

1. 极大似然估计与朴素贝叶斯

2. 半朴素贝叶斯

3. 贝叶斯网

4. EM 算法

首先介绍贝叶斯决策，样本类 i 判断为类 j 落实的损失为 $R(c|c_i|x) = \sum_{j=1}^n \lambda_{ij} p(c_j|x)$

对于所有类别的总体损失，即求期望。分类的目的是 $\min R(c)$ 即 $\min E[R(c|x)]$

当损失用 Acc 表示时 $\lambda_{ij} = \begin{cases} 0 & i=j \\ 1 & i \neq j \end{cases}$ 此时 $R(c|x) = 1 - p(c|x)$

$\therefore h^*(x) = \arg \max_{c \in \mathcal{Y}} p(c|x) = \arg \max_{c \in \mathcal{Y}} \frac{p(x|c) \cdot p(c)}{p(x)}$ 其中 $p(c)$ 为先验概率， $p(x|c)$ 为条件概率，似然值。 $p(x)$ 为归一化因子，这个值与 c 无关

$p(c) = \frac{|D_c|}{|D|}$, $p(x|c)$ 由于 x 属性组合爆炸，无法从有限样本集中获取，其中一种方法

为参数估计，假设其概率分布，基于样本计算参数，最大化 $p(c|D_c|\theta_c)$

$$p(c|D_c|\theta_c) = \prod_{x \in D_c} p(x|\theta_c)$$

由于连乘可能会数值溢出，计算对数似然 $LL(\theta_c) = \sum_{x \in D_c} \log p(x|\theta_c)$

目标：寻找合适的 θ_c 最大化 $LL(\theta_c)$

缺点：对于数据分布的假设严重依赖于经验知识 —————— 这部分就是极大似然估计

下面介绍朴素贝叶斯，假设所有属性相互独立， $p(c|x) = \frac{p(c) \cdot p(x|c)}{p(x)} = \frac{p(c)}{p(x)} \prod_{i=1}^d p(x_i|c)$

此时最优贝叶斯分类器为 $h_{\text{NB}}(x) = \arg \max c p(c|x) = \arg \max c p(c) \cdot \prod_{i=1}^d p(x_i|c)$

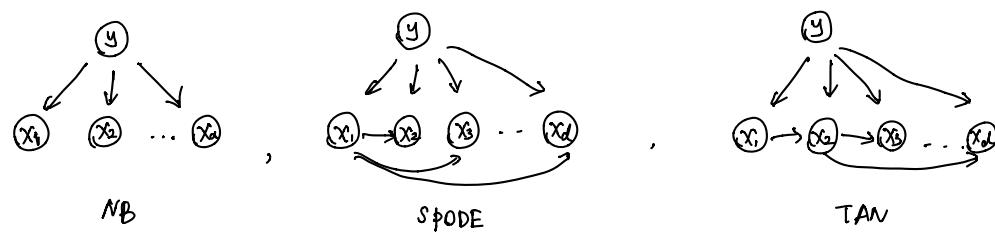
其中 $p(c) = \frac{|D_c|}{|D|}$, $p(x_i|c) = \begin{cases} \frac{|D_{c,x_i}|}{|D_c|}, & \text{离散属性} \\ \frac{1}{\sqrt{2\pi} b_{c,i}} e^{-\frac{1}{2} \frac{(x_i - \mu_{c,i})^2}{b_{c,i}^2}}, & \text{连续属性} \end{cases}$, $\mu_{c,i}, b_{c,i}$ 为 c 类样本 i 属性的均值和方差。

由于样本有限，有些属性不存在，引入拉普拉斯平滑有。

$$\hat{p}(c) = \frac{|D_c| + 1}{|D| + N}, \quad \hat{p}(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N}$$

半朴素贝叶斯分类器：考虑了一部分属性间的相互依赖，相当于折中。

$$p(c|x) \propto p(x) \prod_{i=1}^d p(x_i|c, pa_i), \quad pa_i \text{ 作为父属性}$$



SPODE：

所有属性依赖于同一属性，交叉验证选择方法确定超父属性。

TAN：

通过属性间的条件互信息构建完全图的最大带权生成树，保留强相关的属性依赖。

ADDE：基于集成学习，每个属性都作超父构建 SPODE。

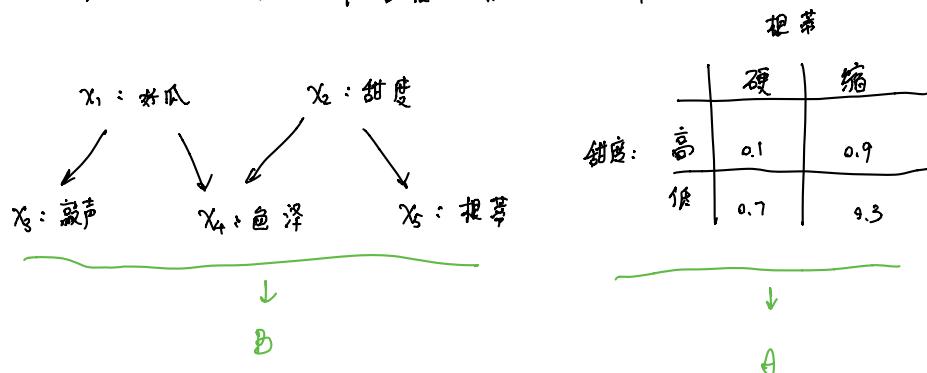
$$p(c|x) \propto \sum_{i=1}^d p(c, x_i) \prod_{j=1}^d p(x_j|c, x_i) \quad , \quad \hat{p}(c|x_i) = \frac{|D_{c,x_i}| + 1}{|D| + N + N_i}$$

$$|D_{c,x_i}| \geq m' \quad \hat{p}(c|x_i, x_j) = \frac{|D_{c,x_i, x_j}| + 1}{|D_{c,x_i}| + N_j}$$

下面是贝叶斯网：有向无环图刻画属性间的依赖关系

\$B = \langle C, \theta \rangle\$, \$C\$ 为网络结构，\$\theta\$ 包含每个属性的条件概率表。

下例。\$p(C \text{ 薯条 = 硬挺} \mid \text{甜度 = 高}) = 0.1\$



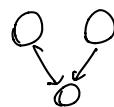
$$\text{此时 } p(x_1, x_2, x_3, x_4, x_5) = p(x_1) p(x_2) p(x_3|x_2) p(x_4|x_1, x_2) p(x_5|x_2)$$

此时

$$x_3 \perp x_4 \mid x_1, \quad x_4 \perp x_5 \mid x_2$$

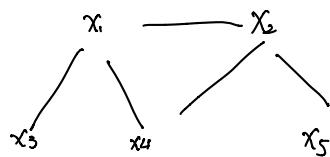
有向图 → 无向图，更好地观察变量的独立性。

1. 找出 V型结构，父节点加无向边



：V型结构

上面有向图转化后的无向图



去掉 $x_1 \rightarrow x_3, x_1 \rightarrow x_4$
 $x_2 \rightarrow x_3, x_2 \rightarrow x_4$
 $x_4 \rightarrow x_5$

去掉 $x_3 \perp x_4 \mid x_1$
 $x_3 \perp x_2 \mid x_1$
 $x_3 \perp x_5 \mid x_1$
 $x_5 \perp x_1 \mid x_2$
 $x_5 \perp x_3 \mid x_2$
 $x_5 \perp x_4 \mid x_2$

学习过程 1. 找出网络结构

2. 计数，算条件概率表。

评分函数 评估网络好坏，与数据符合程度

1. 网络本身 小 2. 用网络描述数据 小

↓

最小描述长度 $S(C|B|D) = f(\theta)|B| - \underline{LL(C|B|D)} + \sum_{i=1}^n \log P_B(x_i)$

1. $f(\theta) = 1$ 时 $AIC(C|B|D) = |B| - LL(C|B|D)$

2. $f(\theta) = \frac{1}{2} \log n$ 时 $BIC(C|B|D) = \frac{\log n}{2} |B| - LL(C|B|D)$

3. $f(\theta) = 0$ 不计算 网络长度。

寻找网络，计算评分在数据集上的

(1) 穷心，基础现有结构每次调整一条边

(2) 增加约束，限定结构

根据 $P(Q|E=e)$ 采样，保证 后验概率收敛于

$$P(Q=q|E=e) \approx \frac{n_q}{\tau}$$

第7章 吉布斯采样法

162

```

输入: 贝叶斯网  $B = (G, \Theta)$ ;
       采样次数  $T$ ;
       证据变量  $E$  及其取值  $e$ ;
       待查询变量  $Q$  及其取值  $q$ 。
待程:
1:  $n_q = 0$ 
2:  $q^{(1)} =$  对  $Q$  随机赋初值
3: for  $t = 1, 2, \dots, T$  do
4:   for  $Q_i \in Q$  do
5:      $Z = E \cup Q \setminus \{Q_i\}$ ;
6:      $z = e \cup q^{(t-1)} \setminus \{q_i^{(t-1)}\}$ ;
7:     根据  $B$  计算分布  $P_B(Q_i | Z=z)$ ;
8:      $q_i^{(t)} =$  根据  $P_B(Q_i | Z=z)$  采样所获  $Q_i$  取值;
9:      $q^{(t)} =$  将  $q^{(t-1)}$  中的  $q_i^{(t-1)}$  用  $q_i^{(t)}$  替换
10:  end for
11:  if  $q^{(t)} = q$  then
12:     $n_q = n_q + 1$ 
13:  end if
14: end for
输出:  $P(Q=q | E=e) \approx \frac{n_q}{T}$ 

```

除去变量 Q_i 外的其他变量。

图 7.5 吉布斯采样算法

EM 算法：

隐变量 为无法观测的变量

$$\mathcal{L}(\theta | x, z) = \ln p(x, z | \theta)$$

$$\mathcal{L}(\theta | x) = \ln p(x | \theta) = \ln \sum_z p(x, z | \theta)$$

初始化 θ^0

1. 基于 θ^t 推断 隐变量 z 的期望

2. 基于 x, z^t 估计 θ^{t+1}

进一步

E 步： θ^t 推断 z 的分布 $p(z | x, \theta^t)$. 计算 $\mathcal{L}(\theta | x, z)$ 关于 z 的期望

$$Q(\theta | \theta^t) = \mathbb{E}_{z|x, \theta^t} \mathcal{L}(\theta | x, z)$$

$$M \text{ 步: } \theta^{t+1} = \underset{\theta}{\operatorname{argmax}} Q(\theta | \theta^{t+1}).$$