



中国科学院南京分院  
Nanjing Branch of Chinese Academy of Sciences

# 人工智能原理与算法

## 6. 朴素贝叶斯模型

夏睿

2023.3.15

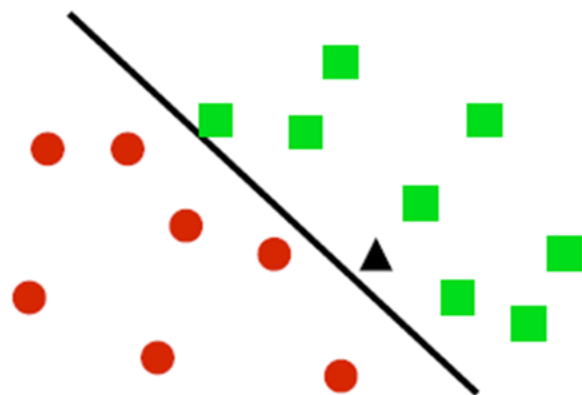
# 目录

- 生成式模型 vs. 判别式模型
- 多项分布朴素贝叶斯
- 多变量伯努利分布朴素贝叶斯
- 过拟合问题与模型正则化
- 高斯分布朴素贝叶斯

# 生成式模型 vs. 判别式模型

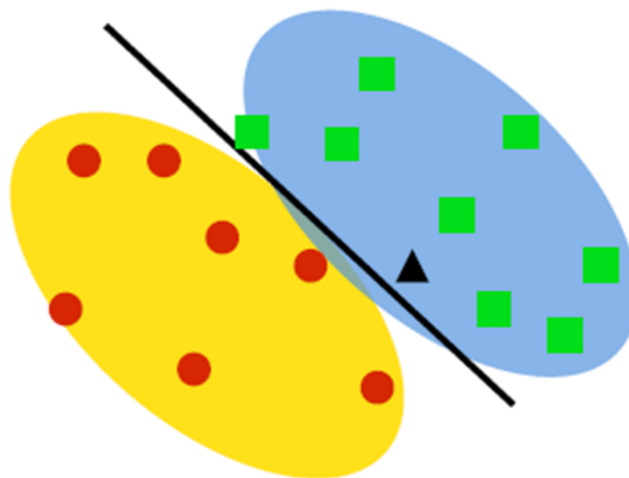
# 生成式 vs. 判别式

- 判别式模型



针对给定输入特征时输出标签的后验概率 $p(y|x)$ 建模，或者直接对预测函数 $y = f(x)$ 建模

- 生成式模型



对输入特征和输出标签的联合分布 $p(x, y)$ 建模，再基于贝叶斯公式 $p(y|x) = p(x, y)/p(x)$ 进行预测

# 假设 – 学习 – 预测

- 判别式模型

- 直接对预测函数建模

$$y = f(\mathbf{x})$$

例子:

Perceptron, SVMs

- 对后验概率建模

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

例子:

Logistic/Softmax Regression

- 生成式模型（对联合分布建模）

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

例子:

Naïve Bayes, Gaussian  
Mixture Model

# 假设 – 学习 – 预测

- 判别式模型
  - 决策函数直接建模

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

学习准则：某种损失函数，如感知机损失、交叉熵损失、最大间隔损失等

- 后验概率建模

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

学习准则：最大似然估计（后验分布）  
⇔ 与某些损失函数等价

- 生成式模型（联合分布建模）

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

学习准则：最大似然估计（联合分布）

# 假设 – 学习 – 预测

- 判别式模型
  - 预测函数

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

- 后验概率

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

- 生成式模型（贝叶斯公式）

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



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

# 生成式模型（以分类为例）

- 模型假设：联合分布

$$p(\mathbf{x}, y = j) = p(y = j)p(\mathbf{x}|y = j)$$

类条件分布

类先验分布

- 类先验分布的模型假设

类别分布：

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

- 类条件分布的模型假设

- 离散的情形，如多项分布

$$p(\mathbf{x}|y = j) = \prod_{i=1}^M \theta_{i,j}^{N(w_i, \mathbf{x})}$$

- 连续的情形，如高斯分布

$$p(\mathbf{x}|y = j) = N(\mathbf{x}|\mu_j, \Sigma_j)$$



# 两种离散分布的朴素贝叶斯模型

$$\mathbf{x} = (w_1, w_2, \dots, w_{|\mathbf{x}|})$$



词袋 (Bag-of-words, BOW)表示

$$p(\mathbf{x}|y = j) = p(w_1, w_2, \dots, w_{|\mathbf{x}|}|y = j) = \prod_{h=1}^{|\mathbf{x}|} p(w_h|y = j)$$



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



一个V面的筛子 (多项分布)

$$p(\mathbf{x}|y = j) = \prod_{i=1}^V p(t_i|y = j)^{I(t_i \in \mathbf{x})} (1 - p(t_i|y = j))^{I(t_i \notin \mathbf{x})}$$



V个不同的硬币 (多变量伯努利分布)

# 抛硬币实验的最大似然估计

- 伯努利分布

$$P(Z = z) = \text{Bern}(z; \mu) = \mu^z (1 - \mu)^{1-z}, z \in \{0, 1\}$$

- 对数似然函数

$$\begin{aligned} \log L &= \log \prod_{k=1}^N \text{Bern}(z^{(k)} | \mu) \\ &= N_1 \log \mu + N_0 \log(1 - \mu) \end{aligned}$$



抛硬币

- 最大似然估计 (MLE)

$$\frac{d \log L}{d \mu} = \frac{N_1}{\mu} - \frac{N_0}{1 - \mu} = 0 \quad \Rightarrow \quad \hat{\mu}_{\text{ML}} = \frac{N_1}{N}$$

# 掷骰子实验的最大似然估计

- 类别分布

$$P(Z = z) = \text{Cate}(z|\theta) = \theta_z, \quad z \in \{1, \dots, V\}$$

$$\text{s. t. } \sum_{z=1}^V \theta_z = 1,$$

- 对数似然函数

$$\begin{aligned} \log L &= \log \prod_{k=1}^N \text{Cate}(z^{(k)}|\theta) \\ &= \sum_{z=1}^V N_z \log \theta_z \end{aligned}$$

- 最大似然估计

$$\widehat{\theta}_{z\text{ML}} = \frac{N_z}{N}$$



掷骰子

# 多项分布 (Multinomial Distribution)

## 朴素贝叶斯

# 模型假设

- 类先验分布和类条件分布

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

$$\begin{aligned} p(\mathbf{x}|y = j) &= p((\omega_1, \omega_2, \dots, \omega_{|\mathbf{x}|})|y = j) = \prod_{h=1}^{|\mathbf{x}|} p(\omega_h|y = j) \\ &= \prod_{i=1}^V p(t_i|c_j)^{N(t_i, \mathbf{x})} = \prod_{i=1}^V \theta_{i|j}^{N(t_i, \mathbf{x})} \end{aligned}$$

- 联合分布

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

模型参数

# 似然函数

- （联合分布）似然函数

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

# 最大似然估计

- 等式约束的最大似然估计

$$\begin{aligned} & \max_{\boldsymbol{\pi}, \boldsymbol{\theta}} L(\boldsymbol{\pi}, \boldsymbol{\theta}) \\ & \text{s. t. } \begin{cases} \sum_{j=1}^C \pi_j = 1 \\ \sum_{i=1}^V \theta_{i|j} = 1, j = 1, \dots, C \end{cases} \end{aligned}$$

- 拉格朗日乘子法

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

# 最大似然估计解析解

- 求导置零

$$\frac{\partial J}{\partial \pi_j} = \sum_{k=1}^N I(y^{(k)} = j) \frac{1}{\pi_j} - \alpha = 0$$
$$\frac{\partial J}{\partial \theta_{i|j}} = \sum_{k=1}^N I(y^{(k)} = j) \frac{N(t_i, \mathbf{x}^{(k)})}{\theta_{i|j}} - \beta_j = 0$$

- 解析解

$$\pi_j = \frac{\sum_{k=1}^N I(y^{(k)} = j)}{\sum_{k=1}^N \sum_{j'=1}^C I(y^{(k)} = j')} = \frac{N_j}{N}$$
$$\theta_{i|j} = \frac{\sum_{k=1}^N I(y^{(k)} = j) N(t_i, \mathbf{x}^{(k)})}{\sum_{k=1}^N I(y^{(k)} = j) \sum_{i'=1}^V N(t_{i'}, \mathbf{x}^{(k)})}$$



# 参数估计推导过程

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

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

$$\pi_j = \sum_{k=1}^N \frac{I(y^{(k)} = j)}{\alpha}$$

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

$$\alpha = \sum_{j'=1}^C \sum_{k=1}^N I(y^{(k)} = j') = N$$

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

$$\theta_{i|j} = \sum_{k=1}^N \frac{I(y^{(k)} = j) N(t_i, \mathbf{x}^{(k)})}{\beta_j}$$

$$\sum_{i'=1}^V \theta_{i|j} = \sum_{k=1}^N \sum_{i'=1}^V I(y^{(k)} = j) \frac{N(t_{i'}, \mathbf{x}^{(k)})}{\beta_j} = 1$$

$$\beta_j = \sum_{k=1}^N \sum_{i'=1}^V I(y^{(k)} = j) N(t_{i'}, \mathbf{x}^{(k)})$$

# 拉普拉斯平滑

- 为防止零概率

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

- 拉普拉斯平滑

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



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

# 一个简单的文本分类数据集

- 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

$c1 = C$

$c2 = J$

- 词表（特征集）

$t1 = \text{Beijing}$

$t2 = \text{Chinese}$

$t3 = \text{Japan}$

$t4 = \text{Macao}$

$t5 = \text{Shanghai}$

$t6 = \text{Tokyo}$

# 多项分布朴素贝叶斯示例

## • 训练

ID	Text	Label
$d_{tr1}$	Chinese Beijing Chinese	C
$d_{tr2}$	Chinese Chinese Shanghai	C
$d_{tr3}$	Chinese Macao	C
$d_{tr4}$	Tokyo Japan Chinese	J

$c1 = C$   
 $c2 = J$

$t1 = \text{Beijing}$   
 $t2 = \text{Chinese}$   
 $t3 = \text{Japan}$   
 $t4 = \text{Macao}$   
 $t5 = \text{Shanghai}$   
 $t6 = \text{Tokyo}$

		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	$\pi_1$	$\theta_{1 1}$	$\theta_{2 1}$	$\theta_{3 1}$	$\theta_{4 1}$	$\theta_{5 1}$	$\theta_{6 1}$
	c2	$\pi_2$	$\theta_{1 2}$	$\theta_{2 2}$	$\theta_{3 2}$	$\theta_{4 2}$	$\theta_{5 2}$	$\theta_{6 2}$

# 多项分布朴素贝叶斯示例

## • 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

## • 测试样本

ID	Text
d <sub>te</sub> 1	Chinese Chinese Chinese Tokyo Japan
d <sub>te</sub> 2	Tokyo Tokyo Japan Shanghai

联合分布	后验概率
$P(d_{te}1, c1)$	$P(c1 d_{te}1)$
$P(d_{te}1, c2)$	$P(c2 d_{te}1)$
$P(d_{te}2, c1)$	$P(c1 d_{te}2)$
$P(d_{te}2, c2)$	$P(c2 d_{te}2)$

# 多变量伯努利分布 (Multi-variate Bernoulli Distribution) 朴素贝叶斯

# 模型假设

- 类先验分布和类条件分布

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

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

- 联合分布

模型参数

$$p(\mathbf{x}, y = j) = \pi_j \prod_{i=1}^V \mu_{i|j}^{I(t_i \in \mathbf{x})} (1 - \mu_{i|j})^{I(t_i \notin \mathbf{x})}$$

# 似然函数

- （联合分布）似然函数

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



# 最大似然估计

- 等式约束下的最大似然估计

$$\begin{aligned} & \max_{\boldsymbol{\pi}, \boldsymbol{\mu}} L(\boldsymbol{\pi}, \boldsymbol{\mu}) \\ & \text{s. t. } \sum_{j=1}^C \pi_j = 1 \end{aligned}$$

- 拉格朗日乘子法

$$\begin{aligned} J &= L(\boldsymbol{\pi}, \boldsymbol{\mu}) + \alpha \left( 1 - \sum_{j=1}^C \pi_j \right) \\ &= \sum_{k=1}^N \sum_{j=1}^C I(y^{(k)} = j) \left( \log \pi_j + \sum_{i=1}^V I(t_i \in \mathbf{x}^{(k)}) \log \mu_{i|j} + I(t_i \notin \mathbf{x}^{(k)}) \log(1 - \mu_{i|j}) \right) + \alpha \left( 1 - \sum_{j=1}^C \pi_j \right) \end{aligned}$$

# 最大似然估计解析解

- 求导置零

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

$$\frac{\partial J}{\partial \mu_{i|j}} = \sum_{k=1}^N I(y^{(k)} = j) \left( \frac{I(t_i \in \mathbf{x}^{(k)})}{\mu_{i|j}} - \frac{I(t_i \notin \mathbf{x}^{(k)})}{1 - \mu_{i|j}} \right) = 0$$

- 解析解

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

$$\mu_{i|j} = \frac{\sum_{k=1}^N I(y^{(k)} = j) I(t_i \in \mathbf{x}^{(k)})}{\sum_{k=1}^N I(y^{(k)} = j)}$$

# 拉普拉斯平滑

- 为防止零概率

$$p(\mathbf{x}, y = j) = \pi_j \prod_{i=1}^V I(t_i \in \mathbf{x}) \mu_{i|j} + I(t_i \notin \mathbf{x}) (1 - \mu_{i|j})$$

- 拉普拉斯平滑

$$\mu_{i|j} = \frac{\sum_{k=1}^N I(y^{(k)} = j) I(t_i \in \mathbf{x}^{(k)})}{\sum_{k=1}^N I(y^{(k)} = j)}$$



$$\mu_{i|j} = \frac{\sum_{k=1}^N I(y^{(k)} = j) I(t_i \in \mathbf{x}^{(k)}) + 1}{\sum_{k=1}^N I(y^{(k)} = j) + 2}$$

# 一个简单的文本分类数据集

- 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

$c1 = C$

$c2 = J$

- 词表（特征集）

$t1 = \text{Beijing}$

$t2 = \text{Chinese}$

$t3 = \text{Japan}$

$t4 = \text{Macao}$

$t5 = \text{Shanghai}$

$t6 = \text{Tokyo}$

# 多变量伯努利分布朴素贝叶斯示例

- 训练

ID	Text	Label
$d_{tr1}$	Chinese Beijing Chinese	C
$d_{tr2}$	Chinese Chinese Shanghai	C
$d_{tr3}$	Chinese Macao	C
$d_{tr4}$	Tokyo Japan Chinese	J

$c1 = C$   
 $c2 = J$

$t1 = \text{Beijing}$   
 $t2 = \text{Chinese}$   
 $t3 = \text{Japan}$   
 $t4 = \text{Macao}$   
 $t5 = \text{Shanghai}$   
 $t6 = \text{Tokyo}$

		Doc	t1	t2	t3	t4	t5	t6
频率	c1	3	1	3	0	1	1	0
	c2	1	0	1	1	0	0	1
概率	c1	$\pi_1$	$\mu_{1 1}$	$\mu_{2 1}$	$\mu_{3 1}$	$\mu_{4 1}$	$\mu_{5 1}$	$\mu_{6 1}$
	c2	$\pi_2$	$\mu_{1 2}$	$\mu_{2 2}$	$\mu_{3 2}$	$\mu_{4 2}$	$\mu_{5 2}$	$\mu_{6 2}$

# 多变量伯努利分布朴素贝叶斯示例

- 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

$c1=C$

$c2=J$

$t1 = \text{Beijing}$

$t2 = \text{Chinese}$

$t3 = \text{Japan}$

$t4 = \text{Macao}$

$t5 = \text{Shanghai}$

$t6 = \text{Tokyo}$

- 测试样本

ID	Text
$d_{te1}$	Chinese Chinese Chinese Tokyo Japan
$d_{te2}$	Tokyo Tokyo Japan Shanghai

联合分布	后验概率
$P(d_{te1}, c1)$	$P(c1 d_{te1})$
$P(d_{te1}, c2)$	$P(c2 d_{te1})$
$P(d_{te2}, c1)$	$P(c1 d_{te2})$
$P(d_{te2}, c2)$	$P(c2 d_{te2})$

# 作业#3

清华文本分类数据集

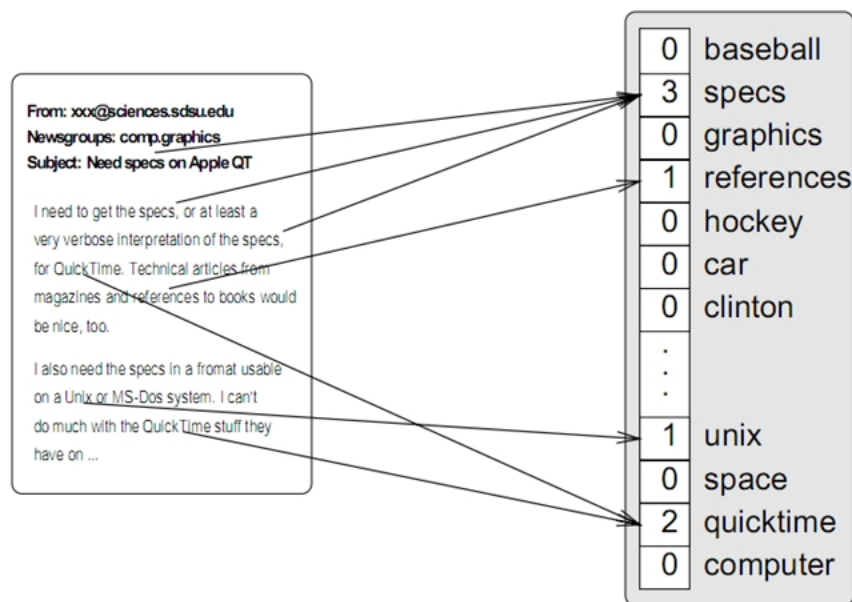
类别	全部	训练样本数	测试样本数
教育	213	150	63
电脑	185	130	55
体育	582	300	282
经济	370	250	120
政治	247	165	82
法律	352	200	152

<http://www.nustm.cn/member/rxia/ml/data/Tsinghua.zip>

- 基于多项分布和多变量伯努利分布两种类条件分布假设实现朴素贝叶斯模型，应用于上述文本分类数据集（不另外分词、不增加停词），报告分类正确率。
- 实现基于向量空间模型进行文本表示，建立文档的特征向量，支持TF、BOOL两种特征权重（具体方法见下页）。基于LibSVM工具包实现支持向量机分类，调节核函数、tradeoff权重C等参数，报告分类正确率。
- 比较下列模型结果（多项分布朴素贝叶斯 vs. 基于TF权重的SVM；多变量伯努利分布模型 vs. 基于BOOL权重的SVM）

# 基于向量空间模型的文本表示

- 向量空间模型



词表  $[t_1, t_2, \dots, t_i, \dots, t_V] =$

[baseball, specs, graphics, ..., quicktime, computer]

- 特征取值计算方法

- BOOL

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

- Term frequency (TF)

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

- Inverse document frequency (IDF)

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

- TF-IDF

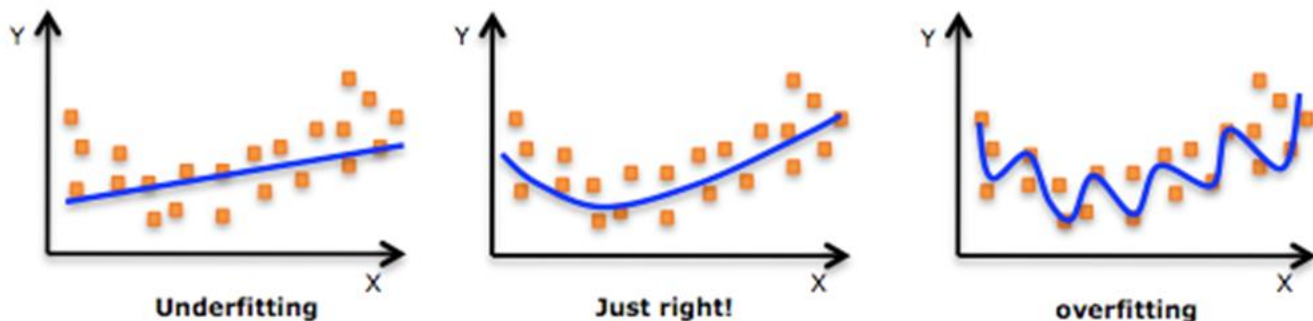
$$\omega_{ki} = tf_{ki} \cdot \log \frac{N}{df_i}$$



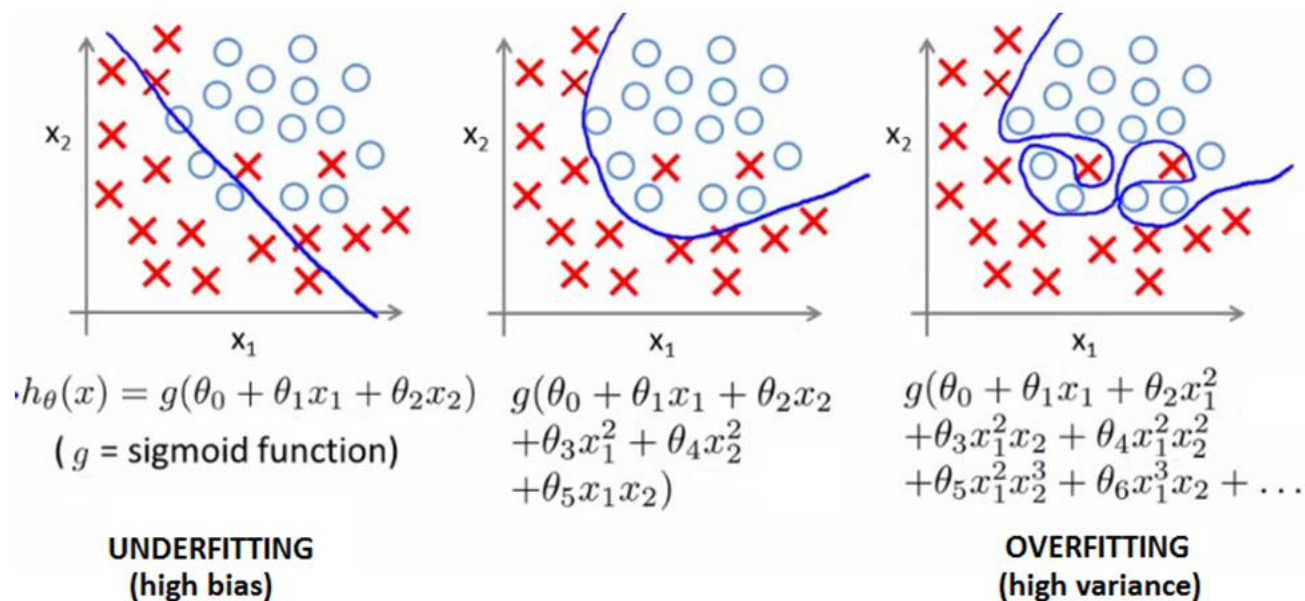
# 过拟合问题与模型正则化

# 过拟合问题

- 回归



- 分类



# 最大似然估计 vs. 最大后验概率估计

- 最大似然Maximum Likelihood (ML)

$$\begin{aligned}\theta_{ML}^* &= \arg \max_{\theta} L(\theta) = \arg \max_{\theta} p(\mathcal{D}|\theta) \\ &= \arg \max_{\theta} \sum_{d \in \mathcal{D}} \log p(d|\theta)\end{aligned}$$

Likelihood

- 最大后验概率Maximum A Posteriori (MAP)

$$\begin{aligned}\theta_{MAP}^* &= \arg \max_{\theta} p(\theta|\mathcal{D}) = \arg \max_{\theta} \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ &= \arg \max_{\theta} p(\mathcal{D}|\theta)p(\theta) \\ &= \arg \max_{\theta} \sum_{d \in \mathcal{D}} \log p(d|\theta) + \log p(\theta)\end{aligned}$$

Likelihood · Prior

# 正则化方法

- 生成式模型

$$\theta_{\text{ML}}^* = \arg \max_{\theta} \sum_{d \in \mathcal{D}} \log p(d|\theta)$$



$$\theta_{\text{MAP}}^* = \arg \max_{\theta} \sum_{d \in \mathcal{D}} \log p(x|\theta) + \log p(\theta)$$

以参数先验分布  
作为正则化项

- 判别式模型

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

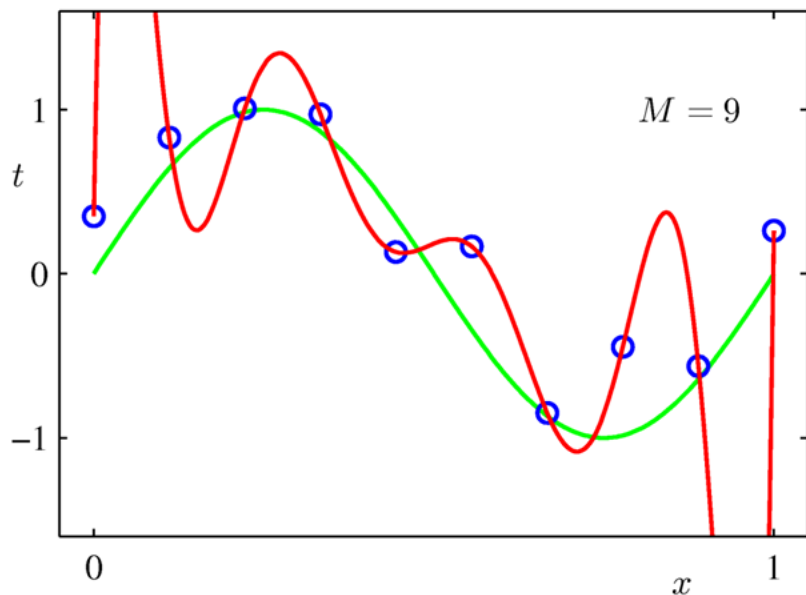


$$\theta^* = \arg \max_{\theta} J(\theta) + \lambda R(\theta)$$

以参数L2、L1范数等  
作为正则化项

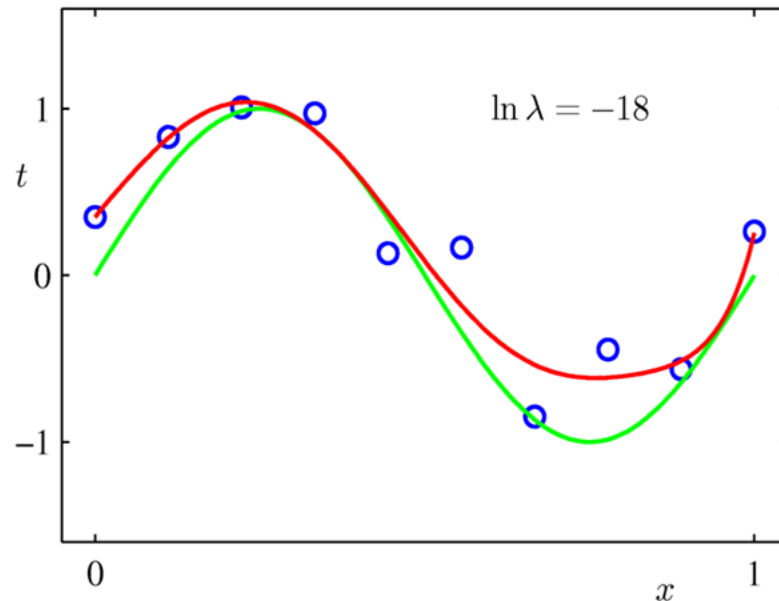
$$\hat{J}_c(\theta) = \sum_{k=1}^N y^{(k)} \log h_{\theta}(x^{(k)}) + (1 - y^{(k)}) \log (1 - h_{\theta}(x^{(k)})) + \frac{\lambda}{2} \|\theta\|^2 \quad \text{Logistic回归}$$

# 正则化方法（多项式回归）



ML

$$L_{\text{ML}}(\mathbf{w}) = -\frac{1}{2} \sum_{k=1}^N (h_{\mathbf{w}}(\mathbf{x}^{(k)}) - \mathbf{y}^{(k)})^2$$



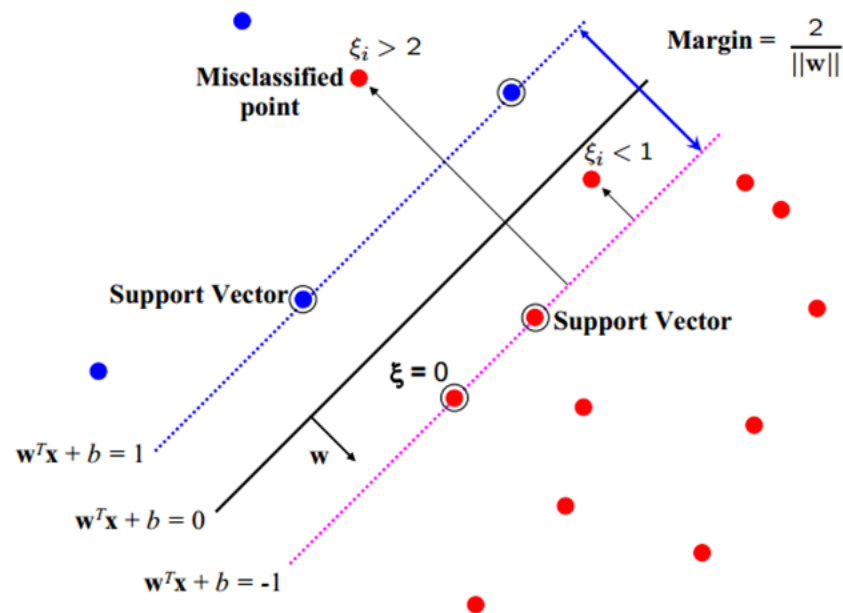
MAP

$$L_{\text{MAP}}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (h_{\mathbf{w}}(\mathbf{x}^{(k)}) - \mathbf{y}^{(k)})^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$$

# 正则化方法（支持向量机）

## 软间隔支持向量机

$$\begin{aligned} \min_{\omega, b, \varepsilon} \quad & \frac{1}{2} \|\omega\|^2 + C \sum_{k=1}^N \varepsilon_k \\ \text{s.t.} \quad & y^{(k)} (\omega^T x^{(k)} + b) \geq 1 - \varepsilon_k, \\ & \varepsilon_k \geq 0 \end{aligned}$$



写成Hinge Loss的形式

$$L_{\text{hinge}} = \sum_{k=1}^N \varepsilon_k + \frac{\lambda}{2} \|\omega\|^2 = \sum_{k=1}^N \max\{0, 1 - y^{(k)} (\omega^T x^{(k)} + b)\} + \boxed{\frac{\lambda}{2} \|\omega\|^2}$$

正则项

# 正则化方法（抛硬币实验）

- 伯努利分布

$$P(Z = z) = \text{Bern}(z|\mu) = \mu^z(1 - \mu)^{1-z}, z \in \{0,1\}$$

- 对数似然函数

$$\begin{aligned}\log L &= \log \prod_{k=1}^N \text{Bern}(z^{(k)}|\mu) \\ &= N_1 \log P(1|\mu) + N_0 \log P(0|\mu) \\ &= N_1 \log \mu + N_0 \log(1 - \mu)\end{aligned}$$



抛硬币

- 最大似然估计 (Maximum likelihood estimation, MLE)

$$\frac{d \log L}{d\mu} = \frac{N_1}{\mu} - \frac{N_0}{1 - \mu} = 0 \quad \Rightarrow \quad \hat{\mu}_{\text{ML}} = \frac{N_1}{N}$$

# 正则化方法（抛硬币实验）


- 引入参数的先验分布

$$p(\mu|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1} \triangleq \text{Beta}(\mu|\alpha, \beta)$$

$$\text{其中 } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$

- 最大后验概率估计值

$$\frac{d}{d\mu} \log P(\mu|Z) = \frac{d}{d\mu} \log P(Z|\mu) + \log P(\mu) = \frac{N_1}{\mu} - \frac{N_0}{1-\mu} + \frac{\alpha-1}{\mu} - \frac{\beta-1}{1-\mu} = 0$$


$$\hat{\mu}_{\text{MAP}} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}$$

$\alpha$ 、 $\beta$ 取特定取值时  
与拉普拉斯平滑等价

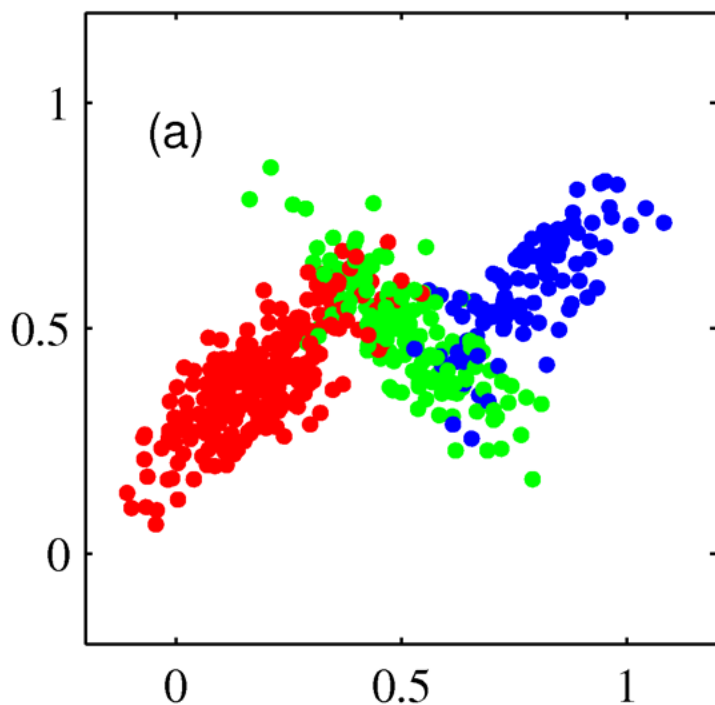


# 高斯分布 (Gaussian Distribution)

## 朴素贝叶斯

# 混合高斯分布数据的分类

三类高斯分布  
的混合数据



- 判别式模型
  - Softmax 回归
  - 多类感知机
  - 前向神经网络
- 生成式模型
  - 朴素贝叶斯?
  - 如何对类条件概率建模?

# 模型假设

- 类先验分布和类条件分布

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

$$p(\mathbf{x}|y = j) = \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_j|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{u}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \mathbf{u}_j)\right)$$

- 联合分布

$$p(\mathbf{x}, y = j) = p(y = j)p(\mathbf{x}|y = j) = \pi_j \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_j|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{u}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \mathbf{u}_j)\right)$$

模型参数

# 最大似然估计

- 似然函数（联合分布）

$$\begin{aligned} L(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{k=1}^N \log p(\mathbf{x}^{(k)}, y^{(k)}) \\ &= \sum_{k=1}^N \log \sum_{j=1}^C I(y^{(k)} = j) p(\mathbf{x}^{(k)}, y^{(k)} = j) \\ &= \sum_{k=1}^N \sum_{j=1}^C I(y^{(k)} = j) \log p(\mathbf{x}^{(k)}, y^{(k)} = j) \\ &= \sum_{k=1}^N \sum_{j=1}^C I(y^{(k)} = j) \log \pi_j \frac{\exp\left(-\frac{1}{2}(\mathbf{x}^{(k)} - \mathbf{u}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}^{(k)} - \mathbf{u}_j)\right)}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_j|}} \end{aligned}$$

# 最大似然估计

- 求导置零得到解析解

$$\frac{\partial J}{\partial \pi_j} = 0 \quad \Rightarrow \quad \pi_j = \frac{\sum_{k=1}^N I(y^{(k)} = j)}{N}$$

$$\frac{\partial J}{\partial \boldsymbol{\mu}_j} = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_j = \frac{\sum_{k=1}^N I(y^{(k)} = j) \mathbf{x}^{(k)}}{\sum_{k=1}^N I(y^{(k)} = j)}$$

$$\frac{\partial J}{\partial \boldsymbol{\Sigma}_j} = 0 \quad \Rightarrow \quad \boldsymbol{\Sigma}_j = \frac{\sum_{k=1}^N I(y^{(k)} = j) (\mathbf{x}^{(k)} - \boldsymbol{\mu}_j)(\mathbf{x}^{(k)} - \boldsymbol{\mu}_j)^T}{\sum_{k=1}^N I(y^{(k)} = j)}$$



**本讲结束 欢迎提问**