4.4 The Laplace Approximation

在贝叶斯方法中，我们经常会求积分。例如对于新数据x，我们求它的预测值的预测分布，我们将要求$p(t|x,w)$和$p(w|X)$对w的积分。但积分在很多情况下会很难求，所以我们想到采用近似算法求它的近似解。

$$p(t|x,X)=\int p(t|x,w)p(w|X)dw$$

近似算法从大的层面看，可以分为解析式近似(*analytical approximations*)和数据采样式近似(*numberical sampling*)两大类别。

解析式近似：是一种确定性近似方法，典型代表：变分法、EM算法
数值采样式近似：是一种随机性近似方法，典型代表：MCMC

在本章中，我们主要介绍拉普拉斯近似(*The Laplace Approximation*)，即用一个高斯分布去近似另一个分布。

拉普拉斯近似（*The Laplace Approximation*)

推导过程

假设有

$$p(z)=\frac{1}{Z}f(z),Z=\int f(z)$$

设$q(z)$为高斯分布，用$q(z)$近似$p(z)$，要求近似分布$q(z)$和目标分布$p(z)$拥有相同的众数，即$q(z)$和$p(z)$在同一点取值最大。

$$\arg\max_z q(z)=\arg\max_z p(z)=z_0$$

or

$$q'(z_0)=p'(z_0)=f'(z_0)=0$$

对$f(x)$取对数后在x_0处二阶泰勒展开

$$lnf(z)\approx lnf(z_0)+lnf(z_0)'(z-z_0)+\frac{lnf(z_0)''}{z_i}(z-z_0)^2$$

有

$$(lnf(z_0))'=\frac{f'(z_0)}{f(z_0)}=0$$

令

$$A=-\frac{d^2}{dz^2}lnf(z)|_{z=z_0}=-(lnf(z_0))''$$

所以有

$$lnf(z) \approx lnf(z_0) - \frac{A}{2}(z - z_0)^2$$
$$e^{lnf(z)} \approx e^{lnf(z_0) - \frac{A}{2}(z-z_0)^2}$$
$$f(z) = f(z_0)e^{-\frac{A}{2}(z-z_0)^2}$$

设高斯分布$N(z|z_0, A^{-1})$，则有

$$\frac{1}{\sqrt{2\pi}}\sqrt{A}\int e^{-\frac{1}{2}(z-z_0)^2 A}dz = 1$$

即

$$\int e^{-\frac{1}{2}(z-z_0)^2 A}dz = \sqrt{\frac{2\pi}{A}}$$

所以

$$Z = \int f(z)dz$$
$$= f(z_0)\sqrt{\frac{2\pi}{A}}$$

所以有

$$p(z) = \frac{1}{Z}f(z)$$
$$\approx \sqrt{\frac{A}{2\pi}}\frac{f(z_0)e^{-\frac{A}{2}(z-z_0)^2}}{f(z_0)}$$
$$= \sqrt{\frac{A}{2\pi}}e^{-\frac{A}{2}(z-z_0)^2}$$

page215

where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. This Gaussian distribution will be well defined provided its precision matrix, given by $\mathbf{A}$, is positive definite, which implies that the stationary point $\mathbf{z}_0$ must be a local maximum, not a minimum or a saddle point.

【这个高斯分布将被很好地定义，前提是它的精度矩阵A是正定的，这意味着平稳点z0必须是局部最大值，而不是最小值或鞍点。】

一般步骤

- 找到极值点z_0
- 求出二阶导在z_0处的取值 并不需要求常数Z

应用条件

因为拉普拉斯近似较为简单，其应用也受到了一定的限制。若想要拉普拉斯近似有较好的效果，目标分布至少需要满足以下两个性质：

- 因为高斯分布是单峰的，我们所求目标分布最好也是单峰的
- 目标分布关于峰值呈轴对称

若目标分布是多峰形态，我们可以对每个峰分别做拉普拉斯近似，再将多个拉普拉斯近似组合成一个混

page216

bution does not need to be known in order to apply the Laplace method. As a result of the central limit theorem, the posterior distribution for a model is expected to become increasingly better approximated by a Gaussian as the number of observed data points is increased, and so we would expect the Laplace approximation to be most useful in situations where the number of data points is relatively large.

【作为中心极限定理的结果，随着观测数据点数量的增加，模型的后验分布预计将越来越接近高斯分布，因此我们预计拉普拉斯近似在数据点数量相对较大的情况下最有用。】

中心极限定理：

- *一般中心极限定理*
  - 即林德伯格—列维(Lindberg-Levy)定理，也叫独立同分布随机变量序列的中心极限定理。
  - 对于任意一个变量x,其N项和或者N项和的均值服从高斯分布。
- *局部中心极限定理*
  - 即棣莫佛—拉普拉斯(de Moivre-Laplace)定理。
  - 设m_A表示n次独立重复实验中事件A的发生次数，P是事件A每次发生的概率，则对于任意的(a,b]恒有

$$\lim_{x\to+\infty} P\{a < \frac{m_A - nP}{\sqrt{nP(1-P)}} \leq b\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

缺点

page216

One major weakness of the Laplace approximation is that, since it is based on a Gaussian distribution, it is only directly applicable to real variables. In other cases it may be possible to apply the Laplace approximation to a transformation of the variable. For instance if $0 \leqslant \tau < \infty$ then we can consider a Laplace approximation of $\ln \tau$. The most serious limitation of the Laplace framework, however, is that it is based purely on the aspects of the true distribution at a specific value of the variable, and so can fail to capture important global properties. In Chapter 10 we shall consider alternative approaches which adopt a more global perspective.

高斯分布的变量x的取值是从负无穷到正无穷，当目标分布的取值不在这个范围内，我们可以对目标分布的变量进行变换，再用拉普拉斯近似。例如，当目标分布的变量τ取值范围为$0 \leq \tau < \infty$，我们可以将τ变为$ln\tau$，再对$ln\tau$取高斯分布。

Model comparison and BIC

令$f(\theta) = p(D|\theta)p(\theta)$

根据拉普拉斯近似有

$$
\begin{aligned}
z &= \int f(\theta)d\theta = \int p(D|\theta)p(\theta)d\theta \\
&\approx f(\theta_{MAP}) \int e^{-\frac{1}{2}(\theta-\theta_{MAP})^T A(\theta-\theta_{MAP})}d\theta \\
&= p(D|\theta_{MAP})P(\theta_{MAP})\frac{(2\pi)^{\frac{M}{2}}}{|A|^{\frac{1}{2}}}
\end{aligned}
$$

所以有

$$
lnp(D) = lnp(z) \approx lnp(D|\theta_{MAP}) + \underbrace{lnp(\theta_{MAP}) + \frac{M}{2}ln2\pi - \frac{1}{2}ln|A|}_{\text{奥卡姆因子}(Occam\,factor)}
$$

因为有

$$
\begin{aligned}
lnp(\theta_{MAP}|D) &= ln\frac{p(D|\theta_{MAP})p(\theta_{MAP})}{p(D)} \\
&= ln\{p(D|\theta_{MAP})p(\theta_{MAP})\} - lnp(D)
\end{aligned}
$$

所以令

$$
\begin{aligned}
A &= -\nabla\nabla ln\{p(D|\theta_{MAP})p(\theta_{MAP})\} = -\nabla\nabla lnp(\theta_{MAP}|D) \\
&= -\nabla\nabla lnp(D|\theta_{MAP}) - \nabla\nabla lnp(\theta_{MAP})
\end{aligned}
$$

记

$$
H = -\nabla\nabla lnp(D|\theta_{MAP})
$$

对于先验,设

$$
p(\theta) = N(\theta|m, V_0)
$$

所以

$$
lnp(\theta) = ln\{\frac{1}{(2\pi)^{\frac{M}{2}}}\frac{1}{(|V_0|)^{\frac{1}{2}}}\} + (-\frac{1}{2}(\theta-m)^T V_0(\theta-m))
$$

所以

$$
\begin{aligned}
\nabla_\theta lnp(\theta) &= -{V_0}^{-1}(\theta-m) \\
\nabla\nabla_\theta lnp(\theta) &= -{V_0}^{-1}
\end{aligned}
$$

所以

$$
A = H + -{V_0}^{-1}
$$

将$p(\theta)$带入$lnp(D)$有

$$
\begin{aligned}
lnp(D) &\approx lnp(D|\theta_{MAP}) - \frac{M}{2}ln2\pi - \frac{1}{2}ln|V_0| - \frac{1}{2}(\theta_{MAP}-m)^T{V_0}^{-1}(\theta_{MAP}-m) + \frac{M}{2}ln2\pi - \frac{1}{2}ln \\
&= lnp(D|\theta_{MAP}) - \frac{1}{2}(\theta_{MAP}-m)^T{V_0}^{-1}(\theta_{MAP}-m) - \frac{1}{2}ln|V_0||A| \\
&= lnp(D|\theta_{MAP}) - \frac{1}{2}(\theta_{MAP}-m)^T{V_0}^{-1}(\theta_{MAP}-m) - \frac{1}{2}ln|V_0A| \\
&= lnp(D|\theta_{MAP}) - \frac{1}{2}(\theta_{MAP}-m)^T{V_0}^{-1}(\theta_{MAP}-m) - \frac{1}{2}ln|V_0H+I| \\
&\approx lnp(D|\theta_{MAP}) - \frac{1}{2}(\theta_{MAP}-m)^T{V_0}^{-1}(\theta_{MAP}-m) - \frac{1}{2}ln|V_0| - \frac{1}{2}ln|H| \\
&= lnp(D|\theta_{MAP}) - \frac{1}{2}(\theta_{MAP}-m)^T{V_0}^{-1}(\theta_{MAP}-m) - \frac{1}{2}ln|H| + C
\end{aligned}
$$

因为有

$$
\begin{aligned}
H &= -\nabla\nabla lnp(D|\theta_{MAP}) \\
&= -\nabla\nabla ln\{p(x_1|\theta_{MAP})p(x_2|\theta_{MAP})\cdots p(x_N|\theta_{MAP})\}
\end{aligned}
$$

所以

$$
H = -\sum_{i=1}^{N}\nabla\nabla lnp(x:|\theta_{MAP}) = -\sum_{i=1}^{N}H_i = -N\hat{H}, \hat{H} = \frac{1}{N}\sum_{i=1}^{N}H_i
$$

$$
-\frac{1}{2}ln|H| = -\frac{1}{2}ln|N\hat{H}| = -\frac{M}{2}lnN - \frac{1}{2}ln|\hat{H}|
$$

当$N \gg 1$时，$ln|\hat{H}|$会非常小，可以忽略不计。所以

$$
lnp(D) \approx lnp(D|\theta_{MAP}) - \frac{1}{2}(\theta_{MAP} - m)^T V_0^{-1}(\theta_{MAP} - m) - \frac{M}{2}lnN + C
$$

$lnp(D) \approx lnp(D|\theta_{MAP}) - \frac{M}{2}lnN$就是我们要介绍的BIC(贝叶斯信息准则*Bayesian Information Criterion*)。从公式中不难看出，BIC既和模型的复杂程度有关(M)，也和数据集的大小有关(N)，是一个简单的带正则化项的误差函数。