概念学习

概念学习

概念学习

- 概念学习：通过描述概念的若干正例和反例（训练样本）归纳出该概念的通用定义。

目标概念*EnjoySport*的正例和反例

Example	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- 学习过程：概念的学习过程就是在假设空间中的搜索过程，即寻找最佳拟合训练样本的假设的过程。
- 为了提高在假设空间中的搜索效率，往往利用假设空间中的一般到特殊的偏序结构。

概念学习

- 问题：下述样本集表示的概念是什么？

目标概念*EnjoySport*的正例和反例

Example	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	**Sunny**	**Warm**	**Normal**	**Strong**	**Warm**	**Same**	**Yes**
2	**Sunny**	**Warm**	**High**	**Strong**	**Warm**	**Same**	**Yes**
3	**Rainy**	**Cold**	**High**	**Strong**	**Warm**	**Change**	**No**
4	**Sunny**	**Warm**	**High**	**Strong**	**Cool**	**Change**	**Yes**

- 假设形式是多种多样的，可以采用属性约束合取式$(p \wedge q)$的形式，但需注意是“有偏”性。
  - 特定值：$Water = Warm$
  - 任意值：$Water = ?$
  - 空值：$Water = \varnothing$

表达形式

- 假设的表达形式
  - 普通假设：表示如果满足约束则有动作，如
    $$\langle Sunny, Warm, Normal, Strong, Warm, Same \rangle$$
  - 最一般假设：表示无论怎样都有动作，如
    $$\langle ?,?,?,?,?,? \rangle$$
  - 最特殊假设：表示无论怎样都没有动作，如
    $$\langle \varnothing,\varnothing,\varnothing,\varnothing,\varnothing,\varnothing \rangle$$

概念学习原型

- 已知
  - 实例空间 X：使用属性描述的实例组成的集合
  - 概念空间 C：目标概念的集合
  - 假设空间 H：假设的集合
  - 训练样本集合 D：描述目标概念的正例和反例

$$\left\{\left\langle \mathbf{x}_1, c\left(\mathbf{x}_1\right)\right\rangle, \left\langle \mathbf{x}_2, c\left(\mathbf{x}_2\right)\right\rangle, \ldots, \left\langle \mathbf{x}_N, c\left(\mathbf{x}_N\right)\right\rangle\right\}$$

- 求解
  - 假设 h，满足 $h \in \mathcal{H}$ 且 $h(\mathbf{x}) = c(\mathbf{x}), \mathbf{x} \in \mathcal{D}$

前提条件

- 学习目标的差异
  - 概念学习的目标

$$\text{find } h \in \mathcal{H} \text{ so that } h(\mathbf{x}) = c(\mathbf{x}) \text{ where } \mathbf{x} \in \mathcal{X}$$

  - 概念学习原型的目标

$$\text{find } h \in \mathcal{H} \text{ so that } h(\mathbf{x}) = c(\mathbf{x}) \text{ where } \mathbf{x} \in \mathcal{D}$$

  - 因为 $\mathcal{D} \subset \mathcal{X}$ 且 $\|\mathcal{D}\| \ll \|\mathcal{X}\|$，故很难保证实例空间 $\boldsymbol{X}$ 的概率分布与训练样本集合 $\boldsymbol{D}$ 的概率分布是一致的。
- 归纳学习的前提条件：对于任意假设，如果在足够大的训练集合中能够很好地拟合目标函数，则在实例空间中也能够很好地拟合目标函数。

集合上的关系

- 定义：令 h_j 和 h_k 为在 X 上定义的布尔函数，则
  - 称 h_j 更加一般或相等于 h_k，记作 $h_j \geq_g h_k$，当且仅当

$$(\forall \mathbf{x} \in \mathcal{X})\left[\left(h_k(\mathbf{x})=1\right) \rightarrow \left(h_j(\mathbf{x})=1\right)\right]$$

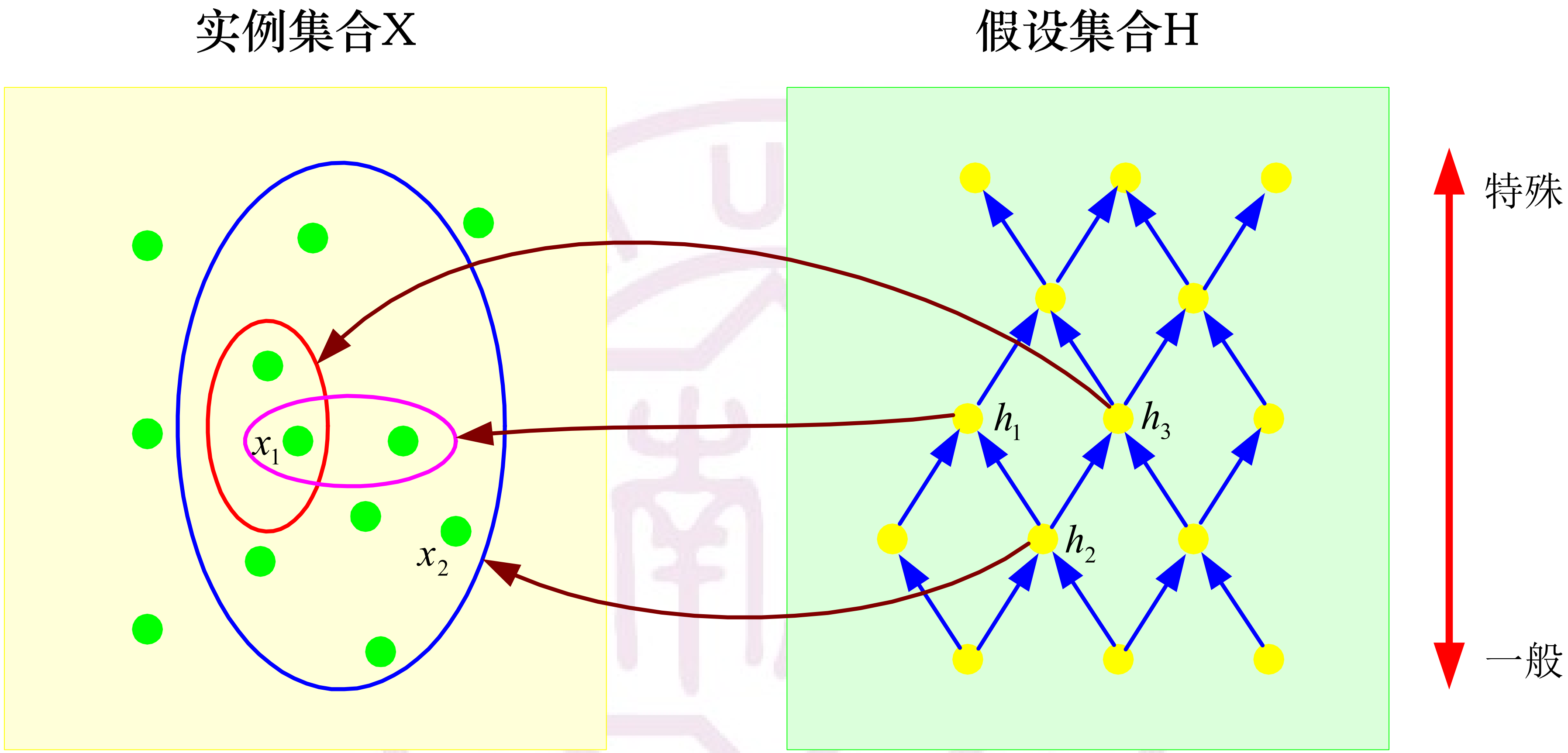
  - 称 h_j 严格地更加一般于 h_k，记作 $h_j >_g h_k$，当且仅当

$$\left(h_j \geq_g h_k\right) \wedge \left(h_k \ngeq_g h_j\right)$$

- $\geq_g$ 关系定义了假设空间 H 上的一个偏序关系。
- 相对于全序关系来说，可能存在着一对假设，满足

$$\left(h_j \ngeq_g h_k\right) \wedge \left(h_k \ngeq_g h_j\right)$$

集合上的关系



$$x_1 = \langle Sunny, Warm, High, Strong, Cool, Same \rangle$$

$$x_2 = \langle Sunny, Warm, high, Light, Warm, Same \rangle$$

$$h_1 = \langle Sunny, ?, ?, Strong, ?, ? \rangle$$

$$h_2 = \langle Sunny, ?, ?, ?, ?, ? \rangle$$

$$h_3 = \langle Sunny, ?, ?, ?, Cool, ? \rangle$$

归纳学习算法

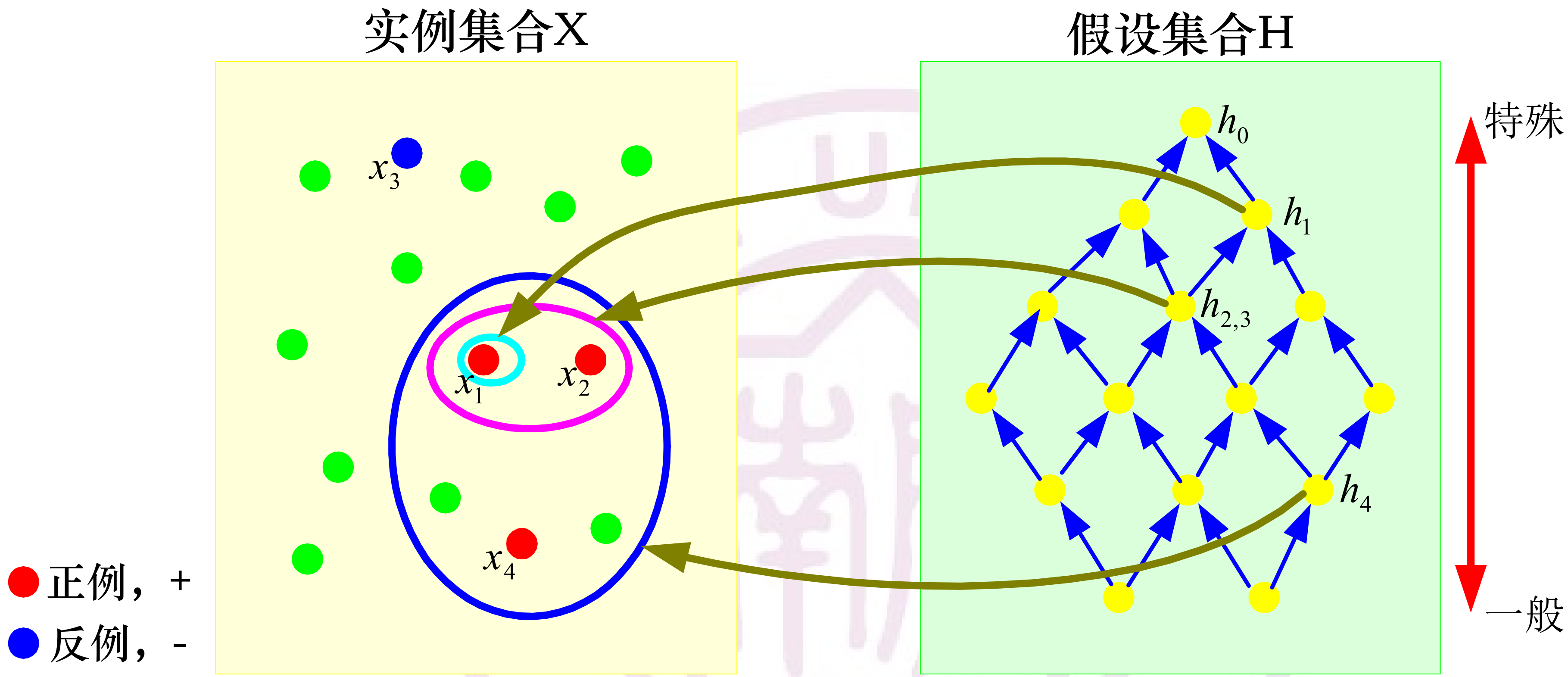
- Find-S算法：以属性约束合取式构造假设空间
  - 将假设 h 初始化为假设空间 $\boldsymbol{H}$ 中的最特殊假设
  - 对每个训练样本正例 $\boldsymbol{x}$

    对假设 h 中的每个属性约束 $a_i = v$

    如果 $\boldsymbol{x}$ 不满足 h 中的约束 $a_i = v$

    则 将 $a_i = v$ 替换为 $\boldsymbol{x}$ 满足的更一般约束
  - 输出假设 h

归纳学习算法



$x_1 = \langle Sunny, Warm, Normal, Strong, Warm, Same \rangle, +$

$x_2 = \langle Sunny, Warm, high, Strong, Warm, Same \rangle, +$

$x_3 = \langle Rainy, Cold, High, Strong, Warm, Change \rangle, -$

$x_4 = \langle Sunny, Warm, High, Strong, Cool, Change \rangle, +$

$h_0 = \langle 0,0,0,0,0,0 \rangle$

$h_1 = \langle Sunny, Warm, Normal, Strong, Warm, Same \rangle$

$h_2 = \langle Sunny, Warm, ?, Strong, Warm, Same \rangle$

$h_3 = \langle Sunny, Warm, ?, Strong, Warm, Same \rangle$

$h_4 = \langle Sunny, Warm, ?, Strong, ?, ? \rangle$

Version Space (VS)

- **定义**：一条假设 h 与训练样本集合 $\boldsymbol{D}$ 是一致的，当且仅当，对 $\boldsymbol{D}$ 中每个样本 $< x, c(x) >$ 都有 $h(x) = c(x)$，即

$$\mathrm{Cons}(h, \mathcal{D}) = \left(\forall \langle x, c(x) \rangle \in \mathcal{D}\right) h(x) = c(x)$$

- **定义**：假设空间 $\boldsymbol{H}$ 和训练样本集合 $\boldsymbol{D}$ 的 Version Space，是 $\boldsymbol{H}$ 中与训练样本集合 $\boldsymbol{D}$ 相一致的所有假设构成的子集合，标记为 $VS_{\mathcal{H},\mathcal{D}}$，即

$$VS_{\mathcal{H},\mathcal{D}} = \left\{ h \in \mathcal{H} \middle| \mathrm{Cons}(h, \mathcal{D}) \right\}$$

归纳学习算法

- 列表消除算法
  - VS $\leftarrow$ 假设空间中所有假设列表
  - 对每个训练样本 $< x, c(x) >$，从 VS 中删除所有满足条件 $h(x) \neq c(x)$ 的假设 h
  - 输出 VS 中的假设列表

VS的简洁表示

- 定义：VS 的一般边界 G， 是最一般成员的集合，即

$$G=\left\{g\in\mathcal{H}\middle|\mathrm{Cons}(g,\mathcal{D})\wedge(\neg\exists g'\in\mathcal{H})\left[\left(g'>_g g\right)\wedge\mathrm{Cons}(g',\mathcal{D})\right]\right\}$$

- 定义：VS 的特殊边界 S，是最特殊成员的集合，即

$$S=\left\{s\in\mathcal{H}\middle|\mathrm{Cons}(s,\mathcal{D})\wedge(\neg\exists s'\in\mathcal{H})\left[\left(s>_g s'\right)\wedge\mathrm{Cons}(s',\mathcal{D})\right]\right\}$$

- 定义：VS 中每个成员都位于这两条边界之间

$$VS_{\mathcal{H},\mathcal{D}}=\left\{h\in\mathcal{H}\middle|(\exists s\in S)(\exists g\in G)\left(g\geq_g h\geq_g s\right)\right\}$$

VS的简洁表示

S: $\{\langle Sunny, Warm, ?, Strong, ?, ? \rangle\}$

$\langle Sunny, ?, ?, Strong, ?, ? \rangle$ $\langle Sunny, Warm, ?, ?, ?, ? \rangle$ $\langle ?, Warm, ?, Strong, ?, ? \rangle$

G: $\{\langle Sunny, ?, ?, ?, ?, ? \rangle, \langle ?, Warm, ?, ?, ?, ? \rangle\}$

自学

- 候选消除算法
  - 算法内容
  - 算法收敛性
- 无偏置学习
  - 定义
  - 如何理解其无用性
- 归纳偏置
  - 定义
  - 意义

总结

- 不要期望学习算法能够准确地学到一个概念
- 不要期望学习结果与目标概念非常接近
- 可以期望学习结果以较高的概率接近目标概念

贝叶斯概念学习

概念学习

- 概念学习的训练样本集合中需要描述概念的正例和反例
- 心理学研究证明人类可以单独从正例样本中学习概念
  - 儿童学习单词涵义的过程
  - 反例只能从主动学习过程中获得
- 学习单词的含义 ⇔ 概念学习 ⇔ 两类分类问题
  - 学习指标函数

$$f(x)=\begin{cases}1 & \text{if } x \text{ is an example of the concept } c\\ 0 & \text{otherwise}\end{cases}$$

  - 两类分类问题需要正例和反例，而概念学习只需要正例。

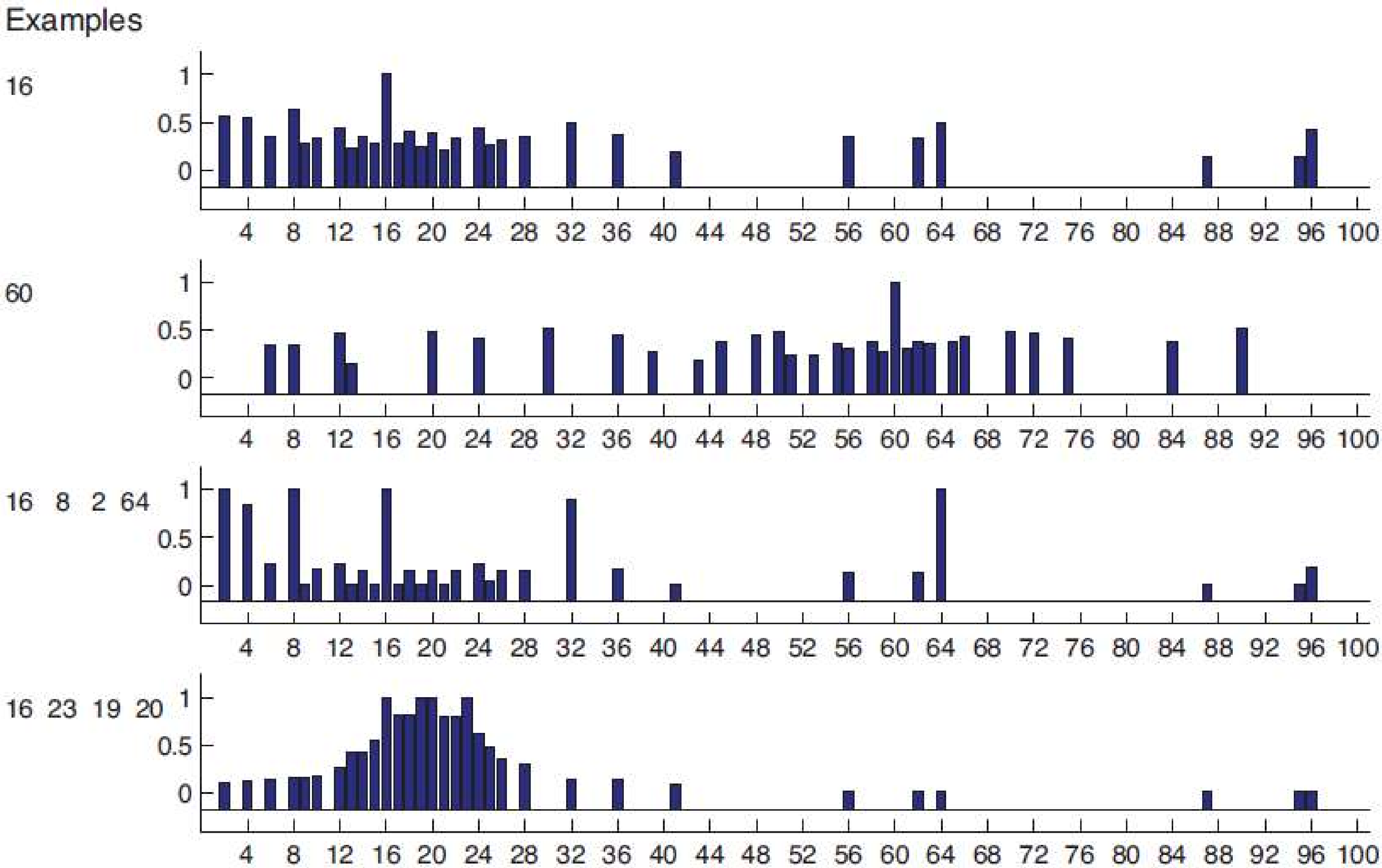
数字游戏

- 简单的概念学习实例
  - 选择一个简单的算术概念 C，如素数
  - 随机选择概念 C 的正例序列 $\mathcal{D}=\{x_1,x_2,\ldots,x_N\}$
  - 询问测试样本 $\tilde{x}$ 是否属于概念 C
- 假设：所有数字为 $1\sim100$ 之间的整数
  - 数字 16 是某个概念的一个正例，请你说出满足该概念的其它正例有
    - 17?
    - 6?
    - 32?
    - 99?
  - 后验概率预测分布 (posterior predictive distribution)

$$p\left(\tilde{x}\in C\middle|\mathcal{D}\right),\tilde{x}\in[1,100]$$

数字游戏

- 平均经验预测分布



数字游戏

- 如果 8、2 和 64 也是这个概念的正例，请说出该概念的其它正例
- 归纳结果：猜测的概念是 2^n

- 课堂讨论题
  - 如何解释上述行为
  - 如何在计算机上仿真实现归纳过程

归纳方法

- 经典归纳方法
  - 假设一个概念的假设空间 $\boldsymbol{H}$：奇数，偶数，1~100间所有整数，2 的幂，所有以 j 结尾的整数 $0 \le j \le 9$，……
  - 与训练样本集合 $\boldsymbol{D}$ 相一致的 $\boldsymbol{H}$ 的子集合(Version Space)
  - 随着 VS 的收缩越来越确定某个概念
- 思考
  - 已知 $\mathcal{D} = \{16\}$，存在着许多与其一致的概念，你如何利用这些概念预测是否 $\tilde{x} \in C$？
  - 已知 $\mathcal{D} = \{16, 8, 2, 64\}$，为什么你选择“2 的幂”而不是“所有偶数”或“除了 32 以外的 2 的幂”？

似然率

- 问题：已知 $\mathcal{D}=\{16,8,2,64\}$，为什么选择 h_{two} 而不是 h_{even}？
  - h_{two} = "powers of two"
  - h_{even} = "even numbers"
- 直觉：避免可能的巧合(suspicious coincidences)
  - 如果概念是偶数，为什么看到的数字都是 2 的幂?
- 定义：概念延伸——属于某个概念的数字集合
  - 偶数：$\{2,4,6,8,\ldots,98,100\}$
  - 2 的幂：$\{2,4,8,16,32,64\}$
- 强采样假设(strong sampling assumption)：样本是从概念延伸中随机均匀采样得到的。
  - 从假设 h 中独立采样 N 项（可置换）的概率为

$$p(\mathcal{D}|h)=\left[\frac{1}{\text{size}(h)}\right]^N=\left[\frac{1}{|h|}\right]^N$$

似然率

- 尺度原则(size principle)：模型喜爱与数据相一致的最简单（最小）假设 $\Leftrightarrow$ Occam's razor （奥坎姆剃刀）
- 问题：选择 h_{two} 还是 h_{even}？
  - 当 $\mathcal{D}=\{16\}$时，有

$$\text{likelihood ratio}=\frac{p(\mathcal{D}|h_{\text{two}})}{p(\mathcal{D}|h_{\text{even}})}=\frac{1/6}{1/50}\approx 8.33$$

  - 当 $\mathcal{D}=\{16,8,2,64\}$ 时，有

$$\text{likelihood ratio}=\frac{p(\mathcal{D}|h_{\text{two}})}{p(\mathcal{D}|h_{\text{even}})}=\frac{\left(\frac{1}{6}\right)^4}{\left(\frac{1}{50}\right)^4}=\frac{7.7\times10^{-4}}{1.6\times10^{-7}}=4812.5$$

先验概率

- 已知 $D=\{16,8,2,64\}$，下述哪个概念更贴近训练样本集合?
  - 概念 1：除 32 之外的 2 的幂
  - 概念 2：2 的幂
- 直观：由于概念 1 不需要解释训练样本集合中缺少 32 的事实，故貌似很贴切。
- 问题：主观感觉概念 1 “不自然”
- 思路：使用先验概率描述概念的主观感觉
- 争议：先验概率和假设空间均会因人而异
  - 小学生和博士生会对同一个问题给出不同的答案（知识背景不同）
- 结论：虽有争议，但很有用。

先验概率

- 例：已知 $\mathcal{D}=\{1200,1500,900,1400\}$
  - 如果归纳的概念属于算术概念，400 和 1183，哪个更像其成员?
  - 如果归纳的概念是胆固醇值，又会如何?
- 背景知识可以通过先验概率这种机制来对问题的求解过程施加影响，否则基于小样本的快速学习就是不可能的。

后验概率

- 定义：后验概率

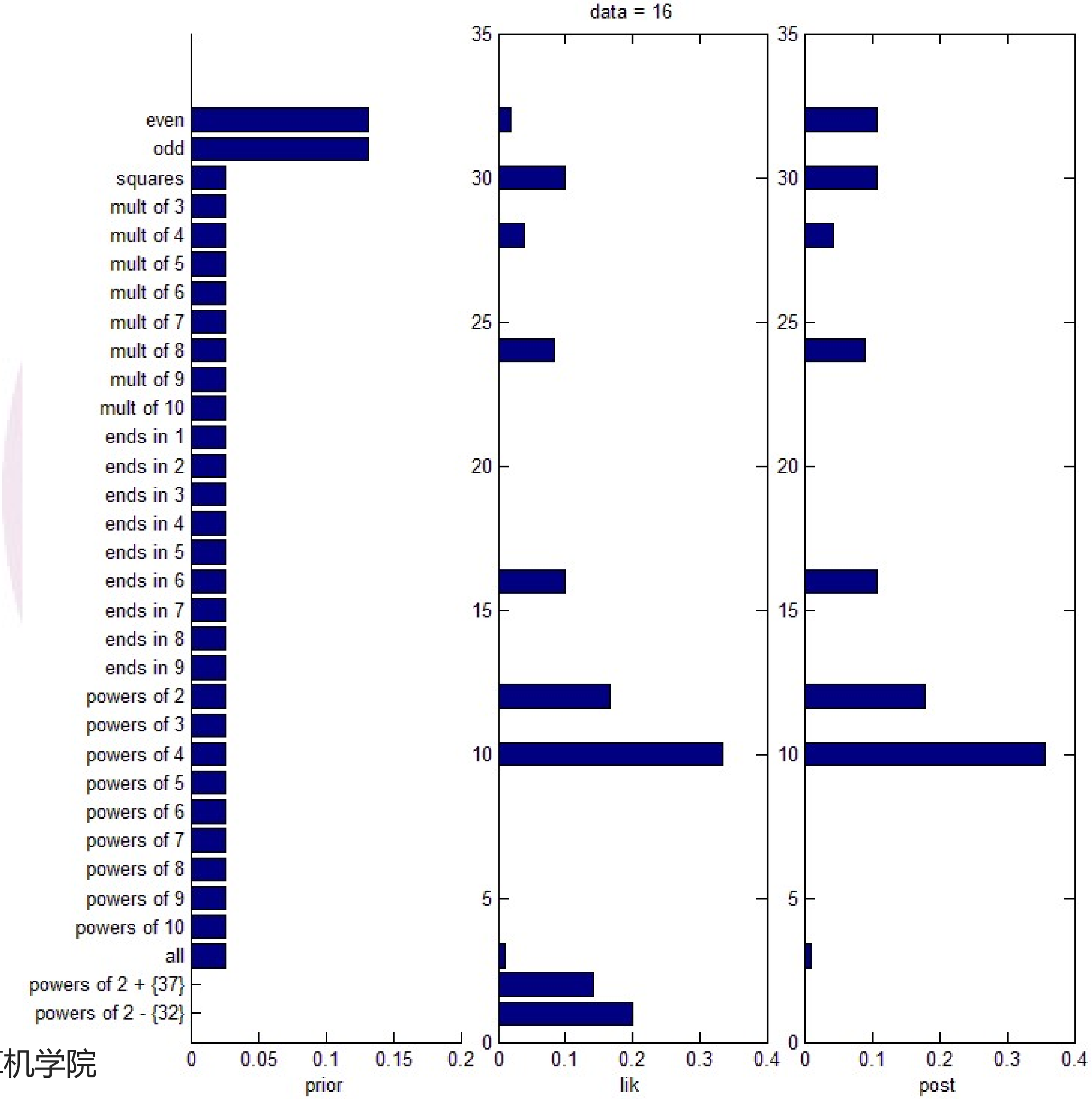
$$p(h|\mathcal{D})=\frac{p(\mathcal{D}|h)p(h)}{\sum_{h'\in\mathcal{H}}p(\mathcal{D},h')}=\frac{p(h)\mathbb{I}(\mathcal{D}\in h)/|h|^N}{\sum_{h'\in\mathcal{H}}p(h')\mathbb{I}(\mathcal{D}\in h')/|h'|^N}$$

其中：当且仅当所有数据 $\mathcal{D}$ 都在假设 h 的概念延伸中时，指示函数 $\mathbb{I}(\mathcal{D}\in h)=1$，否则等于 0。

- 对于绝大多数概念，先验概率是相同的，故后验概率正比于似然值。
- 对于不自然的(unnatural)概念，虽然似然值较高，但是由于先验概率很低，故后验概率也较低。

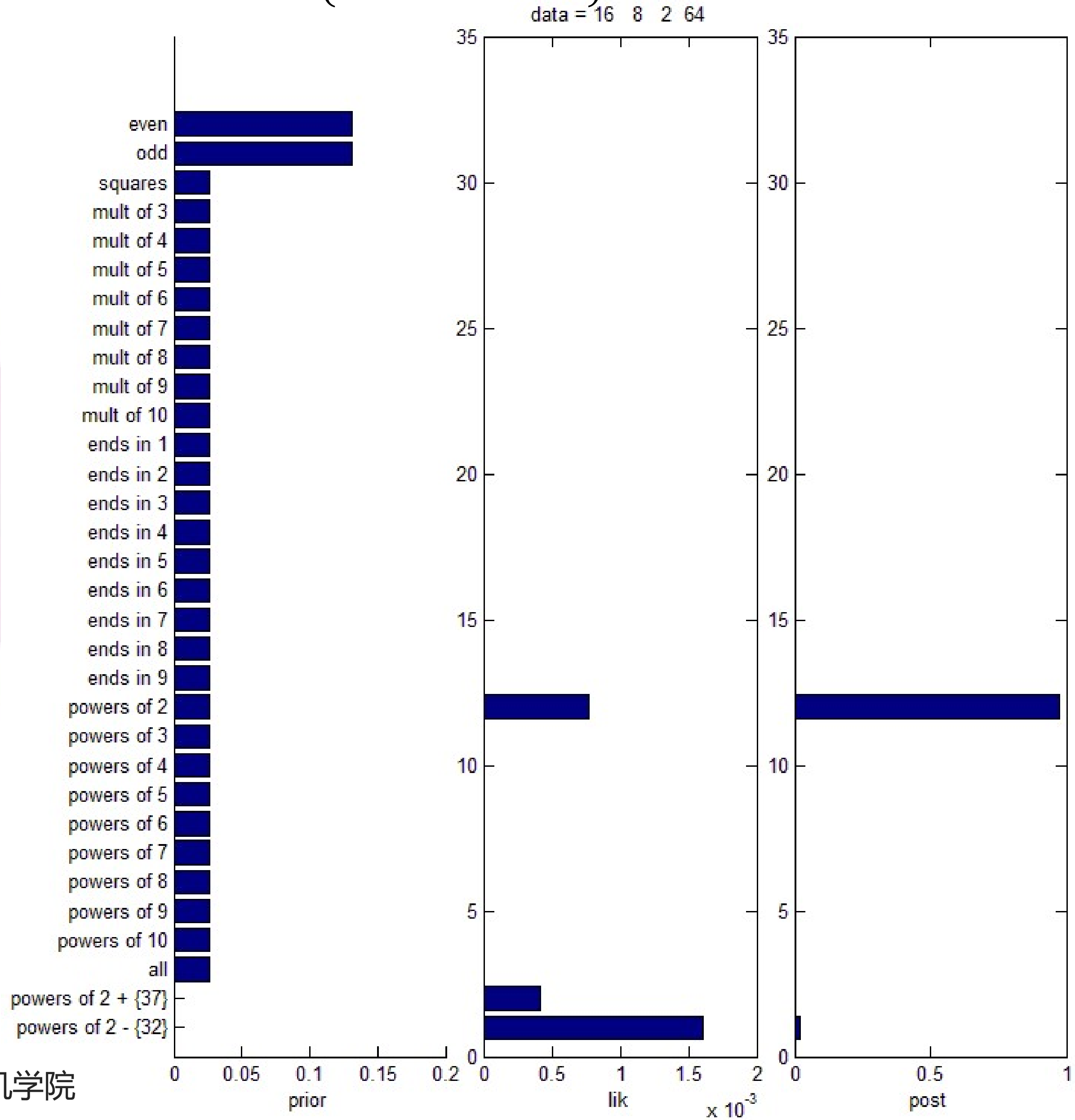
后验概率

❖ 观测到训练样本集合$\mathcal{D}=\{16\}$后



后验概率

- 观测到训练样本集合$\mathcal{D}=\{16,8,2,64\}$后



后验概率

- 一般来说，当拥有足够的训练样本时，后验概率将在某个概念处具有峰值，称为最大后验概率(MAP)估计。

$$p(h|\mathcal{D}) \rightarrow \delta_{\hat{h}^{\text{MAP}}}(h)$$

其中：$\hat{h}^{\text{MAP}}(h) = \arg\max_h p(h|\mathcal{D})$是后验概率的众数，且Dirac测度 δ 定义为

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

- MAP估计也可以表示为

$$\hat{h}^{\text{MAP}} = \arg\max_h p(\mathcal{D}|h)\, p(h)$$

$$\Rightarrow \arg\max_h \left[\log p(\mathcal{D}|h) + \log p(h)\right]$$

后验概率

- 随着训练样本数目越来越多，最大后验概率(MAP)估计收敛到最大似然估计(MLE)。

$$\hat{h}^{\text{MLE}} = \arg\max_h p\left(\mathcal{D}\middle|h\right) \Rightarrow \arg\max_h \log p\left(\mathcal{D}\middle|h\right)$$

- 当训练样本数目足够大时，数据将“淹没”先验概率。
- 假如真实假设在假设空间中，那么 MAP 和 MLE 将收敛到这条假设，故称贝叶斯推理是一致估计。假设空间称为在极限下是可接受的，即：在无限数据量时可以发现真实假设。
- 如果没有充足的训练样本时，假设将收敛到尽可能“接近”真实假设的假设。

朴素贝叶斯分类器

朴素贝叶斯分类器

- 问题：离散数值特征矢量 $\mathbf{x} \in \{1,\ldots,K\}^D$，其中 K 是特征最大值，D 是特征数目。
- 前提：已知类标签，且特征是条件独立的
- 朴素贝叶斯分类器模型

$$p(\mathbf{x}|y=c,\theta)=\prod_{j=1}^{D} p(x_j|y=c,\theta_{jc})$$

- 朴素的含义：不期望特征与类标签是独立的，甚至条件独立的。
- 结论：虽然前提不一定成立，但是可以获得较好的结果。

朴素贝叶斯分类器

- 类条件密度的形式依赖于每个特征的类型
  - 实数值特征 $\mathbf{x} \in \mathbb{R}^D$：可以使用高斯分布

$$p(\mathbf{x}|y=c,\theta)=\prod_{j=1}^{D}\mathcal{N}\left(x_j|\mu_{jc},\sigma_{jc}^2\right)$$

  - 二值特征$x_j \in \{0,1\}$：可以使用Bernoulli分布

$$p(\mathbf{x}|y=c,\theta)=\prod_{j=1}^{D}\mathrm{Ber}\left(x_j|\mu_{jc}\right)$$

其中 μ_{jc} 是在类别 c 中特征 x_j 发生的概率。

  - 范畴特征$x_j \in \{1,\ldots,K\}$：可以使用 Multinoulli 分布

$$p(\mathbf{x}|y=c,\theta)=\prod_{j=1}^{D}\mathrm{Cat}\left(x_j|\mu_{jc}\right)$$

其中 μ_{jc} 是在类别 c 中特征 x_j 的 K 柱直方图。

模型拟合

- 最大似然估计(MLE)
  - 单个数据的概率

$$p(\mathbf{x}_i, y_i|\theta) = p(y_i|\pi)\prod_j p(x_{ij}|\theta_j)$$

$$= \prod_c \pi_c^{\mathbb{I}(y_i=c)} \prod_j \prod_c p(x_{ij}|\theta_{jc})^{\mathbb{I}(y_i=c)}$$

对数似然

$$\log p(\mathcal{D}|\theta) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{D}\sum_{c=1}^{C}\sum_{i:y_i=c} \log p(x_{ij}|\theta_{jc})$$

  - 假设均匀先验概率 $\alpha_k = 1$，则类先验概率的MLE为

$$\hat{\pi}_c = N_c / N$$

其中 $N_c = \sum_i \mathbb{I}(y_i = c)$ 为类别 c 的训练样本数目。

模型拟合

- 似然函数MLE与特征遵循的概率分布形式有关
  - 假设特征是二值特征，则分布为

$$p\left(x_j \middle| y=c\right) \sim \mathrm{Ber}\left(\theta_{jc}\right)$$

相应的MLE为

$$\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$$

  - 分解后验概率

$$\log p\left(\mathcal{D}\middle|\theta\right) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{D}\sum_{c=1}^{C}\sum_{i:y_i=c} \log p\left(x_{ij}\middle|\theta_{jc}\right)$$

有

$$p\left(\theta\middle|\mathcal{D}\right) = p\left(\pi\middle|\mathcal{D}\right)\prod_{j=1}^{D}\prod_{c=1}^{C} p\left(\theta_{jc}\middle|\mathcal{D}\right)$$

$$p\left(\pi\middle|\mathcal{D}\right) = \mathrm{Dir}\left(N_1+\alpha_1,\ldots,N_C+\alpha_C\right)$$

拟合模型

$$p\left(\theta_{jc} \middle| \mathcal{D}\right) = \mathrm{Beta}\left(\left(N_c - N_{jc}\right) + \beta_0, N_{jc} + \beta_1\right)$$

换言之，计算后验概率就是使用似然的实验计数调整先验概率计数。

- Beta分布

$$f\left(x;\alpha,\beta\right) = \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} x^{\alpha-1}\left(1-x\right)^{\beta-1}$$

- 狄利克雷分布(Dirichlet distribution)，或称多元Beta分布

$$\mathrm{Dir}\left(X \middle| \alpha\right) = \frac{1}{\mathrm{B}\left(\boldsymbol{\alpha}\right)} \prod_{i=1}^{d} X_i^{\alpha_i - 1}$$

$$\mathrm{B}\left(\boldsymbol{\alpha}\right) = \frac{\prod_{i=1}^{d} \Gamma\left(\alpha_i\right)}{\Gamma\left(\alpha_0\right)}, \alpha_0 = \sum_{i=1}^{d} \alpha_i, d \geq 3$$

其中：$\boldsymbol{\alpha} \in \left\{\alpha_1, \alpha_2, \ldots, \alpha_d\right\} > 0$ 是分布参数，α_0是分布参数之和，$\mathrm{B}\left(\boldsymbol{\alpha}\right)$是多元Beta函数，$\Gamma\left(\alpha\right)$是Gamma函数。

模型拟合

- 算法2.1：拟合二值特征朴素贝叶斯分类器
  - **set** $N_c = 0, N_{jc} = 0$
  - **for** $i = 1 : N$ **do**

    $c = y_i$

    $N_c = N_c + 1$

    for $j = 1 : D$ **do**

        if $x_{ij} = 1$ **then** $N_{jc} = N_{jc} + 1$
  - **set** $\hat{\pi}_c = N_c / N, \hat{\theta}_{jc} = N_{jc} / N$

预测

❖ 预测需要计算

$$p(y=c|\mathbf{x},\mathcal{D}) \propto p(y=c|\mathcal{D})\prod_{j=1}^{d} p(x_j|y=c,\mathcal{D})$$

$$\propto \left[\int \mathrm{Cat}(y=c|\pi)\,p(\pi|\mathcal{D})\,d\pi\right]$$

$$\prod_{j=1}^{d}\left[\int \mathrm{Ber}(x_j|y=c,\theta_{jc})\,p(\theta_{jc}|\mathcal{D})\,d\theta_{jc}\right]$$

其中： $\mathrm{Cat}(x|\theta)$ 为Multinoulli分布， $\mathrm{Ber}(x|\theta)$为Bernoulli分布。

预测

❖ 如果后验概率为Dirichlet分布，则可简化为

$$p\left(y=c\,|\,\mathbf{x},\mathcal{D}\right)\propto\bar{\pi}_c\prod_{j=1}^{d}\left(\bar{\theta}_{jc}\right)^{\mathbb{I}\left(x_j=1\right)}\left(1-\bar{\theta}_{jc}\right)^{\mathbb{I}\left(x_j=0\right)}$$

$$\bar{\theta}_{jc}=\frac{N_{jc}+\beta_1}{N_c+\beta_0+\beta_1}$$

$$\bar{\pi}_c=\frac{N_c+\alpha_c}{N+\alpha_0}$$

其中 $\alpha_0=\sum_c\alpha_c$。

❖ 如果使用单点近似后验概率 $p\left(\theta|\mathcal{D}\right)\approx\delta_{\hat{\theta}}\left(\theta\right)$，其中 $\hat{\theta}$ 是MLE或MAP估计，则后验概率密度具有相同的形式如下：

$$p\left(y=c\,|\,\mathbf{x},\mathcal{D}\right)\propto\hat{\pi}_c\prod_{j=1}^{d}\left(\hat{\theta}_{jc}\right)^{\mathbb{I}\left(x_j=1\right)}\left(1-\hat{\theta}_{jc}\right)^{\mathbb{I}\left(x_j=0\right)}$$

预测

- 算法2.2：使用二值特征的朴素贝叶斯分类器进行预测
  - **for** $i = 1:N$ **do**
  - **for** $c = 1:C$ **do**
  - $L_{ic} = \log \hat{\pi}_c$
  - **for** $j = 1:d$ **do**
  - **if** $x_{ij} = 1$
  - **then** $L_{ic} = L_{ic} + \log \hat{\theta}_{jc}$
  - **else** $L_{ic} = L_{ic} + \log\left(1 - \hat{\theta}_{jc}\right)$
  - $p_{ic} = \exp\left(L_{ic} - \operatorname{logsumexp}\left(L_{i,:}\right)\right)$
  - $\hat{y}_i = \arg\max_c p_{ic}$

Log-sum-exp计算技巧

- 产生式分类器模型

$$p\left(y=c \middle| \mathbf{x},\theta\right)=\frac{p\left(y=c \middle| \theta\right)p\left(\mathbf{x} \middle| y=c,\theta\right)}{\sum_{c'} p\left(y=c' \middle| \theta\right)p\left(\mathbf{x} \middle| y=c',\theta\right)}$$

- 由于似然值 $p\left(\mathbf{x} \middle| y=c,\theta\right)$ 一般较小（特别是在高维特征空间），故上式计算存在数值下溢问题。
- 算法：取对数计算

$$\log p\left(y=c \middle| \mathbf{x}\right)=b_c-\log\left[\sum_{c'=1}^{C} e^{b_{c'}}\right]$$

$$b_c=\log p\left(\mathbf{x} \middle| y=c\right)+\log p\left(y=c\right)$$

- 由于

$$\log\left[\sum_{c'} e^{b_{c'}}\right]=\log\sum_{c'} p\left(y=c',\mathbf{x}\right)=\log p\left(\mathbf{x}\right)$$

Log-sum-exp计算技巧

- 分解最大项，其它做相对该项计算

$$\log\left[\sum_c e^{b_c}\right] = \log\left[\left(\sum_c e^{b_c - B}\right)e^B\right] = \left[\log\left(\sum_c e^{b_c - B}\right)\right] + B$$

其中$B = \max_c b_c$。

特征选择

- 问题：朴素贝叶斯分类器拟合许多特征的联合概率分布，故
  - 可能造成过度拟合
  - 计算复杂度高
- 解决方法：特征选择——除去无助于分类问题的不相关特征
- 特征选择方法：特征的排序、过滤或筛选
  - 单独评估每个特征的相关性
  - 选择前 K 个特征（K 值的选择需要考虑准确性和复杂性间的平衡）
- 相关性度量：特征 X_j 与类标签 Y 间的互信息

$$\mathrm{I}\left(X_j,Y\right)=\sum_{x_j\in X_j}\sum_{y\in Y}p\left(x_j,y\right)\log\frac{p\left(x_j,y\right)}{p\left(x_j\right)p\left(y\right)}$$

  - 对于二值特征，有

$$\mathrm{I}_j=\sum_c\left[\theta_{jc}\pi_c\log\frac{\theta_{jc}}{\theta_j}+\left(1-\theta_{jc}\right)\log\frac{1-\theta_{jc}}{1-\theta_j}\right]$$

特征选择

其中：

$$\pi_c = p(y=c)$$

$$\theta_{jc} = p(x_j = 1 | y = c)$$

$$\theta_j = p(x_j = 1) = \sum_c \pi_c \theta_{jc}$$

诚信　创新　实践