第5-1节　贝叶斯网络理论及方法

—— 网络结构学习(1)

5.1 结构学习定义

- 结构学习是寻找对数据拟合的最好的贝叶斯网络结构。

假设： 给定变量集合 $X = \{X_1, X_2, \ldots, X_N\}$， D 是其未知分布 $P(X)$ 下的采样，独立同分布。

求取： 贝叶斯网结构 S， 即变量间的直接或间接依赖关系及程度，与 D 的拟合度最高。

- 结构学习存在的问题：
  - 服从分布 $P(X)$的采样数据 D 包含了噪音，不能准确重建分布模型。
  - 添加越多的边，越能较好地拟合数据 D。但由于得到的结构 S 过于复杂，可能会出现过拟合现象，降低贝叶斯分类器的泛化能力。

- 例如：独立抛掷两枚硬币 X 与 Y，共20次。

观察到：3次正面/正面，6次正面/反面，

5次反面/正面，6次反面/反面。

用贝叶斯公式计算，可得到：

$$P(X=\text{正面})=0.45;$$

$$P(Y=\text{正面}\mid X=\text{正面})=1/3;$$

$$P(Y=\text{正面}\mid X=\text{反面})=5/11$$

- 由 Y 的条件概率可以计算得到它的边缘概率为

$$P(Y=\text{正面}) = P(Y=\text{正面} \mid X=\text{正面}) + P(Y=\text{正面} \mid X=\text{反面}) = 1/3 + 5/11 = 26/33 \approx 0.79$$

- 而如果按照 Y 是独立变量来计算的话，

$$P(Y=\text{正面}) = 8/20 = 0.4$$

- 为了拟合20次实验的数据，算法会在 X 和 Y 之间添加一个边，使得学习得到的贝叶斯网结构中 X 和 Y 存在依赖关系。
- 但是，硬币抛掷是独立事件，X 和 Y 之间不应该相关。因此，学习目标应该是得到尽量少的边，又能较好拟合数据的网络结构模型。

5.2 网络结构学习方法概述

无先验知识的结构学习方法有三种：

1. 基于约束的结构学习。通过对数据中的条件依赖或独立关系进行检验，找到解释最好的网络结构。
2. 基于得分的结构学习。定义贝叶斯网的统计得分，然后借助启发式搜索技术找到最高分的结构。
3. 贝叶斯模型平均方法。构建一个可能结构的集合，然后求取它们的平均。

5.3 基于约束的方法

- 基于约束的结构学习算法规则简单、易于实现且学习效果相对较好，其时间复杂度不会随着网络结构中节点数目的增加而上升。

- 通过条件独立检验来确定边的存在与否，从而确定贝叶斯网路的框架。然后根据独立检验中产生的分割集确定边的方向，从而得到贝叶斯网络结构模型。

5.3.1 单边假设检验

- 单边检验用检验统计量的密度曲线和二轴所围成面积中的单侧尾部面积，来构造临界区域进行检验的方法称为单边检验。
- 若采用的显著性水平 α 概率所确定的摒弃区域置于密度曲线的右边，则称为右单边检验；若置于密度曲线的左边，则称为左单边检验。

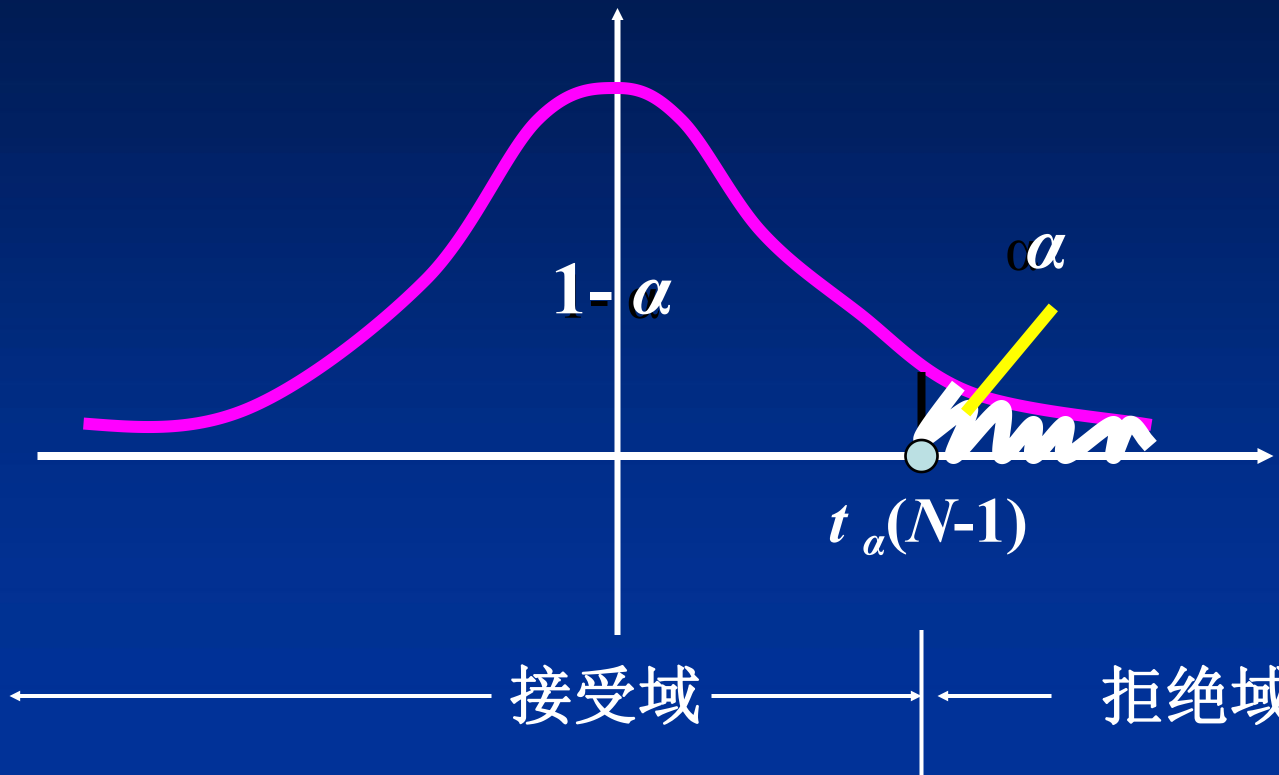
假设：右边检验 H_0：$\mu \le \mu_0$，H_1：$\mu > \mu_0$

左边检验 H_0：$\mu \ge \mu_0$，H_1：$\mu < \mu_0$

拒绝域:

设总体 X 服从 $P(X)$，$X_1, X_2, \ldots, X_N$是来自X的样本。给定显著性水平α

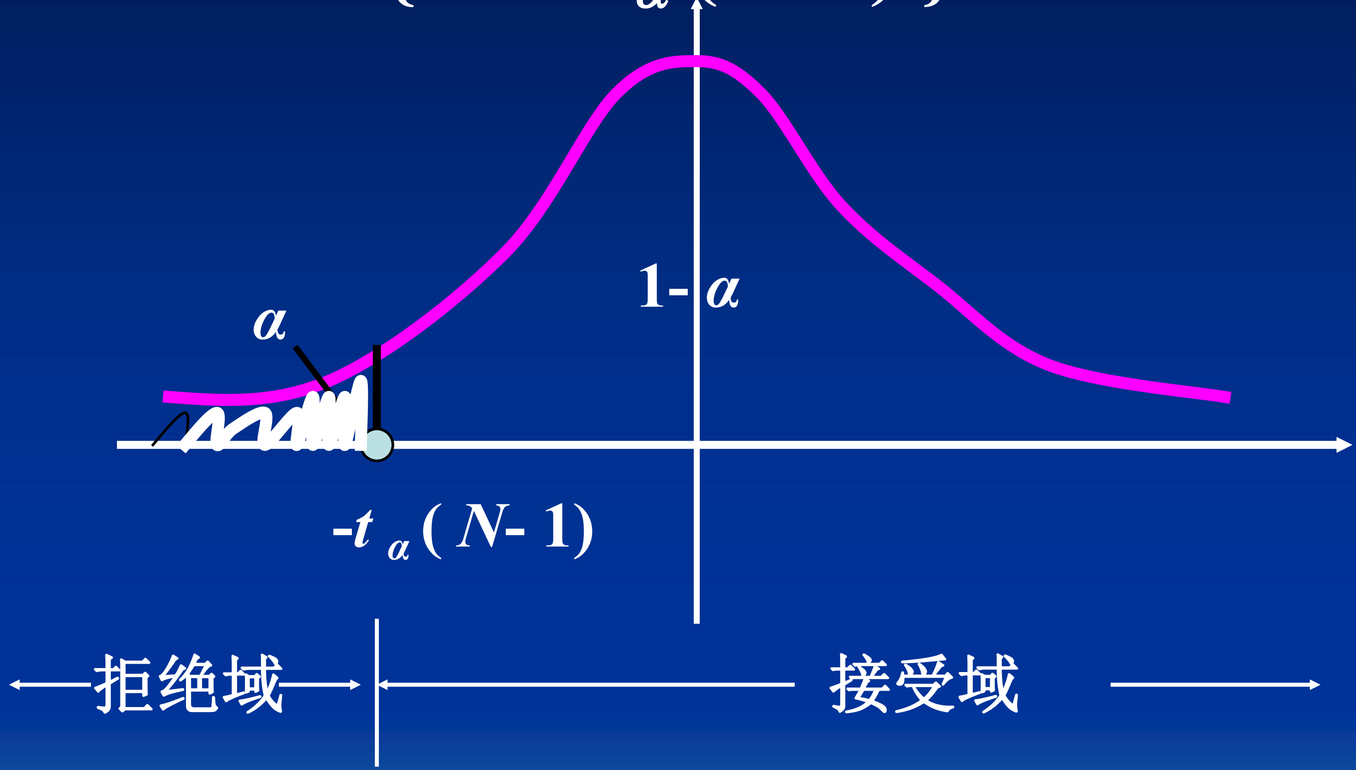
右单边检侧：H_0：$\mu \le \mu_0$, H_1：$\mu > \mu_0$



即 $t \ge t_\alpha(N-1)$时，拒绝H_0, 认为$\mu > \mu_0$

类似地，左单边检侧 H_0：$\mu \geq \mu_0$，H_1：$\mu < \mu_0$

$$P\{T \leq -t_\alpha(N-1)\} = \alpha$$



即 $t \leq t_\alpha(N-1)$时，拒绝H_0，认为 $\mu < \mu_0$

5.3.2 偏差度量

- 在抛掷硬币的实验中，$[X, Y]$的计数值 $M[X, Y]$ 如果接近于$N \cdot \dot{P}(X) \cdot \dot{P}(Y)$ (N是样本总数)，则可以判定 X 和 Y 之间是相互独立的。
- 上述两个值有差异，则说明 X 和 Y 之间存在一定的关联性，当这个关联程度大于可以容忍的范围时，则确认它们之间的联系。
- 有两种计算这两个值差异性的方法。

1. χ^2统计量：

- 计算公式为：

$$d_{\chi^2}(D) = \sum_{X,Y} \frac{\left(M[X,Y] - N \cdot \dot{P}(X) \cdot \dot{P}(Y)\right)^2}{N \cdot \dot{P}(X) \cdot \dot{P}(Y)}$$

对于完全符合独立性检测的数据集，其 $d_{\chi^2}(D) = 0$

2. 互信息量计算：

- 对于离散的随机变量 X，它的熵可以定义为：

$$H(X)=\sum_X P(X)log\frac{1}{P(X)}=-\sum_X P(X)logP(X)$$

- 对于任意两个变量 X 和 Y，它们的联合熵可以为：

$$H(X,Y)=\sum_{X,Y} P(X,Y)log\frac{1}{P(X,Y)}=-\sum_{X,Y} P(X,Y)logP(X,Y)$$

- 给定 Y，则 X 的条件熵可以定义为：

$$H(X|Y) = \sum_X P(X|Y) log \frac{1}{P(X|Y)}$$

意思是在给定 Y 的条件下，X 的不确定性。

- 考虑三个随机变量 X, Y, Z 之间的独立关系，条件熵 $H(X|Z)$ 表示给定 Z，X 的不确定性。如果接着给定Y，X 的不确定性变为 $H(X|Z, Y)$，所以两者之差则是给定 Z 时，观测到Y 的情况下所提供的 X 的信息量，即：

$$I(X;Y|Z) = H(X|Z) - H(X|Z,Y)$$

可以证明：$I(X;Y|Z) = I(Y;X|Z)$

所以它也可以称为给定 Z 下，X 和 Y 的互信息量。

- 上述变量的互信息量测量可写为：

$$d_I(D) = I(X;Y|Z) = \sum_{X,Y} P(X,Y,Z) \cdot log \frac{P(X,Y|Z)}{P(X|Z) \cdot P(Y|Z)}$$

在给定条件 Z 下，变量 X 和 Y 之间的互信息度量了二者之间是否依赖以及依赖的强度。

5.3.3 检验独立性

- 使用显著性水平95%，即允许采样数据的偏差有0.05或更小，大于此偏差即认为假设失败。

- 根据独立性偏差度量公式，有两种不同的独立性检测方法。

1. 基于χ^2统计量：

- 给定变量集合集合 $X = \{X_1, X_2, \ldots, X_N\}$， D 是其未知分布 $P(X)$ 下的采样， 计算独立性假设的 χ^2统计量：

$$\boldsymbol{d}_{\chi^2}(D) = \sum_{X,Y} M(X,Y) log \frac{M(X,Y)}{M(X) \cdot M(Y)}$$

近似地服从自由度为 $(n_X - 1)(n_Y - 1)$ 的卡方分布。

- 基于约束的学习方法之一就是通过卡方统计量做假设检验，通过样本落入的区域来确定是否拒绝独立性假设，若样本统计量掉落在拒绝域内，则拒绝独立性，否则不拒绝。实施步骤如下：

a) 建立假设；

b) 构造的卡方统计值，给出拒绝域；

c) 选择显著性水平；

d) 计算统计量，并通过比较确定是否接受原假设。

通常，显著性水平 α 的值域一般取作[0.0001,0.05]，其意义是衡量了拒绝原假设的正确性为 $1 - \alpha$。

- 卡方统计用于独立性检测实例：

https://wenku.baidu.com/video/course/v/9d4e4f93523b13f51711fd8d1da6fb30

2. 基于互信息量:

- 基于互信息的三阶段算法：

第一阶段是画草图阶段，连通节点的加入会增加节点之间的互信息，而非连通连节点的加入会减少节点之间的互信息。

通过启发式搜索策略，寻找节点之间的分割集，然后通过计算所得到的分割集中两个节点之间的互信息，来确定边的存在与否。

-

最小分割集概念：

- 给定两个变量 X 和 Y，能够将 X 和 Y d-分割的变量集称为分割集，它并不唯一。但是其中最小的变量集合却是唯一的，确定最小分割集的意义在于降低结构学习的复杂度，经常采用启发式方法确定较小的分割集。
-

最小分割集的求取过程：

首先，将连接 X 和 Y 之间的非碰撞节点，且属于 X 父节点的路径储存在 $S(X, Y)$ 中。

然后，初始化分割集 $D(X, Y)$ 为空集合。

接着，将集合 S 中有只包含一个节点的路径，以及能够堵塞最多路径的节点存放入 $D(X, Y)$ 中，将含有这些节点的路径从 S 中删除。

最后，重复执行前一步骤直至 S 为空，可以得到最小分割集。

基于互信息量的变量间独立性判定方法：

- 互信息度量为0，表示上述两个变量间相互独立；
- 互信息度量的值越高，表明两个节点之间的依赖性越强。当 $I(X,Y|Z) >$ 给定的阈值时，说明 X 和 Y 之间有直接的关联，可用一条弧线连接这对节点。

互信息实例

某地二月份天气构成的信源为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1(\text{晴}), x_2(\text{阴}), x_3(\text{雨}), x_4(\text{雪}) \\ \frac{1}{2} \qquad , \frac{1}{4} \quad , \quad \frac{1}{8} \quad , \quad \frac{1}{8} \end{Bmatrix}$$

现有人告诉你今天是不是晴天，作为收到的消息 Y **。**

计算I(X;y1), I(X;Y).

$$p(x_1/y_1) = 0; p(x_2/y_1) = \frac{1}{2}; p(x_3/y_1) = \frac{1}{4}; p(x_4/y_1) = \frac{1}{4}$$

$$p(x_1/y_2) = 1$$

$$p(x_1y_1) = 0; p(x_2y_1) = \frac{1}{4}; p(x_3y_1) = \frac{1}{8}; p(x_4y_1) = \frac{1}{8}, p(x_1y_2) = \frac{1}{2}$$

$$H(X/Y) = \frac{3}{4} bit; I(X;Y) = H(X) - H(X/Y) = 1 bit$$

- 基于互信息的三阶段算法：

　　第二阶段是增厚阶段，通过条件独立检验确定是否需要添加新的边，即添加必要的边；

　　第三阶段是削薄阶段，根据条件独立检验删除冗余的边．最后确定边的方向。

基于互信息量测试的变量独立性学习算法将贝叶斯网络看作是编码了变量间独立性关系的图结构。它的核心思想是：通过样本集 D 验证条件独立性 $I(X, Y \mid Z) < \varepsilon$ 是否成立，若成立，则在网络中节点 X 和 Y 之间不存在边。若不成立，变量 X 和 Y 是依赖的，它们 之间存在边。然后，利用节点集之间的条件独立性，建造一个有向无环图，以尽可能多地覆盖这些条件独立性。

- 基于条件独立性测试的方法比较直观，贴近贝叶斯网络的语义，把条件独立性测试和网络结构的搜索分离开，不足之处是对条件独立性测试产生的误差非常敏感。且在某些情况下条件独立性测试的次数相对于变量的数目成指数级增长。