NUSTM

机器学习

朴素贝叶斯模型

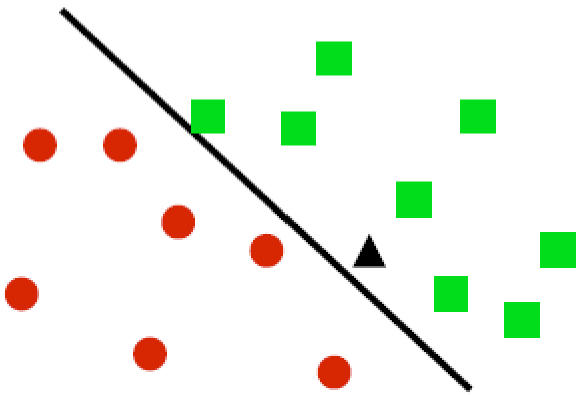
夏 睿
文本挖掘组
南京理工大学
rxia@njust.edu.cn

朴素贝叶斯模型

- 一个概率模型
- 一个生成模型
- 被称为“朴素”假设
- 适用于离散分布
- 广泛应用于文本分类、自然语言处理和模式识别

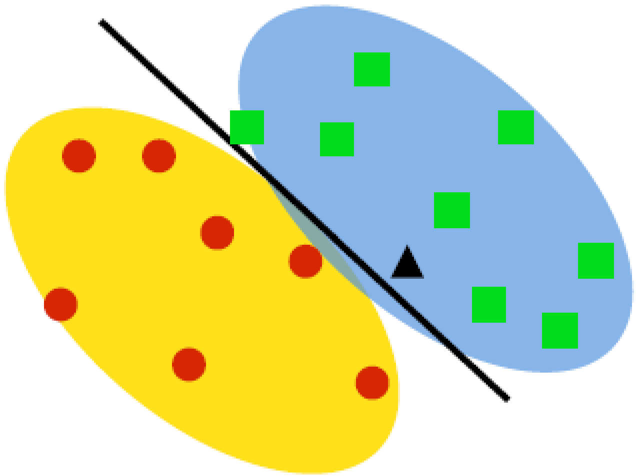
生成式 vs. 判别式

- 判别式模型



它对给定观察p(y|x)类标签的后验概率进行建模

- 生成式模型



它对类标签和观察p(x,y)的联合概率进行建模，然后使用贝叶斯规则(p(y|x)=p(x,y)/p(x))进行预测

朴素贝叶斯假设

- 混合模型

类先验概率

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

类-条件概率

- 词袋 (BOW) 表示

$$x = (\omega_1, \omega_2, \ldots, \omega_{|x|})$$

$$p(x|c_j) = p(\omega_1, \omega_2, \ldots, \omega_{|x|}|c_j) = \prod_{h=1}^{|x|} p(\omega_h|c_j)$$

有两个事件模型

NUSTM

多项式事件模型

模型描述

- 假设

$$p(y = c_j) = \pi_j$$

$$p(x|c_j) = p([\omega_1, \omega_2, \dots, \omega_{|x|}]|c_j) = \prod_{h=1}^{|x|} p(\omega_h|c_j)$$

$$= \prod_{i=1}^{V} p(t_i|c_j)^{N(t_i,x)} = \prod_{i=1}^{V} \theta_{i|j}^{N(t_i,x)}$$

- 联合概率

$$p(x, y = c_j) = p(c_j)p(x|c_j) = \pi_j \prod_{i=1}^{V} \theta_{i|j}^{N(t_i,x)}$$

模型参数

似然函数

- (联合)似然

$$
\begin{aligned}
L(\pi,\theta) &= \log\prod_{k=1}^{N} p(x_k,y_k) \\
&= \log\prod_{k=1}^{N}\sum_{j=1}^{C} I(y_k=c_j)p(y_k=c_j)p(x_k|y_k=c_j) \\
&= \sum_{k=1}^{N}\sum_{j=1}^{C} I(y_k=c_j)\log p(y_k=c_j)p(x_k|y_k=c_j) \\
&= \sum_{k=1}^{N}\sum_{j=1}^{C} I(y_k=c_j)\log \pi_j\prod_{i=1}^{V}\theta_{i|j}^{N(t_i,x_k)} \\
&= \sum_{k=1}^{N}\sum_{j=1}^{C} I(y_k=c_j)\left(\log\pi_j+\sum_{i=1}^{V}N(t_i,x_k)\log\theta_{i|j}\right)
\end{aligned}
$$

最大似然估计

- MLE 公式

$$\max_{\pi,\theta} L(\pi,\theta)$$

$$s.t.\begin{cases}\sum_{j=1}^{C}\pi_j = 1\\ \sum_{i=1}^{V}\theta_{i|j} = 1, j = 1,\dots,C\end{cases}$$

- 应用拉格朗日乘数

$$J = L(\pi,\theta) + \alpha(1-\sum_{j=1}^{C}\pi_j) + \sum_{j=1}^{C}\beta_j\,(1-\sum_{i=1}^{V}\theta_{i|j})$$

$$= \sum_{k=1}^{N}\sum_{j=1}^{C} I(y_k = c_j)[\log\pi_j + \sum_{i=1}^{V} N(t_i,x_k)\log\theta_{i|j}] + \alpha\left(1-\sum_{j=1}^{C}\pi_j\right) + \sum_{j=1}^{C}\beta_j\left(1-\sum_{i=1}^{V}\theta_{i|j}\right)$$

闭式MLE解

- 梯度

$$\frac{\partial J}{\partial \pi_j} = \sum_{k=1}^{N} I(y_k = c_j)\frac{1}{\pi_j} - \alpha = 0$$

$$\frac{\partial J}{\partial \theta_{i|j}} = \sum_{k=1}^{N} I(y_k = c_j)\frac{N(t_i, x_k)}{\theta_{i|j}} - \beta_j = 0$$

- MLE 解

$$\pi_j = \frac{\sum_{k=1}^{N} I(y_k = c_j)}{\sum_{k=1}^{N}\sum_{j'=1}^{C} I(y_k = c_j)} = \frac{N_j}{N}$$

$$\theta_{i|j} = \frac{\sum_{k=1}^{N} I(y_k = c_j) N(t_i, x_k)}{\sum_{k=1}^{N} I(y_k = c_j)\sum_{i'=1}^{V} N(t_{i'}, x_k)}$$

NUSTM

拉普拉斯平滑

- 为了防止零概率

$$p(x, y = c_j) = \pi_j \prod_{i=1}^{V} \theta_{i|j}^{N(t_i, x)}$$

- 拉普拉斯平滑

$$\theta_{i|j} = \frac{\sum_{k=1}^{N} I(y_k = c_j) N(t_i, x_k)}{\sum_{i'=1}^{V} \sum_{k=1}^{N} I(y_k = c_j) N(t_{i'}, x_k)} \qquad \pi_j = \frac{\sum_{k=1}^{N} I(y_k = c_j)}{\sum_{j'=1}^{C} \sum_{k=1}^{N} I(y_k = c_j)}$$

⇓ ⇓

$$\theta_{i|j} = \frac{\sum_{k=1}^{N} I(y_k = c_j) N(t_i, x_k) + 1}{\sum_{i'=1}^{V} \sum_{k=1}^{N} I(y_k = c_j) N(t_{i'}, x_k) + V} \qquad \pi_j = \frac{\sum_{k=1}^{N} I(y_k = c_j) + 1}{\sum_{j'=1}^{C} \sum_{k=1}^{N} I(y_k = c_j) + C}$$

多变量伯努利事件模型

模型描述

- 假设

$$p(y = c_j) = \pi_j$$

$$p(x|y = c_j) = p(t_1, t_2, \dots, t_V|c_j)$$
$$= \prod_{i=1}^{V}[I(t_i \epsilon x)p(t_i|c_j) + I(t_i \notin x)(1 - p(t_i|c_j))]$$
$$= \prod_{i=1}^{V}[I(t_i \epsilon x)\mu_{i|j} + I(t_i \notin x)(1 - \mu_{i|j})]$$

- 联合概率

模型参数

$$p(x, c_j) = \pi_j \prod_{i=1}^{V}[I(t_i \epsilon x)\mu_{i|j} + I(t_i \notin x)(1 - \mu_{i|j})]$$

似然函数

- (联合)似然

$$L(\pi,\mu)=\log\prod_{k=1}^{N}p(x_k,y_k)$$

$$=\sum_{k=1}^{N}\log\sum_{j=1}^{C}I\left(y_k=c_j\right)p(x_k,y_k)$$

$$=\sum_{k=1}^{N}\sum_{j=1}^{C}I\left(y_k=c_j\right)\log p(c_j)\prod_{i=1}^{V}I(t_i\epsilon x)p\left(t_i\middle|c_j\right)+I(t_i\notin x)(1-p\left(t_i\middle|c_j\right))$$

$$=\sum_{k=1}^{N}\sum_{j=1}^{C}I\left(y_k=c_j\right)\left(\log\pi_j+\sum_{i=1}^{V}I(t_i\epsilon x_k)\log\mu_{i|j}+I(t_i\notin x_k)\log(1-\mu_{i|j})\right)$$

NUSTM

最大似然估计

- MLE 公式

$$\max_{\pi,\mu} L(\pi,\mu)$$

$$s.t. \sum_{j=1}^{C} \pi_j = 1$$

- 应用拉格朗日乘数

$$J = L(\pi,\mu) + \alpha\left(1 - \sum_{j=1}^{C} \pi_j\right)$$

$$= \sum_{k=1}^{N} \sum_{j=1}^{C} I(y_k = c_j)\left(log\pi_j + \sum_{i=1}^{V} I(t_i \epsilon x_k)\, log\mu_{i|j} + I(t_i \notin x)\log(1-\mu_{i|j})\right) + \alpha\left(1 - \sum_{j=1}^{C} \pi_j\right)$$

NUSTM

闭式MLE解

- 梯度

$$\frac{\partial J}{\partial \pi_j} = \sum_{k=1}^{N} I(y_k = c_j)\frac{1}{\pi_j} - \alpha = 0$$

$$\frac{\partial J}{\partial \mu_{i|j}} = \sum_{k=1}^{N} I(y_k = c_j)\left(\frac{I(t_i \epsilon x_k)}{\mu_{i|j}} - \frac{I(t_i \notin x_k)}{1-\mu_{i|j}}\right) = 0, \forall j = 1, \ldots, C.$$

- MLE 解

$$\pi_j = \frac{\sum_{k=1}^{N} I(y_k = c_j)}{\sum_{k=1}^{N}\sum_{j'=1}^{C} I(y_k = c_{j'})} = \frac{N_j}{N}$$

$$\mu_{i|j} = \frac{\sum_{k=1}^{N} I(y_k = c_j)\, I(t_i \epsilon x_k)}{\sum_{k=1}^{N} I(y_k = c_j)}$$

拉普拉斯平滑

- 为了防止零概率

$$p(x,c_j)=\pi_j\prod_{i=1}^{V}[I(t_i\epsilon x)\mu_{i|j}+I(t_i\notin x)(1-\mu_{i|j})]$$

- 拉普拉斯平滑

$$\mu_{i|j}=\frac{\sum_{k=1}^{N}I(y_k=c_j)I(t_i\epsilon x_k)}{\sum_{k=1}^{N}I(y_k=c_j)}$$

$$\pi_j=\frac{\sum_{k=1}^{N}I(y_k=c_j)}{\sum_{j'=1}^{C}\sum_{k=1}^{N}I(y_k=c_j)}$$

⇩

$$\mu_{i|j}=\frac{\sum_{k=1}^{N}I(y_k=c_j)I(t_i\epsilon x_k)+1}{\sum_{k=1}^{N}I(y_k=c_j)+2}$$

$$\pi_j=\frac{\sum_{k=1}^{N}I(y_k=c_j)+1}{\sum_{j'=1}^{C}\sum_{k=1}^{N}I(y_k=c_j)+C}$$

文本分类举例

数据集

- 训练数据

ID	Text	Label
$d_{tr}1$	Chinese Beijing Chinese	C
$d_{tr}2$	Chinese Chinese Shanghai	C
$d_{tr}3$	Chinese Macao	C
$d_{tr}4$	Tokyo Japan Chinese	J

- 测试数据

ID	Text
$d_{te}1$	Chinese Chinese Chinese Tokyo Japan
$d_{te}2$	Tokyo Tokyo Japan Shanghai

- 类标签

c1 = C;
c2 = J

- 特征向量

t1 = Beijing
t2 = Chinese
t3 = Japan
t4 = Macao
t5 = Shanghai
t6 = Tokyo

NUSTM

多项式朴素贝叶斯

- 训练

		Doc	t1	t2	t3	t4	t5	t6
Term Frequency	c1	3	1	5	0	1	1	0
	c2	1	0	1	1	0	0	1
Probability	c1	3/4	2/14	(5+1)/(1+5+1+1+6)=6/14	1/14	2/14	2/14	1/14
	c2	1/4	1/9	(1+1)/(1+1+1+6)=2/9	2/9	1/9	1/9	2/9

- 预测

	Un-normalized	Normalized
$P(c1\|d_{te}1)$	(3/4)*(6/14)^3*(1/14)*(1/14)=0.0030121	0.689757
$P(c2\|d_{te}1)$	(1/4)*(2/9)^3*(2/9)*(2/9)=0.0013548	0.310243
$P(c1\|d_{te}2)$	(3/4)*(1/14)^2*(1/14)*(2/14)	0.113547
$P(c2\|d_{te}2)$	(1/4)*(2/9)^2*(2/9)*(1/9)	0.886453

多变量伯努利朴素贝叶斯

- 训练

		Doc	t1	t2	t3	t4	t5	t6
Document Frequency	c1	3	1	3	0	1	1	0
	c2	1	0	1	1	0	0	1
Probability	c1	3/4	2/5	(3+1)/(3+2)=4/5	1/5	2/5	2/5	1/5
	c2	1/4	1/3	(1+1)/(1+2)=2/3	2/3	1/3	1/3	2/3

- 预测

	Un-normalized	Normalized
$P(c1\|d_{te}1)$	(3/4)*(1-2/5)*4/5*1/5*(1-2/5)* (1-2/5)* 1/5=0.005184	0.1911
$P(c2\|d_{te}1)$	(1/4)*(1-1/3)*2/3*2/3*(1-1/3)* (1-1/3)*2/3=0.02195	0.8089
$P(c1\|d_{te}2)$	(3/4)*(1-2/5)*(1-3/5)*1/5*(1-2/5)*2/5*1/5=0.001728	0.2395
$P(c2\|d_{te}2)$	(1/4)*(1-1/3)*(1-2/3)*2/3*(1-1/3)*1/3*2/3=0.005487	0.7605

OpenPR-NB 软件

- 功能
  - 用C++编写
  - 支持多项式和多变量的伯努利事件模型
  - 拉普拉斯平滑
  - 统一的数据格式，如 SVM-light/LibSVM
  - 快速运行与稀疏代表
- 下载

  http://msrt.njust.edu.cn/staff/rxia/index.html#Software

练习

- 实现朴素贝叶斯算法用：
  - 多变量事件模型
  - 多变量伯努利模型
- 运行基于18页给出的训练和测试数据的算法.
- 比较朴素贝叶斯算法和逻辑回归(使用Bag-of-words来表示数据).

Questions?