模型备注——机器学习部分

2017年11月26日

目录

1 导言

本篇文档是对于以下列表中所提到机器学习模型的综述性说明。

一篇综述性说明应该以一种一脉相承的思路揭示各个模型间的关系，对模型形式上的杂多性进行尽可能的消除。在机器学习上，概率论和统计是被使用最多的综述起点。本文的视角不是传统的概率论，而是贝叶斯网络。之所以选择贝叶斯网络是因为：贝叶斯网络对于模型提供了更一般性、更具普遍性的性质，并且概率论视角可以被归纳到贝叶斯网络之中。再者，即便是概率论出发，也很难一以贯之地统一起回归、分类、聚类等等任务。我们很快将说明这些看似完全不同的任务、看似差别很大的模型是如何统一在一个框架下的。

代入数据集的实例在本文中很少，主要罗列模型本身作为更一般的贝叶斯网络特例化的例子。同时，本文不会对于贝叶斯网络和一些模型的过于复杂的性质进行详细论证，可参考材料。

模型的名称可能有不统一的情况（中英文交错），但是不会引起歧义。

文档的组织结构如下：第二章首先介绍贝叶斯网络作为描述变量间因果关系的一般框架，其次第三章介绍如何推断贝叶斯网络中的参数，最后第四章介绍在一般的贝叶斯网络中进行推理的方法。第四章的推理和具体的实用模型之间可能关系不甚紧密，加入文档的主要目的是为了保持理论完整。

本文不保证对于所有涉及模型提供完全的描述，但尽可能介绍了模型之间的关系。

主要参考文献：

Pattern Recognition and Machine Learning, C.M.Bishop，

Machine Learning: A Probabilistic Perspective, K.P.Murphy，

Probabilistic Graphic Models: Principles and Techniques,D.Koller

标（*）的章节可选择略过。

1.1 模型列表

本篇文档涵盖的模型和理论有：

表 1: 模型/算法列表

模型/算法名称	相关章节
贝叶斯网络	2.1
朴素贝叶斯分类器	2.2.3
决策树	2.2.1
随机森林	2.2.1
线性回归	2.2.2，3.1.3，3.2.2
局部加权线性回归	3.1.3
岭回归	3.2.2
lasso回归	3.2.2
逻辑回归	3.1.4
最大似然估计	3.1
最大后验估计	3.2
贝叶斯估计	3.3
主成分分析	3.4.1
概率主成分分析	3.4.1
因子分析	3.4.1
K-均值聚类	3.4.3
期望最大算法	3.4.3
泊松回归	3.1.2

1.2 符号约定

在本篇文档中，如果不进行上下文中特别的说明，一般以下列符号表达对应的语义：

表 2: 符号说明

符号	符号描述
$\mathscr{D}$	一个数据集合
d_i	一个数据集合中的一个条目
$\mathscr{X}$	一个变量集合
X_i	一个变量
$\mathscr{Z}$	隐变量数据集合
$Pa(X_i)$	一个贝叶斯网络中变量X_i的父节点集合
$N(\mathbf{x}\|\mu,\Sigma)$	变量$\mathbf{x}$服从均值为μ，协方差矩阵为Σ的正态分布

其他范数、概率分布和概念均在上下文中有明确定义。

2 变量关系的形式表达

2.1 贝叶斯网络

在建立任何一个模型时，首先考虑模型所涉及的所有变量之间的因果关系。采取这样一种方式记录：构建一个图，其中结点代表变量，如果变量A是变量B的原因，变量B是变量A的结果，就在结点A与结点B间添加一条从A指向B的有向边，即将A做成B的父结点：

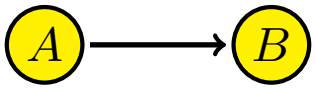

图 1: 因果关系“A为B的原因”的表示

在对于整个模型中的所有变量进行分析时，应该有这样一个直觉上的认知：当已知一个变量所有父节点的情况时，一个变量可以被推断，即它的值可以被表示为一个条件概率分布，其中条件是它的父亲结点。考虑变量$X_i \in \mathscr{X}$，它的父亲结点集合为$Pa(X_i)$，可以认为，给定$Pa(X_i)$的所有值时，可以用$p(X_i|Pa(X_i))$来推断X_i的值。

在变量集合$\mathscr{X}$上的**贝叶斯网络**定义为一个序对(G,θ)，其中图结构G是$\mathscr{X}$中所有因果关系的记录，即每个变量对之间的因果关系都在G中表现为一条有向边。θ用以表征所有$|\mathscr{X}|$个条件概率。给定一个贝叶斯网络时，写出所有变量的联合分布：

$$p(\mathscr{X}|G,\theta)=\prod_{i=1}^{|\mathscr{X}|}p(X_i|Pa(X_i),\theta)$$

例1：考虑如下一贝叶斯网络：

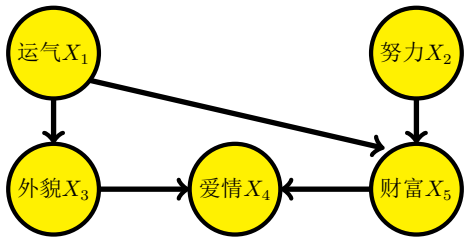

图 2: 一个贝叶斯网络的实例

可以在$\mathscr{X}=\{X_1,X_2,X_3,X_4,X_5\}$上定义联合分布：

$$\begin{aligned}p(\mathscr{X})=&\prod_{i=1}^{5}p(X_i|Pa(X_i))\\=&p(X_1)p(X_2)p(X_3|X_1)p(X_5|X_1,X_2)p(X_4|X_3,X_5)\end{aligned}$$

其中任一条件概率$p(X_i|Pa(X_i))$由网络内置的参数隐式地决定。

贝叶斯网络模型本身是更抽象的概率图模型的一个子类，当变量间的图G是无向时，另有模型**马尔可夫场**。贝叶斯网络中重要的一个特征就是变量间的条件独立性，受篇幅和文档主题所限不展开介绍，可以搜索**d-separation条件**深入阅读。

在本文的讨论中，一般假设G预先给定，即变量间的因果关系已知，从而一个贝叶斯网络的不确定、需推断成分仅是表示条件概率的参数。实际上对于“独立性”“观测数据”“图结构”间的相互推理是一个更为深入的话题。

2.2 局部条件概率的表示

2.2.1 离散关系与决策树

已知贝叶斯网络中变量的联合分布可以写成一系列条件概率分布（Con-ditional Probability Distribution，CPD）的乘积。首先考虑所有变量均取离散值的情况，此时CPD等价于一个在行（或者列）上归一化，且所有元素非负的表格。

考虑图1中的贝叶斯网络，假设变量A和B均为二值变量，且有：

$$\begin{aligned} p(A = a_0) &= 0.4 \\ p(A = a_1) &= 0.6 \\ p(B = b_0|A = a_0) &= 0.2 \\ p(B = b_1|A = a_0) &= 0.8 \\ p(B = b_0|A = a_1) &= 0.9 \\ p(B = b_1|A = a_1) &= 0.1 \end{aligned}$$

容易推断二者的联合概率，或一个变量上的边缘概率：

$$\begin{aligned} p(B = b_0, A = a_0) &= 0.08 \\ p(B = b_1, A = a_0) &= 0.32 \\ p(B = b_0, A = a_1) &= 0.54 \\ p(B = b_1, A = a_1) &= 0.06 \\ p(B = b_0) &= 0.62 \\ p(B = b_1) &= 0.38 \end{aligned}$$

这是一个十分平凡的例子，不过在网络因果结构比较复杂时，推理的难度将呈指数级别上升。类似的例子曾经在机器诊断的模型中被实际使用过，因为在临床病理上许多症状可以被有限状态变量描述。

例2：考虑一个更加不平凡的例子，首先给出贝叶斯网络结构：

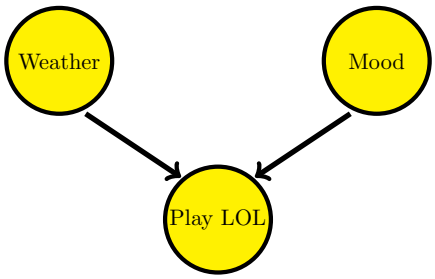

图 3: xxx

同样对变量做二值假设，设：

$$
\begin{aligned}
p(Weather = w_{good}) &= 0.5 \\
p(Weather = w_{bad}) &= 0.5 \\
p(Mood = m_{good}) &= 0.3 \\
p(Mood = m_{bad}) &= 0.7 \\
p(LOL = yes | w_{good}, m_{good}) &= 0.3 \\
p(LOL = no | w_{good}, m_{good}) &= 0.7 \\
p(LOL = yes | w_{good}, m_{bad}) &= 1 \\
p(LOL = no | w_{good}, m_{bad}) &= 0 \\
p(LOL = yes | w_{bad}, m_{good}) &= 0.5 \\
p(LOL = no | w_{bad}, m_{good}) &= 0.5 \\
p(LOL = yes | w_{bad}, m_{bad}) &= 1 \\
p(LOL = no | w_{bad}, m_{bad}) &= 0
\end{aligned}
$$

可以发现一个现象：心情不好时，该人一定会玩LOL，和天气的状况无关；但是当心情好时，该人玩游戏的概率又和天气有关。

虽然贝叶斯网络的图结构本身勾勒了变量间的条件独立关系，但是它所蕴含的独立断言遵从格式：

“给定某些变量时，一些变量和另一些变量相互独立”

然而有时条件独立性并不基于变量，而基于变量的特定赋值，换言之，存在着这样的独立断言：

“给定某些变量符合某种特定赋值时，一些变量和另一些变量相互独立”

这样的情形等价于CPD高维表格中有相同的子块，此时没有必要花费空间存储相同的子表，可以利用**决策树**压缩CPD空间。

决策树每次对于一个条件进行判断，并迭代地选择下一步决策的判断条件，在上例中可以构造如下决策树：

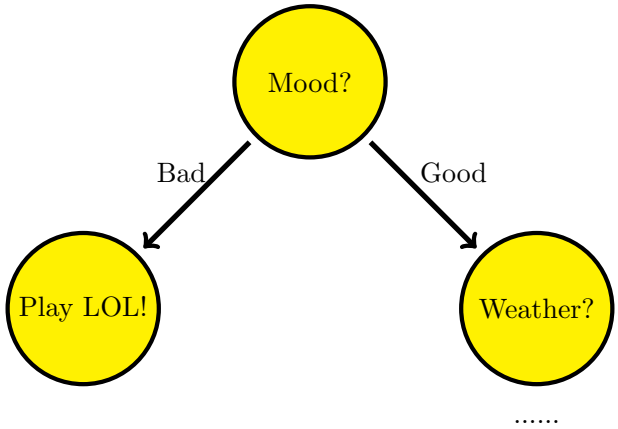

图 4: 图3中CPD表的决策树表达，下一步对于天气的好坏继续迭代

图4中左子树没有出现关于天气的条件，从而描述了“心情不好时，是否玩游戏和天气无关”的上下文特定独立性。出现上下文特定的变量独立性时，决策树可以简化CPD的表达。当没有先验知识时就训练一个决策树模型时，往往不一定能达到理想的效果。

为了数据集上能力的泛化，实际上常常使用训练集的不同子集训练多个决策树，并将多个树的最终输出整合后生成最终方案，这规避了决策树的确定性训练方法带来的过拟合效应，这是**随机森林**算法。

2.2.2 连续关系——线性回归的贝叶斯网络表达

当变量取连续值时，条件概率不能使用CPD表格，需要参数化的分布函数描述。本小节和下一小节介绍含有连续关系的贝叶斯网络中两个最著名的特例。首先虑线性回归问题，直接讨论多元回归。

在回归问题中，数据呈现出$\{(\mathbf{x}_i, y_i)\}$的形式，其中$\mathbf{x}_i$是一个输入的向量，而y_i是对应输出的标量。线性回归的变量因果关系遵从如下的形式：

1、输入变量$\mathbf{x}_i$经过某种变换形成新空间中的向量$\phi_i = \phi(\mathbf{x}_i)$；

2、变换后的输入向量和一个向量$\mathbf{w}$进行内积$\hat{y}_i = \mathbf{w}^T \phi_i$；

3、上一步运算的结果和一个高斯噪声$\epsilon \sim N(0,\sigma^2)$相加得到最终输出$y_i = \hat{y}_i + \epsilon$。

线性回归的贝叶斯网络（统一将需要求的参数用绿色结点表示，确定性的因果关系用红色箭头表示，非确定性的因果关系用黑色箭头表示，圆角矩形内包络的是服从独立同分布的单一数据项模板）如下图所示：

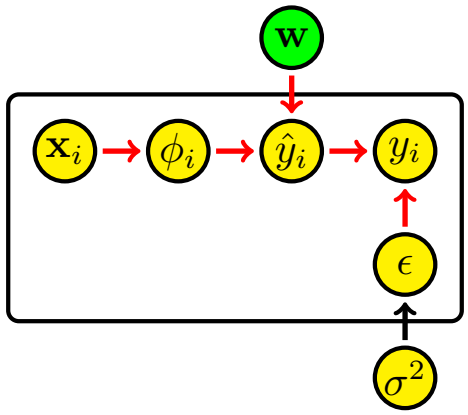

图 5: 线性回归的贝叶斯网络

既然是贝叶斯网络，就可写出所有变量的联合分布：

$$\begin{aligned} p(\mathbf{x}_i, \phi_i, \hat{y}_i, y_i, \mathbf{w}, \sigma^2, \epsilon) =& p(\mathbf{x}_i)p(\phi_i|\mathbf{x}_i)p(\mathbf{w})p(\hat{y}_i|\phi_i, \mathbf{w})p(\sigma^2)p(\epsilon|\sigma^2)p(y_i|\hat{y}_i, \epsilon) \\ =& p(\mathbf{x}_i)p(\mathbf{w})p(\sigma^2)N(\epsilon = y_i - \hat{y}_i|0, \sigma^2) \end{aligned}$$

第二个等式成立是因为确定性关系引导的条件概率为1，因此可以从联合分布中略去$\phi_i, \hat{y}_i, \epsilon$等中间变量。

若对于给定的$\mathbf{w}$和σ^2取条件，就得到了线性回归生成模型：

$$\begin{aligned} p(\mathbf{x}_i, y_i|\mathbf{w}, \sigma^2) =& \frac{p(\mathbf{x}_i, y_i, \mathbf{w}, \sigma^2)}{p(\mathbf{w}, \sigma^2)} \\ =& p(\mathbf{x}_i)N(y_i - \mathbf{w}^T\phi(\mathbf{x}_i)|0, \sigma^2) \end{aligned}$$

如果认为$\mathbf{x}_i$也是给定的，没有生成的必要，得到：

$$p(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) = N(y_i - \mathbf{w}^T\phi(\mathbf{x}_i)|0, \sigma^2)$$

结合数据项独立同分布的假设，就得到**标准线性回归模型**。

本节仅导出线性回归的贝叶斯网络表示，具体求解方法和变形放到下一章中讨论。

2.2.3 连续关系——线性分类的贝叶斯网络表达

本节考虑分类问题，数据格式$\{(\mathbf{x}_i, t_i)\}$，其中$\mathbf{x}_i$对应一个表示特征的向量，而t_i编码该特征向量的分类标签。这样一个数据项的生成因果过程可以理解为如下两个步骤：

1、根据某种条件选取t_i；

2、给定t_i，从该类对应的分布函数中取得$\mathbf{x}_i$。

一个不失一般性又易于处理的分布选择是：认为生成t_i的分布是一个参数为π的多元伯努利分布，而给定类别标签时的$\mathbf{x}_i$生成分布满足一个参数为μ_{t_i}, Σ_{t_i}的多元正态分布。此时的模型称为**线性分类**模型。

对应如下的贝叶斯网络：

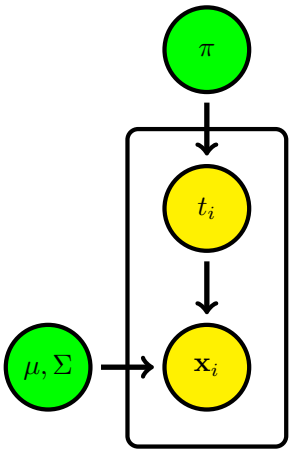

图 6: 线性分类的贝叶斯网络

类似处理线性回归问题，写出此时所有变量的联合分布，虽然图6看起来比图5简单，但是图6中的随机因果关系更多，所以得出的形式也更复杂。为了形式上的方便，使用热洞方式编码t_i，即对于总共C个类型，t_i对应一个分量非0即1的C维列向量，当且仅当$t_i = c$时，$t_{ic} = 1$：

$$\begin{aligned} p(\pi, \mu, \Sigma, t_i, \mathbf{x}_i) =& p(\pi)p(\mu, \Sigma)p(t_i|\pi)p(\mathbf{x}_i|\mu, \Sigma, t_i) \\ =& p(\pi)p(\mu, \Sigma)\prod_{c=1}^{C}\pi_c^{t_{ic}}\prod_{c=1}^{C}(N(\mathbf{x}_i|\mu_c, \Sigma_c))^{t_{ic}} \end{aligned}$$

类似的，在该式中将需要求值的变量置于条件中：

$$p(t_i, \mathbf{x}_i|\pi, \mu, \Sigma) = \prod_{c=1}^{C}(\pi_c N(\mathbf{x}_i|\mu_c, \Sigma_c))^{t_{ic}}$$

这是**线性分类的生成模型**。

具体的求解方法在下一章介绍，本章最后说明该模型的两种直接变形：

1、如果取$\Sigma_c = \Sigma, c = 1, ..., C$，即所有类别对应的协方差矩阵全部相同，此时模型变化为**逻辑回归（Logistic Regression）**，它有区别于一般线性分类生成模型的参数训练方法，并且保证特征空间中分类决策平面的线性；

2、如果假设所有类别对应的协方差矩阵全是对角矩阵，则模型变化为**朴素贝叶斯分类器（Naive Bayesian Classifier，NBC）**，因为朴素贝叶斯假设是：给定分类标签时，各个特征分量相互独立，所以协方差矩阵除对角元素外均为零。当然NBC中特征的生成分布可以采取正态分布以外的分布，但是使用拉普拉斯近似成正态分布后仍可以得到协方差矩阵为对角阵的结论。

3 参数估计

3.1 最大似然估计

考虑这样一个情景：给定一个数据集$\mathscr{D}$，并且选定一个模型$M(\theta)$，则模型的某种参数θ选取策略都能导出一个似然概率：

$$p(\mathscr{D}|\theta)$$

最大似然估计是这样一种策略，选择θ_{ML}：

$$\theta_{ML} = \arg\max_{\theta} \{p(\mathscr{D}|\theta)\}$$

换言之，最大似然估计选择这样一个参数θ_{ML}，和其他的参数选择相比，θ_{ML}有更大的概率生成数据集。

从最优化的角度，自然希望$p(\mathscr{D}|\theta)$是关于θ的凹函数，一个常用的技巧是最小化似然概率的对数的相反数（Negative Logarithm Likelihood，NLL）：

$$NLL(\theta) = -\log p(\mathscr{D}|\theta)$$

$$\theta_{ML} = \arg\min_{\theta} \{-\log p(\mathscr{D}|\theta)\}$$

3.1.1 最大似然估计的导出

极大似然求解技巧里面的对数运算并不是十分符合直觉，但这并不仅仅是出于最优化的便利，这里给出一个信息论角度的解释。

设生成数据集$\mathscr{D}$的真实概率分布为p_r，现试图寻找一个尽可能接近p_r的概率分布p_θ，这等价于最小化某种度量下p_r和p_θ的差值。根据信息理论选择相对熵作为两个分布之间差值的度量，相对熵的定义为（连续情况转化为积分运算）：

$$D(p_r||p_\theta) = \sum_x p_r(x) \log \frac{p_r(x)}{p_\theta(x)}$$

有：

$$\begin{aligned} D(p_r||p_\theta) =& \mathbb{E}_{p_r}[\log \frac{p_r}{p_\theta}] \\ =& \mathbb{E}_{p_r}[\log p_r] - \mathbb{E}_{p_r}[\log p_\theta] \\ =& -H(p_r) - \frac{1}{|\mathscr{D}|} \sum_{i=1}^{|\mathscr{D}|} \log p_\theta(d_i) \end{aligned}$$

其中最后一个等式成立的原因是弱大数定律，这一推导结果中的第一项是真实分布的负熵，和优化目标无关，所以只需要最大化最后的合式，它就是数据集上对数似然的平均值。当数据集$\mathscr{D}$服从独立同分布假设时，明显有：

$$\log p_\theta(\mathscr{D}) = \log p(\mathscr{D}|\theta) = \log \prod_{i=1}^{|\mathscr{D}|} p(d_i|\theta) = \sum_{i=1}^{|\mathscr{D}|} \log p(d_i|\theta)$$

当最小化对数似然的相反数以求得一个估计时，我们实际上在做这样一件事：寻找真实分布在当前模型定义的分布族中的一种投影。

这种投影是矩投影（Moment Projection），还有另一种信息投影（Information Projection），信息投影最小化$D(p_\theta||p_r)$。

倘若假设数据集中的所有分量互相独立，那么实际上是在一个没有任何边（因果关系）的空图贝叶斯网络中进行原始分布的投影，此时可投影分布空间为：

$$p_\theta(\mathscr{X}) = \prod_{i=1}^{|\mathscr{X}|} p_{\theta i}(X_i)$$

可以证明，此时最大似然估计得出的分布是：

$$p_{\theta_{ML}}(\mathscr{X}) = \prod_{i=1}^{|\mathscr{X}|} p_r(X_i)$$

因为（记$\mathscr{X}_{-i} = \mathscr{X}/X_i$）：

$$\begin{aligned}
D(p_r||p_\theta) =& \mathbb{E}_{p_r}[\log \frac{p_r}{p_\theta}] \\
=& \mathbb{E}_{p_r}[\log \frac{p_r}{p_{\theta_{ML}}} \frac{p_{\theta_{ML}}}{p_\theta}] \\
=& D(p_r||p_{\theta_{ML}}) + \mathbb{E}_{p_r}[\sum_{i=1}^{|\mathscr{X}|} \log \frac{p_r(X_i)}{p_\theta(X_i)}] \\
=& D(p_r||p_{\theta_{ML}}) + \sum_{i=1}^{|\mathscr{X}|} \mathbb{E}_{p_r}[\log \frac{p_r(X_i)}{p_\theta(X_i)}] \\
=& D(p_r||p_{\theta_{ML}}) + \sum_{i=1}^{|\mathscr{X}|} \sum_{X_i} p_r(X_i) \log \frac{p_r(X_i)}{p_\theta(X_i)} \sum_{\mathscr{X}_{-i}} p(\mathscr{X}_{-i}|X_i) \\
=& D(p_r||p_{\theta_{ML}}) + \sum_{i=1}^{|\mathscr{X}|} D(p_r(X_i)||p_\theta(X_i)) \\
\geq& D(p_r||p_{\theta_{ML}})
\end{aligned}$$

换言之，一个分布向空图投影时只需要将每个分量的一阶矩（期望）对齐即可，这是矩投影的命名原因之一。该思路在指数分布族上有更广泛的意义，这里不加展开，可参考exponential family。

3.1.2 概率分布的最大似然估计

本小节中通过一些对于概率分布进行最大似然估计的例子，说明使用该方法的步骤。

例3：伯努利分布的最大似然估计。将可能出现的两种情况记为0和1，在数据集中分别出现次数为α_0和α_1，θ是模型预测出现情况为1的概率。

首先给出似然概率：

$$p(\mathscr{D}|\theta) = \prod_{i=1}^{|\mathscr{D}|} \theta^{d_i}(1-\theta)^{1-d_i}$$

给出NLL：

$$\begin{aligned} NLL(\theta) &= -\sum_{i=1}^{|\mathscr{D}|} d_i \log\theta + (1-d_i)\log(1-\theta) \\ &= -\alpha_0 \log(1-\theta) - \alpha_1 \log\theta \end{aligned}$$

由于：

$$\frac{d^2 NLL(\theta)}{d\theta^2} = \frac{\alpha_0}{(1-\theta)^2} + \frac{\alpha_1}{\theta^2} > 0$$

所以凸性成立，对NLL的一阶导数置零可得最大似然估计：

$$\theta = \frac{\alpha_1}{\alpha_0 + \alpha_1}$$

例4：多元（设总共有K种可能情况）伯努利分布的最大似然估计。将每个数据项d_i采用热洞编码，待学习的参数向量为θ，仍然记第k种情况出现的次数为α_k，此时的似然概率为：

$$p(\mathscr{D}|\theta) = \prod_{i=1}^{|\mathscr{D}|} \prod_{k=1}^{K} \theta_k^{d_{ik}}$$

给出NLL：

$$NLL(\theta) = -\sum_{i=1}^{|\mathscr{D}|} \sum_{k=1}^{K} d_{ik} \log\theta_k$$

凸性易证，这是一个带约束的优化问题，等式约束为（不等式约束隐式成立）：

$$\sum_{k=1}^{K}\theta_k = 1$$

引入拉格朗日乘子v，此时拉格朗日函数为：

$$L(\theta, v) = -\sum_{i=1}^{|\mathscr{D}|}\sum_{k=1}^{K} d_{ik}\log\theta_k + v(\sum_{k=1}^{K}\theta_k - 1)$$

对于一个特定的k^*有：

$$\frac{d}{dk^*}L(\theta, v) = -\frac{\sum_{i=1}^{|\mathscr{D}|} d_{ik^*}}{\theta_{k^*}} + v = -\frac{\alpha_{k^*}}{\theta_{k^*}} + v$$

故：

$$\theta_k = \frac{\alpha_k}{v}$$

对于所有的k指标求和可得：

$$v = |\mathscr{D}|$$

故：

$$\theta_k = \frac{\alpha_k}{|\mathscr{D}|} = \frac{\alpha_k}{\sum_{k'=1}^{K}\alpha_{k'}}$$

在$k = 2$时该解退化到例3的情况。

例5：泊松分布的最大似然估计。泊松分布的概率密度函数为：

$$Poi(x|\lambda) = \exp\{-\lambda\}\frac{\lambda^x}{x!}$$

先得到似然概率：

$$\begin{aligned} p(\mathscr{D}|\lambda) =& \prod_{i=1}^{|\mathscr{D}|} Poi(d_i|\lambda) \\ =& \exp\{-|\mathscr{D}|\lambda\}\,\lambda^{\sum_{i=1}^{|\mathscr{D}|} d_i}\frac{1}{\prod_{i=1}^{|\mathscr{D}|} d_i!} \end{aligned}$$

故：

$$NLL(\lambda) = |\mathscr{D}|\lambda - (\sum_{i=1}^{|\mathscr{D}|} d_i)\ln\lambda$$

凸性易证，对一阶导置零得到：

$$\lambda_{ML} = \frac{\sum_{i=1}^{|\mathscr{D}|} d_i}{|\mathscr{D}|}$$

这里介绍一个利用泊松分布的回归方法——**泊松回归**，首先声明泊松回归中的变量因果关系：

1、给定输入$\mathbf{x}_i$和未知参数向量θ做出$\lambda_i = \exp\left\{\mathbf{x}_i^T\theta\right\}$；

2、输出y_i服从参数λ_i的泊松分布。

如图所示：

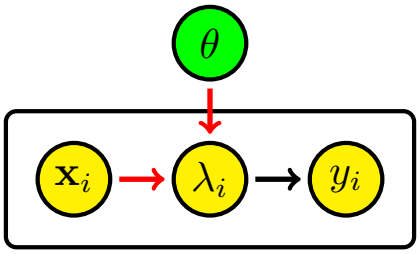

图 7: 泊松回归的贝叶斯网络

可以直接写出全变量集的联合分布：

$$p(\mathbf{x}_i, \lambda_i, \theta, y_i) = p(\mathbf{x}_i)p(\theta)Poi(y_i|\exp\left\{\mathbf{x}_i^T\theta\right\})$$

以待求参数为条件，并将$\mathbf{x}_i$视为已知即得似然概率：

$$p(y_i|\mathbf{x}_i, \theta) = \exp\left\{-\exp\left\{\mathbf{x}_i^T\theta\right\}\right\}\frac{\exp\left\{y_i\mathbf{x}_i^T\theta\right\}}{y_i!}$$

故：

$$NLL(\theta) = \exp\left\{(\sum_{i=1}^{|\mathscr{D}|}\mathbf{x}_i)^T\theta\right\} - (\sum_{i=1}^{\mathscr{D}}y_i\mathbf{x}_i)^T\theta$$

因为（广义不等式定义在半正定矩阵锥上）：

$$\nabla^2 NLL(\theta) = (\sum_{i=1}^{|\mathscr{D}|}\mathbf{x}_i)(\sum_{i=1}^{|\mathscr{D}|}\mathbf{x}_i)^T\exp\left\{(\sum_{i=1}^{|\mathscr{D}|}\mathbf{x}_i)^T\theta\right\} \geq 0$$

所以可以取一阶导为零的条件为解：

$$(\sum_{i=1}^{|\mathscr{D}|}\mathbf{x}_i)^T\theta = \log(\frac{(\sum_{i=1}^{|\mathscr{D}|}y_i\mathbf{x}_i)^T(\sum_{i=1}^{|\mathscr{D}|}\mathbf{x}_i)}{(\sum_{i=1}^{|\mathscr{D}|}\mathbf{x}_i)^T(\sum_{i=1}^{|\mathscr{D}|}\mathbf{x}_i)})$$

例6：多元高斯分布的最大似然估计。多元高斯分布（Multivariate Normal，MVN）的密度函数也即似然概率为：

$$p(\mathbf{x}|\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

本例分析中记数据全集为$\mathbf{X},|\mathbf{X}|=N$，单个数据项为$\mathbf{x}_n$，先求NLL：

$$NLL(\mu,\Sigma)=\frac{N}{2}\log|\Sigma|+\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)^T\Sigma^{-1}(\mathbf{x}_n-\mu)$$

该函数是关于μ的二次型，对于μ是凸的。另一方面可以证明函数$f(X)=\log|X^{-1}|$与$f(x)=\mathbf{tr}(X^{-1})$都是凸的，所以NLL是凸的。

此处举例证明$f(X)=\log|X^{-1}|$的凸性，令$X=Z+tV$，其中Z,V为n阶对称矩阵，设$g(t)=log|Z+tV|$，有：

$$\begin{aligned}g(t)=&\log|Z+tV|\\=&log|Z^{\frac{1}{2}}(I+tZ^{-\frac{1}{2}}VZ^{-\frac{1}{2}})Z^{\frac{1}{2}}|\\=&\log|Z|+\sum_{i=1}^{n}\log(1+t\lambda_i)\end{aligned}$$

得到（其中λ为$Z^{-\frac{1}{2}}VZ^{-\frac{1}{2}})Z^{\frac{1}{2}}$的特征值）：

$$g''(t)=-\sum_{i=1}^{n}\frac{\lambda_i^2}{(1+t\lambda_i)^2}\leq 0$$

所以$\log|X|$为凹函数，得证原命题。

首先来求μ_{ML}：

$$\frac{\partial}{\partial\mu}NLL(\mu,\Sigma)=\sum_{n=1}^{N}(\mu-\mathbf{x}_n)^T\Sigma^{-1}=0$$

所以：

$$\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

求Σ_{ML}时，实际上求$\Lambda_{ML}=\Sigma_{ML}^{-1}$，因为NLL是关于$\Lambda$的凸函数，过程中需要利用以下矩阵微分公式：

$$\begin{aligned}\frac{\partial}{\partial\mathbf{A}}\log|\mathbf{A}|=&\mathbf{A}^{-T}\\\frac{\partial}{\partial\mathbf{A}}\mathbf{tr}(\mathbf{BA})=&\mathbf{B}^T\end{aligned}$$

再定义：

$$\mathbf{S}_n = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^T$$

因为：

$$\begin{aligned}\frac{\partial}{\partial\Lambda}NLL =& \frac{\partial}{\partial\Lambda}\left\{-\frac{N}{2}\log|\Lambda| + \frac{N}{2}\mathbf{tr}(\mathbf{S}_n\Lambda)\right\}\\ =& \frac{N}{2}\left\{-\Lambda^{-1} + \mathbf{S}_n\right\}\\ =& \frac{N}{2}\left\{\mathbf{S}_n - \Sigma\right\}\end{aligned}$$

所以：

$$\Sigma_{ML} = \mathbf{S}_n$$

在一维的情况下，这一解退化成：

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n = \bar{x}$$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 = \frac{1}{N}\sum_{n=1}^{N}x_i^2 - (\bar{x})^2$$

例7：均匀分布的最大似然估计。本节最后考虑一个看起来很平凡的例子，假设一个分布在$[0,\theta]$上的均匀分布，现试图从$\mathscr{D}$中推断θ，似然概率此时为（$\mathbb{I}$为示性函数）：

$$p(d|\theta) = \frac{1}{\theta}\mathbb{I}(d \le \theta)$$

则：

$$NLL(\theta) = |\mathscr{D}|\log\theta - \sum_{i=1}^{|\mathscr{D}|}\log\mathbb{I}(d \le \theta)$$

显然，倘若θ小于$\mathscr{D}$中的任何一个元素，NLL都会直接到达无穷大，所以必有：

$$\theta \ge \max\{d_i\}, i = 1, ..., |\mathscr{D}|$$

从NLL的第一项来看，θ应该尽可能取小的值，所以只能有：

$$\theta_{ML} = \max\{d_i\}, i = 1, ..., |\mathscr{D}|$$

3.1.3 线性回归的最大似然估计

本节给出线性回归的最大似然估计和一些变形。

在2.2.2节最后已经得出了线性回归的似然概率：

$$\begin{aligned} p(y_i|\mathbf{x}_i,\mathbf{w},\sigma^2) =& N(y_i-\mathbf{w}^T\phi(\mathbf{x}_i)|0,\sigma^2) \\ =& \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_i-\mathbf{w}^T\phi(\mathbf{x}_i))^2\right\} \end{aligned}$$

不难发现线性回归的最大似然估计类似于一个一元正态分布的最大似然估计：

$$NLL(\mathbf{w}) = C\cdot\sum_{i=1}^{|\mathscr{D}|}(y_i-\mathbf{w}^T\phi(\mathbf{x}_i))^2$$

这是一个二次型函数的最小化问题，最大似然估计取值为一阶导的零点：

$$\nabla_{\mathbf{w}}NLL(\mathbf{w}) = \sum_{i=1}^{|\mathscr{D}|}(y_i-\mathbf{w}^T\phi(\mathbf{x}_i))\phi(\mathbf{x}_i)^T$$

故：

$$\mathbf{w}_{ML}^T(\sum_{i=1}^{|\mathscr{D}|}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T) = \sum_{i=1}^{|\mathscr{D}|}y_i\phi(\mathbf{x}_i)^T$$

$$\mathbf{w}_{ML}^T = (\sum_{i=1}^{|\mathscr{D}|}y_i\phi(\mathbf{x}_i)^T)(\sum_{i=1}^{|\mathscr{D}|}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T)^{-1}$$

利用矩阵可以写成更紧凑的形式，这里不再介绍。

既然线性回归的贝叶斯网络给出了似然概率的形式，当然也可以将σ^2视为待定变量并进行最大似然估计，此时等价于在这样一个贝叶斯网络中进行分析，其中代表σ^2的结点被染成绿色：

为保凸性，令$\sigma^{-2}=v$：

$$\nabla_v NLL(v) = \nabla_v\left\{-\frac{|\mathscr{D}|}{2}\log v+\frac{v}{2}\sum_{i=1}^{|\mathscr{D}|}(y_i-\mathbf{w}^T\phi(\mathbf{x}_i))^2\right\}$$

故：

$$\sigma_{ML}^2 = v_{ML}^{-1} = \frac{\sum_{i=1}^{|\mathscr{D}|}(y_i-\mathbf{w}^T\phi(\mathbf{x}_i))^2}{|\mathscr{D}|}$$

不难发现该式和一元高斯分布中对于方差的估计是很类似的，均是一个平方误差在数据集上的均值。

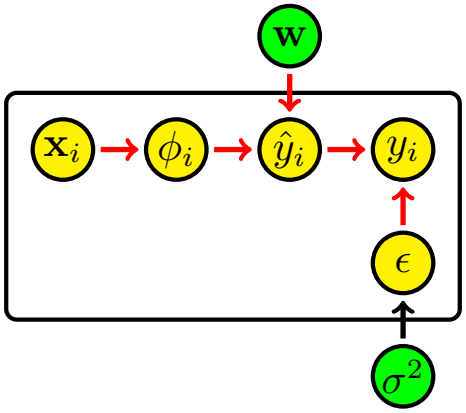

图 8: 线性回归的贝叶斯网络——将σ^2视为待求参数

再考虑另一种扩展，先前一直假设σ^2是固定不随着$\mathscr{D}$变化的值，现在假设σ^2随着$\mathscr{D}$变换，换言之，当$i \neq j$时，不一定有$\sigma_i^2 = \sigma_j^2$，此时的网络结构为：

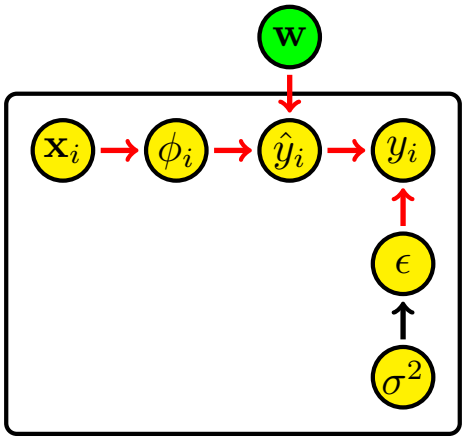

图 9: 线性回归的贝叶斯网络——将σ_i^2视为每次变换的量

单个数据项的联合分布与似然概率不变：

$$p(y_i|\mathbf{x}_i, \mathbf{w}, \sigma_i^2) = N(y_i - \mathbf{w}^T\phi(\mathbf{x}_i)|0, \sigma_i^2)$$

考虑从中推断参数$\mathbf{w}$的过程，此时不能直接将σ^2合并到常数项中消除，

记$v_i = \sigma_i^{-2}$（最优化的求解方法并没有很大差别）：

$$NLL(\mathbf{w}) = \sum_{i=1}^{|\mathscr{D}|} v_i (y_i - \mathbf{w}^T \phi(\mathbf{x}_i))^2$$

进一步地，σ^2描述的是第i个回归项中y_i上的方差，一般可以假设这一方差正相关于于$\mathbf{x}_i$与$\hat{\mathbf{x}}$的差值，因为这代表$\mathbf{x}_i$是一个离群点，所以它对于真实的NLL应该有很小的贡献，当取（认为这种正相关服从一个方差为τ^2的正态分布时）：

$$v_i = \exp\left\{-\frac{(\mathbf{x}_i - \hat{\mathbf{x}})^T(\mathbf{x}_i - \hat{\mathbf{x}})}{2\tau^2}\right\}$$

时，得到模型称为**局部加权线性回归**，它是一种对于离群异常数据鲁棒的线性回归方法。

虽然图9描述的是一般的重新加权的线性回归方法，但是标准局部加权线性回归的联合分布形式与图9所导出的一致，因为局部加权线性回归仅仅在图9中引入了额外的确定性因果关系（红色边），额外的因果关联因子积为1，并不影响联合分布。

3.1.4 线性分类的最大似然估计

本节给出线性分类的最大似然估计和一些变形。

利用2.2.3节给出的似然概率：

$$p(t_i, \mathbf{x}_i | \pi, \mu, \Sigma) = \prod_{c=1}^{C} (\pi_c N(\mathbf{x}_i | \mu_c, \Sigma_c))^{t_{ic}}$$

可得（以α_c记$\mathscr{D}$出现分类为c的数据项数量）：

$$\begin{aligned} NLL(\pi, \mu, \Sigma) &= -\sum_{i=1}^{|\mathscr{D}|}\sum_{c=1}^{C} t_{ic}(\log \pi_c + \log N(\mathbf{x}_i | \mu_c, \Sigma_c)) \\ &= -\sum_{c=1}^{C}\sum_{i=1}^{|\mathscr{D}|} t_{ic} \log \pi_c - \sum_{c=1}^{C}\sum_{i=1}^{|\mathscr{D}|} t_{ic} \log N(\mathbf{x}_i | \mu_c, \Sigma_c) \\ &= -\sum_{c=1}^{C} \alpha_c \log \pi_c - \sum_{c=1}^{C} \alpha_c \log N(\mathbf{x}_i | \mu_c, \Sigma_c) \end{aligned}$$

NLL分解成了两个独立项的和，可以对两项分别最优化，对于第一项的优化等价于例4，得到：

$$\pi_{c,ML} = \frac{\alpha_c}{\sum_{c'=1}^{C} \alpha_{c'}} = \frac{\alpha_c}{|\mathscr{D}|}$$

对于第二项的优化等价于对C个独立的MVN进行最大似然估计，套用例6的结论即可（其中$\mathscr{D}_c$是$\mathscr{D}$中满足$t=c$的子集）：

$$\mu_{c,ML}=\frac{1}{\mathscr{D}_c}\sum_{i=1}^{|\mathscr{D}_c|}d_{ci}$$

$$\Sigma_{c,ML}=\frac{1}{\mathscr{D}_c}\sum_{i=1}^{|\mathscr{D}_c|}(d_{ci}-\mu_c)(d_{ci}-\mu_c)^T$$

以上给出的是线性分类生成模型的最大似然估计参数推断。整个过程等同于：先对于各个类型数据项的数量进行多元伯努利最大似然估计，再对于每一个类型所包含的特征变量子集进行MVN的最大似然估计。

进一步假设$\mathbf{x}$的值一直是给定的，那么模型就从一个生成模型转化为一个**判别模型**。对$\mathbf{x}$取条件，利用贝叶斯公式：

$$p(t_i=c|\mathbf{x}_i,\pi,\mu,\Sigma)=\frac{p(t_i=c,\mathbf{x}_i|\pi,\mu,\Sigma)}{\sum_{c'=1}^{C}p(t_i=c',\mathbf{x}_i|\pi,\mu,\Sigma)}$$

分子分母中的各项都是显式的似然概率：

$$\begin{aligned}p(t_i=c|\mathbf{x}_i,\pi,\mu,\Sigma)=&\frac{\pi_cN(\mathbf{x}_i|\mu_c,\Sigma_c)}{\sum_{c'=1}^{C}\pi_{c'}N(\mathbf{x}_i|\mu_{c'},\Sigma_{c'})}\\=&\frac{\pi_c\frac{1}{|\Sigma_c|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_c)^T\Sigma_c^{-1}(\mathbf{x}_i-\mu_c)\right\}}{\sum_{c'=1}^{C}\pi_{c'}\frac{1}{|\Sigma_{c'}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_{c'})^T\Sigma_{c'}^{-1}(\mathbf{x}_i-\mu_{c'})\right\}}\end{aligned}$$

寻求这样一种近似：

$$\Sigma_c=\Sigma,c=1,...C$$

此时：

$$p(t_i=c|\mathbf{x}_i,\pi,\mu,\Sigma)=\frac{\exp\left\{\mathbf{x}_i^T\beta_c+\gamma_c\right\}}{\sum_{c'=1}^{C}\exp\left\{\mathbf{x}_i^T\beta_c+\gamma_c\right\}}$$

其中：

$$\beta_c=\Sigma^{-1}\mu_c$$

$$\gamma_c=-\mu_c^T\Sigma^{-1}\mu_c+\ln\pi_c$$

可继续认为$\mathbf{x}:=(\mathbf{x}^T,1)^T$，并做成：

$$p(t_i=c|\mathbf{x}_i,\pi,\mu,\Sigma)=p(t_i=c|\mathbf{x}_i,\mathbf{W})=\frac{\exp\left\{\mathbf{x}_i^T\mathbf{w}_c\right\}}{\sum_{c'=1}^{C}\exp\left\{\mathbf{x}_i^T\mathbf{w}_{c'}\right\}}$$

幂函数中的二次项已经统一消除，这个条件概率函数的形式是sotfmax函数。

当$C=2$时：

$$p(t_i=0|\mathbf{x}_i,\pi,\mu,\Sigma)=\frac{1}{1+\exp\{\mathbf{w}^T\mathbf{x}_i+\gamma\}}$$

给出可以用Sigmoid函数表达的**逻辑回归**。

为了求解这种形式的判别模型中的参数，写出此时的NLL：

$$\begin{aligned}NLL(\mathbf{W})&=-\log\prod_{i=1}^{|\mathscr{D}|}\prod_{c=1}^{C}p(t_i=c|\mathbf{x}_i,\mathbf{W})^{t_{ic}}\\&=-\sum_{i=1}^{|\mathscr{D}|}\sum_{c=1}^{C}t_{ic}\log p(t_i=c|\mathbf{x}_i,\mathbf{W})\\&=-\sum_{i=1}^{|\mathscr{D}|}\left\{\sum_{c=1}^{C}t_{ic}\mathbf{w}_c^T\mathbf{x}_i-\log(\sum_{c'=1}^{C}\exp\left\{\mathbf{w}_{c'}^T\mathbf{x}_i\right\})\right\}\end{aligned}$$

此时NLL的海森矩阵半正定性涉及张量积运算，可参考MLAPP习题8.4。

最后给出二类线性分类判别模型（即逻辑回归模型）的NLL，有：

$$p(t_i=1|\mathbf{x}_i,\mathbf{w})=\frac{1}{1+\exp\{-\mathbf{w}^T\mathbf{x}_i\}}$$

其中经过了$\mathbf{x}_i$多一项1的扩展，和参数合并成一个线性项的过程，记两种类型为$\{-1,+1\}$，并选择t_i直接编码为类型的值。

$$p(\mathscr{D}|\mathbf{w})=\prod_{i=1}^{|\mathscr{D}|}p(t_i=1|\mathbf{x}_i,\mathbf{w})^{\frac{t_i+1}{2}}(1-p(t_i=1|\mathbf{x}_i,\mathbf{w}))^{\frac{1-t_i}{2}}$$

$$NLL(\mathbf{w})=\sum_{i=1}^{|\mathscr{D}|}\log(1+\exp\left\{-t_i\mathbf{x}_i^T\mathbf{w}\right\})$$

凸性由：

$$\nabla^2 NLL(\mathbf{w})=\sum_{i=1}^{|\mathscr{D}|}t_i^2\mathbf{x}_i\mathbf{x}_i^T\frac{\exp\left\{-t_i\mathbf{x}_i^T\mathbf{w}\right\}}{(1+\exp\left\{-t_i\mathbf{x}_i^T\mathbf{w}\right\})^2}$$

是一系列半正定矩阵的非负加权和得到。

线性回归的判别模型一般一个没有解析解的无约束凸优化问题，由于目标函数可微，使用下降方法必然得到唯一的全局最优解。使用Newton下降法于Logistic回归上时特别地被称为Iterative Reweighted Least Square(IRLS)

3.1.5 一般贝叶斯网络的最大似然估计

一般贝叶斯网络的最大似然估计是最抽象的情况，但是其实践意义并不高。因为只有所有变量取离散情况，或者连续变量的因果关系取线性高斯关系时，贝叶斯网络的参数估计和推理有良好的封闭形式。

但是本节仍将说明，将所有变量视作由因果关系维系的网络，比将所有变量的联合分布视作一个任意的黑箱，可以简化计算复杂性。

2.1节给出贝叶斯网络中所有变量的联合分布：

$$p(\mathscr{X}|G,\theta)=\prod_{i=1}^{|\mathscr{X}|}p(X_i|Pa(X_i),\theta)$$

认为图结构，即变量间的因果关系是预先给定的，并把θ视作待求的参数随机变量，则上式的意义就是确定贝叶斯网络骨架G时，一组合法参数θ的似然概率。

贝叶斯网络中的最大似然估计（仅对于参数而言，不对于图结构而言）是：

$$\theta_{ML}=\arg\max_{\theta}\{p(\mathscr{X}|G,\theta)\}$$

类似地使用NLL来变换问题，则有：

$$\theta_{ML}=\arg\min_{\theta}\left\{-\sum_{i=1}^{|\mathscr{X}|}\log p(X_i|Pa(X_i,\theta))\right\}$$

如果再进一步假设所有的因果关系之间没有共享参数，则上式可以简化成：

$$\theta_{X_i|Pa(X_i),ML}=\arg\min_{\theta_{X_i|Pa(X_i)}}\left\{-\log p(X_i|Pa(X_i),\theta_{X_i|Pa(X_i)})\right\}$$

换言之，可以把表征所有CPD的集合的参数集合θ分别地最优化，且优化一个局部CPD时只需要参考给定数据集合中和$\{X_i,Pa(X_i)\}$有关的部分即可。

即贝叶斯网络的最大似然估计，一般可以转化为对于每个局部的条件概率模型进行独立的最大似然估计，并组合所有参数形成整个网络的最大似然估计参数集合。和直接处理所有变量的联合分布相比，贝叶斯网络允许整个任务的最优解分解成规模上更简单的子问题的最优解之组合。

回顾线性分类生成模型的例子，在该问题的贝叶斯网络图6中，可知似然概率仅由两个CPD相乘所得，所以进行最大似然估计时仅需要独立地进

行两个CPD的最大似然估计即可，这也就是3.1.4节中第一个NLL可以分解成两个独立项的和的原理。

例8：离散贝叶斯网络的最大似然估计解析形式。考虑一个所有变量均为离散随机变量的贝叶斯网络，此时对应的所有CPD均为表格，给定数据集合$\mathscr{D}$，现试图求所有的具体CPD表值的最大似然估计。

此时所有的CPD都呈多项伯努利分布形式，按照刚才的讨论：

$$\theta_{X_i|Pa(X_i),ML} = \arg \min_{\theta_{X_i|Pa(X_i)}} \left\{ -\log p(X_i|Pa(X_i), \theta_{X_i|Pa(X_i)}) \right\}$$

由任一组父节点的联合取值$v \in Val(Pa(X_i))$都能定义一个在X_i上归一化的多元伯努利分布，据此进一步分割并约简上式：

$$\theta_{X_i|v,ML} = \arg \min \left\{ -\log p(X_i|Pa(X_i) = v, \theta_{X_i|v}) \right\}$$

因为这是一个多元伯努利模型，利用例4中的结论，记$\mathscr{D}$中符合$Val(Pa(X_i)) = v$的元素数量为α，对于X_i的任意取值x_{ij}，记$\mathscr{D}$中符合$Val(X_i, Pa(X_i)) = (x_{ij}, v)$的元素数量为$\alpha_j$，那么对应着$(x_{ij}, v)$处的表值即是：

$$\frac{\alpha_j}{\alpha}$$

它是$\mathscr{D}$中，一个局部因果变量集合均符合某特定赋值的数量，与仅仅因变量集合符合特定赋值的数量之比。故可以每次只考虑极大似然估计一个果变量的所有因变量构成的局部因果集合。

3.2 最大后验估计

最大后验估计（MAP）的形式是：

$$\theta_{MAP} = \arg\max_{\theta} \{p(\theta|\mathscr{D})\}$$

用负对数的形式改写：

$$\theta_{MAP} = \arg\min_{\theta} \{-\log p(\theta|\mathscr{D})\}$$

再利用贝叶斯公式：

$$p(\theta|\mathscr{D}) = \frac{p(\mathscr{D}|\theta)p(\theta)}{p(\mathscr{D})} \propto p(\mathscr{D}|\theta)p(\theta)$$

最终：

$$\theta_{MAP} = \arg\min_{\theta} \{-\log p(\mathscr{D}|\theta) - \log p(\theta)\}$$

即最大后验估计的目标函数是最大似然估计的目标函数增加一项，这一项被称作**正则项**。当$p(\theta)$很平坦时，可以近似地认为最大后验估计等价于最大似然估计。同时因为$\log p(\mathscr{D}|\theta)$本身随着$|\mathscr{D}|$增长，所以最大后验估计在数据量足够庞大时收敛于最大似然估计。

理论上可以选择任何形式的$p(\theta)$来表示对于参数的先验知识，但是这可能会破坏问题的解析性质。一般而言对于某个NLL，我们希望找到一个$-\log p(\theta)$，使得后验估计的解析形式与似然估计的解析形式相仿，此时$p(\theta)$称作原似然概率分布的**共轭先验分布，Conjugate Prior**。

对于一般的分布而言，共轭先验分布没有固定的寻找方法。

3.2.1 概率分布的后验估计

本小节枚举一些对于概率分布的参数进行后验估计的例子。

例9：多元伯努利分布的最大后验估计。多元伯努利分布的NLL为：

$$\begin{aligned} NLL(\theta) &= -\sum_{i=1}^{|\mathscr{D}|}\sum_{k=1}^{K} d_{ik}\log\theta_k \\ &= -\sum_{k=1}^{K}\alpha_k\log\theta_k \end{aligned}$$

多元伯努利分布的共轭先验分布为狄利克雷分布：

$$Dir(\mathbf{x}|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K} x_k^{\alpha_k-1}\mathbb{I}(\mathbf{1}^T\mathbf{x}=1)$$

令：

$$p(\theta) = Dir(\theta|\pi) \propto \prod_{k=1}^{K} \theta_k^{\pi_k - 1}$$

则：

$$NLP(\theta) = NLL(\theta) - \log p(\theta) = -\sum_{k=1}^{K} (\alpha_k + \pi_k - 1) \log \theta_k$$

此时求导数置零得MAP很容易，还可以发现：后验分布与先验分布一样，同为狄利克雷分布：

$$p_{posterior}(\theta) = Dir(\theta|\pi + \alpha)$$

MAP的解就是该分布的模。这就是一个“合适的”先验分布（共轭先验分布）良好的解析性质，它使得后验分布和先验分布在分布族上统一，所以对于新增的数据集可以累积地将后验知识融入先验知识之中。

多元伯努利分布的先验分布参数可以理解为：在读入真实实验数据之前，已经观察到k变量出现了π_k次。这个模型叫做**Dirichet-Multinomial Model**，$K = 2$时退化为**Beta-Binomial Model**。Beta分布是二元伯努利分布的共轭先验分布。

例10：泊松分布的最大后验估计。泊松分布的NLL为：

$$NLL(\lambda) = |\mathscr{D}|\lambda - (\sum_{i=1}^{|\mathscr{D}|} d_i) \ln \lambda$$

这里直接指出泊松分布的共轭先验分布为伽马分布：

$$Ga(T|a, b) = \frac{b^a}{\Gamma(a)} T^{a-1} \exp\{-Tb\}$$

则：

$$p(\lambda) = Ga(\lambda|a, b) \propto \lambda^{a-1} \exp\{-b\lambda\}$$

$$\begin{aligned} NLP(\lambda) =& NLL(\lambda) - \log p(\lambda) \\ =& (|\mathscr{D}| + b)\lambda - (a - 1 + \sum_{i=1}^{|\mathscr{D}|} d_i) \ln \lambda \end{aligned}$$

即：

$$p_{posterior}(\lambda) = Ga(\lambda|a + \sum_{i=1}^{|\mathscr{D}|} d_i, b + |\mathscr{D}|)$$

故最大后验估计为一个伽马分布的模，在读入一个新数据项时，只需要把伽马分布的第一个参数递增读入的数据项，第二个参数递增1，既得到吸纳新知识后的后验分布。该先验分布也等同于引入了总共b个先验观测项，其均值为$\frac{a}{b}$。

例11：均匀分布的后验估计。均匀分布的共轭先验分布是帕累托分布：

$$Pareto(x|k,m) = km^k x^{-(k+1)}\mathbb{I}(x \geq m)$$

设均匀分布在$[0,\theta]$上，且：

$$p(\theta) = Pareto(\theta|K,b)$$

则：

$$\begin{aligned}
p(\mathscr{D},\theta) =& p(\theta)p(\mathscr{D}|\theta) \\
=& Kb^K\theta^{-(K+1)}\mathbb{I}(\theta \geq b)\prod_{i=1}^{|\mathscr{D}|}\mathbb{I}(\theta \geq d_i)\theta^{-1} \\
=& Kb^K\theta^{-(K+1+|\mathscr{D}|)}\mathbb{I}(\theta \geq b)\mathbb{I}(\theta \geq \max\{d\}) \\
=& Kb^K\theta^{-(K+1+|\mathscr{D}|)}\mathbb{I}(\theta \geq \mu)
\end{aligned}$$

其中$\mu = \max\{b, d_i\}, i = 1, ..., |\mathscr{D}|$。

计算**边缘似然**：

$$\begin{aligned}
p(\mathscr{D}) =& \int p(\mathscr{D},\theta)d\theta \\
=& \int_\mu^\infty Kb^K\theta^{-(K+1+|\mathscr{D}|)}d\theta \\
=& \frac{Kb^K}{N+K}\mu^{-N-K}
\end{aligned}$$

最后由：

$$\begin{aligned}
p(\theta|\mathscr{D}) =& \frac{p(\mathscr{D},\theta)}{p(\mathscr{D})} \\
=& \frac{N+K}{\theta^{N+K+1}}\mu^{N+K}\mathbb{I}(\theta \geq \mu) \\
=& Pareto(\theta|N+K,\mu)
\end{aligned}$$

后验与先验分布取相同形式，可见帕累托分布确实是均匀分析的共轭先验分布。

3.2.2 线性回归的最大后验估计

现在讨论线性回归模型的后验估计，目标变量为$\mathbf{w}$时，首先需要定义一个$p(\mathbf{w}|\theta)$，暂时认为θ是固定的，此时的贝叶斯网络为：

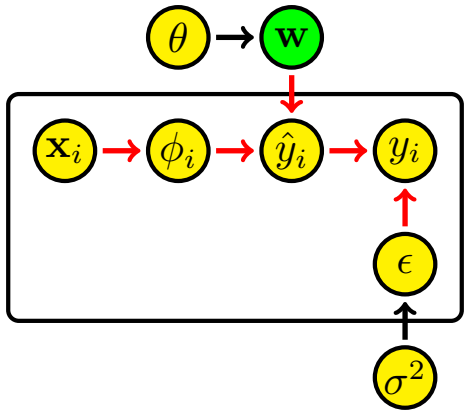

图 10: 线性回归的贝叶斯网络——后验估计

注意到区别于图5中的情况，虽然$\mathbf{w}$仍是一个待求参数（绿色结点），但是它是由一个概率分布定义的，所以$\mathbf{w}$需要由一个参数分布给出。当更高阶的参数θ确定时，通过最大后验估计可以给出$\mathbf{w}$的一个确定值，如果θ也被做成待定参数，则可以通过最大似然估计方式求出θ_{ML}，并用$p(\mathbf{w}|\theta_{ML})$给出$\mathbf{w}$的分布。

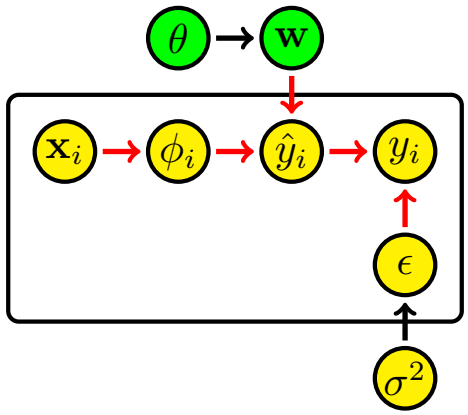

图 11: 线性回归的贝叶斯网络——贝叶斯估计

从图中完全重新导出联合分布的形式或者直接利用NLP的式子都可以

得到同样的：

$$NLP(\mathbf{w}) = NLL(\mathbf{w}) - \log p(\mathbf{w}|\theta)$$

考虑两种情况：先验分布为正态分布、先验分布为拉普拉斯分布。

1、当先验分布为正态分布时，设：

$$p(\mathbf{w}|\theta = \tau^2) = N(\mathbf{w}|0, \tau^2)$$

则：

$$NLP(\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{i=1}^{|\mathscr{D}|}(y_i - \mathbf{w}^T\phi(\mathbf{x}_i))^2 + \frac{1}{2\tau^2}\mathbf{w}^T\mathbf{w}$$

最大后验估计是最小化一个正定二次型的问题，可以解析求解。这样一个线性回归的变形叫做**岭回归，Ridge Regression**。

2、当先验分布为拉普拉斯分布时，设：

$$p(w_i|\theta = \lambda^{-1}) = \frac{\lambda}{2}\exp\{-\lambda|w_i|\}$$

$$p(\mathbf{w}) \propto \prod_{i=1}^{|\mathbf{w}|}\exp\{-\lambda|w_i|\}$$

故：

$$NLP(\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{i=1}^{|\mathscr{D}|}(y_i - \mathbf{w}^T\phi(\mathbf{x}_i))^2 + \lambda\|\mathbf{w}\|_1$$

此时的最优化目标函数是不可微的，最优化时需要利用到一些问题变换的技术和亚微分的知识，这里不做介绍。这一线性回归变形叫做**套索回归，Lasso(least absolute shrinkage and selection operator) Regression**

套索回归的命名原因是它往往能给出一个稀疏的解，即$\mathbf{w}$中有一些分量会被置零，因此也可以用作自然的关联判别算法。它倾向于提供一个稀疏解是因为它最小化了1-范数，而岭回归最小化2-范数，在高维空间中1-范数球和2-范数球相比更加贴近坐标轴。不过二者都是图10所引导的，唯一的区别在于由θ指向$\mathbf{w}$的箭头蕴含的因果关系不同。

3.2.3 一般贝叶斯网络的最大后验估计

一般贝叶斯网络的后验估计和3.1.5节的思路一脉相承，譬如对于离散贝叶斯网而言，只需要利用例9的结论代入先验分布为狄利克雷分布即可。

因为贝叶斯网络蕴含的独立性质，整个网络的最大后验估计等同于合并各个局部的最大后验解。

3.3 *贝叶斯估计

无论最大似然估计还是最大后验估计都抛弃了关于参数的不确定性的信息，贝叶斯估计认为必须存储完整的后验分布以作为参数估计的结果。最大似然或者后验估计返回参数的一个确定性的取值，而贝叶斯估计返回一个定义在参数可行空间上的后验分布。

记一个模型中有一个待定参数θ，以某种因果关系和其他的变量受一张贝叶斯网络联系。现在有至少三种方法处理θ的学习情况：

1、最大似然估计：认为θ有一个最合适的值θ_{ML}，最大化了似然概率。

2、最大后验估计：认为θ由一个先验分布$p(\theta)$生成，它有一个最合适的值θ_{MAP}，最大化了后验概率。

3、贝叶斯估计：认为θ由一个参变先验分布$p(\theta|\pi)$生成，对于π而言，认为π有一个最合适的值π_{ML}，最大化了似然概率。而θ服从$p(\theta|\pi_{ML})$。

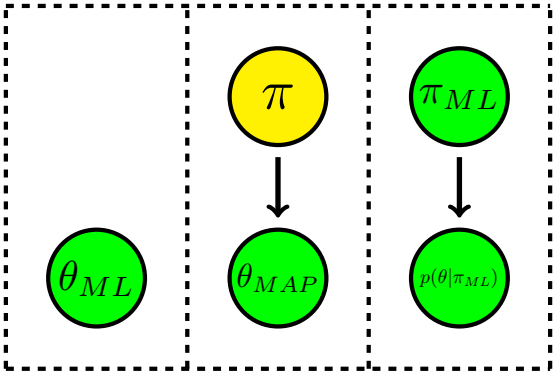

图 12: 最大似然估计；最大后验估计；贝叶斯估计（局部贝叶斯网络）

从上图中可得到将任何一个最大似然估计过程转化为最大后验估计的过程：即给所求参数增加一个额外的因变量作为父节点，并给父节点赋予一个值。将最大似然估计改造成贝叶斯估计的过程即将增加的父节点理解为一个待求参数。

原则上可以将待定参数上的因变量链延长：

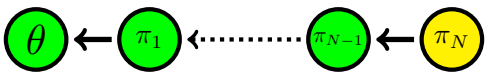

图 13: 经验贝叶斯

这样延长链后理论上可以在θ上引导任意复杂的分布。可以通过对于整条因果链的中间变量求和来消除除了第一因与θ以外的所有变量：

$$\begin{aligned}
p(\theta|\pi_N) &= \int ... \int p(\theta, \pi_1, ...\pi_{N-1}|\pi_N) d\pi_1 ... d\pi_{N-1} \\
&= \int ... \int \left\{ \prod_{i=1}^{N-1} p(\pi_i|\pi_{i+1}) \right\} p(\theta|\pi_1) d\pi_1 ... d\pi_{N-1} \\
&= \int ... \int \left\{ \prod_{i=1}^{N-3} p(\pi_i|\pi_{i+1}) \right\} \\
&\quad \cdot \left\{ \int p(\pi_{N-2}|\pi_{N-1}) p(\pi_{N-1}|\pi_N) d\pi_{N-1} \right\} d\pi_1 ... d\pi_{N-2}
\end{aligned}$$

可利用以上最后一个等式来简化计算，这是**经验贝叶斯方法，Empirical Bayesian**的思路。

在贝叶斯统计理论中，一个模型是由其所定义的似然概率和其先验分布形成的二元序对，模型间进行比较的基准是**边缘似然，Marginal Likelihood**：

$$p(\mathscr{D}|M) = \int p(\mathscr{D}|\theta, M) p(\theta|M) d\theta$$

可以选择一个具有最大边缘似然的模型来作为结论，不过按照贝叶斯统计的手段是赋予高边缘似然的模型更高的权重，并最终在所有的模型上加权求和来形成最终的决策。

在边缘似然的计算中参数渐进学习的过程蕴含在$p(\mathscr{D}|\theta, M)$中：

$$p(\mathscr{D}|\theta, M) = \prod_{i=1}^{|\mathscr{D}|} p(d_i|\theta, M, d_1, ...d_{i-1})$$

当$|\mathscr{D}|$很大时，可由弱大数定律得到：

$$\frac{1}{|\mathscr{D}|} \log p(\mathscr{D}|M) \approx \mathbb{E}[\log p(d|M, \mathscr{D})]$$

其中右侧的期望是对于真实分布而言的。

所以渐进意义上，基于边缘似然的贝叶斯模型选择和3.1.1节描述了投影是一致的。

完全的贝叶斯方法理论价值大于实践效用（因为模型空间本身无法微分），所以此处不再深入介绍。

3.4 隐变量模型

隐含变量模型在贝叶斯网络中引入“隐含”变量，即没有体现在$\mathscr{D}$中，但是在构成$\mathscr{D}$的因果关系中起到明确作用的变量。

3.4.1 连续隐变量模型

之前已经说过，在贝叶斯网络中，当变量为连续，且因果关系为线性高斯关系时，整个贝叶斯网络的所有变量的联合分布或者局部的边缘分布、条件分布均为高斯分布。

据此考虑一个如下的贝叶斯网络（隐元用白色背景表示），其中$\mathbf{z}_i$的分布服从一个单位正态分布，$\mathbf{d}_i$的条件分布服从一个正态分布：

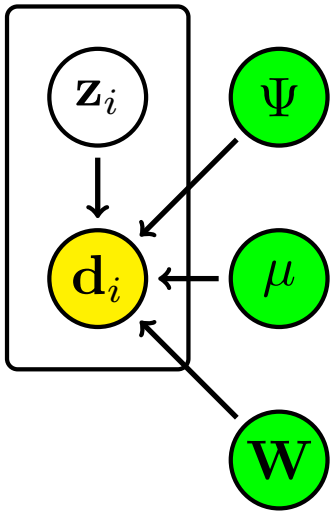

图 14: 一个连续隐元模型的贝叶斯网络

此时有：

$$\begin{aligned} p(\mathbf{z}_i, \mathbf{d}_i, \mathbf{W}, \mu, \Psi) =& p(\mathbf{z}_i)p(\mathbf{W})p(\mu)p(\Psi)p(\mathbf{d}_i|\mathbf{z}_i, \mathbf{W}, \mu, \Psi) \\ =& N(\mathbf{z}_i|\mathbf{0}, \mathbf{I})p(\mathbf{W})p(\mu)p(\Psi)N(\mathbf{d}_i|\mathbf{W}\mathbf{z} + \mu, \Psi) \end{aligned}$$

可以给出**完备数据项**$\{\mathbf{z}_i, \mathbf{d}_i\}$的似然概率：

$$p(\mathbf{z}_i, \mathbf{d}_i|\mathbf{W}, \mu, \Psi) = N(\mathbf{z}_i|\mathbf{0}, \mathbf{I})N(\mathbf{d}_i|\mathbf{W}\mathbf{z} + \mu, \Psi)$$

利用高斯分布的性质，可以消去隐变量$\mathbf{z}_i$来求得显变量$\mathbf{d}_i$的边缘分布（参考PRML，2.3.3高斯变量的贝叶斯理论，或者MLAPP，4.3联合高斯分

布中的推理）：

$$\begin{aligned} p(\mathbf{d}_i|\mathbf{W},\mu,\Psi) =& \int p(\mathbf{z}_i,\mathbf{d}_i|\mathbf{W},\mu,\Psi)d\mathbf{z}_i \\ =& N(\mathbf{d}_i|\mu,\Psi+\mathbf{W}\mathbf{W}^T) \end{aligned}$$

此时在给定$\mathscr{D}$时，可以直接用最大似然估计推断μ，隐变量方法的优势在协方差矩阵的估计上。一般有$|\mathbf{z}| < |\mathbf{d}|$，故$\Psi$是一个$|\mathbf{d}| * |\mathbf{d}|$的矩阵，而$\mathbf{W}$是一个$|\mathbf{d}| * |\mathbf{z}|$的矩阵。如果再进一步假设$\Psi$具有某种简单的形式，譬如对角阵，那么就可能将$\mathscr{D}$协方差矩阵的$\frac{|\mathbf{d}|*(1+\mathbf{d})}{2}$个参数压缩到$|\mathbf{d}| + |\mathbf{d}| * |\mathbf{z}|$个。

假定Ψ作为对角阵的几种情况，演化出了几种不同的变形：

1、Ψ为一般的对角阵，此时模型称为**Factor Analysis,FA**；

2、$\Psi = \sigma^2\mathbf{I}$，此时模型称为**Probabilistic Principal Components Analysis,PPCA**；

3、$\Psi = \mathbf{0}$，此时模型称为**Principal Components Analysis,PCA**；

这里只说明一下PCA的情况。

PCA等价于将给定数据集$\mathscr{D}$的生成概率视作一个MVN，并用SVD分解该分布的协方差矩阵，最后将分解形式中的对角矩阵中较小的值置为零。假设PCA处理后对角阵非零元素为$|\mathbf{z}|$个。则此时的协方差矩阵就可以化约为$\mathbf{W}\mathbf{W}^T$的形式，$\mathbf{W}$由SVD分离出的正交矩阵中对应非零奇异值的分量乘以对角矩阵中非零分量的平方根做成。

3.4.2 离散隐变量模型

考虑如下一个贝叶斯网络，其中z_i服从一个多元伯努利分布，$\mathbf{d}_i$的条件分布服从一个正态分布。设z_i取C个可能离散值中的一个，C种状态引导了C种不同的正态分布：

图12引导联合分布：

$$\begin{aligned} p(\mathbf{d}_i,z_i,\pi,\mu,\Sigma) =& p(\pi)p(\mu)p(\Sigma)p(z_i|\pi)p(\mathbf{d}_i|z_i,\mu,\Sigma) \\ =& p(\pi)p(\mu)p(\Sigma)\pi_{z_i}N(\mathbf{d}_i|\mu_{z_i},\Sigma_{z_i}) \end{aligned}$$

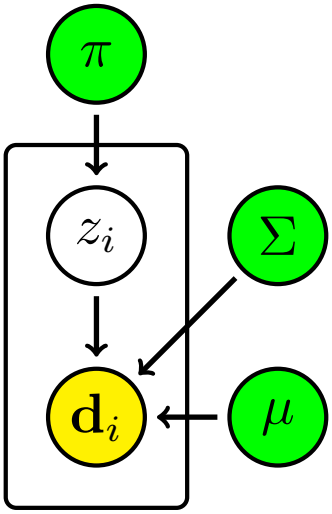

图 15: 一个离散隐元模型的贝叶斯网络

对参数变量取条件，在对z_i求和消去之可得显变量的似然概率：

$$\begin{aligned} p(\mathbf{d}_i|\pi,\mu,\Sigma) &= \sum_{c=1}^{C} p(\mathbf{d}_i, z_i = c|\pi,\mu,\Sigma) \\ &= \sum_{c=1}^{C} \pi_c N(\mathbf{d}_i|\mu_c,\Sigma_c) \end{aligned}$$

这是一个**混合高斯分布**，可以将$p(\mathbf{d}_i|z_i,\theta)$替换成其他的分布，构成其他概率分布的混合模型。任何形如：

$$p(\mathbf{d}_i|\theta) = \sum_{c=1}^{C} \pi_c p(\mathbf{d}_i|c,\theta)$$

的混合分布都可以建模成图12的贝叶斯网络。注意到图12和图6实质上是一致的，所以一般线性分类中特征向量的分布就是一个混合高斯分布，但是区别在于分类标签并非可见的（z_i为隐元）。

3.4.3 隐变量模型的参数推断方法

含有隐元的模型可以在可观测数据集上刻画出比一般的朴素模型复杂得多的分布。譬如如果没有隐元而试图刻画混合分布时，则因果关系就将呈现出相当复杂的模式。一般将隐元变量和显变量的集合$\{\mathscr{Z},\mathscr{D}\}$视作完备

的数据集合，在完备的数据集合上求参数的似然概率往往是很容易的，因为这就是在一个良好定义的贝叶斯网络中求似然概率。

然而问题在于，既然$\mathscr{Z}$没有被观测到，那么即便能写出完备数据集合的似然概率也没有意义。

一个替代方式就是：首先推断$\mathscr{Z}$的分布，再基于$\mathscr{Z}$的分布求完备数据集合的似然，整个算法可以分成两个步骤迭代进行：

1、求：

$$p(\mathscr{Z}|\mathscr{D},\theta^{old})$$

2、最大似然估计：

$$\theta^{new}=\theta_{ML}=\arg\min_{\theta}\left\{\mathbb{E}_{p(\mathscr{Z})}[-\log p(\mathscr{Z},\mathscr{D}|\theta)]\right\}$$

因为最大似然估计中需要对于$p(\mathscr{Z})$的分布取期望，所以在求$p(\mathscr{Z}|\mathscr{D},\theta^{old})$时仅需要寻找一些期望值即可，故上述的两个步骤也被称为**期望步骤（E-step）**和**最大步骤（M-step）**，整个算法称为**期望最大算法（EM algorithm）**。

任何含有隐元的模型都可以使用EM进行参数估计，它最终收敛于参数集合的最大似然估计以及此时隐元的条件分布。

例13：混合高斯分布的最大似然估计。考虑图12引导的混合高斯分布，现试图进行参数估计。首先尝试性地直接对显变量的似然概率：

$$p(\mathbf{d}_i|\pi,\mu,\Sigma)=\sum_{c=1}^{C}\pi_c N(\mathbf{d}_i|\mu_c,\Sigma_c)$$

此时的损失函数形为：

$$NLL(\pi,\mu,\Sigma)=-\sum_{i=1}^{|\mathscr{D}|}\log\sum_{c=1}^{C}\pi_c N(\mathbf{d}_i|\mu_c,\Sigma_c)$$

此时和式出现在对数函数以内，故无法直接展开成一系列独立项的和，导致了形式上的复杂性。下面使用EM算法进行参数推断，首先已知完备数据集的似然概率，使用热洞编码$\mathbf{z}_i$：

$$p(\mathbf{d}_i,\mathbf{z}_i|\pi,\mu,\Sigma)=\prod_{c=1}^{C}\pi_c^{z_{ic}}N(\mathbf{d}_i|\mu_c,\Sigma_c)^{z_{ic}}$$

下面求E-step中隐元的分布：

$$\begin{aligned}
p(z_i|\mathbf{d}_i,\pi,\mu,\Sigma) =& \frac{p(z_i,\mathbf{d}_i|\pi,\mu,\Sigma)}{\sum_{c=1}^{C} p(z_i=c,\mathbf{d}_i|\pi,\mu,\Sigma)} \\
=& \frac{p(z_i|\pi,\mu,\Sigma)p(\mathbf{d}_i|z_i,\pi,\mu,\Sigma)}{\sum_{c=1}^{C} p(z_i=c|\pi,\mu,\Sigma)p(\mathbf{d}_i|z_i=c,\pi,\mu,\Sigma)} \\
=& \frac{\pi_{z_i} N(\mathbf{d}_i|\mu_{z_i},\Sigma_{z_i})}{\sum_{c=1}^{C} \pi_c N(\mathbf{d}_i|\mu_c,\Sigma_c)} \\
=& \gamma_{i,z_i}
\end{aligned}$$

在M-step中，对$\mathscr{D}$的分布最小化完备数据集NLL的期望：

$$\begin{aligned}
\mathbb{E}[NLL(\theta)] =& \mathbb{E}[-\sum_{i=1}^{|\mathscr{D}|} \log p(\mathbf{d}_i,z_i|\pi,\mu,\Sigma)] \\
=& \mathbb{E}[-\sum_{i=1}^{|\mathscr{D}|}\sum_{c=1}^{C} z_{ic}\left\{\log\pi_c + \log N(\mathbf{d}_i|\mu_c,\Sigma_c)\right\}] \\
=& -\sum_{i=1}^{|\mathscr{D}|}\sum_{c=1}^{C} \mathbb{E}[z_{ic}]\left\{\log\pi_c + \log N(\mathbf{d}_i|\mu_c,\Sigma_c)\right\} \\
=& -\sum_{i=1}^{|\mathscr{D}|}\sum_{c=1}^{C} \gamma_{i,c}\log\pi_c - \sum_{i=1}^{|\mathscr{D}|}\sum_{c=1}^{C} \gamma_{i,c}\log N(\mathbf{d}_i|\mu_c,\Sigma_c)
\end{aligned}$$

此时NLL的期望如我们所愿拆分成了两项，参考多元伯努利分布的参数估计和第一项的形式，可以认为NLL的第一部分等同于一个“软热洞编码”状态下的多元伯努利分布的最大似然估计，而第二项等同于“软热洞编码”分类下的线性分类生成模型中一个类别的最大似然参数估计。

整个混合高斯分布的EM算法中M-step优化的就是线性分类生成模型的NLL的“软热洞编码”形式，在这种编码中，归一化仍有保证$\mathbf{1}^T\gamma_i = 1$。原本的热洞编码中t_{ic}非零即一，而γ_{ic}可以理解为$\mathbf{d}_i$属于类型c的概率。

M-step解的封闭形式可泛化3.1.4节的结论得到。

考虑到$\gamma_{i,c}$可以被看做数据项$\mathbf{d}_i$由类别c生成的概率，那么可以把推断$\gamma_{i,c}$的过程看做一个非监督（事先未知类别信息）地分类的过程，这也就是**聚类算法**的原型。

将上述混合高斯模型的最大似然估计进行三个方面的改动即获得**K-means聚类算法**：

1、取聚类标签$\hat{\gamma}_i$为一个热洞向量，其中为1的下标是γ_i各分量中最大元素的下标；

2、取$\pi_c = \pi, c = 1, ..., C$；

此时可以把K-means聚类理解为类似EM的两个步骤：

1、E-step：将每个数据项$\mathbf{d}_i$硬分类成其最有可能的聚类；

2、M-step：根据E-step的分类，重新定位每个聚类群的均值μ_c。

即K-means聚类算法是混合高斯分布EM算法的一种离散近似。

4 *贝叶斯网络推理

4.1 精确推理

贝叶斯网络中的推理之这样一种任务：在给定了网络结构G以及CPD参数θ后，对于所有变量集合$\mathscr{X}$的一个子集$\mathscr{X}'$，求（本节讨论中(G,θ)一直处在条件内，统一略去）：

$$p(\mathscr{X}')$$

对于给定的贝叶斯网络(G,θ)，可以直接得出的是$\mathscr{X}$的联合分布，利用贝叶斯公式：

$$\begin{aligned} p(\mathscr{X}') &= \sum_{\mathscr{X}/\mathscr{X}'} p(\mathscr{X}) \\ &= \sum_{\mathscr{X}/\mathscr{X}'} \prod_{i=1}^{|\mathscr{X}|} p(X_i|Pa(X_i)) \end{aligned}$$

因为$p(X_i|Pa(X_i))$是关于$\{X_i, Pa(X_i)\}$这个变量集合的函数，所以不妨令：

$$p(X_i|Pa(X_i)) = \phi(X_i, Pa(X_i)) = \phi_j(scope[\phi_j])$$

并令：

$$scope[\phi_j] = \{X_i, Pa(X_i)\}$$

利用$X \notin scope[\phi_A]$时可以交换和-积顺序：

$$\sum_X \phi_A \phi_B = \phi_A \sum_X \phi_B$$

不妨设$\mathscr{X}/\mathscr{X}' = X_1, X_2, ...X_N$：

$$\begin{aligned} p(\mathscr{X}') &= \sum_{\mathscr{X}/\mathscr{X}'} \prod_{j=1}^{M} \phi_j(scope[\phi_j]) \\ &= \sum_{X_1}\sum_{X_2}...\sum_{X_{N-1}}\sum_{X_N}\prod_{j=1}^{M}\phi_j(scope[\phi_j]) \\ &= \sum_{X_1}\sum_{X_2}...\sum_{X_{N-1}}\prod_{l}\phi_l(scope[\phi_l])\sum_{X_N}\prod_{k}\phi_k(scope[\phi_k]) \\ &= \sum_{X_1}\sum_{X_2}...\sum_{X_{N-1}}\prod_{l}\phi_l(scope[\phi_l])\psi(scope[\psi]) \end{aligned}$$

在上式第三个等式中使用了交换和-积顺序的法则，其中$X_N \in scope[\phi_l]$而且$X_N \notin scope[\phi_k]$，第四个等式中遍历X_N的所有可能情况将该变量消去，形成新的因子ψ，易知：

$$scope[\psi] = \left\{ \bigcup_k scope[\phi_k] \right\} / X_N$$

如此递归地消除非$\mathscr{X}'$的变量比直接遍历$\mathscr{X}/\mathscr{X}'$并求和的方法可能节约指数次数的加法和查表运算。该算法称作贝叶斯网络推理的**和-积算法，Sum-Product algorithm**。

同样可以在给定证据时进行推理：

$$p(\mathscr{X}'|\mathscr{X}'' = e) = \frac{p(\mathscr{X}', \mathscr{X}'' = e)}{p(\mathscr{X}'' = e)}$$

只需执行两次和-积算法求出$p(\mathscr{X}', \mathscr{X}'')$和$p(\mathscr{X}'')$即可。（另一种方法是利用证据$\mathscr{X}'' = e$预先约简贝叶斯网络的CPD表，再进行和-积推理，这等价于在一个新的贝叶斯网络(G, θ_e)中执行一般的推理）

考虑一个表达症状、病因等变量的的贝叶斯网络，在已经完成了因果关系的搭建以及CPD参数的估计以后，在该网络上执行推理就可以得出某种病症的先验概率，也可以机械地在给定症状时推断患者身患各种疾病的概率。

在含有隐元的模型中，也可以对于其他参数和显元边缘化来推理隐元的分布，当隐元有实际语义时，这一推理是有意义的。一个例子就是语音识别中隐含马尔可夫模型的隐元推断，它包含了夹杂噪音的语音信号中的语义成分（不过严格来讲，用于语音识别的隐含马尔可夫的推理并不是精确贝叶斯推理，而是最大似然推理）。

在一般的机器学习模型中，可以将所有的模型参数做成一个随机变量，则在训练完网络参数以后，对除参数以外的所有变量边缘化来推理模型参数的分布，这是贝叶斯估计的本质。如果按照“非机器学习”的思路把所有模型参数设置成一个固定的值，则它们在执行完参数估计后就已经确定，但在贝叶斯统计中理论上还需要多执行一次边缘化。

4.2 关于推理的其他主题

贝叶斯网络的推理还有很多有关的主题，譬如将推理视作一种向特定图结构投影的**变分推理，Variational Inference**，工作在由一个贝叶斯网络基本结构生成的高层结构上的**置信传播，Belief Propagation**，**置信更新，Belief Update**等等。本文不对于这些算法进行介绍的理由是：它们主要用于克服图结构不确定的困难，而本文在一开始就暂且将讨论范围限制在因果关系确定，即图结构确定的贝叶斯网络上。

其他还有基于采样的推理方法，在给定(G, θ)时进行采样是很直接的，从：

$$p(\mathscr{X}) = \prod_{i=1}^{|\mathscr{X}|} p(X_i|Pa(X_i))$$

只需要找到一种i的遍历顺序，使得进行每一个X_i的采样时，其父节点$Pa(X_i)$均已被采样即可。换言之，i的遍历顺序需要是有向图G中的一个拓扑序，这个顺序可以通过在G上运行深度优先搜索得到（可见《算法导论》相关章节）。