta |D|_u}{|D|} p(w|C)$$

基于翻译模型的IR模型

- 基本的QLM模型不能解决词语失配(word mismatch)问题, 即查询中的用词和文档中的用词不一致, 如: 电脑 vs. 计算机
- 假设 Q 通过一个有噪声的香农信道变成 D , 从 D 估计原始的 Q

翻译概率 生成概率



$$P(Q | D) = \prod_i P(q_i | D) = \prod_i \sum_j P(q_i | w_j) P(w_j | M_D)$$

- 翻译概率 $P(q_i | w_j)$ 在计算时可以将词项之间的关系融入。
 - 基于词典来计算(人工或者自动构造的同义词/近义词/翻译词典)
 - 基于语料库来计算(标题、摘要 vs. 文本; 文档锚文本 vs. 文档)

KL距离(相对熵)模型

$$Score(Q, D) = \log \frac{P(Q | M_D)}{P(Q | M_C)}$$

$$Score(Q, D) = \sum_{q_i \in Q} tf(q_i, Q) * \log \frac{P(q_i | M_D)}{P(q_i | M_C)}$$

$$P(Q | M_D) = \frac{|Q|!}{\prod_{q_i \in Q} tf(q_i, Q)!} \prod_{q_i \in Q} P(q_i | D)^{tf(q_i, Q)}$$

$$P(Q | M_C) = \frac{|Q|!}{\prod_{q_i \in Q} tf(q_i, Q)!} \prod_{q_i \in Q} P(q_i | C)^{tf(q_i, Q)}$$



多项分布

$$\begin{aligned} &\propto \sum_{q_i \in Q} P(q_i | M_Q) * \log \frac{P(q_i | M_D)}{P(q_i | M_C)} \\ &= \sum_{q_i \in Q} P(q_i | M_Q) * \log \frac{P(q_i | M_D)}{P(q_i | M_Q)} - \sum_{q_i \in Q} P(q_i | M_Q) * \log \frac{P(q_i | M_C)}{P(q_i | M_Q)} \\ &= -KL(M_Q, M_D) + \boxed{KL(M_Q, M_C)} \end{aligned}$$

对同一Q，为常数

$$\propto -KL(M_Q, M_D) = \sum_{q_i \in Q} P(q_i | M_Q) * \log P(q_i | M_D) - \boxed{\sum_{q_i \in Q} P(q_i | M_Q) * \log P(q_i | M_Q)}$$

$$\propto \sum_{q_i \in Q} P(q_i | M_Q) * \log P(q_i | M_D)$$

负的交叉熵

本讲内容

- 文本分类的概念及其与IR的关系
- 朴素贝叶斯分类器(朴素贝叶斯)
- 文本分类的评价

提纲

- ① 上一讲回顾
- ② 文本分类
- ③ 朴素贝叶斯
- ④ 朴素贝叶斯理论
- ⑤ 文本分类评价

文本分类

- Text classification或者 Text Categorization: 给定分类体系，将一篇文本分到其中一个或者多个类别中的过程。
- 按类别数目: binary vs. multi-class
- 按每篇文档赋予的标签数目: sing label vs. multi label

一个文本分类任务：垃圾邮件过滤

From: ''' <takworl1d@hotmail.com>
Subject: real estate is the only way... gem oalvgkay
Anyone can buy real estate with no money down
Stop paying rent TODAY !
There is no need to spend hundreds or even thousands for
similar courses
I am 22 years old and I have already purchased 6 properties
using the
methods outlined in this truly INCREDIBLE ebook.
Change your life NOW !
=====
Click Below to order:
<http://www.wholesaledaily.com/sales/nmd.htm>
=====

如何编程实现对上类信息的识别和过滤？

文本分类的形式化定义： 训练

给定：

- 文档空间 X
 - 文档都在该空间下表示—通常都是某种高维空间
- 固定的类别集合 $C = \{c_1, c_2, \dots, c_j\}$
 - 类别往往根据应用的需求来认为定义 (如, 相关类 vs. 不相关类)
- 训练集 D , 文档 d 用 c 来标记, $\langle d, c \rangle \in X \times C$

利用学习算法, 可以学习一个分类器 γ , 它可以将文档映射成类别:

$$\gamma: X \rightarrow C$$

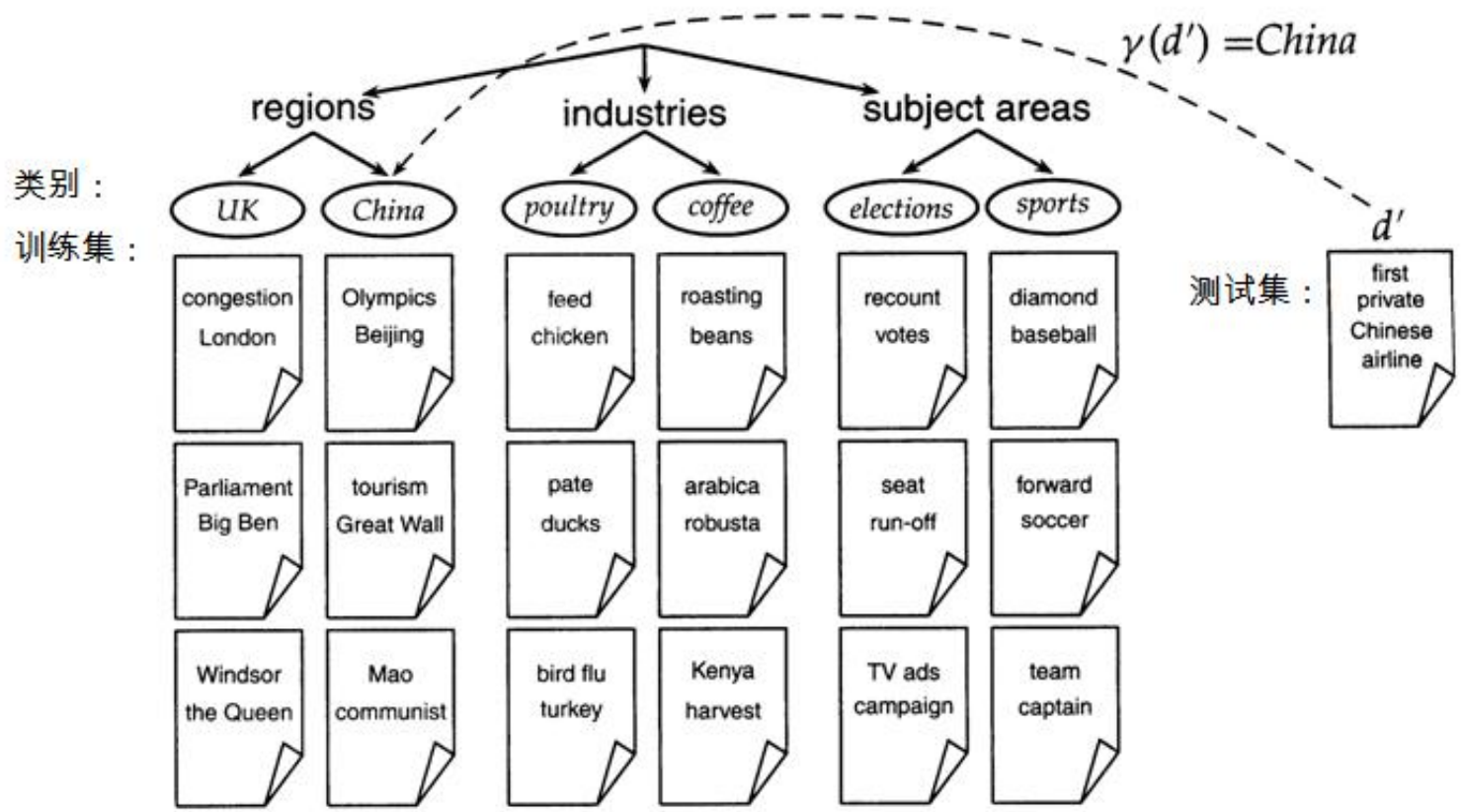
文本分类的形式化定义：应用/测试

给定： $d \in X$

确定： $\gamma(d) \in C,$

即确定 d 最可能属于的类别

主题分类



课堂练习

- 试举出文本分类在信息检索中的应用例子

搜索引擎中的文本分类应用

- 语言识别 (类别: English vs. French等)
- 垃圾网页的识别 (垃圾网页 vs. 正常网页)
- 是否包含淫秽内容 (色情 vs. 非色情)
- 领域搜索或垂直搜索 – 搜索对象限制在某个垂直领域（如健康医疗）（属于该领域 vs. 不属于该领域）
- 静态查询 (如, Google Alerts)
- 情感识别: 影评或产品评论是贬还是褒 (褒评 vs. 贬评)

分类方法: 1. 手工方法

- Web发展的初期, Yahoo使用人工分类方法来组织Yahoo目录, 类似工作还有: ODP, PubMed
- 如果是专家来分类精度会非常高
- 如果问题规模和分类团队规模都很小的时候, 能否保持分类结果的一致性
- 但是对人工分类进行规模扩展将十分困难, 代价昂贵
- → 因此, 需要自动分类方法

分类方法: 2. 规则方法

- Google Alerts的例子是基于规则分类的
- 存在一些IDE开发环境来高效撰写非常复杂的规则 (如Verity)
- 通常情况下都是布尔表达式组合 (如Google Alerts)
- 如果规则经过专家长时间的精心调优，精度会非常高
- 建立和维护基于规则的分类系统非常繁琐，开销也大

一个Verity主题 (一条复杂的分类规则)

comment line	# Beginning of art topic definition		
top-level topic	art ACCRUE		
topic definition modifiers	/author = "fsmith" /date = "30-Dec-01" /annotation = "Topic created by fsmith"		
subtopic	* 0.70 performing-arts ACCRUE	subtopic	* 0.70 film ACCRUE
evidencetopic	** 0.50 WORD		** 0.50 STEM
topic definition modifier	/wordtext = ballet	subtopic	/wordtext = film
evidencetopic	** 0.50 STEM		** 0.50 motion-picture PHRASE
topic definition modifier	/wordtext = dance		*** 1.00 WORD
evidencetopic	** 0.50 WORD		/wordtext = motion
topic definition modifier	/wordtext = opera		*** 1.00 WORD
evidencetopic	** 0.30 WORD		/wordtext = picture
topic definition modifier	/wordtext = symphony		** 0.50 STEM
subtopic	* 0.70 visual-arts ACCRUE	subtopic	/wordtext = movie
	** 0.50 WORD		* 0.50 video ACCRUE
	/wordtext = painting		** 0.50 STEM
	** 0.50 WORD		/wordtext = video
	/wordtext = sculpture		** 0.50 STEM
			/wordtext = vcr
			# End of art topic

分类方法: 3. 统计/概率方法

- 文本分类被定义为一个学习问题，这也是本书中的定义，包括：
 - (i) 通过有监督的学习，得到分类函数 γ ，然后将其
 - (ii) 应用于对新文档的分类
- 后面将介绍一系列分类方法: 朴素贝叶斯, Rocchio, kNN, SVM
- 当然，没有免费的午餐：需要手工构建训练集
- 但是，该手工工作一般人就可以完成，不需要专家。

提纲

- ① 上一讲回顾
- ② 文本分类
- ③ 朴素贝叶斯**
- ④ 朴素贝叶斯理论
- ⑤ 文本分类评价

朴素贝叶斯分类器

- 朴素贝叶斯是一个概率分类器
- 文档 d 属于类别 c 的概率计算如下：

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

- n_d 是文档的长度(词条的个数)
- $P(t_k | c)$ 是词项 t_k 出现在类别 c 中文档的概率
- $P(t_k | c)$ 度量的是当 c 是正确类别时 t_k 的贡献
- $P(c)$ 是类别 c 的先验概率
- 如果文档的词项无法提供属于哪个类别的信息，那么我们直接选择 $P(c)$ 最高的那个类别

具有最大后验概率的类别

- 朴素贝叶斯分类的目标是寻找“最佳”的类别
- 最佳类别是具有最大后验概率(maximum a posteriori -MAP)的类别 c_{map} :

$$c_{\text{map}} = \arg \max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg \max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

对数计算

- 很多小概率的乘积会导致浮点数下溢出
- 由于 $\log(xy) = \log(x) + \log(y)$, 可以通过取对数将原来的乘积计算变成求和计算
- 由于 \log 是单调函数, 因此得分最高的类别不会发生改变
- 因此, 实际中常常使用的是:

$$c_{\text{map}} = \arg \max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k | c)]$$

朴素贝叶斯分类器

- 分类规则:

$$c_{\text{map}} = \arg \max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k | c)]$$

- 简单说明:
 - 每个条件参数 $\hat{P}(t_k | c)$ 是反映 t_k 对 c 的贡献高低的一个权重
 - 先验概率 $\hat{P}(c)$ 是反映类别 c 的相对频率的一个权重
 - 因此，所有权重的求和反映的是文档属于类别的可能性
 - 选择最具可能性的类别

参数估计 1: 极大似然估计

- 如何从训练数据中估计 $\hat{P}(c)$ 和 $\hat{P}(t_k|c)$?

- 先验:

$$\hat{P}(c) = \frac{N_c}{N}$$

- N_c : 类 c 中的文档数目; N : 所有文档的总数

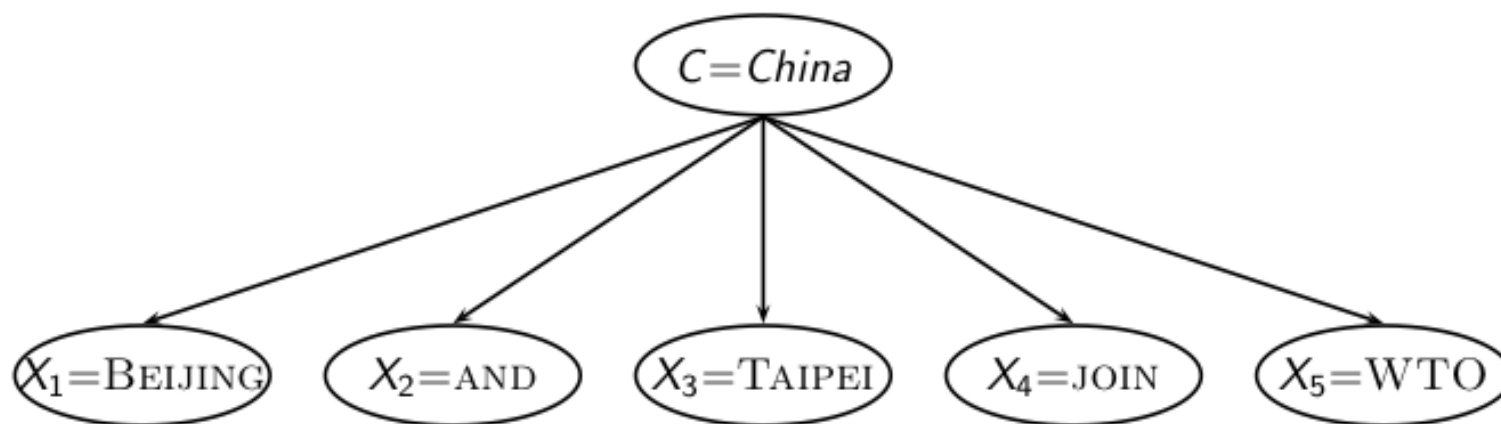
- 条件概率:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

- T_{ct} 是训练集中类别 c 中的词条 t 的个数 (多次出现要计算多次)
- 给定如下的 朴素贝叶斯 独立性假设(independence assumption):

$$\hat{P}(t_{k_1}|c) = \hat{P}(t_{k_2}|c)$$

MLE估计中的问题：零概率问题



$$P(China|d) \propto P(China) \cdot P(\text{BEIJING}|China) \cdot P(\text{AND}|China) \\ \cdot P(\text{TAIPEI}|China) \cdot P(\text{JOIN}|China) \cdot$$

$$P(\text{WTO}|China)$$

$$\hat{P}(\text{WTO}|China) = \frac{T_{China, \text{WTO}}}{\sum_{t' \in V} T_{China, t'}} = \frac{0}{\sum_{t' \in V} T_{China, t'}} = 0$$

MLE估计中的问题：零概率问题（续）

- 如果 WTO 在训练集中没有出现在类别 China 中，那么就会有如下的零概率估计：

$$\hat{P}(\text{WTO}|\text{China}) = \frac{T_{\text{China}, \text{WTO}}}{\sum_{t' \in V} T_{\text{China}, t'}} = 0$$

- → 那么，对于任意包含WTO的文档， $P(\text{China}|\text{d}) = 0$ 。
- 一旦发生零概率，将无法判断类别

避免零概率: 加一平滑

- 平滑前:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

- 平滑后: 对每个量都加上1

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B}$$

- B 是不同的词语个数 (这种情况下词汇表大小 $|V| = B$)

避免零概率: 加一平滑 (续)

- 利用加1平滑从训练集中估计参数
- 对于新文档，对于每个类别，计算 (i) 先验的对数值之和以及 (ii) 词项条件概率的对数之和
- 将文档归于得分最高的那个类

朴素贝叶斯: 训练过程

TRAINMULTINOMIALNB(\mathbb{C}, \mathbb{D})

```
1   $V \leftarrow \text{EXTRACTVOCABULARY}(\mathbb{D})$ 
2   $N \leftarrow \text{COUNTDOCS}(\mathbb{D})$ 
3  for each  $c \in \mathbb{C}$ 
4  do  $N_c \leftarrow \text{COUNTDOCSINCLASS}(\mathbb{D}, c)$ 
5       $\text{prior}[c] \leftarrow N_c / N$ 
6       $\text{text}_c \leftarrow \text{CONCATENATETEXTOFALLDOCSINCLASS}(\mathbb{D}, c)$ 
7      for each  $t \in V$ 
8      do  $T_{ct} \leftarrow \text{COUNTTOKENSOFTERM}(\text{text}_c, t)$ 
9      for each  $t \in V$ 
10     do  $\text{condprob}[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'} (T_{ct'}+1)}$ 
11 return  $V, \text{prior}, \text{condprob}$ 
```

朴素贝叶斯: 测试

```
APPLYMULTINOMIALNB( $\mathbb{C}$ ,  $V$ ,  $prior$ ,  $condprob$ ,  $d$ )  
1   $W \leftarrow \text{EXTRACTTOKENSFROMDOC}(V, d)$   
2  for each  $c \in \mathbb{C}$   
3  do  $score[c] \leftarrow \log prior[c]$   
4      for each  $t \in W$   
5      do  $score[c] + = \log condprob[t][c]$   
6  return  $\arg \max_{c \in \mathbb{C}} score[c]$ 
```

课堂练习

	docID	words in document	in $c = \textit{China}$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

- 估计朴素贝叶斯分类器的参数
- 对测试文档进行分类

例子: 参数估计

Priors: $\hat{P}(c) = 3/4$ and $\hat{P}(\bar{c}) = 1/4$ Conditional probabilities:

$$\hat{P}(\text{CHINESE}|c) = (5 + 1)/(8 + 6) = 6/14 = 3/7$$

$$\hat{P}(\text{TOKYO}|c) = \hat{P}(\text{JAPAN}|c) = (0 + 1)/(8 + 6) = 1/14$$

$$\hat{P}(\text{CHINESE}|\bar{c}) = (1 + 1)/(3 + 6) = 2/9$$

$$\hat{P}(\text{TOKYO}|\bar{c}) = \hat{P}(\text{JAPAN}|\bar{c}) = (1 + 1)/(3 + 6) = 2/9$$

上述计算中的分母分别是 $(8 + 6)$ 和 $(3 + 6)$ ，这是因为 $text_c$ 和 $text_{\bar{c}}$ 的大小分别是8和3，词汇表大小是6。

例子: 分类

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$

$$\hat{P}(\bar{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

因此, 分类器将测试文档分到 $c = \text{China}$ 类, 这是因为 d_5 中起正向作用的 CHINESE 出现 3 次的权重高于起反向作用的 JAPAN 和 TOKYO 的权重之和。

朴素贝叶斯的时间复杂度分析

mode	time complexity
training	$\Theta(\mathbb{D} L_{ave} + \mathbb{C} V)$
testing	$\Theta(L_a + \mathbb{C} M_a) = \Theta(\mathbb{C} M_a)$

- L_{ave} : 训练文档的平均长度, L_a : 测试文档的平均长度, M_a : 测试文档中不同的词项个数 \mathbb{D} : 训练文档, V : 词汇表, \mathbb{C} : 类别集合
- $\Theta(|\mathbb{D}|L_{ave})$ 是计算所有数字的时间
- $\Theta(|\mathbb{C}||V|)$ 是从上述数字计算参数的时间
- 通常来说: $|\mathbb{C}||V| < |\mathbb{D}|L_{ave}$
- 测试时间也是线性的 (相对于测试文档的长度而言).
- 因此: 朴素贝叶斯 对于训练集的大小和测试文档的大小而言是线性的。这是最优的

提纲

- ① 上一讲回顾
- ② 文本分类
- ③ 朴素贝叶斯
- ④ 朴素贝叶斯理论**
- ⑤ 文本分类评价

朴素贝叶斯: 分析

- 接下来对朴素贝叶斯的性质进行更深层次的理解
- 包括形式化地推导出分类规则...
- ... 然后介绍在推导中的假设

朴素贝叶斯规则

给定文档的条件下，我们希望得到最可能的类别

$$C_{\text{map}} = \arg \max_{c \in \mathbb{C}} P(c|d)$$

应用贝叶斯定律 $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$:

$$C_{\text{map}} = \arg \max_{c \in \mathbb{C}} \frac{P(d|c)P(c)}{P(d)}$$

由于分母 $P(d)$ 对所有类别都一样，因此可以去掉:

$$C_{\text{map}} = \arg \max_{c \in \mathbb{C}} P(d|c)P(c)$$

过多参数/稀疏性问题

$$\begin{aligned}C_{\text{map}} &= \arg \max_{c \in \mathbb{C}} P(d|c)P(c) \\ &= \arg \max_{c \in \mathbb{C}} P(\langle t_1, \dots, t_k, \dots, t_{n_d} \rangle | c)P(c)\end{aligned}$$

- 上式中存在过多的参数 $P(\langle t_1, \dots, t_k, \dots, t_{n_d} \rangle | c)$ ，每个参数都是一个类别和一个词语序列的组合
- 要估计这么多参数，必须需要大量的训练样例。但是，训练集的规模总是有限的
- 于是出现数据稀疏性（ data sparseness ）问题

朴素贝叶斯条件独立性假设

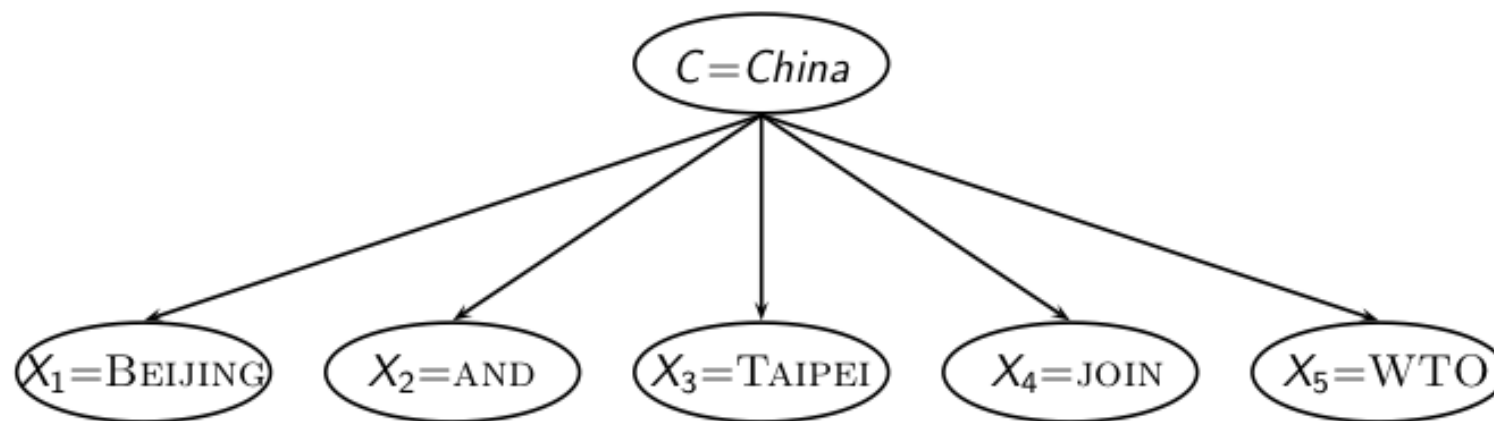
为减少参数数目，给出朴素贝叶斯条件独立性假设：

$$P(d|c) = P(\langle t_1, \dots, t_{n_d} \rangle | c) = \prod_{1 \leq k \leq n_d} P(X_k = t_k | c)$$

假定上述联合概率等于某个独立概率 $P(X_k = t_k | c)$ 的乘积。前面我们提到可以通过如下方法来估计这些先验概率和条件概率：

$$\hat{P}(c) = \frac{N_c}{N} \text{ and } \hat{P}(t|c) = \frac{T_{ct}+1}{(\sum_{t' \in V} T_{ct'})+B}$$

生成式(Generative)模型



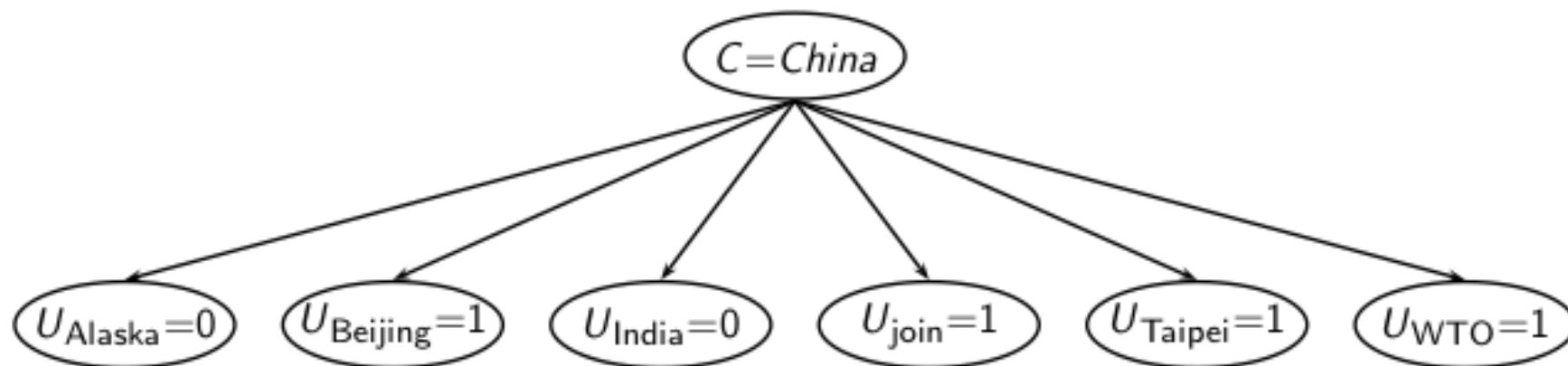
$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

- 利用概率 $P(c)$ 产生一个类
- 以该类为条件，（在各自位置上）基于概率 $P(t_k|c)$ 产生每个词语，这些词语之间相互独立
- 对文档分类时，找出最有可能生成该文档的类别

第二个独立性假设

- $\hat{P}(t_{k_1}|c) = \hat{P}(t_{k_2}|c)$
- 例如，对于 *UK* 类别中的一篇文档，在第一个位置上生成 QUEEN 的概率和在最后一个位置上生成它的概率一样
- 上述两个独立性假设实际上是词袋模型([bag of words model](#))

另一个朴素贝叶斯模型： 贝努利模型



朴素贝叶斯独立性假设不成立的情况

- 自然语言文本中，上述独立性假设并不成立
- 条件独立性假设：

$$P(\langle t_1, \dots, t_{n_d} \rangle | c) = \prod_{1 \leq k \leq n_d} P(X_k = t_k | c)$$

- 位置独立性假设：

$$\hat{P}(t_{k_1} | c) = \hat{P}(t_{k_2} | c)$$

- 课堂练习

- 给出条件独立性假设不成立的例子
 - 给出位置独立性假设不成立的例子
- 在这些假设都不成立的情况下，为什么朴素贝叶斯方法有用？

朴素贝叶斯方法起作用的原因

- 即使在条件独立性假设严重不成立的情况下，朴素贝叶斯方法能够高效地工作
- 例子

	c_1	c_2	class selected
true probability $P(c d)$	0.6	0.4	c_1
$\hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k c)$	0.00099	0.00001	
NB estimate $\hat{P}(c d)$	0.99	0.01	c_1

- 概率 $P(c_2|d)$ 被过低估计(0.01)，而概率 $P(c_1|d)$ 被过高估计(0.99)。
- 分类的目标是预测正确的类别，并不是准确地估计概率
- 准确估计 \Rightarrow 精确预测
- 反之并不成立！

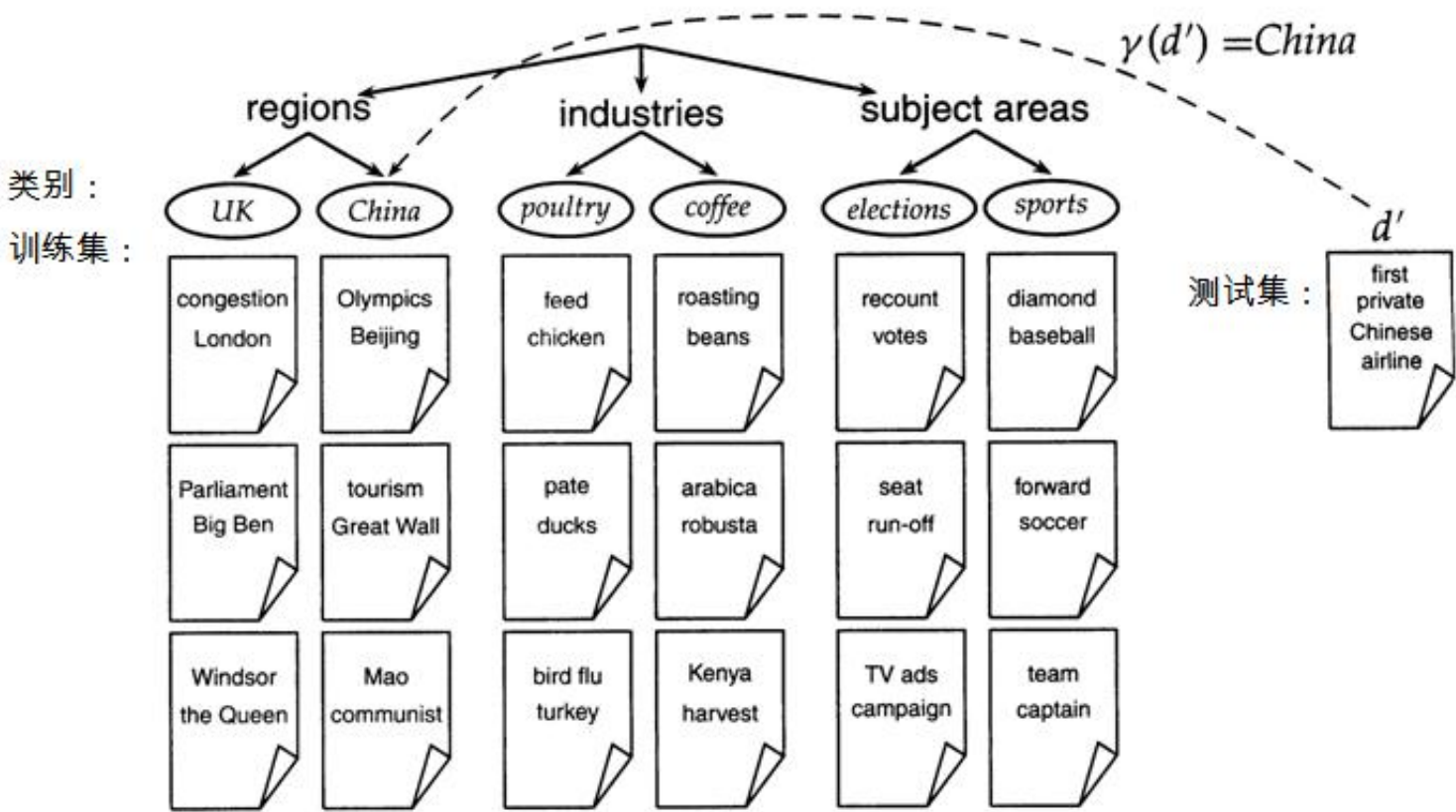
朴素贝叶斯 并不朴素

- 朴素贝叶斯在多次竞赛中胜出 (比如 KDD-CUP 97)
- 相对于其他很多更复杂的学习方法，朴素贝叶斯对非相关特征更具鲁棒性
- 相对于其他很多更复杂的学习方法，朴素贝叶斯对概念漂移 (concept drift) 更鲁棒 (概念漂移是指类别的定义随时间变化)
- 当有很多同等重要的特征时，该方法由于类似于决策树的方法
- 一个很好的文本分类基准方法 (当然，不是最优的方法)
- 如果满足独立性假设，那么朴素贝叶斯是最优的 (文本当中用于成立，但是对某些领域可能成立)
- 非常快
- 存储开销少

提纲

- ① 上一讲回顾
- ② 文本分类
- ③ 朴素贝叶斯
- ④ 朴素贝叶斯理论
- ⑤ 文本分类评价

Reuters语料上的评价



例子：Reuters语料

symbol	statistic	value
<i>N</i>	documents	800,000
<i>L</i>	avg. # word tokens per document	200
<i>M</i>	word types	400,000
	avg. # bytes per word token (incl. spaces/punct.)	6
	avg. # bytes per word token (without spaces/punct.)	4.5
	avg. # bytes per word type	7.5
	non-positional postings	100,000,000
type of class	number	examples
region	366	UK, China
industry	870	poultry, coffee
subject area	126	elections, sports

一篇Reuters文档



You are here: [Home](#) > [News](#) > [Science](#) > [Article](#)

Go to a Section: [U.S.](#) [International](#) [Business](#) [Markets](#) [Politics](#) [Entertainment](#) [Technology](#) [Sports](#) [Oddly Enough](#)

Extreme conditions create rare Antarctic clouds

Tue Aug 1, 2006 3:20am ET

[Email This Article](#) [Print This Article](#) [Reprints](#)

[\[-\]](#) Text [\[+\]](#)



SYDNEY (Reuters) - Rare, mother-of-pearl colored clouds caused by extreme weather conditions above Antarctica are a possible indication of global warming, Australian scientists said on Tuesday.

Known as nacreous clouds, the spectacular formations showing delicate wisps of colors were photographed in the sky over an Australian

分类评价

- 评价必须基于测试数据进行，而且该测试数据是与训练数据完全独立的 (通常两者样本之间无交集)
- 很容易通过训练可以在训练集上达到很高的性能 (比如记忆所有的测试集合)
- 指标: 正确率、召回率、 F_1 值、分类精确率(classification accuracy)等等

正确率 P 及召回率 R

	in the class	not in the class
predicted to be in the class	true positives (TP)	false positives (FP)
predicted to not be in the class	false negatives (FN)	true negatives (TN)

$$P = TP / (TP + FP)$$

$$R = TP / (TP + FN)$$

F 值

- F_1 允许在正确率和召回率之间达到某种均衡

- $$F_1 = \frac{1}{\frac{1}{2} \frac{1}{P} + \frac{1}{2} \frac{1}{R}} = \frac{2PR}{P + R}$$

- 也就是 P 和 R 的调和平均值：
$$\frac{1}{F} = \frac{1}{2} \left(\frac{1}{P} + \frac{1}{R} \right)$$

微平均 vs. 宏平均

- 对于一个类我们得到评价指标 F_1
- 但是我们希望得到在所有类别上的综合性能
- 宏平均(Macroaveraging)
 - 对类别集合 C 中的每个类都计算一个 F_1 值
 - 对 C 个结果求平均Average these C numbers
- 微平均(Microaveraging)
 - 对类别集合 C 中的每个类都计算TP、FP和FN
 - 将 C 中的这些数字累加
 - 基于累加的TP, FP, FN计算P、R和 F_1

朴素贝叶斯 vs. 其他方法

(a)	NB	Rocchio	kNN	SVM
micro-avg-L (90 classes)	80	85	86	89
macro-avg (90 classes)	47	59	60	60

(b)	NB	Rocchio	kNN	trees	SVM
earn	96	93	97	98	98
acq	88	65	92	90	94
money-fx	57	47	78	66	75
grain	79	68	82	85	95
crude	80	70	86	85	89
trade	64	65	77	73	76
interest	65	63	74	67	78
ship	85	49	79	74	86
wheat	70	69	77	93	92
corn	65	48	78	92	90
micro-avg (top 10)	82	65	82	88	92
micro-avg-D (118 classes)	75	62	n/a	n/a	87

Evaluation measure: F_1 Naive Bayes does pretty well, but some methods beat it consistently (e.g., SVM).

本讲小结

- 文本分类的概念及其与IR的关系
- 朴素贝叶斯分类器(朴素贝叶斯)
- 文本分类的评价

参考资料

- 《信息检索导论》第13 章
- <http://ifnlp.org/ir>
- Weka: 一个包含了 朴素贝叶斯在内的数据挖掘工具包
- Reuters-21578 – 最著名的文本分类语料 (当然, 当前已经显得规模太小)

课后习题

- 习题13-2
- 习题13-9