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.

注意:ベイズでのΘの扱い

尤度では、Θは確率変数ではない

事前分布が確率法則に従うのであれば、Θは確率変数となる

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

ベイズの推定での利点

ベイズでは、厳密な確率推論がパラメータ推定にも適用できる。

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.

EAP 推定値

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.

ベイズ推定の一致性

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.

無情報事前分布のルール

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.

Proper prior

このimproper prior distribution は, ベイズ統計学の整合性を壊す.
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.

Jefferys prior (Box and Tiao 1973)

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, $l(\phi|x) = l(\hat{\phi}|\boldsymbol{x}) - \frac{n}{2}(\phi - \hat{\phi})^2(-\frac{1}{n}\frac{\partial^2 l(\phi|x)}{\partial\phi^2})_{\hat{\phi}}$
θとϕが1 対1変数 変換であるとき

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

$$= J(\hat{\theta})\left(\frac{\partial\theta}{\partial\phi}\right)^2_{\theta=\hat{\theta}}. \quad \text{このとき,} \quad \left|\frac{\partial\theta}{\partial\phi}\right|_{\theta=\hat{\theta}} \propto J^{-1/2}(\hat{\theta})$$

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. ϕに関して近似的データ移動型(データに対しても尤度の形は同じまま移動する). このとき、無情報事前分布は $p(\theta) \propto \left|\frac{\partial\theta}{\partial\phi}\right|_{\theta=\hat{\theta}} \propto J^{-1/2}(\hat{\theta})$

Jefferys prior (Box and Tiao 1973)

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, $l(\phi|x) = l(\hat{\phi}|\boldsymbol{x}) - \frac{n}{2}(\phi - \hat{\phi})^2(-\frac{1}{n}\frac{\partial^2 l(\phi|x)}{\partial\phi^2})_{\hat{\phi}}$ (定数)
θとϕが1 対1変数 変換であるとき

$$J(\hat{\phi}) = J(\hat{\theta})\left(\frac{\partial\theta}{\partial\phi}\right)^2_{\theta=\hat{\theta}}. \quad \text{このとき,} \quad \left|\frac{\partial\theta}{\partial\phi}\right|_{\theta=\hat{\theta}} \propto J^{-1/2}(\hat{\theta})$$

となるように変換ϕを選べば,

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

となり、形状は$l(\hat{\phi}|\boldsymbol{x})$のまま移動させる ϕに関して近似的データ移動型となる.

一般には

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

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（最も一般的！！）

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．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

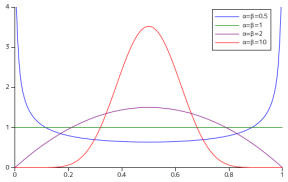
$$\hat{\theta} = \frac{x+\alpha-1}{n+\alpha+\beta-2}$$

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．



EAP推定量

$$\hat{\theta} = \frac{x+\alpha}{n+\alpha+\beta}$$

となり、例えば，事前分布が一様となる場合($Beta(1, 1)$)の推定値は

$$\hat{\theta} = \frac{x+1}{n+2}$$

データがない場合は、$\hat{\theta} = \frac{1}{2}$となり、データが増えるごとに真値に近づく。

EAP推定量でジェフリーズ事前分布

$$\hat{\theta} = \frac{x+\alpha}{n+\alpha+\beta}$$

となり、例えば，事前分布が一様となる場合($Beta(1, 1)$)の推定値は

$$\hat{\theta} = \frac{x+1/2}{n+1}$$

データがない場合は、一様分布同様に$\hat{\theta} = \frac{1}{2}$となるが、一様分布よりもデータに速く影響を受ける。

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，$\sigma_0^2 = \frac{\sigma^2}{n_0}$とする。

$$p(\mu) = N(\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \propto \left(\frac{\sigma^2}{n_0}\right)^{-\frac{1}{2}}\exp\left\{-\frac{n_0(\mu-\mu_0)^2}{2\sigma^2}\right\}$$

$$p(\sigma^2) = Ig(\nu_0, \lambda_0) = \frac{(\lambda_0/2)^{\frac{1}{2}\nu_0}}{\Gamma(\frac{1}{2}\nu_0)}(\sigma^2)^{-\frac{1}{2}\nu_0-1}\exp\left(-\frac{\lambda_0}{2\sigma^2}\right) \quad (\text{逆ガンマ分布})$$

$$\propto (\sigma^2)^{-\frac{1}{2}\nu_0-1}\exp\left(-\frac{\lambda_0}{2\sigma^2}\right)$$

事前分布はこれらの積の形で以下のように表される．自由度$\nu_0 = n_0 - 1$ とすると

$$p(\mu, \sigma^2) = p(\mu|\mu_0, \sigma_0^2)p(\sigma^2|\nu_0, \lambda_0)$$

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

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

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

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．

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

$\nu_0 = n_0 - 1$より

$$\propto (\sigma^2)^{-\frac{1}{2}(n+n_0)-1}\exp\left\{-\frac{\lambda_0+S^2+n_0(\mu-\mu_0)^2+n(\mu-\bar{x})^2}{2\sigma^2}\right\}$$

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，

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

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

この事後分布もまた，正規分布と逆ガンマ分布の積となり，

$$N\times IG(n_0+n, \mu_*, \nu_0+n, \lambda_*)$$

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．

μの周辺事後分布

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．

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

μ の周辺事後分布は，t 分布$t(\nu_*, \mu_*, \lambda_*/n_*)$ に従う．

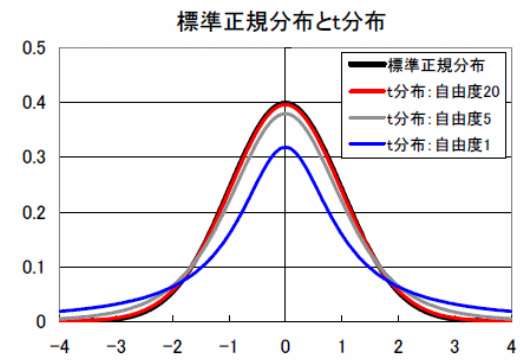
MAP推定値

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MAP推定値は，

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

正規分布とt分布



σ^2の周辺事後分布

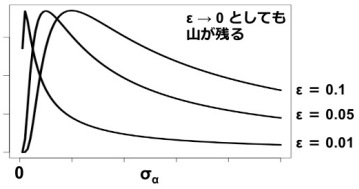
σ^2についての周辺事後分布は

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

となり，σ^2の周辺事後分布は，逆ガンマ分布 $IG(\nu_*/2, \lambda_*/2)$ に従うことがわかる．

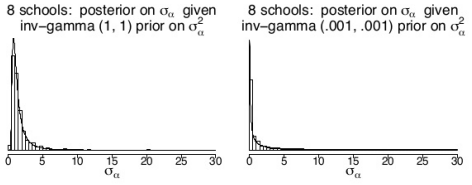
逆ガンマ分布

- $\sigma_\alpha \sim$ InvGamma(ε, ε)



8-schools 逆ガンマ分布

- 左： ε = 1, 右： ε = 0.001
- ε によって事後分布が大きく異なる



37

MAP推定値

σ^2のベイズ推定値は，逆ガンマ分布のモードが$\frac{\lambda_{*/2}}{\nu_{*/2}+1} = \frac{\lambda_*}{\nu_*+2}$であることより，$\sigma^2$ のMAP推定値は，

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

EAP推定値

μ のEAP推定値は，平均値とモードが同一なので

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

σ^2 のMAP推定値は，逆ガンマ分布のモードが$\frac{\lambda_{*/2}}{\nu_{*/2}-1} = \frac{\lambda_*}{\nu_*-2}$であることより，

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

EAPの分散のほうがMAPの分散が大きく推定される。-→ データの分散が大きいことはデータの特徴をよく理解できることを示し、よい推定量であることを示す。

注意

データの分散を大きくすることはよい

しかし、

推定値の分散を大きくすることは　推定値の悪さを示しているので良くないことに注意。

例題

以下のどちらのかけを選ぶと得か？

1. 50個の赤玉と50個の白玉が入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。

2. 赤玉と白玉が合わせて100個入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。

赤玉の出る確率は

1. 50個の赤玉と50個の白玉が入った壺から一つ玉を取り出し，それが赤玉(A)の確率

$$P(A) = \frac{50}{50+50}$$

2. 赤玉と白玉が合わせて100個入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。

$P(A) = \psi$とする。

$$E\big(P(\psi)\big) = \int_0^1 P(\psi)d\psi = \frac{1}{2}$$

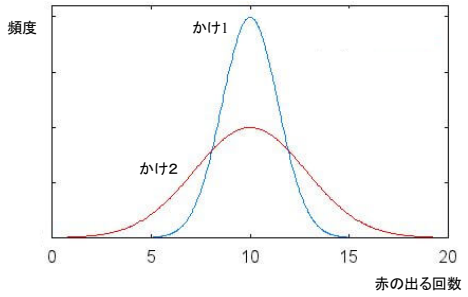
例題

以下のどちらのかけを選ぶと得か？

1. 50個の赤玉と50個の白玉が入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。これを100回繰り返す。

2. 赤玉と白玉が合わせて100個入った壺から一つ玉を取り出し，それが赤玉であったら1万円もらえる。白玉であったら1万円支払う。これを100回繰り返す。

賭け1は博打性大



18. 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，複数のモデルからど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？

→

モデル選択基準

AIC (1973)

Akaike Information Criterion

$$AIC = -2\mathrm{E}[\ln L] = -2\ln L + 2k$$

ここで，$\ln L$は対数最大尤度、kはモデルのパラメータ数

Akaike, H., "Information theory and an extension of the maximum likelihood principle", *Proceedings of the 2nd International Symposium on Information Theory*, Petrov, B. N., and Caski, F. (eds.), Akadimiai Kiado, Budapest: 267-281 (1973).

AICは一致性を持たない

尤度はモデルを複雑にするといくらでも大きくなってしまう。そこでその平均を考えるとモデルの複雑さ(パラメータ数)をペナルティとして考えないといけないことがわかる。

しかし、AICはデータ数を増やしても真のモデルを選択する確率が1.0に収束しない。

ベイズではモデルの確率を考える

m:モデル，M:モデル候補集合，x:データ

$$p(m|x) = \frac{p(x|m)p(m)}{\sum_{i=1}^{M} p(x|m_i)p(m_i)}$$

今、すべての$p(m)$が同一だと考えると

$p(x|m)$ が最大となるモデルを選択すればよい。

ここで

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

を周辺尤度と呼ぶ。

19 周辺尤度

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.

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

BIC(Schwarz 1978)

周辺尤度は、モデルごとにパラメータ空間を積分消去しなければならない。より、簡単に用いるために周辺尤度の漸近近似としてBICが求められた。これは漸近一致性を持つ。

$$\mathrm{BIC} = \ln(L) - \frac{1}{2}k\ln(n)$$

ここで，$\ln L$は対数最大尤度、kはモデルのパラメータ数，nはデータ数.

Schwarz, Gideon E. (1978), "Estimating the dimension of a model", *Annals of Statistics*, **6** (2): 461–464

20. 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. $p(y|x) = \int_\Theta p(y|\theta)p(\theta|x)d\theta$

$$= \int_\Theta \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\mathrm{n}+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \times\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta$$

$$\propto \binom{n}{y}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}\frac{\Gamma(x+\alpha)(n-x+\beta)}{\Gamma(n+\alpha+\beta)}$$

$$= \frac{n!}{y!\,(n-y)!}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}\frac{\Gamma(x+\alpha)(n-x+\beta)}{\Gamma(n+\alpha+\beta)}$$

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

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$$

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

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

とすると，事後分布は

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

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

予測分布は

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

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

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

$$\propto \int\int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{(x_{n+1}-\mu)^2+S^2+n(\mu-\bar{x})^2}{2\sigma^2}\right]d\mu d\sigma^2$$

$$= \int\int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{1}{2\sigma^2}\left\{(n+1)(\mu-\bar{\mu})^2+S^2 + \frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}\right]d\mu d\sigma^2$$

$$\propto \int (\sigma^2)^{-\frac{\nu+1}{2}-2}\exp\left[-\frac{1}{2\sigma^2}\left\{S^2+\frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}\right]d\sigma^2$$

$$\propto \left\{S^2+\frac{n}{n+1}(x_{n+1}-\bar{x})^2\right\}^{-\frac{\nu+1}{2}}$$

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

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.

21. (従来手法)統計的仮説検定

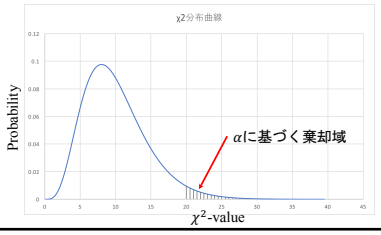
- ある仮説が正しいかどうかを標本（データ）から判定する手法.
- 統計的仮説（Statistical Hypothesis）：
  - 帰無仮説（null hypothesis）：棄却されることを前提とした仮説を表し H_0とする.
  - 対立仮説（alternative hypothesis）：帰無仮説が棄却されたときの採用される仮説を表し H_1とする.
- 有意水準α：ユーザが設定する帰無仮説を棄却する基準であり，誤って帰無仮説を棄却してしまう確率を表す.

仮説検定の手順

1. 帰無仮説H_0，対立仮説H_1を決める.
2. 得られたデータから統計量を求める.
   - 用いる統計量：T（T分布），F（F分布），χ^2（χ^2分布）
3. 用いる統計量が確率分布にどれだけ従っているかを表す確率p値を求める．p値は帰無仮説が正しい確率とも言われ，有意水準αより小さければ，帰無仮説H_0は棄却し対立仮説H_1を採用する.

独立性検定

- 帰無仮説：2変数間が独立
- 対立仮説：2変数間が従属
- 一般的にχ^2 統計量を用いて自由度dfのχ^2分布との適合度により独立性を検定する
- 例：自由度10のχ^2検定



- 検定方法：
  p値 $<\alpha$ → 従属
  p値 $>\alpha$ → 独立

仮説検定法の問題点

- 検定の精度：p値と有意水準αに依存する.

これによって引き起こされる問題

- 真に帰無仮説が正しいが，誤って棄却してしまう.
  →第一種の過誤（Type Ⅰ error）と呼ばれる.
- 真に対立仮説が正しいが，帰無仮説を棄却しない.
  →第二種の過誤（Type Ⅱ error）と呼ばれる.
- ベイジアンネットワークの学習では，誤差によって帰無仮説を棄却してしまうため，過学習を起こし，漸近的に真の構造を推定できる保証がない.

ベイズ的アプローチによる検定

- Bayes factor：
  二つのモデルの周辺尤度の比により検定する.
- 漸近的に真の独立性検定が可能である.
- データセットを$\mathbf{X}$，独立なモデルを$g_1: p(x_1,x_2)=p(x_1)p(x_2)$，従属なモデルを$g_2: p(x_1,x_2)=p(x_1|x_2)p(x_2)$としたときの周辺尤度の比 Bayes factor（BF）：

$$BF=\frac{p(\mathbf{X}\mid g_1)}{p(\mathbf{X}\mid g_2)}$$

$BF>1$：独立
$BF<1$：従属 と判定する

シミュレーション実験1
-Type Ⅰ errorの検証-

- 2ノード間が真に独立である構造を用いて実験を行う.
- χ^2 統計量を用いた検定ではデータ数を増やしたとしても Type Ⅰ errorが発生するが, Bayes factorでは漸近的に収束することを示す.

- 実験方法

1. 条件付き確率パラメータを$(p(x_1=1|x_2=1)=0.8, p(x_1=1|x_2=0)=0.2)$, $(p(x_1=1|x_2=1)=0.7, p(x_1=1|x_2=0)=0.3)$, $(p(x_1=1|x_2=1)=0.6, p(x_1=1|x_2=0)=0.4)$の3つの条件で, データ数10, 50, 100, 500, 1000, 5000, 10000, 20000, 50000でランダムにデータを発生する.
2. 各手法を用いて各データ数で100回検定を行う.

- 比較手法
  - ✓ Bayes factor, χ^2 検定 $(\alpha=0.05)$, G^2 検定 $(\alpha=0.05)$

実験結果 – 確率パラメータ:0.8

表：Type Ⅰ errorの発生率

	10	50	100	500	1000	5000	10000	20000	50000
BF	0.26	0.07	0.02	0.0	0.0	0.0	0.0	0.0	0.0
χ^2	0.16	0.0	0.0	0.03	0.08	0.07	0.03	0.05	0.01
G^2	0.17	0.05	0.02	0.03	0.08	0.06	0.03	0.05	0.01

表：p値平均

	10	50	100	500	1000	5000	10000	20000	50000
χ^2	0.0	-	-	0.036064	0.026768	0.023187	0.027279	0.020648	0.030637
G^2	0.001127	0.038292	0.02177	0.032469	0.026032	0.019153	0.026245	0.021414	0.030028

※BF: Bayes factor

実験結果 – 確率パラメータ:0.7

表：Type Ⅰ errorの発生率

	10	50	100	500	1000	5000	10000	20000	50000
BF	0.23	0.09	0.03	0.02	0.0	0.0	0.0	0.0	0.0
χ^2	0.08	0.08	0.07	0.07	0.05	0.02	0.04	0.08	0.09
G^2	0.14	0.11	0.08	0.07	0.05	0.03	0.04	0.08	0.1

表：p値平均

	10	50	100	500	1000	5000	10000	20000	50000
χ^2	0.0082498	0.019822	0.028824	0.019722	0.025156	0.035909	0.024547	0.026415	0.021641
G^2	0.020995	0.017964	0.030004	0.019436	0.025627	0.041014	0.02444	0.026468	0.024468

※BF: Bayes factor

実験結果 – 確率パラメータ:0.6

表：Type Ⅰ errorの発生率

	10	50	100	500	1000	5000	10000	20000	50000
BF	0.12	0.04	0.01	0.01	0.0	0.0	0.0	0.0	0.0
χ^2	0.02	0.06	0.04	0.14	0.03	0.07	0.05	0.04	0.04
G^2	0.08	0.06	0.04	0.14	0.03	0.06	0.05	0.04	0.04

表：p値平均

	10	50	100	500	1000	5000	10000	20000	50000
χ^2	0.015722	0.025221	0.03339	0.027574	0.014073	0.03214	0.018765	0.022972	0.035721
G^2	0.03172	0.02506	0.033419	0.027464	0.014107	0.029259	0.018804	0.022961	0.035726

※BF: Bayes factor

シミュレーション実験2
-Type Ⅱ errorの検証-

- 2ノード間が真に従属である構造を用いて実験を行い, Type Ⅱ errorの発生率とp値を検証する.
- 実験方法

1. 条件付き確率パラメータを$(p(x_1=1|x_2=1)=0.8, p(x_1=1|x_2=0)=0.2)$, $(p(x_1=1|x_2=1)=0.7, p(x_1=1|x_2=0)=0.3)$, $(p(x_1=1|x_2=1)=0.6, p(x_1=1|x_2=0)=0.4)$の3つの条件で, データ数10, 20, 30, 40, 50, 100, 200, 500, 1000でランダムにデータを発生する.
2. 各手法を用いて各データ数で100回検定を行う.

- 比較手法
  - ✓ Bayes factor, χ^2 検定 $(\alpha=0.05)$, G^2 検定 $(\alpha=0.05)$

実験結果 – 確率パラメータ:0.8

表：Type Ⅱ errorの発生率

	10	20	30	40	50	100	200	500	1000
BF	0.3	0.29	0.19	0.11	0.02	0	0	0	0
χ^2	0.61	0.42	0.19	0.1	0.02	0	0	0	0
G^2	0.49	0.32	0.18	0.11	0.02	0	0	0	0

表：p値平均

	10	50	100	500	1000	100	200	500	1000
χ^2	0.27575	0.20121	0.22199	0.11874	0.27337	-	-	-	-
G^2	0.27454	0.23282	0.24667	0.12579	0.28718	-	-	-	-

※BF: Bayes factor

実験結果 – 確率パラメータ:0.7

表：Type II errorの発生率

	10	20	30	40	50	100	200	500	1000
BF	0.62	0.61	0.45	0.44	0.31	0.09	0	0	0
χ^2	0.89	0.75	0.43	0.39	0.23	0.02	0	0	0
G^2	0.72	0.65	0.43	0.37	0.24	0.02	0	0	0

表：p値平均

	10	20	30	40	50	100	200	500	1000
χ^2	0.42438	0.29754	0.31684	0.25316	0.25494	0.07056	-	-	-
G^2	0.43542	0.32035	0.31742	0.26442	0.2483	0.069456	-	-	-

※BF: Bayes factor

実験結果 – 確率パラメータ:0.6

表：Type II errorの発生率

	10	20	30	40	50	100	200	500	1000
BF	0.83	0.87	0.83	0.77	0.7	0.65	0.33	0.05	0
χ^2	0.98	0.88	0.81	0.73	0.62	0.54	0.19	0.01	0
G^2	0.88	0.87	0.79	0.73	0.61	0.54	0.19	0.01	0

表：p値平均

	10	20	30	40	50	100	200	500	1000
χ^2	0.44253	0.42958	0.35118	0.44008	0.34622	0.24334	0.19523	0.051237	-
G^2	0.44174	0.42751	0.35762	0.43911	0.35013	0.24296	0.19504	0.051323	-

※BF: Bayes factor

仮説検定の比較

従来の仮説検定では、結果が不安定で必ず誤差が残るのに対して、ベイズ検定では漸近的に正しい仮説を選ぶことができる。

レポート3

ベイズ手法と 従来の統計手法を比較し、それぞれの長所、欠点を分析せよ。