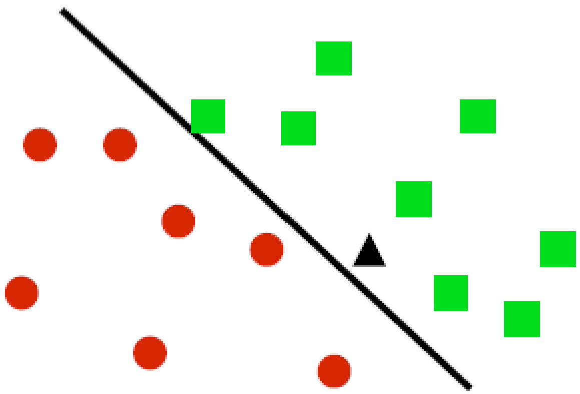
第七讲
朴素贝叶斯模型

夏睿
计算机科学与工程学院
南京理工大学
rxia@njust.edu.cn

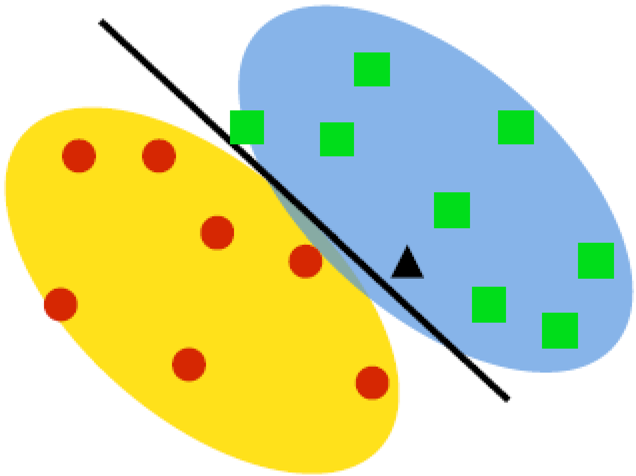
生成式 vs. 判别式

- 判别式模型



对给定观测值的标签的后验概率$p(y|x)$建模

- 生成式模型



对观测值和标签的联合概率$p(x,y)$建模，然后用贝叶斯法则$p(y|x)=p(x,y)/p(x)$进行预测

假设 – 学习 – 决策

- 判别式模型
  - 决策函数直接建模

$$h = f(\boldsymbol{x})$$

例子:
Perceptron, SVMs

  - 对后验概率建模

$$h = p(y|\boldsymbol{x})$$

例子:
Logistic/Softmax Regression

- 生成式模型（对联合分布建模）

$$h = p(\boldsymbol{x}, y) = p(y)p(\boldsymbol{x}|y)$$

例子:
Naïve Bayes, GMM

假设 – 学习 – 决策

- 判别式模型
  - 决策函数直接建模

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

学习准则：某些损失函数，如感知机损失、交叉熵损失、最大间隔损失等

  - 后验概率建模

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_k \log p\left(y^{(k)} \middle| \boldsymbol{x}^{(k)}\right)$$

学习准则：最大似然估计
Maximum Likelihood (条件分布) ⇔ 与某些损失函数等价

- 生成式模型（联合分布建模）

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_k \log p\left(\boldsymbol{x}^{(k)}, y^{(k)}\right)$$

学习准则：最大似然估计
Maximum Likelihood （联合分布）

假设－学习－决策

- 判别式模型
  - 决策函数

$$y = h = f(\boldsymbol{x})$$

  - 后验概率

$$\arg\max_y p(y|\boldsymbol{x})$$

- 生成式模型（贝叶斯公式）

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})}$$

$$\Downarrow$$

$$\arg\max_y p(y|\boldsymbol{x}) = \arg\max_y p(\boldsymbol{x}, y) = \arg\max_y p(\boldsymbol{x}|y)p(y)$$

朴素贝叶斯模型

- 概率模型
- 生成式模型
- “朴素”的类条件分布假设
- 适用于离散分布
- 广泛应用于自然语言处理和词袋表示的模式识别

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

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

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

对于文本分类问题，有两种分布假设

多项式分布假设

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

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

最大似然估计

- 等式约束的最大似然估计

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

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

- 拉格朗日乘子法

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

$$= \sum_{k=1}^{N}\sum_{j=1}^{C} I\left(y^{(k)} = c_j\right)\left(\log\pi_j + \sum_{i=1}^{V} N\left(t_i,\boldsymbol{x}^{(k)}\right)\log\boldsymbol{\theta}_{i|j}\right) + \alpha\left(1-\sum_{j=1}^{C}\pi_j\right) + \sum_{j=1}^{C}\beta_j\left(1-\sum_{i=1}^{V}\theta_{i|j}\right)$$

最大似然估计解析解

- 梯度置零

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

$$\frac{\partial J}{\partial \theta_{i|j}} = \sum_{k=1}^{N} I\left(y^{(k)} = c_j\right)\frac{N\left(t_i, \boldsymbol{x}^{(k)}\right)}{\theta_{i|j}} - \beta_j = 0$$

- 闭式解

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

$$\theta_{i|j} = \frac{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) N\left(t_i, \boldsymbol{x}^{(k)}\right)}{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) \sum_{i'=1}^{V} N\left(t_{i'}, \boldsymbol{x}^{(k)}\right)}$$

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

$$\theta_{i|j} = \frac{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) N\left(t_i, \boldsymbol{x}^{(k)}\right)}{\sum_{i'=1}^{V} \sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) N\left(t_{i'}, \boldsymbol{x}^{(k)}\right)}$$

$$\Downarrow$$

$$\theta_{i|j} = \frac{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) N\left(t_i, \boldsymbol{x}^{(k)}\right) + 1}{\sum_{i'=1}^{V} \sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) N\left(t_{i'}, \boldsymbol{x}^{(k)}\right) + V}$$

多变量伯努利分布假设

模型描述

- 假设

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

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

- 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

$$
\begin{aligned}
L(\boldsymbol{\pi}, \boldsymbol{\mu}) &= \log \prod_{k=1}^{N} p(\boldsymbol{x}^{(k)}, y^{(k)}) \\
&= \sum_{k=1}^{N} \log \sum_{j=1}^{C} I\left(y^{(k)} = c_j\right) p\left(\boldsymbol{x}^{(k)}, y^{(k)}\right) \\
&= \sum_{k=1}^{N} \sum_{j=1}^{C} I\left(y^{(k)} = c_j\right) \log p(c_j) \prod_{i=1}^{V} Ip\left(t_i \epsilon \boldsymbol{x}^{(k)}\right)\left(t_i \middle| c_j\right) + I\left(t_i \notin \boldsymbol{x}^{(k)}\right)\left(1 - p\left(t_i \middle| c_j\right)\right) \\
&= \sum_{k=1}^{N} \sum_{j=1}^{C} I\left(y^{(k)} = c_j\right) \left(\log \pi_j + \sum_{i=1}^{V} I(t_i \epsilon \boldsymbol{x}_k) \log \mu_{i|j} + I(t_i \notin \boldsymbol{x}_k) \log(1 - \mu_{i|j}) \right)
\end{aligned}
$$

最大似然估计

- 等式约束下的最大似然估计

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

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

- 拉格朗日乘子法

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

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

最大似然估计解析解

- 梯度置零

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

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

- 闭式解

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

$$\mu_{i|j} = \frac{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) I\left(t_i \epsilon \boldsymbol{x}^{(k)}\right)}{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right)}$$

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

$$\mu_{i|j} = \frac{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) I\left(t_i \epsilon \boldsymbol{x}^{(k)}\right)}{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right)}$$

$$\Downarrow$$

$$\mu_{i|j} = \frac{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) I\left(t_i \epsilon \boldsymbol{x}^{(k)}\right) + 1}{\sum_{k=1}^{N} I\left(y^{(k)} = c_j\right) + 2}$$

朴素贝叶斯文本分类举例

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

- 类别标签

c1 = C
c2 = J

- 特征向量

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

多项式朴素贝叶斯

- 训练

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

c1 = C
c2 = J

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

		Doc	t1	t2	t3	t4	t5	t6
频率	c1	3	1	5	0	1	1	0
	c2	1	0	1	1	0	0	1
概率	c1	$\pi_1 = \frac{3}{4}$	$\theta_{1\|1} = \frac{2}{14}$	$\theta_{2\|1} = \frac{5+1}{1+5+1+1+6} = \frac{6}{14}$	$\theta_{3\|1} = \frac{1}{14}$	$\theta_{4\|1} = \frac{2}{14}$	$\theta_{5\|1} = \frac{2}{14}$	$\theta_{6\|1} = \frac{1}{14}$
	c2	$\pi_1 = \frac{1}{4}$	$\theta_{1\|2} = \frac{1}{9}$	$\theta_{2\|2} = \frac{1+1}{1+1+1+1+6} = \frac{2}{9}$	$\theta_{3\|2} = \frac{2}{9}$	$\theta_{4\|2} = \frac{1}{9}$	$\theta_{5\|2} = \frac{1}{9}$	$\theta_{6\|2} = \frac{2}{9}$

多项式朴素贝叶斯

- 训练结果

	Doc	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**
c1	3/4	2/14	6/14	1/14	2/14	2/14	1/14
c2	1/4	1/9	2/9	2/9	1/9	1/9	2/9

c1=C
c2=J

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

- 测试

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

联合分布	**后验概率**
$P(d_{te}1,c1)=(3/4)*(6/14)^3*(1/14)*(1/14)=0.003012$	$P(c1\|d_{te}1)=0.689718$
$P(d_{te}1,c2)=(1/4)*(2/9)^3*(2/9)*(2/9)=0.001355$	$P(c2\|d_{te}1)=0.310282$
$P(d_{te}2,c1)=(3/4)*(1/14)^2*(1/14)*(2/14)=0.000039$	$P(c1\|d_{te}2)=0.113372$
$P(d_{te}2,c2)=(1/4)*(2/9)^2*(2/9)*(1/9)=0.000305$	$P(c2\|d_{te}2)=0.886628$

多变量伯努利朴素贝叶斯

- 训练

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

c1 = C
c2 = J

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

		Doc	t1	t2	t3	t4	t5	t6
频率	c1	3	1	5	0	1	1	0
	c2	1	0	1	1	0	0	1
概率	c1	$\pi_1 = \frac{3}{4}$	$\mu_{1\|1} = \frac{2}{5}$	$\mu_{2\|1} = \frac{3+1}{3+2} = \frac{4}{5}$	$\mu_{3\|1} = \frac{1}{5}$	$\mu_{4\|1} = \frac{2}{5}$	$\mu_{5\|1} = \frac{2}{5}$	$\mu_{6\|1} = \frac{1}{5}$
	c2	$\pi_1 = \frac{1}{4}$	$\mu_{1\|2} = \frac{1}{3}$	$\mu_{2\|2} = \frac{1+1}{1+2} = \frac{2}{3}$	$\mu_{3\|2} = \frac{2}{3}$	$\mu_{4\|2} = \frac{1}{3}$	$\mu_{5\|2} = \frac{1}{3}$	$\mu_{6\|2} = \frac{2}{3}$

多变量伯努利朴素贝叶斯

- 训练结果

	Doc	t1	t2	t3	t4	t5	t6
c1	3/4	2/5	4/5	1/5	2/5	2/5	1/5
c2	1/4	1/3	2/3	2/3	1/3	1/3	2/3

c1=C
c2=J

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

- 测试

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

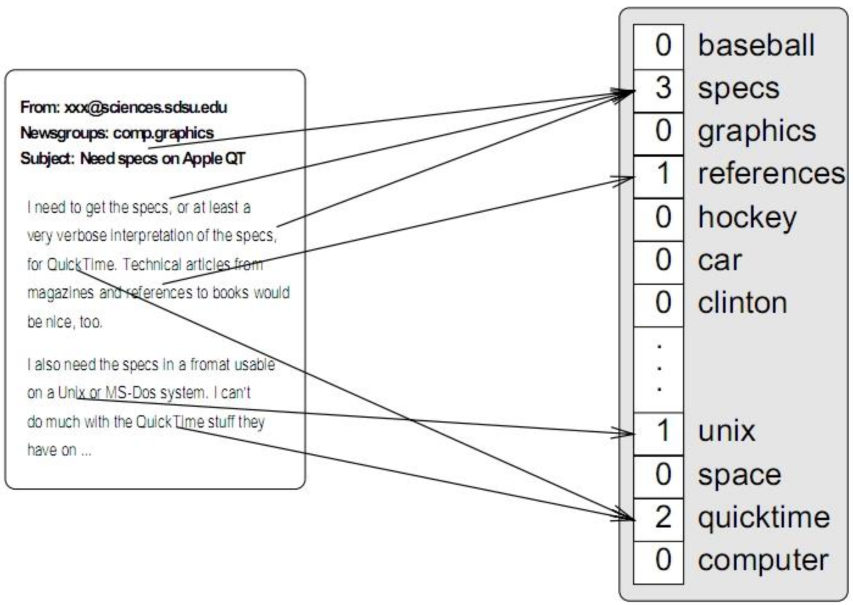
联合分布	后验概率
$P(d_{te}1,c1)=(3/4)*(1-2/5)*(4/5)*(1/5)*(1-2/5)*(1-2/5)*(1/5)=0.005184$	$P(c1\|d_{te}1)=0.191066$
$P(d_{te}1,c2)=(1/4)*(1-1/3)*(2/3)*(2/3)*(1-1/3)*(1-1/3)*(2/3)=0.021948$	$P(c2\|d_{te}1)=0.808934$
$P(d_{te}2,c1)=(3/4)*(1-2/5)*(1-4/5)*(1/5)*(1-2/5)*(2/5)*(1/5)=0.000864$	$P(c1\|d_{te}2)=0.136042$
$P(d_{te}2,c2)=(1/4)*(1-1/3)*(1-2/3)*(2/3)*(1-1/3)*(1/3)*(2/3)=0.005487$	$P(c2\|d_{te}2)=0.863958$

分组作业#6：基于朴素贝叶斯的文本分类

- 基于以下类条件分布假设实现朴素贝叶斯模型
  - 多项分布
  - 多变量伯努利分布
- 基于上述实现，在以下清华文本分类数据集上继续宁模型训练和测试，并词表大小、报告分类正确率。
  http://www.nustm.cn/member/rxia/ml/data/Tsinghua.zip
- 实现基于向量空间模型作为文本表示方法的softmax回归模型，支持TF、BOOL两种特征权重（具体方法见下页），并在上述数据上进行训练和测试，报告分类正确率、并绘制损失函数下降动态曲线。
- 在上述数据上比较朴素贝叶斯和softmax回归模型，其中多项分布朴素贝叶斯与基于TF权重的softmax回归模型比较，多变量伯努利分布模型与基于BOOL权重的softmax回归模型比较。

基于向量空间模型的文本表示

- 向量空间模型



词表 $[t_1, t_2, \cdots, t_i, \cdots, t_V] =$
[baseball, specs, graphics, …, quicktime, computer]

- 特征权重方法
  - BOOL (presence)

$$\omega_{ki} = \begin{cases} 1, & if\ t_i\ exists\ in\ \boldsymbol{d}_k \\ 0, & otherwise \end{cases}$$

  - Term frequency (TF)

$$\omega_{ki} = tf_{ki}$$

  - Inverse document frequency (IDF)

$$\omega_i = \log \frac{N}{df_i}$$

  - TF-IDF

$$\omega_{ki} = tf_{ki} \cdot \log \frac{N}{df_i}$$

欢迎提问！