半朴素贝叶斯

前言

我们在前边介绍的朴素贝叶斯，是基于强假设下的推论，也就是属性符合条件独立假设，从而分解条件概率。但是这显然在现实生活中难以实现，因为就像我们之前提到的：对于“如果天气是热、高温、晴”三种情况下，我们不出门。它的联合概率为 $P(热，高温，晴|不出门)$，但是如果属性条件独立，那么就可以写为 $P(热|不出门) \times P(高温|不出门) \times P(晴|不出门)$，但是其中的“热”和“高温”实际上很难条件独立，必然存在联系。所以朴素贝叶斯就败在这里。

为了解决这一问题，我们希望基于属性条件假设允许一定程度的宽松，由此诞生了“半朴素贝叶斯”(semi-naive Bayes)。

半朴素贝叶斯

关于半朴素贝叶斯如何实现宽松的，我们需要详细介绍一下。它的基本思想是适当的考虑一部分属性间的依赖信息，从而既不需要进行完全联合概率计算，也不至于彻底忽略比较强的属性依赖关系。“独依赖估计”(One-Dependent Estimator, 简称 ODE)是半朴素贝叶斯最常用的一种策略。顾名思义，其“独依赖”就是假设每个属性在类别之外最多依赖于一个属性，即：

$$P(c|x) \propto P(c) \prod_{i=1}^{d} P(x_i|c, pa_i), \quad (7.21)$$

其中，pa_i 为属性 x_i 所依赖的属性，成为 x_i 的父属性（父子关系）。此时，对于每个属性 x_i，若其父属性 pa_i 已知，则可通过拉普拉斯修正后的式(7.20)估计概率值 $P(x_i|c, pa_i)$。于是，问题的关键就转化为如何确定每个属性的父属性，不同的做法产生不同的独依赖分类器。

SPODE

这些分类器中，最直接的做法是假设所有属性都依赖于同一个属性，称为“超父”(super parent)，然后通过交叉检验等模型选择方法来确定超父属性，由此就形成了SPODE(Super Parent ODE)方法。例如下图b，x_1 是超父属性（来自西瓜书）

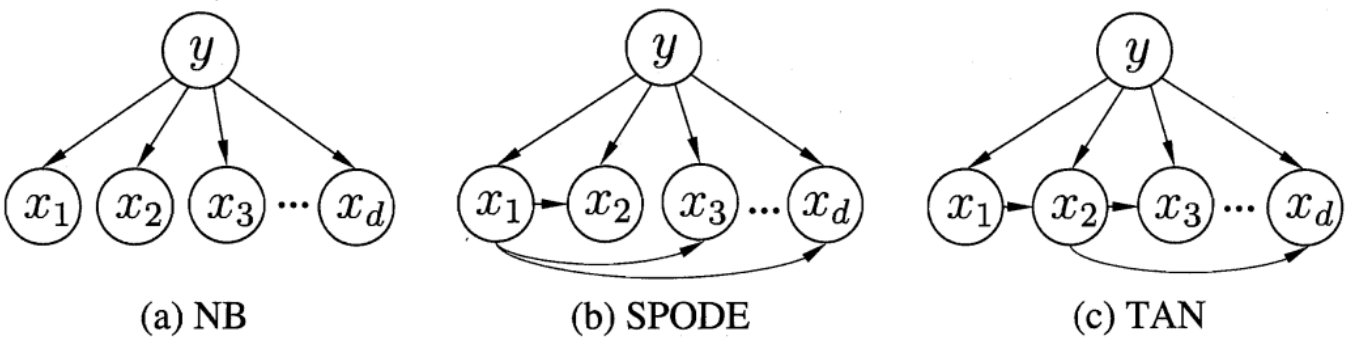

图 7.1 朴素贝叶斯与两种半朴素贝叶斯分类器所考虑的属性依赖关系

TAN

其他的，像 TAN (Tree Augmented naive Bayes)则是在最大带权生成树(maximum weighted spanning tree)算法的基础上，通过以下步骤将属性间关系约简为上图(c)所示结构：

怎么算呢？还记得我们在决策树中学过的信息论么？没错，我们希望估计两个事件的依赖性，我们可以使用互信息，但是不再是简单的互信息，而是条件互信息：

我们有数据集 $\mathcal{D}=\{(x^{(n)},y^{(n)})\}_{n=1}^{N}$，其中特征集合为 $X_1,\dots,X_m$，类别 Y。

TAN 的模型结构：每个特征 X_i 的父结点有两个：类别 Y（始终有）和至多一个其它特征 Pa_i（由树结构决定）。联合分布形式为：

$$P_{TAN}(X_1,\dots,X_m,Y)=P(Y)\prod_{i=1}^{m}P(X_i\mid Y,,Pa_i).$$

对数似然（训练数据）：

$$\mathcal{L}=\sum_{n=1}^{N}\log P_{TAN}\left(x^{(n)},y^{(n)}\right)=\sum_{n=1}^{N}\left[\log P(y^{(n)})+\sum_{i=1}^{m}\log P\left(x_i^{(n)}\mid y^{(n)},,pa_i^{(n)}\right)\right].$$

需要注意：$\log P(y^{(n)})$ 与特征间边的选择无关（只依赖类别边），因此在比较不同树结构时可以忽略常数项。我们需要选择一棵特征之间的树，使得 $\sum_{n,i}\log P(x_i^{(n)}\mid y^{(n)},pa_i^{(n)})$ 最大。

把经验分布写入（用经验频率 $\hat{p}$ 表示），模型对数似然的可改写成期望形式。关键技巧：比较任意给定有向树模型和“条件独立”基线（即朴素贝叶斯，只有 $Y\to X_i$ 无其它父节点）的对数似然差异。令朴素贝叶斯模型为 NB，则对数似然差：

$$\Delta\mathcal{L}=\mathcal{L}_{TAN}-\mathcal{L}_{NB}=\sum_{n=1}^{N}\sum_{i=1}^{m}\left[\log P(x_i^{(n)}\mid y^{(n)},pa_i^{(n)})-\log P(x_i^{(n)}\mid y^{(n)})\right].$$

把经验概率代入并将和合并为经验期望（以频率表示），可以得到（省去常数、N 因子仅影响尺度）：

$$\Delta\mathcal{L} \propto \sum_{i} \sum_{pa_i=j} \sum_{x_i,x_j,y} \hat{p}(x_i,x_j,y), \log \frac{\hat{p}(x_i \mid x_j, y)}{\hat{p}(x_i \mid y)}.$$

上式对每个有父 X_j 的 X_i 都有一项。把对数项重写成互信息形式，发现每条边 (i,j) 的贡献为

$$\sum_{x_i,x_j,y} \hat{p}(x_i,x_j,y) \log \frac{\hat{p}(x_i,x_j \mid y)}{\hat{p}(x_i \mid y)\hat{p}(x_j \mid y)} = I_{\hat{p}}(X_i; X_j \mid Y),$$

也就是**关于经验分布的条件互信息**。因此总体增益大致等于树中所有被选边的条件互信息之和。

于是，将对数似然差的边权写明后，单条特征对边的权重就得到了西瓜书同款式子，我们说的条件互信息：

$$w_{ij} = I_{\hat{p}}(X_i; X_j \mid Y) = \sum_{x_i,x_j,y} \hat{p}(x_i,x_j,y) \log \frac{\hat{p}(x_i,x_j \mid y)}{\hat{p}(x_i \mid y)\hat{p}(x_j \mid y)}.$$

西瓜书：

$$I(x_i, x_j \mid y) = \sum_{x_i,x_j;\, c\in\mathcal{Y}} P(x_i, x_j \mid c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c)P(x_j \mid c)}, \quad (7.22)$$

若要最大化 $\mathcal{L}_{TAN}$ ，等价于在完全图上为每对 (i,j) 计算权重 w_{ij} ，然后选择使边权总和最大的生成树——即**最大生成树（Maximum spanning tree）**，权重取 $I(x_i, x_j \mid y)$ 。

可以很容易地看出，条件互信息 $I(x_i, x_j \mid y)$ 刻画了属性 x_i 和 x_j 在已知类别下的相关性。因此，直接通过最大化生成树算法，TAN 实际上仅保留了强相关属性之间的依赖性

AODE（来自西瓜书）

AODE (Averaged One-Dependent Estimator) 是一种基于集成学习机制、更为强大的独依赖分类器。与 SPODE 通过模型选择确定超父属性不同，AODE 尝试将每个属性作为超父来构建 SPODE，然后将那些具有足够训练数据支撑的 SPODE 集成起来作为最终结果，即：

$$P(c \mid x) \propto \sum_{\substack{i=1 \\ |D_{x_i}| \geq m'}}^{d} P(c, x_i) \prod_{j=1}^{d} P(x_j \mid c, x_i), \quad (7.23)$$

其中 D_{x_i} 是在第 i 个属性上取值为 x_i 的样本的集合， m' 为阈值常数。显然，AODE 需估计 $P(c, x_i)$ 和 $P(x_j|c, x_i)$ 。类似式(7.20)拉普拉斯修正，有：

$$\hat{P}(c, x_i) = \frac{|D_{c,x_i}| + 1}{|D| + N_i}, \quad (7.24)$$

$$\hat{P}(x_j \mid c, x_i) = \frac{|D_{c,x_i,x_j}| + 1}{|D_{c,x_i}| + N_j}, \quad (7.25)$$

其中 N_i 是第 i 个属性可能的取值数， D_{c,x_i} 是类别为 c 且在第 i 个属性上取值为 x_j 的样本集合。不难看出，与朴素贝叶斯分类器类似，AODE 的训练过程也是“计数”，即在训练数据集上对符合条件的样本进行计数的过程，与朴素贝叶斯分类器相似，AODE 无需模型选择，既能通过预计算节省预测时间，也能采取懒惰学习方式在预测时再进行计数，并且易于实现增量学习。若训练数据非常充

分，泛化性能有可能提升；但在有限样本条件下，则又陷入估计高阶联合概率（“多个随机变量的联合分布函数”）的泥沼。