Bayes統計入門

黒木 玄 2019-09-03

Kullback-Leibler情報量

以下，確率密度関数を確率分布もしくは単に分布と呼んだりする．

分布 $q(x)$ と分布 $p(x)$ に対して，

$$D(q\|p)=\int q(x)\log\frac{q(x)}{p(x)}\,dx = G(q\|p)-S(q) \quad (\text{ただし } 0\log 0=0 \text{ とする.})$$

を Kullback-Leibler 情報量と呼ぶ．ここで，

$$G(q\|p)=-\int q(x)\log p(x)\,dx,\quad S(q)=G(q\|q)=-\int q(x)\log q(x)\,dx.$$

$G(q\|p)$ は汎化誤差 (generalization error)，$S(q)$ は Shannon 情報量と呼ばれる

練習問題 $D(q\|p)\geq 0$ を示せ．

解答例 $f(x)=x\log x$ $(x\geqq 0)$ とおく．$f'(x)=\log x+1$，$f''(x)=\frac{1}{x}>0$ なので f は下に凸な関数である．ゆえに，Jensenの不等式より

$$D(q\|p)=\int\frac{q(x)}{p(x)}\log\frac{q(x)}{p(x)}\,p(x)\,dx=\int f\left(\frac{q(x)}{p(x)}\right)p(x)\,dx \geqq f\left(\int\frac{q(x)}{p(x)}\,p(x)\,dx\right)$$

（↑ Jensenの不等式．）

$$=f\left(\int q(x)\,dx\right)=f(1)=1\log 1=0.$$

□

練習問題 分布 $q(x)$ の平均，分散を μ_0,σ_0^2 と書くとき，$p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ について $D(q\|p)$ を求め，$D(q\|p)$ を最小化する μ,σ^2 を求めよ．

解答例 $S(q)=-\int q(x)\log q(x)\,dx$ とおくと，

$$(x-\mu)^2=((x-\mu_0)-(\mu-\mu_0))^2$$
$$=(x-\mu_0)^2-2(\mu-\mu_0)(x-\mu_0)+(\mu-\mu_0)^2$$

$$D(q\|p)=\int q(x)\log\frac{q(x)}{p(x)}\,dx=-\int q(x)\log p(x)\,dx-S(q)=\frac{1}{2}\int q(x)\left(\log(2\pi\sigma^2)+\frac{(x-\mu)^2}{\sigma^2}\right)dx-S(q)$$

$$=\frac{1}{2}\left(\log(2\pi\sigma^2)+\frac{\sigma_0^2+(\mu-\mu_0)^2}{\sigma^2}\right)-S(q)$$

ゆえに，これを最小化する μ,σ^2 は $\mu=\mu_0$，$\sigma^2=\sigma_0^2$ となる．

□

注意 この事実は，分布 $q(x)$ をもっともよく予測する正規分布を求めることと，$q(x)$ の平均と分散を求めることが同じであることを意味する．

□

練習問題 分布 $q(x)$ は $\int_{-\infty}^{a} q(x)\,dx = \int_{a}^{\infty} q(x)\,dx \Leftrightarrow a = a_0$ と $\int_{-\infty}^{\infty} |x-a|\, q(x)\,dx = b_0$ をみたしているとする

$p(x) = \frac{1}{2b} e^{-\frac{|x-a|}{b}}$ (Laplace分布) とおく。 $D(q\|p)$ を最小化する a, b を求めよ。

解答例 $S(q) = -\int_{-\infty}^{\infty} q(x) \log q(x)\,dx$ とおくと,

$$D(q\|p) = -\int_{-\infty}^{\infty} q(x) \log p(x)\,dx - S(q) = \int_{-\infty}^{\infty} q(x)\left(\log(2b) + \frac{|x-a|}{b}\right) dx - S(q)$$

$$= \log(2b) + \frac{1}{b}\int_{-\infty}^{\infty} |x-a|\, q(x)\,dx - S(q).$$

$$\frac{\partial}{\partial a}\int_{-\infty}^{\infty} |x-a|\, q(x)\,dx = \int_{a}^{\infty} q(x)\,dx - \int_{-\infty}^{a} q(x)\,dx = 0 \text{ より},$$

$D(q\|p)$ を最小化する a, b は $a = a_0, b = b_0$ となることがわかる. □

注意 この事実は分布 $q(x)$ をもっともよく予測するLaplace分布を求めることと,
分布 $q(x)$ の中央値 (median) と中央値との差の絶対値の平均を求めることが同じであることを意味する. □

このように, 分布 $q(x)$ の平均, 分散, 中央値などの記述統計量を求めることと,
特殊な分布 (正規分布, Laplace分布) でもっとも $q(x)$ をよく予測する (フィットする) ものを
求めることは同じことになる

練習問題 $w = (\alpha, \theta)$, $p(x|w) = \begin{cases} \frac{1}{\Gamma(\alpha)\theta^{\alpha}} e^{-x/\theta} x^{\alpha-1} & (x>0) \\ 0 & (x \leqq 0) \end{cases}$ (← ガンマ分布) のとき, $D(q\|p)$ を最小化する
α, θ を求めよ。

解答例 以下, $q(x) = 0$ $(x \leqq 0)$ と仮定する. $\frac{d}{d\alpha}\log\Gamma(\alpha) = \psi(\alpha)$ を digamma 関数と呼ぶ. ($\alpha>0$ で単調増加)

$$-\log p(x|w) = \log\Gamma(\alpha) + \alpha\log\theta + \frac{x}{\theta} - (\alpha-1)\log x, \quad D(q\|p) = G(q\|p) - S(q),$$

$$G(q\|p) = -\int q(x)\log p(x|w)\,dx = \log\Gamma(\alpha) + \alpha\log\theta + \frac{1}{\theta}\underbrace{\int q(x)\,x\,dx}_{=:\mu_0} - (\alpha-1)\underbrace{\int q(x)\log x\,dx}_{=:\log\nu_0}$$

$$\frac{\partial}{\partial\theta} G(q\|p) = \frac{\alpha}{\theta} - \frac{\mu_0}{\theta^2} = \frac{\theta\alpha - \mu_0}{\theta^2} = 0, \quad \therefore \theta\alpha = \mu_0$$

$$\frac{\partial}{\partial\alpha} G(q\|p) = \psi(\alpha) + \log\theta - \log\nu_0 = \psi(\alpha) - \log\alpha + \log\mu_0 - \log\nu_0 = 0,$$

$$\therefore \log\alpha - \psi(\alpha) = \log\mu_0 - \log\nu_0 \geqq 0 \quad (\text{← Jensenの不等式}, \ (\log\alpha - \psi(\alpha) \text{は非負単調減少}))$$

(← これで α が決まる.)

□

Sanovの定理 (rough version)

$$\begin{pmatrix}\text{分布 } p(x) \text{ にしたがう乱数列}\\ X_1, X_2, \dots, X_n \text{ がまるで}\\ \text{分布 } q(x) \text{ にしたがう乱数列}\\ \text{に見える確率}\end{pmatrix} = \exp\left(-nD(q\|p) + o(n)\right) \quad \text{as } n \to \infty. \qquad \square$$

有限集合上の確率測度の場合の証明 $p_1, \dots, p_r, q_1, \dots, q_r \geq 0$, $\sum_{i=1}^r p_i = 1$, $\sum_{i=1}^r q_i = 1$ と仮定する. p は i の確率 p_i である $\{1, 2, \dots, r\}$ 上の確率測度であるとし, q も同様とする. このとき,

$$D(q\|p) = \sum_{i=1}^r q_i \log \frac{q_i}{p_i} \quad (0 \log 0 = 0 \text{ と約束}).$$

p にしたがう乱数列 $X_1, X_2, \dots, X_n$ の中の i の個数を K_i と書く. このとき,

$$\begin{pmatrix}X_1, \dots, X_n \text{ がまるで分布 } q\\ \text{の乱数列に見える確率}\end{pmatrix} = \left(\frac{K_i}{n} \approx q_i \text{ のときの } \frac{n!}{K_1! \cdots K_r!} p_1^{K_1} \cdots p_r^{K_r}\right)$$

$$\approx n^n e^{-n} \sqrt{2\pi n} \frac{p_1^{nq_1} \cdots p_r^{nq_r}}{(nq_1)^{nq_1} e^{-nq_1} \sqrt{2\pi n q_1} \cdots (nq_r)^{nq_r} e^{-nq_r} \sqrt{2\pi n q_r}}$$

$$= \frac{1}{\sqrt{(2\pi n)^{r-1} q_1 \cdots q_r}} \left(\left(\frac{q_1}{p_1}\right)^{q_1} \cdots \left(\frac{q_r}{p_r}\right)^{q_r}\right)^{-n}$$

$$= \exp\left(-n \underbrace{\sum_{i=1}^r q_i \log \frac{q_i}{p_i}}_{= D(q\|p)} + O(\log n)\right) \quad \text{as } n \to \infty. \qquad \square$$

Sanovの定理

$$\begin{pmatrix}\text{分布 } p(x) \text{ にしたがう乱数列}\\ X_1, X_2, \dots, X_n \text{ がまるで}\\ \text{分布 } q(x) \text{ にしたがう乱数列}\\ \text{に見える確率}\end{pmatrix} = \exp\left(-nD(q\|p) + o(n)\right) \quad \text{as } n \to \infty$$

は KL情報量 $D(q\|p)$ が"分布 $p(x)$ にしたがう長さ n の乱数列が分布 $q(x)$ にしたがう乱数列に見える確率が指数関数的に減少する速さ"であることを意味する

KL情報量の使い方1 Sanovの定理は分布 $p(x)$ による分布 $q(x)$ のシミュレーションの誤差(予測誤差 (prediction error) と呼ぶ)が Kullback-Leibler 情報量 $D(q\|p)$ で表わされることを意味しているとみなせる.

(モデル $p(x)$ による真の分布 $q(x)$ の予測誤差) $= D(q\|p)$.

KL情報量の使い方2 分布 $p(x)$ が与えられているとき，特殊な形の分布 $q(x)$ で分布 $p(x)$ からもっとも出て来やすい分布を得たければ" KL情報量 $D(q\|p)$ を最小化する分布 $q(x)$ を特殊な形の分布の中から見付ければ"よい．

- $p(x)$ を事前分布とする canonical 分布の導出
- 事後分布 $p(x)$ の平均場近似 (いわゆる変分 Bayes 推論に一般化)

Canonical 分布の導出 分布 $p(x)$ からもっとも出て来やすい分布 $q(x)$ で" 条件

$$\int f(x)\,q(x)\,dx = c$$

をみたすものを求めよう．Lagrange の未定乗数法を使う．

$L := L[q,\beta,\lambda] := \int q(x)\log\frac{q(x)}{p(x)}\,dx + \beta\left(\int f(x)q(x)\,dx - c\right) + (\lambda-1)\left(\int q(x)\,dx - 1\right)$ とおく．

$$\begin{cases} \dfrac{\partial L}{\partial \beta} = \displaystyle\int f(x)q(x)\,dx - c = 0 \\ \dfrac{\partial L}{\partial \lambda} = \displaystyle\int q(x)\,dx - 1 = 0 \\ \dfrac{\delta L}{\delta q(x)} = \log\dfrac{q(x)}{p(x)} + 1 + \beta f(x) + \lambda - 1 = \log q(x) - \log p(x) + \beta f(x) + \lambda = 0 \end{cases}$$

以上を解こう，$q(x) = e^{-\lambda-\beta f(x)}p(x)$，$e^{-\lambda}\int e^{-\beta f(x)}p(x)\,dx = 1$，

$Z := e^{\lambda} = \int e^{-\beta f(x)}p(x)\,dx$，$\boxed{q(x) = \dfrac{1}{Z}e^{-\beta f(x)}p(x)}$ ← 事前分布 $p(x)$ の canonical 分布

$\dfrac{1}{Z}\int f(x)e^{-\beta f(x)}p(x)\,dx = c$ を解いて c から β を得る． □

以上の結果は，分布 $p(x)$ にしたがう乱数列 $X_1,\dots,X_n$ を条件

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) = c$$

で制限して得られる条件付き確率分布における X_1 の分布が $n\to\infty$ で

$$q(x) = \frac{1}{Z}e^{-\beta f(x)}p(x)$$

に収束することを意味する．

最尤法（さいゆうほう, maximum likelihood estimation）

分布 $q(x)$ にしたがう乱数の列 $X_1, \dots, X_n$（分布 $q(x)$ にしたがう独立同分布確率変数列）を母集団分布 $q(x)$ の**サイズ n のサンプル（標本）**と呼ぶ。

- 大数の法則： $\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \int f(x) q(x)\, dx$.
- 中心極限定理： $\int\cdots\int f\left(\frac{x_1+\cdots+x_n-n\mu_0}{\sqrt{n}\,\sigma_0}\right) q(x_1)\cdots q(x_n)\, dx_1\cdots dx_n \to \int f(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx$

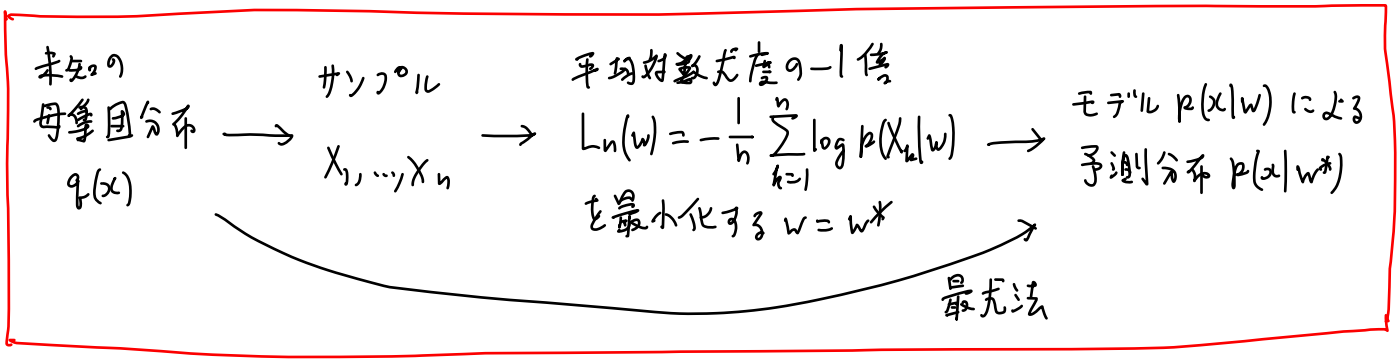
統計学の基本問題 未知の分布 $q(x)$ のサンプル $X_1, \dots, X_n$（の実現値）のみが得られるとき、サンプルだけから未知の分布についてどこまで知ることができるか？ □

以下、未知の分布 $q(x)$ のサンプル $X_1, \dots, X_n$ のみが得られていると仮定する。

最尤法 パラメーター w を持つ分布 $p(x|w)$ が与えられているとする。このとき、
w の関数 $w \mapsto p(X_1|w)\cdots p(X_n|w)$ をサンプルの**尤度関数（ゆうどかんすう）**と呼ぶ。
さらに、$L_n(w) = -\frac{1}{n}\sum_{k=1}^{n} \log p(X_k|w)$ を**平均対数尤度**の -1 倍と呼ぶ。
大数の法則より、$n\to\infty$ のとき、

$$L_n(w) = -\frac{1}{n}\sum_{k=1}^{n} \log p(X_k|w) \to -\int q(x)\log p(x|w)\, dx = G(q\|p(\cdot|w))$$

$D(q\|p) = G(q\|p) - S(q)$ だったので、分布 $p(x|w)$ による未知の分布 $q(x)$ の予測誤差を最小にするためには汎化誤差 $G(q\|p(\cdot|w))$ を最小にする w を求めればよい。
しかし、$q(x)$ は未知だと仮定したのでそれ自体は不可能である。
そこで汎化誤差 $G(q\|p(\cdot|w))$ の代わりに、$L_n(w) = -\frac{1}{n}\sum_{k=1}^{n} \log p(X_k|w)$ を最小にする $w = w^*$ を考え、未知の $q(x)$ が予測分布 $p(x|w^*)$ で近似されると期待する。
この方法を**最尤法**と呼ぶ。（注）モデル $p(x|w)$ の形に制限された分布による $q(x)$ の近似になっていると期待する。



練習問題 $w=(\mu,\sigma^2)$, $p(x|w)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ のとき, $L_n(w)$ を最小化する μ,σ^2 を求めよ.

解答例 $-\log p(x|w)=\frac{1}{2}\left(\log(2\pi\sigma^2)+\frac{(x-\mu)^2}{\sigma^2}\right)$. $\overline{X}:=\frac{1}{n}\sum_{k=1}^n X_k$, $V(X)=\frac{1}{n}\sum_{k=1}^n(X_k-\overline{X})^2$ とおく.

$$L_n:=L_n(w)=-\frac{1}{n}\sum_{k=1}^n\log p(X_k|w)=\frac{1}{2}\left(\log(2\pi\sigma^2)+\frac{1}{\sigma^2}\frac{1}{n}\sum_{k=1}^n(X_k-\mu)^2\right).$$

$$\frac{1}{2}\frac{\partial}{\partial\mu}\frac{1}{n}\sum_{k=1}^n(X_k-\mu)^2=\mu-\frac{1}{n}\sum_{k=1}^n X_k=\mu-\overline{X}=0,\quad\therefore\ \mu=\overline{X}.$$

$$\frac{\partial}{\partial\sigma^2}\left(\log(2\pi\sigma^2)-\frac{1}{\sigma^2}V(X)\right)=\frac{1}{\sigma^2}-\frac{V(X)}{\sigma^4}=\frac{\sigma^2-V(X)}{\sigma^4}=0\quad\therefore\ \sigma^2=V(X).$$ □.

練習問題 $w=(a,b)$, $p(x|w)=\frac{1}{2b}e^{-\frac{|x-a|}{b}}$ のとき, $L_n(w)$ を最小化する a,b を求めよ.

解答例 $X_1,\dots,X_n$ の中央値を $\hat{a}$ と書き, $\hat{b}=\frac{1}{n}\sum_{k=1}^n|X_k-\hat{a}|$ とおく.

$-\log p(x|w)=\log(2b)+\frac{|x-a|}{b}$, $L_n:=L_n(w)=\log(2b)+\frac{1}{b}\frac{1}{n}\sum_{k=1}^n|X_k-a|$.

$\sum_{k=1}^n|X_k-a|$ を最小化する a は $X_1,\dots,X_n$ の median $\hat{a}$ である. $\therefore\ a=\hat{a}$.

$a=\hat{a}$ のとき, $L_n=\log(2b)+\frac{\hat{b}}{b}$ でこれを最小化する b は $b=\hat{b}$. □

要するにサンプルの記述統計量である

サンプル平均 $\overline{X}=\frac{1}{n}\sum_{k=1}^n X_k$, サンプル分散 $V(X)=\frac{1}{n}\sum_{k=1}^n(X_k-\overline{X})^2$,

サンプル中央値 $\hat{a}$, サンプルの中央値との差の絶対値の平均 $\hat{b}=\frac{1}{n}\sum_{k=1}^n|X_k-\hat{a}|$

を求めることは, それぞれ正規分布モデル, Laplace分布モデルの最尤法に一致する.

練習問題 $p(x|w) = \begin{cases} \frac{1}{\Gamma(\alpha)\theta^\alpha} e^{-x/\theta} x^{\alpha-1} & (x>0) \\ 0 & (x\le 0) \end{cases}$ のとき，$L_n(w)$ を最小化する $w=(\alpha,\theta)$ を求めよ

解答例

$$-\log p(x|w) = \log\Gamma(\alpha) + \alpha\log\theta + \frac{x}{\theta} - (\alpha-1)\log x.$$

$$L_n := L_n(w) = \log\Gamma(\alpha) + \alpha\log\theta + \frac{1}{\theta}\underbrace{\frac{1}{n}\sum_{k=1}^n X_k}_{=:\overline{X}} - (\alpha-1)\underbrace{\frac{1}{n}\sum_{k=1}^n \log X_k}_{=:\overline{\log X}}.$$

$$\frac{\partial L_n}{\partial\theta} = \frac{\alpha}{\theta} - \frac{\overline{X}}{\theta^2} = \frac{\theta\alpha - \overline{X}}{\theta^2} = 0. \quad \therefore\ \theta\alpha = \overline{X}.$$

$$\frac{\partial L_n}{\partial\alpha} = \psi(\alpha) + \log\theta - \overline{\log X} = \psi(\alpha) - \log\alpha + \log\overline{X} - \overline{\log X} = 0.$$

$$\therefore\ \log\alpha - \psi(\alpha) = \log\overline{X} - \overline{\log X} \ge 0. \quad \leftarrow \text{これで } \alpha \text{ が決まる.}$$

($\log$ 相加平均 $\ge$ 相乗平均 の対数)

□.

$\alpha \mapsto \log\alpha - \psi(\alpha)$ の逆関数の計算を除けば，ガンマ分布モデルの最尤法は
サンプルの相加平均と相乗平均（の対数）を計算することと同じである.
このような形でも相加・相乗平均が出て来る.

最尤法の場合と同様にパラメーター w を持つ分布 $p(x|w)$ を用意する。
Bayes統計の基本は「よくわからないことはサイコロをふって決める仮想世界を考え，その仮想世界で偶然に現実世界で観測されたサンプルが生成された場合に制限した条件付き確率分布を考え，その条件付き確率分布が現実世界を近似していると期待すること」である。

パラメーター w を持つ分布 $p(x|w)$ の形で未知の分布 $q(x)$ を近似したいのだが，最適なパラメーター w の値はよくわからないので，パラメーター w の確率分布 $\varphi(w)$ を準備する。$\varphi(w)$ をよく事前分布 (prior) という。

"現実世界" のサンプル $X_1, \cdots, X_n$ は未知の母集団分布 $q(x)$ にしたがう独立試行で生成されており，その確率密度関数は未知の

$$Q(x_1, \cdots, x_n) = q(x_1) \cdots q(x_n)$$

になる。

一方，"仮想世界" (Bayes' world) では，"サイコロ" $\varphi(w)$ をふってランダムにパラメーター w の値 W を決めた得た母集団分布 $p(x|W)$ にしたがう独立試行でサンプルが生成されていると仮定する。サンプルの分母の確率密度関数は

$$Z(x_1, \cdots, x_n) = \int \varphi(w)\, p(x_1|w) \cdots p(x_n|w)\, dw$$

になる。

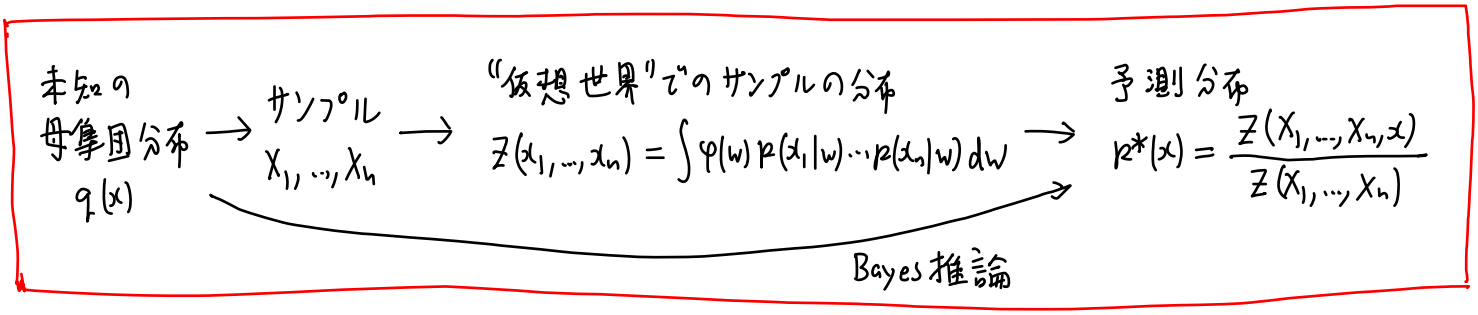
"仮想世界" の確率分布はそのままでは "現実世界" の確率分布を近似しているとはとても期待できない。

しかし，もしも "仮想世界" において偶然に "現実世界" と同じサンプル $X_1, \cdots, X_n$ が生成された場合に制限したならば，その次に "仮想世界" で生成される X_{n+1} の分布

$$p^*(x_{n+1}) = \frac{(\text{制限された } x_{n+1} \text{ の確率(密度)})}{(\text{制限された全確率(密度)})} = \frac{Z(X_1, \cdots, X_n, x_{n+1})}{Z(X_1, \cdots, X_n)}$$

は "現実世界" の X_{n+1} の分布 $q(x_{n+1})$ を近似していると期待できるのではないか？

実際にそのように考えて，$p^*(x) = \dfrac{Z(X_1, \cdots, X_n, x)}{Z(X_1, \cdots, X_n)}$ を $q(x)$ の予測分布として採用するのが Bayes推論 (Bayes' inference) である。



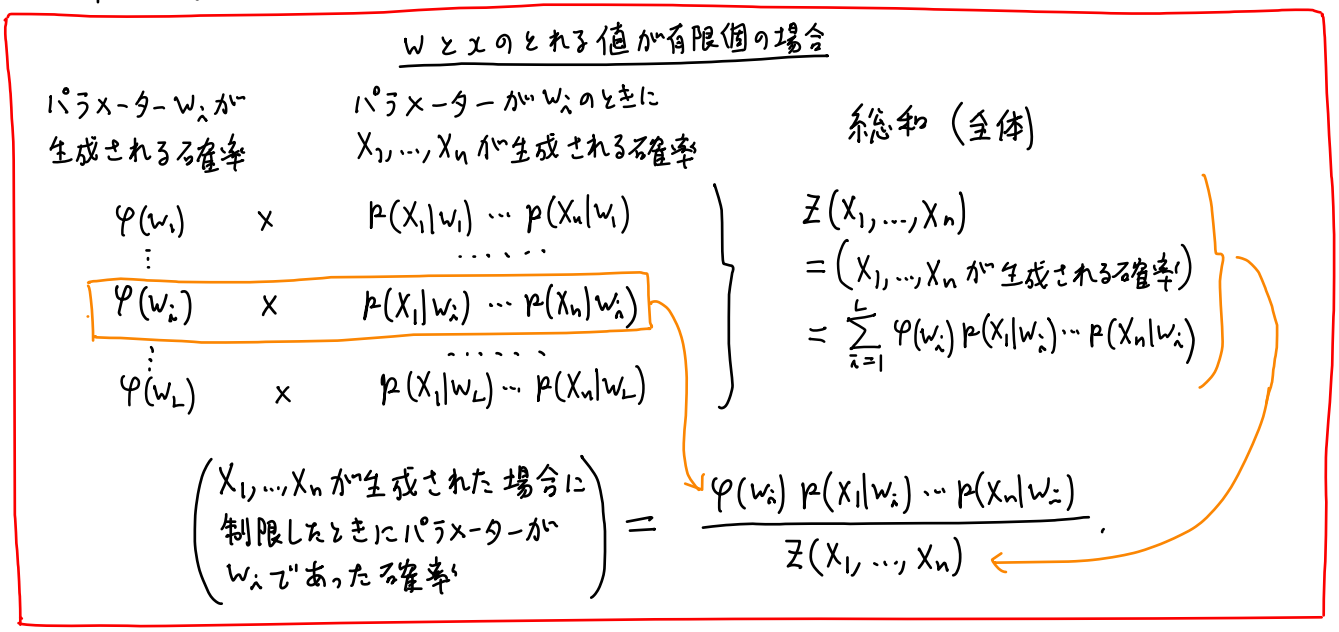
事後分布　以上のストーリーには事後分布 (posterior) が出て来なかった.
事後分布はBayesの定理を使うまでもなく, 予測分布を書き直せば自然に出て来る.

$$p^*(x)=\frac{1}{Z(X_1,\dots,X_n)}\int \varphi(w)p(X_1|w)\cdots p(X_n|w)p(x|w)\,dw=\int \frac{\varphi(w)p(X_1|w)\cdots p(X_n|w)}{Z(X_1,\dots,X_n)}p(x|w)\,dw.$$

すなわち,

$$\varphi^*(w)=\frac{\varphi(w)p(X_1|w)\cdots p(X_n|w)}{Z(X_1,\dots,X_n)},\quad Z(X_1,\dots,X_n)=\int \varphi(w)p(X_1|w)\cdots p(X_n|w)\,dw,$$

この$\varphi^*(w)$を事後分布と呼ぶ. 事後分布は, "仮想世界"で偶然に"現実世界"と同じサンプル$X_1,\dots,X_n$が生成された場合に制限して得られるパラメーターwの条件付き確率分布である



定理 (渡辺澄夫) ゆるい条件もとでBayes推論で得られる予測分布$p^*(x)$はモデル$p(x|w)$の範囲で$q(x)$をもっとも良く予測する分布にサンプルサイズ$n\to\infty$で近付く、さらに, $n\to\infty$での分配関数$Z(X_1,\dots,X_n)$や予測分布$p^*(x)$の予測誤差$D(q\|p^*)$の漸近挙動も(原理的には)詳しくわかる. □

↖ 広中の特異点解消 (のdivisor版) を本質的に使う.

詳しくは, 渡辺澄夫著『ベイズ統計の理論と方法』(2012) を参照.

統計力学との類似 $L_n(w) = -\frac{1}{n}\sum_{k=1}^{n} \log p(X_i|w)$ を使うと,

$$Z(X_1,\cdots,X_n) = \int \varphi(w)\, p(X_1|w)\cdots p(X_n|w)\, dw = \int e^{-n L_n(w)} \varphi(w)\, dw.$$

これは統計力学での分配函数

$$Z(\beta) = \int e^{-\beta H(w)} \varphi(w)\, dw \qquad \left(\begin{array}{l}\text{等確率の原理を仮定すると}\\ \varphi(w)\text{は省略されるが残してある}\end{array}\right)$$

と非常に似ている.

そこで, $Z(X_1,\cdots,X_n)$ を逆温度 β を入れて次のように一般化する:

$$Z(X_1,\cdots,X_n|\beta) = \int \varphi(w)\, p(X_1|w)^{\beta}\cdots p(X_n|w)^{\beta}\, dw = \int e^{-\beta n L_n(w)} \varphi(w)\, dw.$$

これは Hamiltonian が対数尤度の -1 倍

$$n L_n(w) = -\sum_{k=1}^{n} \log p(X_k|w)$$

である場合の canonical 分布の分配函数の形をしている.

統計学	物理学
対数尤度函数の -1 倍 $n L_n(w) = -\sum_{k=1}^{n} \log p(X_k\|w)$	確率変数 $X_1,\cdots,X_n$ を含む Hamiltonian $H(w)$
$Z(X_1,\cdots,X_n\|\beta) = \int e^{-\beta n L_n(w)} \varphi(w)\, dw$	$Z(\beta) = \int e^{-\beta H(w)} \varphi(w)\, dw$

$\beta\to\infty$ 絶対零度への極限 $\beta\to\infty$ で物理系は Hamiltonian を最小化する基底状態で凍り付く. Bayes統計での Hamiltonian は対数尤度函数の -1 倍なので, "系" は対数尤度函数の -1 倍 $nL_n(w)$ を最小化する $w=w^*$ で凍り付く.

これは Bayes 推測の $\beta\to\infty$ での極限が最尤法であることを意味している. □

$\beta=0$	$\beta=1$	β	$\beta=\infty$
事前分布	通常のベイズ推論	逆温度 β でのベイズ推論	最尤法

逆温度 β 付きの事後分布と予測分布

$$Z(X_1,\dots,X_n|\beta) := \int \varphi(w)\, p(X_1|w)^\beta \cdots p(X_n|w)^\beta\, dw = \int e^{-\beta n L_n(w)} \varphi(w)\, dw,$$

$$\varphi^*(w|\beta) := \frac{\varphi(w)\, p(X_1|w)^\beta \cdots p(X_n|w)^\beta}{Z(X_1,\dots,X_n|\beta)} = \frac{e^{-\beta n L_n(w)} \varphi(w)}{Z(X_1,\dots,X_n|\beta)},$$

$$p^*(x|\beta) := \int \varphi^*(w|\beta)\, p(x|w)\, dw = \frac{\int \varphi(w)\, p(X_1|w)^\beta \cdots p(X_n|w)^\beta\, p(x|w)\, dw}{\int \varphi(w)\, p(X_1|w)^\beta \cdots p(X_n|w)^\beta\, dw}.$$

このとき

$$Z\begin{pmatrix} X_1 & \cdots & X_n \\ \beta_1 & \cdots & \beta_n \end{pmatrix} := \int \varphi(w)\, p(X_1|w)^{\beta_1} \cdots p(X_n|w)^{\beta_n} dw$$

とおくと,

$$Z(X_1,\dots,X_n|\beta) = Z\begin{pmatrix} X_1 & \cdots & X_n \\ \beta & \cdots & \beta \end{pmatrix}, \qquad p^*(x|w) = \frac{Z\begin{pmatrix} X_1 & \cdots & X_n & x \\ \beta & \cdots & \beta & 1 \end{pmatrix}}{Z\begin{pmatrix} X_1 & \cdots & X_n \\ \beta & \cdots & \beta \end{pmatrix}}.$$

分配関数 $Z\begin{pmatrix} x_1 & \cdots & x_n \\ \beta_1 & \cdots & \beta_n \end{pmatrix}$ ですべてを表わせる.

指数型分族と共役事前分布（例外的に容易な場合）

$$\begin{cases} p(x|w) = v(x)\, e^{-\langle f(w), g(x)\rangle} & \left(f(w), g(x) \in \mathbb{R}^r,\ \langle , \rangle : \text{内積}\right) \\ \varphi(w|c) = Z(c)^{-1} e^{-\langle f(w), c\rangle}, \quad Z(c) = \int e^{\langle f(w), c\rangle} dw & \left(c \in \mathbb{R}^r\right) \end{cases}$$

指数型分布族

事前分布 $\varphi(w)$ がパラメータ c による場合 $\varphi(w) = \varphi(w|c)$

共役事前分布

$$\varphi(w|c)\, p(x_1|w)^{\beta_1} \cdots p(x_n|w)^{\beta_n} = Z(c)^{-1} v(x_1)^{\beta_1} \cdots v(x_n)^{\beta_n} e^{-\langle f(w),\, c + \sum_{k=1}^n \beta_k g(x_k)\rangle}$$

$$Z\begin{pmatrix} x_1 & \cdots & x_n \\ \beta_1 & \cdots & \beta_n \end{pmatrix} = Z(c)^{-1} v(x_1)^{\beta_1} \cdots v(x_n)^{\beta_n} Z\left(c + \sum_{k=1}^n \beta_k g(x_k)\right).$$

$$\varphi^*(w|c) = \frac{\varphi(w|c)\, p(X_1|w)^{\beta_1} \cdots p(X_n|w)^{\beta_n}}{Z\begin{pmatrix} x_1 & \cdots & x_n \\ \beta_1 & \cdots & \beta_n \end{pmatrix}} = \varphi\left(w \,\middle|\, c + \sum_{k=1}^n \beta_k g(X_k)\right).$$

$$p^*(x) = \frac{Z\begin{pmatrix} X_1 & \cdots & X_n & x \\ \beta_1 & \cdots & \beta_n & 1 \end{pmatrix}}{Z\begin{pmatrix} X_1 & \cdots & X_n \\ \beta_1 & \cdots & \beta_n \end{pmatrix}} = v(x) \frac{Z\left(c + \sum_{k=1}^n \beta_k g(X_k) + g(x)\right)}{Z\left(c + \sum_{k=1}^n \beta_k g(X_k)\right)}.$$

すべてこのように計算できてしまう！

Bayes統計の“仮想世界”の確率分布をサンプル $X_1, \ldots, X_n$ が生成された場合に制限することは，共役事前分布のパラメータ c を

$$c \longmapsto c + \sum_{k=1}^n \beta_k g(X_k)$$

と更新することに対応している．

(注) $n=0$ の場合の予測分布を

$$p^*(x|c) = v(x) \frac{Z(c + g(x))}{Z(c)}$$

と書くと，一般の n での予測分布 $p^*(x)$ は

$$p^*(x) = p^*\left(x \,\middle|\, c + \sum_{k=1}^n \beta_k X_k\right)$$

と表わされる．共役事前分布を使用できる場合には，予測分布の explicit formula を求めるためにはサンプルサイズ $n=0$ の予測分布の公式を求めれば十分である．

例 (正規分布モデル) $w=(\mu,\lambda)$, $c=(a,b,c,d)$

$$p(x|w)=\sqrt{\frac{\lambda}{2\pi}}\,e^{-\frac{\lambda}{2}(x-\mu)^2}=\frac{1}{\sqrt{2\pi}}e^{-\left(\frac{\lambda}{2}x^2-\lambda\mu x+\frac{\lambda\mu^2}{2}-\frac{\log\lambda}{2}\right)}$$

$$\begin{cases} f(w)=\left(\frac{\lambda}{2},-\lambda\mu,\frac{\lambda\mu^2}{2},-\frac{\log\lambda}{2}\right) \\ g(x)=(x^2,x,1,1) \end{cases}$$

$$\varphi(w|c)=z(c)^{-1}e^{-\left(\frac{\lambda}{2}a-\lambda\mu b+\frac{\lambda\mu^2}{2}c-\frac{\log\lambda}{2}d\right)}$$

$$=z(c)^{-1}e^{-\frac{\lambda c}{2}\left(\mu-\frac{b}{c}\right)^2-\frac{\lambda}{2}\frac{ac-b^2}{c}}\lambda^{\frac{d}{2}}$$

$$\left(a-2b\mu+c\mu^2=c\left(\mu-\frac{b}{c}\right)^2+\frac{ac-b^2}{c}\right)$$

$$\int e^{-\frac{\lambda c}{2}(\mu-\frac{b}{c})^2}d\mu=\sqrt{\frac{2\pi}{\lambda c}}$$

$$\frac{d-1}{2}=\frac{d+1}{2}-1$$

$$\int_0^\infty e^{-\frac{ac-b^2}{2c}\lambda}\lambda^{\frac{d-1}{2}}d\lambda=\Gamma\left(\frac{d+1}{2}\right)\left(\frac{2c}{ac-b^2}\right)^{\frac{d+1}{2}}$$

$$\therefore\quad z(c)=\sqrt{\frac{2\pi}{c}}\,\Gamma\left(\frac{d+1}{2}\right)\left(\frac{2c}{ac-b^2}\right)^{\frac{d+1}{2}}$$

$$(x^2+a)(c+1)-(x+b)^2$$
$$=(c+1)x^2+(c+1)a-x^2-2bx-b^2$$
$$=cx^2-2bx+(c+1)a-b^2$$
$$=c\left(x-\frac{b}{c}\right)^2+(c+1)a-\left(1+\frac{1}{c}\right)b^2$$
$$=c\left(x-\frac{b}{c}\right)^2+\frac{c+1}{c}(ac-b^2)$$
$$=\frac{c+1}{c}(ac-b^2)\left(\frac{c}{c+1}\frac{c}{ac-b^2}\left(x-\frac{b}{c}\right)^2+1\right)$$

$$c+g(x)=(a+x^2,\ b+x,\ c+1,\ d+1),$$

$$z(c+g(x))=\begin{pmatrix}x\text{によらない}\\ \text{定数}\end{pmatrix}\times\left(1+\frac{c}{c+1}\frac{c}{ac-b^2}\left(x-\frac{c}{b}\right)^2\right)^{-\frac{d+2}{2}}$$

サンプルサイズ $n=0$ の予測分布は

$$p^*(x|c)=\begin{pmatrix}x\text{によらない}\\ \text{定数}\end{pmatrix}\times\left(1+\frac{c}{c+1}\frac{c}{ac-b^2}\left(x-\frac{c}{b}\right)^2\right)^{-\frac{d+2}{2}}\quad\leftarrow\text{自由度}\ \frac{d+2}{2}\ \text{の}t\text{分布}$$

$\beta_k=\beta$ のときのサンプルサイズ n の予測分布は

$$p^*(x|\tilde{c})=p^*\left(x|c+\beta\sum_{k=1}^n g(X_k)\right)=\begin{pmatrix}x\text{によらない}\\ \text{定数}\end{pmatrix}\times\left(1+\frac{\tilde{c}}{\tilde{c}+1}\frac{\tilde{c}}{\tilde{a}\tilde{c}-\tilde{b}^2}\left(x-\frac{\tilde{c}}{\tilde{b}}\right)^2\right)^{-\frac{\tilde{d}+2}{2}},$$

$$\begin{cases} \tilde{a}=a+n\beta\frac{1}{n}\sum_{k=1}^n X_k^2=a+n\beta\left(V(X)+\bar{X}^2\right), \\ \tilde{b}=b+n\beta\frac{1}{n}\sum_{k=1}^n X_k=b+n\beta\bar{X}, \\ \tilde{c}=c+n\beta \\ \tilde{d}=d+n\beta \end{cases}$$

$$\begin{cases} \bar{X}:=\frac{1}{n}\sum_{k=1}^n X_k \\ V(X):=\frac{1}{n}\sum_{k=1}^n(X_k-\bar{X})^2=\frac{1}{n}\sum_{k=1}^n X_k^2-\bar{X}^2 \end{cases}$$

サンプルを生成した分布 $q(x)$ の平均と分散をそれぞれ μ_0,σ_0^2 と書くと, $n\to\infty$ のとき,

$$\frac{\tilde{c}+1}{\tilde{c}}\sim 1,\quad\frac{\tilde{a}\tilde{c}-\tilde{b}^2}{\tilde{c}}\sim n\beta V(X)\sim n\beta\sigma_0^2,\quad\frac{\tilde{b}}{\tilde{c}}\sim\bar{X}\sim\mu_0,\quad\frac{\tilde{d}+2}{2}=\frac{n\beta}{2}+\frac{d+1}{2}$$

$$p^*(x|\tilde{c})\sim\begin{pmatrix}x\text{によらない}\\ \text{定数}\end{pmatrix}\left(1+\frac{1}{n\beta\sigma_0^2}(x-\mu_0)^2\right)^{-\frac{n\beta}{2}-\frac{d+1}{2}}\sim\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}}.$$

たしかに, $n\to\infty$ で予測分布は正規分布の範囲で分布 $q(x)$ をもっとも予測する分布に収束する. □

一般の指数型分布族の場合

$$p(x|w) = v(x)\, e^{-\langle f(w), g(x)\rangle}, \quad w=(w_1,\dots,w_d), \quad f(w)=(f_1(w),\dots,f_r(w))$$
$$\varphi(w|c) = z(c)^{-1} e^{-\langle f(w), c\rangle}$$

これ(予測誤差)を最小化したい

$$-\log p(x|w) = -\log v(x) + \langle f(w), g(x)\rangle$$

(汎化誤差) $= D(q\|p(\cdot|w)) + S(q)$

$G(q\|p(\cdot|w)) = -\int q(x)\log p(x|w)\,dx = -\log v(x) + \langle f(w), \int g(x)q(x)\,dx\rangle$ を最小にする $w=w_0$ が一意に存在すると仮定する. このとき, $p(x|w_0)$ は $p(x|w)$ 型の分布で $q(x)$ をもっともよく予測する分布になる.

$\beta_k=\beta$ のとき, $n\to\infty$ において,

$$\tilde{c} := c + \sum_{k=1}^{n} \beta g(X_k) = c + n\beta \underbrace{\frac{1}{n}\sum_{k=1}^{n} g(X_k)}_{n\to\infty \to \int g(x)q(x)\,dx \ \text{(大数の法則)}} = n\beta\left(\int g(x)q(x)\,dx + o(1)\right)$$

$$\therefore \langle f(w), \tilde{c}\rangle = n\beta\left(\underbrace{\langle f(w), \int g(x)q(x)\,dx\rangle}_{\text{これは } w=w_0 \text{ で最小になる.}} + o(1)\right)$$

Laplace 近似

ゆえに, 事後分布 $\varphi(w|\tilde{c}) = z(\tilde{c})^{-1} e^{-\langle f(w), \tilde{c}\rangle}$ は $n\to\infty$ で $w=w_0$ に集中する.

ゆえに, $n\to\infty$ とすると, $p^*(x) = \int p(x|w)\,\varphi(w|\tilde{c})\,dw \to \int p(x|w)\,\delta(w-w_0)\,dw = p(x|w_0)$,

指数型分布族とその共役事前分布を用いた Bayes 推測では, サンプルサイズ $n\to\infty$ で予測分布は用いた指数型分布族の範囲内でサンプルを生成した分布をもっともよく予測する分布に収束する.

注意 指数型分布族と共役事前分布の組み合わせは, 議論が特別にシンプルになる非常に特殊な場合である.

- 確率モデルとして指数型分布族を採用していても, 事前分布として共役事前分布を使用するとは限らない.
- 確率モデル全体の中で指数型分布族は非常に特殊なものであり, より適切な複雑なモデルを採ると指数型分布族以外になることが大部分である.
- 複雑なモデルでも上に書いたようなことがうまく行くことは渡辺澄夫氏によって示された. □

練習問題 Bernoulli分布（コイン投げ分布）モデル $p(x|w) = w^x (1-w)^{1-x}$ $(x=0,1,\ 0\leqq w\leqq 1)$

の共役事前分布を作って，Bayes推測の様子を調べよ．

（$x=1$ の確率が $p(1|w)=w$，$x=0$ の確率が $p(0|w)=1-w$）

解答例 $p(x|w) = w^x(1-w)^{1-x} = e^{x\log w + (1-x)\log(1-w)}$

$\varphi(w|a,b) = Z(a,b)^{-1} e^{a\log w + b\log(1-w)} = Z(a,b)^{-1} w^a (1-w)^b$．← 共役事前分布

$$Z(a,b) = \int_0^1 w^a(1-w)^b dw = \int_0^1 w^{(a+1)-1}(1-w)^{(b+1)-1} dw = B(a+1,b+1) = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)},$$

分布 $q(x) = w_0^x (1-w_0)^{1-x} = p(x|w_0)$ にしたがう0と1の乱数列 $X_1, X_2, \cdots, X_n$ を考える．

$X_1,\cdots,X_n$ の中で1であるものの個数を K_n と書く：$K_n = X_1 + \cdots + X_n$．

大数の法則より，$\frac{K_n}{n} \to w_0$ $(n\to\infty)$．

事後分布は $\varphi(w|a+K_n, b+n-K_n) = B(a+K_n+1, b+n-K_n+1)\, w^{a+K_n}(1-w)^{b+n-K_n}$．

対数尤度関数の−1倍は，$-\log p(x|w) = \begin{cases} -\log w & (x=1) \\ -\log(1-w) & (x=0) \end{cases}$ なので，

$$nL_n(w) = -\sum_{k=1}^n \log p(X_k|w) = -K_n\log w - (n-K_n)\log(1-w) = n\left(-\frac{K_n}{n}\log w - \left(1-\frac{K_n}{n}\right)\log(1-w)\right)$$

$$= n\Big(\underbrace{-w_0\log w - (1-w_0)\log(1-w)}_{\text{これは } w=w_0 \text{ で最小になる．}} + O(1)\Big)$$

（$-\log\varphi(w|a,b) = -a\log w - b\log(1-w) = n\left(-\frac{a}{n}\log w - \frac{b}{n}\log(1-w)\right)$）

$w = w_0 + \Delta w$ とおく．

$$-w_0\log w - (1-w_0)\log(1-w) = -w_0\underbrace{\log(w_0+\Delta w)}_{=\log w_0 + \log\left(1+\frac{\Delta w}{w_0}\right)} - (1-w_0)\underbrace{\log(1-w_0-\Delta w)}_{=\log(1-w_0)+\log\left(1-\frac{\Delta w}{1-w_0}\right)}$$

$$\log\left(1+\frac{\Delta w}{w_0}\right) = \frac{\Delta w}{w_0} - \frac{\Delta w^2}{2w_0^2} + O(\Delta w^3), \qquad \log\left(1-\frac{\Delta w}{1-w_0}\right) = -\frac{\Delta w}{1-w_0} - \frac{\Delta w^2}{2(1-w_0)^2} + O(\Delta w^3)$$

$$= -w_0\log w_0 - (1-w_0)\log(1-w_0) + \underbrace{\frac{1}{2}\left(\frac{1}{w_0}+\frac{1}{1-w_0}\right)\Delta w^2}_{=\frac{\Delta w^2}{2w_0(1-w_0)}} + O(\Delta w^3)$$

$$\therefore\ nL_n(w_0+\Delta w) = n\left(-w_0\log w_0 - (1-w_0)\log(1-w_0) + \frac{\Delta w^2}{2w_0(1-w_0)} + O(\Delta w^3)\right)$$

これより，事後分布は w_0 の近くで $\propto e^{-n\beta\frac{\Delta w^2}{2w_0(1-w_0)}}\varphi(w_0+\Delta w)$ のようにふるまう．

すなわち $n\to\infty$ で事後分布は $\propto e^{-n\beta\frac{\Delta w^2}{2w_0(1-w_0)}}$（正規分布）のようにふるまう．（中心極限定理）

$n\to\infty$ で事後分布はデルタ分布 $\delta(w-w_0)$ に収束する． □

注意 これ例は有限集合上の確率をあつかっているので非常に特殊である． □

最尤法とBayes推論の予測分布のちがい

パラメーター w を含む確率分布 $p(x|w)$ とパラメーター w の事前分布 $\varphi(w)$ が与えられているとし，未知の分布 $q(x)$ のサンプル $X_1, \cdots, X_n$ が得られているとする．

平均対数尤度関数の -1 倍を $L_n(w) = -\frac{1}{n}\sum_{k=1}^{n} \log p(X_k|w)$ と書く．

最尤法 $L_n(w)$ を最小化するパラメーター $w = \hat{w}$ に対する $\hat{q}(x) = p(x|\hat{w})$ が最尤法の予測分布になる．

Bayes推論 事後分布 $\varphi^*(w) = \dfrac{e^{-n L_n(w)}\varphi(w)}{Z(X_1, \cdots, X_n|\beta)}$ に関する $p(x|w)$ の平均 $p^*(x) = \int p(x|w)\varphi^*(w)\,dw$ がBayes推論の予測分布になる．

注意 もしも事後分布が $L_n(w)$ を最小化する $w = \hat{w}$ に集中していれば"最尤法の予測分布とBayes推論の予測分布には大きなちがいはない．

しかし，事後分布が大きく広がった分布になっていると，最尤法とBayes推測の予測分布はかなりちがったものになる．そのようなとき，Bayes推論の予測分布は広がった「ぼやけた」分布になる．$\longrightarrow$ コンピューターによる計算結果を参照．

事後分布が広がった分布になることは，尤度関数の形が広がった形になるのとほぼ同じことである．そのような状況はサンプルサイズ n が小さいせいで，Bayes推論および最尤推定が十分に収束していないことを意味している．

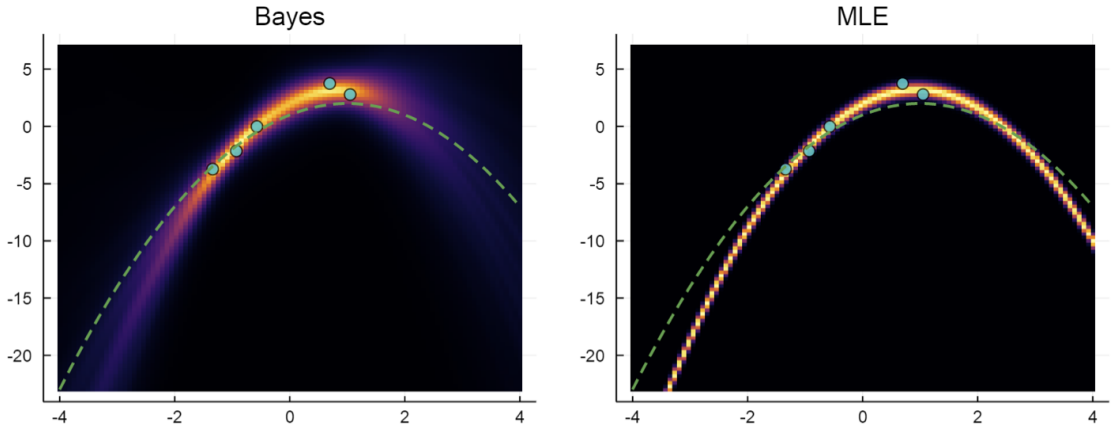
そのような推定・推論が十分に収束していないときに，Bayes推論の予測分布にはそのことが反映されて，予測はぼやけたものになる．

しかし，最尤法では尤度関数を最大化するパラメーター $w = \hat{w}$ のみを考え，尤度関数が持っている推定の収束し具合の情報を捨ててしまうので，予測分布を見ても推定の収束し具合はよくわからない．

だから，最尤法を使う場合には収束し具合の情報を"信頼区間"の形で示すことが多い．しかし，"信頼区間"と尤度関数や事後分布の広がり具合の情報は別であることには注意しなければいけない．

覚えておくべきこと 最尤法が使える場合には(正則モデルの場合には)，尤度関数や事後分布の広がり具合や"信頼区間"の幅は推定・推論の収束し具合を表わしている． □

例 次のグラフは $y = 1 + 2x - x^2 + \varepsilon$, $\varepsilon \sim \mathrm{Normal}(0,1)$, $x \sim \mathrm{Uniform}(-2,2)$ という法則で生成したサイズ $n=5$ のサンプルを $y = \beta_0 + \beta_1 x + \beta_2 x + \varepsilon$, $\varepsilon \sim \mathrm{Normal}(0,1)$ というモデルで推定した結果である.



シアンのドットがサンプルで点線がノイズ ε を除いた真の法則を表わしている.
ヒートマップでプロットされているのが予測分布である.
Bayes推論の予測分布はかなりぼやけており, 最尤法 (maximum likelihood estimation, MLE, 実質的に最小二乗法) の予測分布は細くくっきりしている.
さらに, Bayes推論の予測分布ではサンプルの x が生成される範囲の $-2 < x < 2$ の外側では非常にぼけているが, 最尤法 (MLE) の予測分布ではサンプルの x の範囲の情報は見えなくなってしまっている.

モデルのパラメーターが $\beta_0, \beta_1, \beta_2, \sigma$ の4個でサンプルサイズは $n=5$ なので, これは相当に苦しい推定・推論である. Bayes推論の側の予測分布にはその苦しさが, 分布がぼやける (予測があいまいになる) という形で現れている.

Bayes推論の予測には「予測の分からなさ加減」がよく表われており, その意味でBayes推論の予測は「正直な予測」と言える.

それに対して, 最尤法の予測分布には, 尤度関数全体の形が持っているはずの「予測の分からなさ加減」が反映されていない. 最尤法では予測分布以外に"信頼区間"などの情報も示す必要があるだろう.

Bayes推論に似て非なるものたち

($p^*(x) = \frac{Z(X_1,\dots,X_n,x)}{Z(X_1,\dots,X_n)}$ と条件付き確率分布の形で自然に書ける.)

Bayes推論 事後分布 $\varphi^*(w) = Z^{-1}\varphi(w)\prod_{k=1}^n p(X_k|w)$ に関する確率モデル $p(x|w)$ の平均を予測分布 $p^*(x) = \int \varphi^*(w)p(x|w)dw$ とする. □

最尤法 (MLE) 尤度関数 $w \mapsto \prod_{k=1}^n p(X_k|w)$ を最大化する $w=\hat{w}$ を確率モデル $p(x|w)$ に代入してできる $p(x|\hat{w})$ を予測分布とする. □

MAP法 (最大事後確率法, maximum a posteriori) 事後分布 $\varphi^*(w)$ または同じことだが (事前分布)×(尤度関数) $\varphi(w)\times\prod_{k=1}^n p(X_k|w)$ を最大化する $w=\tilde{w}$ を確率モデル $p(x|w)$ に代入してできる $p(x|\tilde{w})$ を予測分布とする.

MAP法は正則化項付きの最適化とみなされる. すなわち,

$$L(w) = \underbrace{-\sum_{k=1}^n \log p(X_k|w)}_{\text{損失関数 (loss function)}} \underbrace{-\log\varphi(w)}_{\text{正則化項 (regularization term)}}$$

← 罰則項ともよばれる

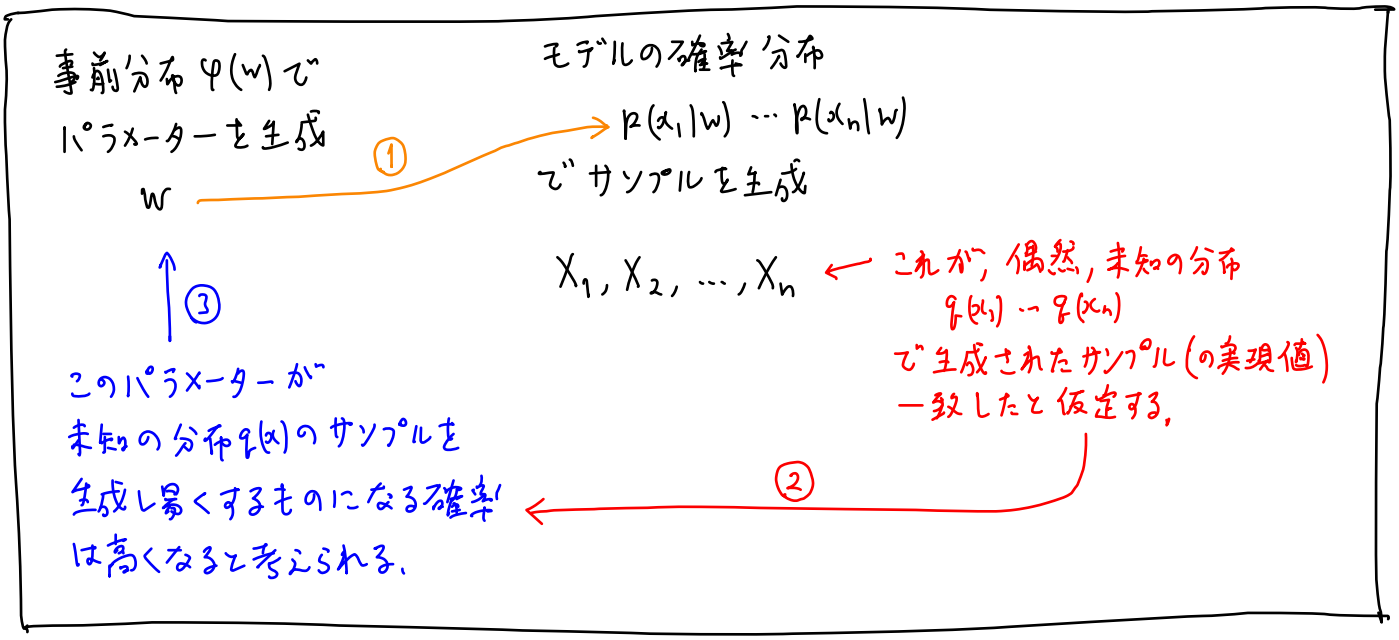
の最小化する $w=\tilde{w}$ を求めることと本質的に同じである. $-\log\varphi(w)$ が $\lambda\frac{1}{p}\sum_{i=1}^d |w_i|^p$ の形のとき, L^p-regularization という. L^2-regularization は正規分布の事前分布の場合に対応し, Ridge正則化と呼ばれる. L^1-regularization は Laplace分布の事前分布の場合に対応し, Lasso正則化と呼ばれる. このような形で MAP法もよく使われている. □

他の変種 MAP法のパラメーター推定値は事後分布 $\varphi^*(w)$ を最大化する $w=\tilde{w}$ であった. MAP法は事後分布の最頻値 (major) を解とする推定法である. 最頻値の代わりに平均値 $\overline{w} = \int w\varphi^*(w)dw$ (や適切に定義された中央値) を代入して予測分布 $p(x|\overline{w})$ を作ることもできる. しかし, モデルが複雑な場合での使用はおすすめできない. □

注意 事後分布を求めて, 事後分布の記述統計量を求めることだけで (例:最頻値 → MAP法), 「Bayes統計」と呼んでいる場合が非常に多い. Bayes統計の本質は予測分布を事後分布の平均 (もしくはそれと同値な条件) で定義することなので, 誤解しないように注意した方がよい. 「似て非なるもの」との区別は重要である. □

補足 すでに説明したことだが，パラメーター w を分布 $\varphi(w)$ によってランダムに決めて，確率分布 $p(x|w)$ にしたがう独立試行でサンプル $X_1, \dots, X_n$ をランダムに生成する"仮想世界"を，サンプル $X_1, \dots, X_n$ が現実の調査・観察で得たデータ（＝サンプル）に一致した場合に制限する条件付き確率分布を考えることが Bayes 推論の基本的な考え方であった．"仮想世界"で生成されたサンプルが現実世界で得たサンプルに一致している場合には，「よくわからない」という理由で事前分布 $\varphi(w)$ によってランダマイズされたパラメーター w が現実世界で得たサンプルを生成し易いものになっている確率が高くなっていると考えられ，さらに現実世界のサンプルを生成し易いパラメーター w に対する分布 $p(x|w)$ は現実の確率法則を近似したものになると期待される．そして，実際にそうであることを相当に一般的な仮定のもとで示すことができる！ これが Bayes 統計の数学的な基礎になっている．

一方，Bayes 推論と似て非なるものたちについては，以上のような確率に関する自然な直観は適用できず，（特異モデルを含む）一般的な設定で似て非なるものたちによる推定・推測がうまく行くことを示すのは難しい．似て非なるものとの区別は重要である．



特異モデル

統計学では，未知の分布 $q(x)$ に対して，確率モデル $p(x|w)$ の範囲内の分布で予測誤差

$$D(w) = D(q\|p(\cdot|w)) = G(q\|p(\cdot|w)) - S(q) \quad \left(G(q\|p) = -\int q(x)\log p(x)\,dx,\ S(q) = G(q\|q)\right)$$

を最小にするものを近似的に求める問題がよく出て来る．

$p(x|w)$ による $q(x)$ の予測誤差 $D(w)$ を最小にする w を w_0 と書くことにする．（唯一つと仮定する．）このとき，関数 $D(w)$ は $w=w_0$ において次のように Taylor 展開される：

$$D(w) = D(w_0) + \frac{1}{2}\sum_{i,j=1}^{d} D_{w_i w_j}(w_0)(w_i - w_{0i})(w_j - w_{0j}) + O(\|w-w_0\|^3).$$

ここで，$D_{w_i w_j}(w)$ は $D(w)$ の2階の偏導関数である．$D(w_0)$ の最小性より，$D_{w_i}(w_0) = 0$．

$D(w_0)$ の最小性より，実対称行列 $\left[D_{w_i w_j}(w_0)\right]_{i,j=1}^{d}$ の固有値はすべて0以上になる．

その固有値がすべて正のとき，モデル $p(x|w)$ は（真の分布 $q(x)$ について）正則であるという．

（そうでないとき特異であるという．w_0 が唯一つでなくても $p(x|w_0)$ がすべて同じになる十分条件も仮定．）

注意 典型的な場合に，$q(x) = p(x|w_0)$ について $p(x|w)$ が特異モデルになる w の集合は測度零集合（全次元が正の部分集合）になるのだが，「測度零集合なので現実の問題では無視してよい」ということにはならない．モデルが厳密には正則モデルであったとしても Hessian $\left[D_{w_i w_j}(w_0)\right]_{i,j=1}^{d}$ の正の固有値の中に0に非常に近いものがあれば，現実に得られるサンプルサイズの解像度では実質的に特異モデルだとみなすのが正しい場合も存在するからである． □

注意 （$\left[D_{w_i w_j}(w_0)\right]_{i,j=1}^{d}$ と Fisher 情報行列の関係）上で $q(x) = p(x|w_0)$ であると仮定する．

このとき，$D(w) = -\int p(x|w_0)\log p(x|w)\,dx - S(p(\cdot|w_0))$．$\int p(x|w)\,dx = 1$ より $\int p_{w_i}(x|w)\,dx = 0, \ldots$ ゆえに

$$D_{w_i}(w_0) = -\int p(x|w_0)\frac{p_{w_i}(x|w_0)}{p(x|w_0)}\,dx = 0 \qquad \int p_{w_i w_j}(x|w)\,dx = 0$$

$$D_{w_i w_j}(w_0) = -\int p(x|w_0)\left(\frac{\partial^2}{\partial w_i \partial w_j}\Big|_{w=w_0}\log p(x|w)\right)dx \ \downarrow$$

$$= \int p(x|w_0)\left(\frac{p_{w_i}(x|w_0)\,p_{w_j}(x|w_0)}{p(x|w_0)^2} - \frac{p_{w_i w_j}(x|w_0)}{p(x|w_0)}\right)dx$$

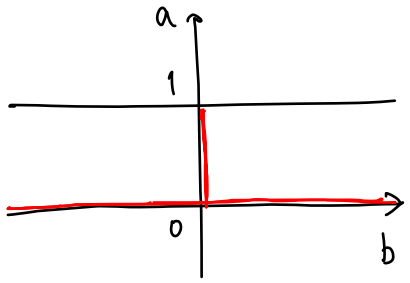
$$= \int p(x|w_0)\left(\frac{\partial}{\partial w_i}\Big|_{w=w_0}\log p(x|w)\right)\left(\frac{\partial}{\partial w_j}\Big|_{w=w_0}\log p(x|w)\right)dx$$

$\left[D_{w_i w_j}(w_0)\right]_{i,j=1}^{n}$ はモデル $p(x|w)$ の Fisher 情報量行列の $w=w_0$ での値に一致．

Fisher 情報量行列は本質的に予測誤差の Taylor 展開の2次の項のことである． □

例 混合正規分布モデル $p(x|a,b) = (1-a)\dfrac{e^{-x^2/2}}{\sqrt{2\pi}} + a\dfrac{e^{-(x-b)^2/2}}{\sqrt{2\pi}}$ $(0 \le a \le 1,\ b \in \mathbb{R})$

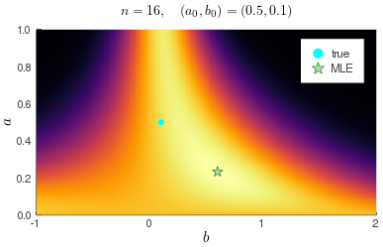
このとき、$q(x) = \dfrac{e^{-x^2/2}}{\sqrt{2\pi}}$ はモデル $p(x|a,b)$ の $p(x|a,0)$ と $p(x|0,b)$ で実現可能である

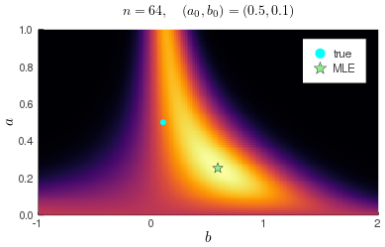


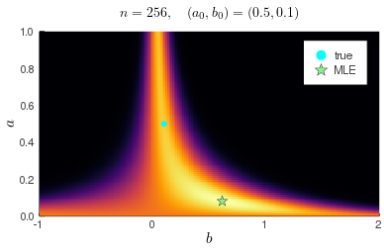
← 赤の線上のパラメーター (a,b) に対して
$p(x|a,b) = q(x)$ となる.

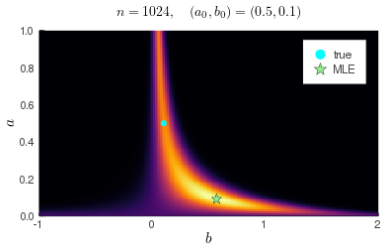
$p(x|a,b)$ は $q(x)$ について特異モデルになっている. $p(x|a,b)$ の確率密度関数は2つの正規分布の山を持つ形になるのだが2つの山が近付いて区別できなくなる極限で特異モデルになっている.

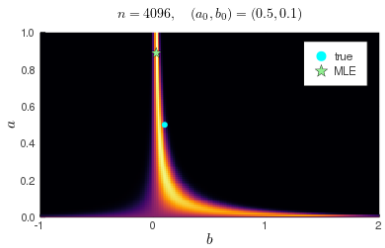
それでは $q(x)$ が $\dfrac{e^{-x^2/2}}{\sqrt{2\pi}}$ ではなくそれに近い分布だったらどうなるだろうか?
たとえば $q(x) = p(x|a_0,b_0)$ で b_0 が0に近い値だったらどうなるだろうか?
そのような場合に $q(x)$ のサンプル $X_1,\dots,X_n$ から尤度関数 $(a,b) \mapsto \prod_{k=1}^{n} p(X_k|a,b)$ をプロットしたものが以下である.

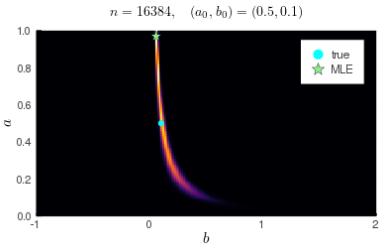












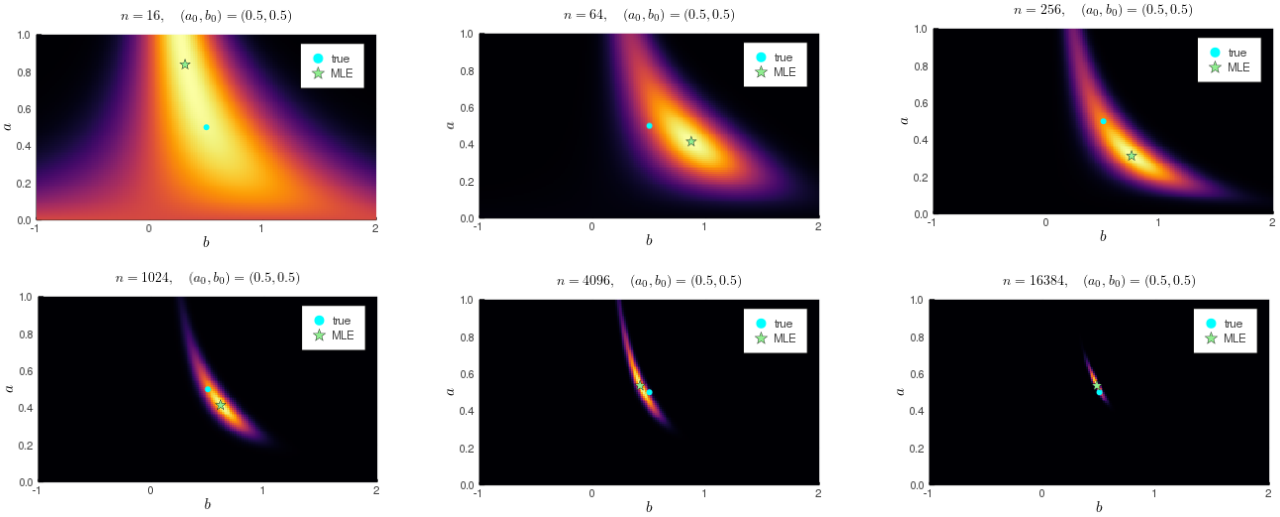
これを見ると $(a,0),(0,b)$ に沿って尤度関数の台が拡がっていることがわかる.
サンプルを生成した分布(真の分布)は $q(x) = p(x|0.5,0.1)$ である.
モデル $p(x|a,b)$ は $q(x)$ について正則モデルである.
しかし, 特異モデルに非常に近く, サンプルサイズ n を増やしてもなかなか真のパラメーター値に集中する単峰型にならない.
MLEは最尤法の解である(星印).
サンプルサイズ n を大きくしても星印の位置は全然収束していない.

例 (正則モデルの尤度関数の典型的な漸近挙動の例)

$p(x|a,b) = (1-a)\dfrac{e^{-x^2/2}}{\sqrt{2\pi}} + a\dfrac{e^{-(x-b)^2/2}}{\sqrt{2\pi}}$ $(0 \leq a \leq 1,\ b \in \mathbb{R})$ とおく.

分布 $q(x) = p(x|0.5, 0.5)$ でサイズ n のサンプル $X_1, \ldots, X_n$ を生成して,

モデル $p(x|a,b)$ の尤度関数 $(a,b) \mapsto \prod_{k=1}^{n} p(X_k|a,b)$ をプロットしてみた,



サンプルサイズ n を大きくしていくと, 尤度関数は単峰型になり,

その台がサンプルを生成した母集団分布 $q(x) = p(x|0.5, 0.5)$ のパラメータ値(真の値)に集中して行っていることがわかる.

これが, 正則モデルの尤度関数の漸近挙動の典型例である.

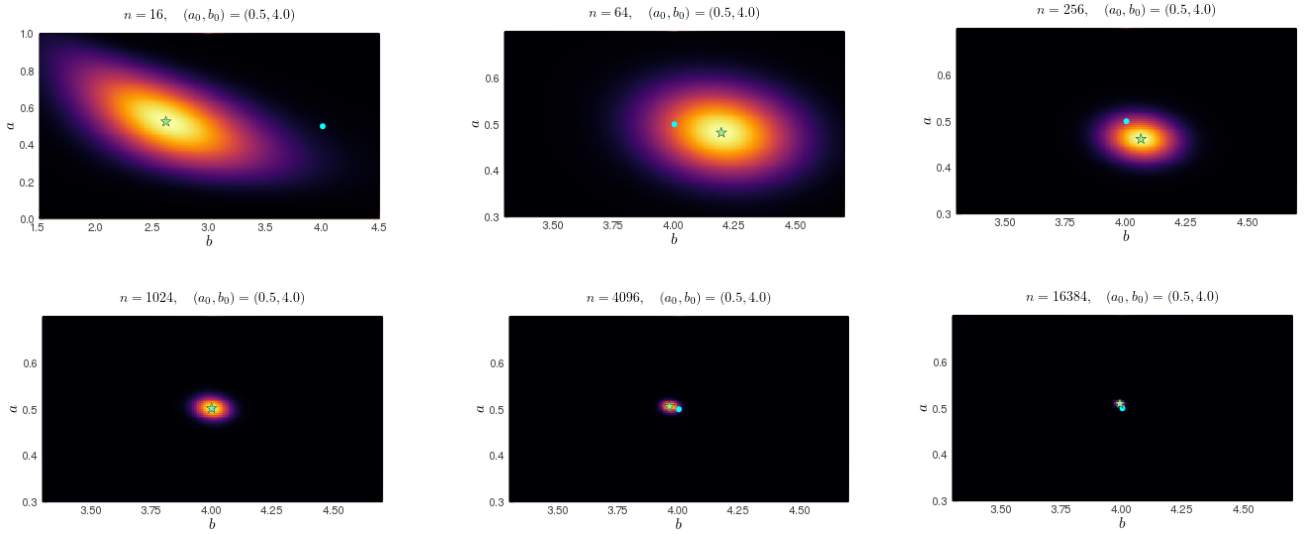
最尤法の解の星印の位置は $n \to \infty$ で真の値 $(0.5, 0.5)$ に収束する. □

注意 渡辺澄夫氏によって, 特異モデルの場合も含めて Bayes 推測の $n \to \infty$ での漸近挙動がわかっている. 本質的に広中の特異点解消定理を使う. そこでの特異点解消の使い方は M. Atiyah 氏による「特異点解消は具体的な積分の計算にも役に立つ」という指摘で拡まったものである.

M. F. Atiyah, Resolution of Singularities and Division of Distributions, 1970.
https://doi.org/10.1002/cpa.3160230202

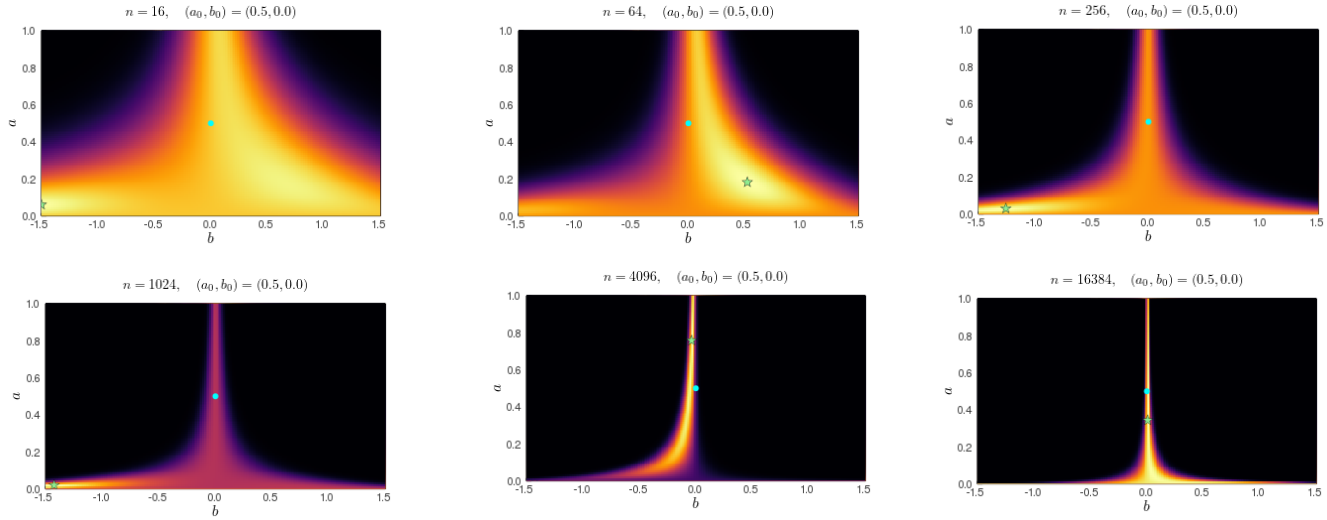
□

例（より完全に正則モデルな場合）



最尤法の解の星印は正則モデルでは $n \to \infty$ で真の値に収束する。 □

例（ぴったり特異モデルの場合）



$n \to \infty$ で尤度が $ab = 0$ の近くに集中して来ている。 □

情報量規準：自由エネルギー

Bayes推論で用いる数学的モデル内で，サンプル $X_1, \cdots, X_n$ は密度函数

$$Z_n(x_1, \cdots, x_n) = \int \varphi(w)\, p(x_1|w) \cdots p(x_n|w)\, dw$$

を持つ確率分布で生成されるのであった．一方，母集団分布 $q(x)$ のサンプルは密度函数

$$Q_n(x_1, \cdots, x_n) = q(x_1) \cdots q(x_n)$$

を持つ確率分布で生成される．分布 Z_n による分布 Q_n の予測誤差は

$$D(Q_n \| Z_n) = G(Q_n \| Z_n) - S(Q_n),$$

と表される，ここで，

$$G(Q_n \| Z_n) = \int\cdots\int Q_n\, (-\log Z_n)\, dx_1 \cdots dx_n = (-\log Z_n \text{の} Q_n \text{に関する平均}), \quad S(Q_n) = G(Q_n \| Q_n).$$

$G(Q_n \| Z_n)$ が小さな Z_n ほど Q_n の予測における誤差が小さくなる．

しかし，統計学を使用する状況では母集団分布 $q(x)$ は未知であり，分布 Q_n を用いた平均を計算することはできない，

分布 Q_n で生成されたサンプル（の実現値）$X_1, \cdots, X_n$ のみを我々は使用できる，

$G(Q_n \| Z_n) = (-\log Z_n \text{の} Q_n \text{に関する平均})$ の推定値として，$-\log Z_n$ の実現値

$$F_n = -\log Z_n(X_1, \cdots, X_n) \quad (X_1, \cdots, X_n \text{は} Q_n \text{のサンプルの実現値})$$

を用いることが考えられる．平均（期待値）の推定値として実現値を用いることは，かなり大胆な発想だが，F_n の漸近挙動より，誤差の小さなモデルの選択のために十分に役に立つことが知られている．この F_n を Bayes 自由エネルギー と呼ぶ，

定理（渡辺澄夫） $D(q \| p(\cdot|w))$ を最小にする w を w_0 と書き，いくつか適当な条件を仮定すると，ある正実数 λ と正の整数 m が存在して，

$$F_n = nL_n(w_0) + \lambda \log n - (m-1) \log\log n + O(1),$$

ここで，

$$L_n(w) = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i|w) \quad (\text{対数尤度の} -\tfrac{1}{n} \text{倍}).$$

□

注意 $q(x)$ のサンプル $X_1, \cdots, X_n$ に依存するのは F_n，$nL_n(w_0)$，$O(1)$ の部分である．

仮に，2つの確率モデル $p_1(x|w)$ と $p_2(x|w)$ の w_0 をそれぞれ $w_0^{(1)}$，$w_0^{(2)}$ と書くとき，$p_1(x|w_0^{(1)}) = p_2(x|w_0^{(2)})$ となっているならば，$nL_n(w_0)$ の項は2つのモデルで同じになり，n が十分大きければ，F_n の大小関係は，モデルから決まる定数 λ, m だけで決まることになる．そして，$nL_0(w_0)$ の項は $O(n)$ のオーダーなので，$p_1(x|w_0^{(1)}) \neq p_2(x|w_0^{(2)})$ ならば $nL_0(w_0)$ の項の差は大きくなる可能性が高い，以上が自由エネルギーがモデル選択に役に立つ理由である．□

注意 しかし，自由エネルギーの数値計算は一般に容易ではない！ □

情報量規準：WAIC

Bayes推論では分布$q(x)$のサンプル$X_1, \cdots, X_n$から予測分布

$$p_n^*(x) = \int \varphi_n^*(w)\, p(x|w)\, dw, \quad \varphi_n^*(w) = \frac{\varphi(w)\, p(X_1|w) \cdots p(X_n|w)}{Z(X_1, \cdots, X_n)}$$

が得られる．このとき，予測分布$p^*(x)$による$q(x)$の予測の誤差は

$$D(q\|p_n^*) = G(q\|p_n^*) - S(q) \quad \left(G(q\|p_n^*) = -\int q(x) \log p_n^*(x)\, dx,\ S(q) = G(q\|q)\right)$$

と表わされる．汎化誤差$G(q\|p_n^*)$が小さいほど予測誤差は小さくなる．しかし，$G(q\|p_n^*)$は未知の分布$q(x)$に関する$-\log p_n^*(x)$の平均であり，直接計算することができない．大数の法則を使えば汎化誤差$G(q\|p_n^*)$を学習誤差 (training error)

$$T_n = -\frac{1}{n}\sum_{k=1}^{n} \log p_n^*(X_k) = -\frac{1}{n}\sum_{k=1}^{n} \log \int \varphi_n^*(w)\, p(X_k|w)\, dw$$

で近似できる．この近似による誤差は$O(1/n)$のオーダーになるが，$O(1/n)$の誤差は(少なくとも正則モデルの場合には)確率モデル$p(x|w)$のパラメーター$w = (w_1, \cdots, w_d)$の個数dに比例して大きくなる．ゆえに，パラメーターの個数が異なるモデルの選択をより正しく行いたければ$O(1/n)$の項の推定量をうまく見付けてやる必要がある．

一意実現可能正則モデルのケースにおける赤池情報量規準AICでは$O(1/n)$の項の推定量を$\frac{d}{n}$と設定した：

$$aic_n = -\frac{1}{n}\sum_{k=1}^{n} \log p(X_k|\hat{w}_n) + \frac{d}{n}, \quad \hat{w}_n \text{は最尤法の解．}$$

ただし，統計学ではχ^2検定との相性の良さを優先して，これの$2n$倍をAICとすることが多い：

$$AIC_n = -2\sum_{k=1}^{n} \log p(X_k|\hat{w}) + 2d.$$

(パラメーターの個数の分だけペナルティーを課すだけなのでわかりやすい．)

特異モデルのBayes統計でも使える$O(1/n)$の項の推定量を見付けたのが，渡辺澄夫による widely applicable information criterion WAIC である：

$$waic_n = T_n + \frac{V_n}{n}.$$

ここで，

$$V_n = \sum_{k=1}^{n}\left(\int \varphi_n^*(w) (\log p(X_k|w))^2 dw - \left(\int \varphi_n^*(w) \log p(X_k|w)\, dw\right)^2\right)$$

$$= \sum_{k=1}^{n} \left(\text{事後分布 } \varphi_n^*(w) \text{に関する } \log p(X_k|w) \text{の分散}\right).$$

$WAIC_n = 2n\, waic_n$とおくと，

$$WAIC_n = -2\sum_{k=1}^{n}\Big(\log\left(\text{事後分布 } \varphi_n^*(w) \text{に関する } p(X_k|w) \text{の平均}\right) + \left(\text{事後分布 } \varphi_n^*(w) \text{に関する } \log p(X_k|w) \text{の分散}\right)\Big).$$

定理 (渡辺澄夫) 分布 $q(x)$ のサンプルを動かす平均を $E[\ ]$ と書くと,

$$E[G(q\|p_n^*)] = E[\mathrm{waic}_n] + O\left(\frac{1}{n^2}\right).$$

サンプルを動かす平均を取れば汎化誤差と WAIC は $O(1/n^2)$ の誤差を除いて一致する. □

注意 waic_n は直接 $G(q\|p_n^*)$ の推定量として使われ, その場合の推定誤差は $n\to\infty$ としても $O\left(\frac{1}{n^2}\right)$ で小さくはならない. $O\left(\frac{1}{n}\right)$ の誤差が残る! □

定理 (渡辺澄夫) $L_n(w) = -\frac{1}{n}\sum_{k=1}^{n}\log p(X_k|w)$ とおくと,

$$\left(G(q\|p_n^*) - G(q\|p(\cdot|w_0))\right) + \left(\mathrm{waic}_n - L_n(w_0)\right) = \frac{2\lambda}{n} + o\left(\frac{1}{n}\right).$$

他にνという実数有理不変量もある.

λ はモデルの実数有理不変量,

両辺を $2n$ 倍すると,

$$2n\left(G(q\|p_n^*) - G(q\|p(\cdot|w_0))\right) + \left(\mathrm{WAIC}_n - 2\sum_{k=1}^{n}(-\log p(X_k|w_0))\right) = 4\lambda + o(1).$$ □

注意 おおざっぱに言えば, これは予測分布 $p_n^*(x)$ の汎化誤差とその推定量である waic_n の $O\left(\frac{1}{n}\right)$ の項が逆相関するということである!

WAIC は, 2つのモデルの waic_n が小さな方が予測分布の予測誤差が小さいと判断するために使われるのだが, 上の逆相関性のせいで $n\to\infty$ としてもその判断が間違っている確率が 0 に収束しない場合が出て来てしまう. AIC についても同様であり, AIC や WAIC の利用者はそれらが非常に繊細な道具であることを理解して使う必要がある. □

注意 $\mathrm{waic}_n - L_n(w_0)$ の部分は次のように書き直される.

$$\mathrm{waic}_n - L_n(w_0) = \begin{pmatrix}\text{パラメーター } w \text{ を持つ}\\ \text{確率モデル } p(x|w)\\ \text{の } \mathrm{waic}_n\end{pmatrix} - \begin{pmatrix}\text{パラメーターが固定されていて}\\ \text{パラメーターを何も持たない}\\ \text{確率モデル } p(x|w_0)\\ \text{の } \mathrm{waic}_n\end{pmatrix}.$$

パラメーターが固定されているとき, 事後分布による平均は平均を取る前の値そのものになり, 分散は 0 になるので $L_n(w_0)$ はモデル $p(x|w_0)$ の waic_n とみなされる. □

例 $D(q|p(\cdot|w))$ を最小化する $w=w_0$ が1つで $w=w_0$ でその Hessian が退化していないとき (正則モデルの場合に), $w=(w_1,\dots,w_d)$ (パラメーターが d 個) ならば $\lambda=\frac{d}{2}$, $m=1$ となる. 特異モデルで λ は $\frac{d}{2}$ より小さくなる. 特異モデルのケースの方が Bayes 推論の収束は速くなる. □

汎化誤差とWAICの"逆相関"は次の定理から導かれる.

定理 (渡辺澄夫) サンプルに依存するある確率変数 E_n が存在して,

$$G(q\|p_n^*) - G(q\|p(\cdot|w_0)) = \frac{1}{n}\left(\lambda + \frac{1}{2}E_n - \frac{1}{2}V_n\right) + o\left(\frac{1}{n}\right),$$
$$\mathrm{waic}_n - L_n(w_0) = \frac{1}{n}\left(\lambda - \frac{1}{2}E_n + \frac{1}{2}V_n\right) + o\left(\frac{1}{n}\right). \qquad \square$$

この定理はパラメーター z, w を持つ確率モデル $p(x|z,w)$ について, $G(q\|p(\cdot|z,w))$ を最小化するパラメーターが z_0, w_0 のとき, $p^{(1)}(x|z,w) = p(x|z,w)$, $p^{(0)}(x|z) = p(x|z,w_0)$ とおくと,

$$G(q\|p_n^{(1)*}) - G(q\|p_n^{(0)*}) = \frac{1}{n}\left(\lambda + \frac{1}{2}E_n - \frac{1}{2}V_n\right) + o\left(\frac{1}{n}\right)$$
$$\mathrm{waic}_n(p^{(1)}) - \mathrm{waic}(p^{(0)}) = \frac{1}{n}\left(\lambda - \frac{1}{2}E_n + \frac{1}{2}V_n\right) + o\left(\frac{1}{n}\right)$$

$\begin{cases} z = (z_1, \cdots, z_{d_0}) \\ w = (w_{d_0+1}, \cdots, w_{d_1}) \end{cases}$

の形に拡張されるように思われる. $p^{(1)}$ が正則モデルならば $\lambda = \frac{d_1 - d_0}{2}$ となるだろう.

情報量規準 LOOCV

$X_1, \dots, X_n$ から X_m を除いたサンプルから作った予測分布に X_m を代入したものは

$$p^*_{n,m}(X_m) = \frac{Z(X_1, \dots, X_m, \dots, X_n)}{Z(X_1, \dots, \not{X_m}, \dots, X_n)} = \frac{\int \varphi(w) p(X_1|w) \cdots p(X_n|w)\, dw}{\int \varphi(w) p(X_1|w) \cdots p(X_n|w)\, p(X_m|w)^{-1}\, dw}$$

$$= \left(X_1, \dots, X_n \text{から作った事後分布に関する } p(X_m|w)^{-1} \text{の平均} \right)^{-1}$$

$$= \left(\int p(X_m|w)^{-1} \varphi^*_n(w)\, dw \right)^{-1}$$

これの対数の -1 倍の m を動かした平均を LOOCV (leave-one-out cross validation) と呼ぶ：

$$\mathrm{loocv}_n = -\frac{1}{n} \sum_{m=1}^{n} \log p_{n,m}(X_m) = \frac{1}{n} \sum_{m=1}^{n} \log \int p(X_m|w)^{-1} \varphi^*_n(w)\, dw.$$

これの $X_1, \dots, X_n$ を動かした平均は $G(q \| p^*_{n-1})$ の $X_1, \dots, X_n$ を動かした平均と一致している.

定理 (渡辺澄夫) $\mathrm{loocv}_n = \mathrm{waic}_n + O\left(\frac{1}{n^2}\right)$. □

loocv_n は waic_n と同様の使い方をできる.