2. 確率とビリーフ(2)

植野真臣
電気通信大学
大学院情報システム学研究科

16. ベイズ原理

定義15 (事後分布)
X=($X_1, \cdots, X_m$)が独立同一分布$f(x|\theta)$ に従うn 個の確率変数とする. n 個の確率変数に対応したデータ$\boldsymbol{x} = (x_1, \cdots, x_n)$ が得られたとき,

$$p(\theta|x) = \frac{p(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int_\Theta p(\theta)\prod_{i=1}^{n} f(x_i|\theta)\, d\theta}$$

を**事後分布**(posterior distribution)と呼び,
$p(\theta)$ を**事前分布**(prior distribution)と呼ぶ.

事後分布最大化推定量

定義16 (MAP推定値)
データx を所与として, 以下の事後分布最大となるパラメータを求めるとき,

$$\hat{\theta} = \arg max\{p(\theta): \theta \in C\}$$

$\hat{\theta}$をベイズ推定値(Bayesian estimator)または,
事後分布最大化推定値(maximum a posterior estimator, **MAP** 推定値)と呼ぶ.

Note: ベイズ推定は, すべての確率空間で成り立つわけではない. パラメータの事前確率が確率の公理を満たすときにのみ成立する.

ベイズ推定の一致性

定義17 (EAP 推定値)
データ$\boldsymbol{x}$ を所与として, 以下の事後分布によるパラメータの期待値を求めるとき,

$$\hat{\theta} = E(\theta\{p(\theta): \theta \in C\})$$

$\hat{\theta}$ を期待事後推定値(expected a posterior estimator , **EAP** 推定値)と呼ぶ.
ベイズ推定値も強一致性をもつ.

定理11 (ベイズ推定の一致性)
ベイズ推定において推定値$\hat{\theta}$が真のパラメータ θ^*の強一致推定値となるような事前分布が設定できる.

また, ベイズ推定値も漸近的正規性をもち, 誤差を計算できる.

定理12 (ベイズ推定の漸近正規性)
事後確率密度関数が正則条件(regular condition)の下で微分可能のとき, ベイズ推定値が漸近分散 $I(\theta^*)^{-1}$をもつ漸近正規推定値となる事前分布を設定できる.

17. 無情報事前分布

1.8.1 ジェフリーズの事前分布
事後分布

$$p(\theta|x) = \frac{p(\theta)\prod_{i=1}^{n} f(x_i|\theta)}{\int_\Theta p(\theta)\prod_{i=1}^{n} f(x_i|\theta)\, d\theta}$$

を求めるための事前分布$p(\theta)$ の設定について, どのように設定するかが問題となる. 通常, データを採取するまで, われわれはデータについての情報をもたない.

そのために, $p(\theta)$ は無知を表す分布でなくてはならない.

このような無知を示す事前分布を**無情報事前分布**(non–informative prior distribution)と呼ぶ. 無知の状態を示す事前分布の選択のルールとして, Jeffreys(1961)は, つぎの二つの提案をしている.

まず, 一つの母数θについて考えると,

1.母数θについて, $\theta \in (-\infty, \infty)$ のみの情報があるとき, 事前分布は一様分布となる.

$$p(\theta) \propto const$$

2. 母数θについて, $\theta \in (0, \infty)$のみの情報があるとき, θの対数が一様であるような事前分布を考える. すなわち, $p(\log\theta) \propto const$であるから, 変数変換すれば,

$$p(\theta) \propto \frac{1}{\theta}$$

ルール1 を選択する場合, 事後分布= 尤度となるが, $\int_{-\infty}^{\infty} p(\theta) \neq 1$となり, 事前分布$p(\theta)$は確率の公理を満たさない. このような事前分布を**improper prior distribution** と呼ぶ.

しかし, このimproper prior distribution は, ベイズ統計学の整合性を壊すという意味で, 議論を招いた. そこで, 閉区間に局所的一様分布を考える**principle of stable estimation**(Edwards et al.1963)が提案されている. 例えば, $\theta \in [a, b]$ であれば, $p(\theta) = \frac{1}{b-a}$となり, $\int_{-\infty}^{\infty} p(\theta) = 1$ と確率の公理を満たす.

また, 確率変数の定義を満たしたところで, この一様分布の事前分布には問題がある.
例えば, $\theta \in [a, b]$ では, $p(\theta) = const$であるが, $\kappa = \theta^{10}$ としても, ジェフリーズのルールに従えば, $p(\kappa) = const$ となる. しかし, 変数変換すれば, そのようにならないことがわかる. このようなことを考慮して, Box and Tiao(1973)は, ある母数ϕの尤度が, データが変わってもその形状は変わらず, その位置のみを変更させるとき, その母数をデータ移動型母数と呼んだ

以下は，データ移動型母数を見出す方法である．対数尤度$l(\phi|x)$ は，最尤推定値$\hat{\phi}$のまわりでテイラー展開すると，

$l(\phi|x)= l(\hat{\phi}|x) - \frac{n}{2}(\phi-\hat{\phi})^2(-\frac{1}{n}\frac{\partial^2 l(\phi|x)}{\partial\phi^2})_{\hat{\phi}}$

いま，x のデータ発生モデルが指数形分布族であることを仮定して$J(\hat{\phi})$ とおく．これは，$(-\frac{1}{n}\frac{\partial l(\phi|x)}{\partial\phi^2})_{\hat{\phi}}$が$\hat{\phi}$ のみの関数を仮定するのと同値である．

θとϕが1 対1 変換であるとき

$$J(\hat{\phi}) = (-\frac{1}{n}\frac{\partial^2 l(\phi|x)}{\partial\phi^2})_{\phi=\hat{\phi}} = \left(-\frac{1}{n}\frac{\partial^2 l(\phi|x)}{\partial\theta^2}\right)_{\theta=\hat{\theta}}\left(\frac{\partial\theta}{\partial\phi}\right)^2_{\theta=\hat{\theta}}$$

$$= J(\hat{\theta})\left(\frac{\partial\theta}{\partial\hat{\phi}}\right)^2_{\theta=\hat{\theta}}$$

このとき，

$$\left|\frac{\partial\theta}{\partial\phi}\right|_{\theta=\hat{\theta}} \propto J^{-1/2}(\hat{\theta})$$

となるように変換ϕを選べば，$J(\hat{\phi})$は定数となり，尤度は$(\phi-\hat{\phi})^2$ の関数となる．すなわち，ϕに関して近似的データ移動型となる．このとき，無情報事前分布は

$$p(\theta) \propto \left|\frac{\partial\theta}{\partial\phi}\right|_{\theta=\hat{\theta}} \propto J^{-1/2}(\hat{\theta})$$

となる．

また，指数型分布の仮定を抜いた場合，

$$\left|\frac{\partial\theta}{\partial\phi}\right|_{\theta=\widehat{\theta}} \propto I^{-1/2}(\theta)$$

となる．$I(\theta)$ はフィッシャー情報量を示す．
すなわち，母数θの事前分布は，フィッシャー情報量$I(\theta)$に比例させるというルールである．これが，ジェフリーズが提唱した母数の変換の不変性から導いた分布に一致するので，

ジェフリーズの事前分布と呼ばれる．

データ情報最大化事前分布

データ情報最大化事前分布Zellner（1971）は，
データのもつ情報と比較して，事前情報のもつ情報を最小にするような分布を無情報事前分布としている．
情報を情報理論の枠組みで定義すると，事前分布における情報量と事後分布における情報量との差として伝達情報量で定義できる．
すなわち，

$$G = -\int_{\Theta} p(\theta)\log p(\theta)d\theta + \int_{\Theta}\int_{x} p(\theta|x)\log p(\theta|x)\,p(x)dx\,p(\theta)d\theta$$

を最大化させる事前分布を，データ情報最大化事前分布（maximum data information distribution）と呼ぶ．

自然共役事前分布

ベイズ統計の中で最も一般的で，ベイズ的な有効性を発揮できると考えられるのが，この自然共役事前分布である．

これまでの事前分布では，データを得る前の事前分布とデータを得た後の事後分布は，分布の形状が変化する．

しかし，データの有無にかかわらず，分布の形状は同一のほうが自然であろう．

そこで，事前分布と事後分布が同一の分布族に属するとき，その事前分布を自然共役事前分布（natural conjugate prior distribution）と呼ぶ．

ここでは，特にこの自然共役事前分布を中心にベイズ的推論を行うようにする．

自然共役事前分布を用いた推定例

例7 (二項分布)

$$f(x|\theta,n) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

コインを投げてn 回中x 回表が出たときの確率θをベイズ推定しよう．

尤度関数は，$\binom{n}{x}\theta^x(1-\theta)^{n-x}$であり，

二項分布の自然共役事前分布は，以下のベータ分布（$Beta(\alpha,\beta)$）である．

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

事後分布は，

$$p(\theta|n,x,\alpha,\beta) = \frac{\Gamma(\mathrm{n}+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(\mathrm{n}-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

とやはりベータ分布となる．対数をとり，以下の対数事後分布を最大化すればよい．

$$\log p(\theta|n,x,\alpha,\beta) = \log\frac{\Gamma(\mathrm{n}+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(\mathrm{n}-x+\beta)} + (x+\alpha-1)\log\theta + (n-x+\beta-1)\log(1-\theta)$$

以下の対数事後分布を最大化すればよい.

$$\log p(\theta|n,x,\alpha,\beta) = \log\frac{\Gamma(\mathrm{n}+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(\mathrm{n}-x+\beta)} + (x+\alpha-1)\log\theta + (n-x+\beta-1)\log(1-\theta)$$

$\frac{\partial \log p(\theta|n,x,\alpha,\beta)}{\partial\theta} = 0$のとき，対数事後分布は最大となるので，

$$\frac{\partial \log p(\theta|n,x,\alpha,\beta)}{\partial\theta} = \frac{(x+\alpha-1)}{\theta} - \frac{(n-x+\beta-1)}{1-\theta}$$

$$= \frac{x+\alpha-1-x\theta-\alpha\theta+\theta-n\theta+x\theta-\beta\theta+\theta}{\theta(1-\theta)}$$

$$= \frac{x+\alpha-1-(n+\alpha+\beta-2)\theta}{\theta(1-\theta)} = 0$$

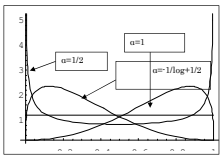
$\theta(1-\theta) \neq 0$とすると

$$\theta = \frac{x+\alpha-1}{n+\alpha+\beta-2}$$

がベイズ推定値となる．さて，α, β は事前分布のパラメータであるが，これをハイパーパラメータ(hyper parameter)と呼ぶ．このハイパーパラメータによって，事前分布はさまざまな形状をとる(図)．
例えば，事前分布が一様となる場合($Beta(1,1)$)の推定値は，

$$\hat{\theta} = \frac{x}{n}$$

となり，最尤解に一致する．



例題

例8 (正規分布)

$$P(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\}$$

$(x_1,\cdots,x_n)$ を得たときのμ,σ^2 を求めよう.

尤度は，

$$L = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

このとき，自然共役事前分布は，

$$p(\mu) = N(\mu_0,\sigma_0^2),$$

$$p(\sigma^2) = \chi^{-2}(\nu_0,\lambda_0),(\text{逆カイ二乗分布})$$

すなわち，事前分布はこれらの積の形で以下のように表される．

$$p(\mu,\sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$$
$$\propto \left(\frac{\sigma^2}{n_0}\right)^{-\frac{1}{2}}\exp\left\{-\frac{n_0(\mu-\mu_0)^2}{2\sigma^2}\right\}(\sigma^2)^{-\frac{1}{2}\nu_0-1}\exp\left(-\frac{\lambda_0}{2\sigma^2}\right)$$
$$= (\sigma^2)^{-\frac{1}{2}(\nu_0+1)-1}\exp\left\{-\frac{\lambda_0+n_0(\mu-\mu_0)^2}{2\sigma^2}\right\}$$

ここで，$n_0,\mu_0,\nu_0,\lambda_0$はハイパーパラメータであり，$n_0 = \nu_0+1$ という関係にある．
一方，これを尤度に掛け合わせて事後分布を導くのであるが，計算の簡便さのために，以下のように尤度を変形させる．

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

ここで指数部分$\exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$を三平方の定理により，推定平均$\bar{x}$を介して，以下のように分解する．

$$\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{2\sigma^2} + \frac{(\bar{x}-\mu)^2}{2\sigma^2}$$

これより，尤度Lは，

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{2\sigma^2}\right\}\exp\left\{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right\}$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{S^2+n(\mu-\bar{x})^2}{2\sigma^2}\right\}$$

ただし，ここで，

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, S^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

と書き換えられる．

さて，この尤度Lと先の事前分布を掛け合わせることによって，以下のような事後分布が得られる．

ここで，$\nu_0 = n_0 - 1$とおいて，

$$p(\mu,\sigma^2|x) \propto L \times p(\mu,\sigma^2)$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{S^2+n(\mu-\bar{x})^2}{2\sigma^2}\right\}$$
$$\times (\sigma^2)^{-\frac{1}{2}(\nu_0+1)-1}\exp\left\{-\frac{\lambda_0+n_0(\mu-\mu_0)^2}{2\sigma^2}\right\}$$
$$\propto (\sigma^2)^{-\frac{1}{2}(n+n_0)-1}\exp\left\{-\frac{\lambda_0+S^2+n_0(\mu-\mu_0)^2+n(\mu-\bar{x})^2}{2\sigma^2}\right\}$$

さらに，指数部分のうち，λ_0+S^2以外の部分に，平方完成を行うと，結局，

$$p(\mu,\sigma^2|x) \propto (\sigma^2)^{-\frac{1}{2}(n+n_0)-1}\exp\left\{-\frac{\lambda_*+(n_0+n)(\mu-\mu_*)^2}{2\sigma^2}\right\}$$

ただし，

$$\lambda_* = \lambda_0 + S^2 + \frac{n_0 n(\bar{x}-\mu_0)^2}{n_0+n}, \mu_* = \frac{n_0\mu_0+n\bar{x}}{n_0+n}$$

となる．
この事後分布もまた，正規分布と逆カイ二乗分布の積となり，

$$N \times \chi^{-2}(n_0+n,\mu_*,\nu_0+n,\lambda_*)$$

と略記する．
さて，これらの事後分布は，μとσ^2の同時事後確率分布であることがわかる．

このように，複数のパラメータを同時に最大化させる場合，つぎのような周辺化（marginalization）を行い，個々のパラメータの分布を導く．

このような分布を周辺事後分布（marginal posterior distribution）と呼ぶ．

すなわち，パラメータμ についての周辺事後分布は以下のようにして求められる．

$$\begin{aligned} p(\mu|x) &= \int_0^\infty p(\mu,\sigma^2|x)p(\sigma^2)d\sigma^2 \\ &\propto \frac{\Gamma\left[\frac{(\nu_*+1)}{2}\right]}{\sqrt{\frac{\nu_*\pi\lambda_*}{n_*}}\Gamma\left(\frac{\nu_*}{2}\right)}\left\{1+\frac{(\mu-\mu_*)^2}{\mu_*}\right\}^{-\frac{1}{2}(\nu_*+1)} \\ &\equiv t(\nu_*,\mu_*,\lambda_*/n_*) \end{aligned}$$

このようにμ の周辺事後分布は，t 分布$t(\nu_*,\mu_*,\lambda_*/n_*)$ に従うことがわかる．
また，パラメータ σ^2についての周辺事後分布も同様にして，以下のように求められる．

$$\begin{aligned} p(\sigma^2|x) &= \int_0^\infty p(\mu,\sigma^2|x)p(\mu)d\mu \\ &\propto \frac{\frac{\nu_*}{\lambda_*^2}}{2^{\frac{\nu_*}{2}}\Gamma\left(\frac{\nu_*}{2}\right)}(\sigma^2)^{-\frac{\nu_*}{2}-1}\exp\left(-\frac{\lambda_*}{2\sigma^2}\right) \\ &\equiv \chi^{-2}(\nu_*,\lambda_*) \end{aligned}$$

となり，σ^2の周辺事後分布は，逆カイ二乗分布$\chi^{-2}(\nu_*,\lambda_*)$ に従うことがわかる．

また，事後確率最大化によるベイズ推定値は，t 分布のモードがμ_*であることより，

μ の推定値は，

$$\hat{\mu} = \frac{n_0\mu_0 + n\bar{x}}{n_0+n}$$

となり，σ^2 のベイズ推定値は，逆カイ二乗分布のモードが$\frac{\lambda_*}{\nu_*-2}$ であることより，σ^2 の推定値は，

$$\widehat{\sigma^2} = \frac{\left\{\lambda_0 + S^2 + \frac{n_0 n(\bar{x}-\mu_0)^2}{n_0+n}\right\}}{\nu_*-2}$$

となる．

18. 予測分布

データやモデルを用いて推論を行う重要な目的の一つに，未知の事象の予測が挙げられる．
この予測問題のためには，最もよく用いられるのは，

$$p(y|\hat{\theta})$$

で示されるplug–in distribution と呼ばれる分布である．しかし，$\hat{\theta}$ は推定値であるためにそのサンプルのとり方によってこの分布は大きく変化する．ベイズ的アプローチでは，この$\hat{\theta}$ のばらつき（$\hat{\theta}$ の事後分布）を考慮し，以下のように予測分布を定義する．

定義18
モデルm から発生されるデータ$\boldsymbol{x}$ により，未知の変数y の分布を予測するとき，以下の分布を予測分布（predictive distribution）と呼ぶ．

$$p(y|\boldsymbol{x},m) = \int_\Theta p(y|\theta,m)p(\theta|\boldsymbol{x},m)d\theta$$

例9 (二項分布) ベータ分布を事前分布とした二項分布の予測分布は，以下のようになる．

$$\begin{aligned} p(y|x) &= \int_\Theta p(y|\theta)p(\theta|x)d\theta \\ &= \int_\Theta \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\mathrm{n}+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\times\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta \\ &\propto \binom{n}{y}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}\frac{\Gamma(x+\alpha)(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \\ &= \frac{n!}{y!\,(n-y)!}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}\frac{\Gamma(x+\alpha)(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \end{aligned}$$

特に，α, β が整数のとき

$$p(y|x) \propto \frac{n!}{y!\,(n-y)!}\frac{y!\,(n-y)!}{(n+1)!}\frac{(x+\alpha-1)!\,(n-x+\beta-1)!}{(n+\alpha+\beta-1)!}$$

例10 (正規分布) 事前分布を$N(\mu,\sigma^2)$ 分布

$$\begin{aligned} p(\mu,\sigma^2) &= p(\mu|\sigma^2)p(\sigma^2) \\ &\propto \left(\frac{\sigma^2}{n_0}\right)^{-\frac{1}{2}}\exp\left\{-\frac{n_0(\mu-\mu_0)^2}{2\sigma^2}\right\}(\sigma^2)^{-\frac{1}{2}\nu_0-1}\exp\left(-\frac{\lambda_0}{2\sigma^2}\right) \\ &= (\sigma^2)^{-\frac{1}{2}(\nu_0+1)-1}\exp\left\{-\frac{\lambda_0+n_0(\mu-\mu_0)^2}{2\sigma^2}\right\} \end{aligned}$$

とすると，事後分布は

$$p(\mu,\sigma^2|x) \propto (\sigma^2)^{-\frac{1}{2}(n+n_0)-1}\exp\left\{-\frac{\lambda_* + (n_0+n)(\mu-\mu_*)^2}{2\sigma^2}\right\}$$

である．ただし，

$$\lambda_* = \lambda_0 + S^2 + \frac{n_0 n(\bar{x}-\mu_0)^2}{n_0+n}, \mu_* = \frac{n_0\mu_0+n\bar{x}}{n_0+n}$$

となる．予測分布は

$$p(x_{n+1}|x) = \int\int p(x_{n+1}|\mu,\sigma^2)p(\mu,\sigma^2|x_1,\cdots,x_n)d\mu d\sigma^2$$

ここで，

$$p(x_{n+1}|\mu,\sigma^2) \propto (\sigma^2)^{-1}\exp\left\{-\frac{(x_{n+1}-\mu)^2}{2\sigma^2}\right\}$$

より，

$$\begin{aligned} p(x_{n+1}|\boldsymbol{x}) &= \int\int p(x_{n+1}|\mu,\sigma^2)p(\mu,\sigma^2|x_1,\cdots,x_n)d\mu d\sigma^2 \\ &\propto \int\int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{(x_{n+1}-\mu)^2+S^2+n(\mu-\bar{x})^2}{2\sigma^2}\right]d\mu d\sigma^2 \\ &= \int\int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{1}{2\sigma^2}\left\{(n+1)(\mu-\bar{\mu})^2+S^2 \right.\right. \\ &\quad \left.\left. +\frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}\right]d\mu d\sigma^2 \\ &\propto \int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{1}{2\sigma^2}\left\{S^2+\frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}\right]d\sigma^2 \\ &\propto \left\{S^2+\frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}^{-\frac{\nu+1}{2}} \end{aligned}$$

$$\propto \left[1 + \left\{\frac{x_{n+1} - \bar{x}}{\sqrt{\frac{n+1}{n\nu}S^2}}\right\}^2 / \nu\right]^{-\frac{\nu+1}{2}}$$

ただし,ここで

$$\bar{\mu} = \frac{n\bar{x} + x_{n+1}}{n+1}$$

ここで,

$$t = \frac{x_{n+1} - \bar{x}}{\sqrt{\frac{n+1}{n\nu}S^2}}$$

とおくとき,tは自由度νのt分布に従う.

1.8. データから統計モデルを選択

統計モデルのパラメータ(母数)をデータから推定するには，尤度最大化により漸近的一致性が得られた.

ひとつのデータに対して，複数のモデルから度のモデルが一番よいかを決定するときに，尤度最大化は使えるのであろうか？

答えはNOである. 尤度は「モデルのデータへのあてはまり」を示しており，モデルはパラメータ数を多くすればするほどあてはまりがよくなるので結果として，複雑なモデルを選ぶだけである. これをオーバーフィッティングと呼ぶ.

パラメータ推定に対して，モデル選択は一つ上の階層の学習であり，このとき推定されるパラメータを周辺化した周辺尤度がモデル選択に用いられる.

1.9 周辺尤度

モデルの候補が複数ある場合に,

データx からモデルmを選択することをモデル選択(model selection)と呼ぶ.

ベイズ統計では，一般的に，モデル選択のために以下の周辺尤度を最大にするモデルを選択する.

定義19
データxを所与としたモデルmの尤度を周辺化して周辺尤度(marginal likelihood)，**ML**と呼ぶ.

$$p(x|m) = \int_{\Theta} p(x|\theta, m)p(\theta|m)d\theta$$

ベイジアンネットワークの構造を学習するために，周辺尤度を最大にする構造を選択すればよい.

1.10 予測分布情報量基準

m^*を真のモデル,xをデータ，x_{N+1}を予測データとする。

$\sum_{x_{N+1}} p(x_{N+1}|x, m^*) \log \frac{p(x_{N+1}|x, m)}{p(x_{N+1}|x, m^*)}$

$=\sum_{x_{N+1}} p(x_{N+1}|x, m^*) \log p(x_{N+1}|x, m) -$

$\sum_{x_{N+1}} p(x_{N+1}|x, m^*) \log p(x_{N+1}|x, m^*)$

ここで

$\sum_{x_{N+1}} p(x_{N+1}|x, m^*) \log p(x_{N+1}|x, m^*)$

は定数なので

$\sum_{x_{N+1}} p(x_{N+1}|x, m^*) \log p(x_{N+1}|x, m)$

を最大化するmを求めればよい。

ただし、$p(x_{N+1}|x, m^*) \approx \sum_m p(x_{N+1}|x, m)p(m|x)$

周辺尤度
データ数が大きい時のみに一致性がある。

予測分布情報量基準
データ数が少ない時にもよく予測する。