ベイズ統計の基礎定理

渡辺澄夫
東京工業大学

このPDFをごらんいただき、ありがとうございます。
ここではベイズ統計の基礎定理をまとめています。

◎　ベイズ統計には数学的な法則があります。この法則は普遍的なものであり、人間の主観や思惑には依存しません。法則は、どんな（真の分布、確率モデル、事前分布）に対しても成り立ちます。ベイズ統計において確率モデルや事前分布の客観性を知りたいときには、このPDFに記載されている定理を参考にしてみてください。

◎　ここに記載されていることは、すぐに実問題へ応用できます。コーディングも容易です。ぜひ一度、お試しください。研究の世界的な状況をお知りになりたいかたはキーワード「 waic statistics 」で検索をしてみてください。

記号

基本的な記号の意味は次の通りです。

真の分布　$q(x)$　$(x \in R^N)$

独立データ　$X^n=(X_1,X_2,\ldots,X_n) \sim q(x)$

確率モデル　$p(x|w)$　$(w \in R^d)$

事前分布　$\varphi(w)$

ベイズ事後分布の定義　および　事後分布による平均と分散の記号　は次の通りです。

事後分布　$p(w|X^n) = \frac{1}{Z} \varphi(w) \prod_{i=1}^{n} p(X_i|w)$

事後分布による平均と分散を　$E_w[\ ]$, $V_w[\ ]$ で表す。

事後分布の作り方

このPDFでは、事後分布による平均　$E_w[\ ]$　を算出する方法についての詳細は説明しません。

平均の計算ができればそれを用いて分散 $V_w[\ ]$ の計算も行うことができます。

（事後平均の計算法） パラメータの集合 $\{ w_k ; k=1,2,3,\ldots,K \}$ で、任意の関数 f(w) について

$$E_w[f(w)] \fallingdotseq \frac{1}{K} \sum_{k=1}^{K} f(w_k)$$

が成り立つものを得るために、マルコフ連鎖モンテカルロ(MCMC)法がしばしば利用されます。代表的なMCMC法として、メトロポリス法、ギブスサンプリング法、ハミルトニアン法、ランジュバン法などがあります。いくつかの方法の組み合わせについてもよく研究されています。またそれらを実装したソフトウエアとして代表的なものに BUGS, STAN, JAGS などがあります。ネットワークを検索すると見つかると思います。

確率モデルと事前分布の客観評価

(1) 現実のほとんどの問題において真の分布 q(x) は不明です。真の分布から得られたサンプル X^n の実現値に対して、データ分析者は確率モデルと事前分布(p,φ)を準備します。データに関連がある様々な予備知識（自然科学、人文社会科学、関係ある統計量など）をできる限り活用して(p,φ)を検討します。その際に（p,φ）が未知の分布 q(x) に対して適切であるかどうかを客観的に調べたくなることがあります。

(2) 確率モデルと事前分布 (p,φ) の客観評価に用いられる指標として代表的なものに予測誤差と対数周辺尤度があります。その定義は次ページ以下に書いてあります。

(3) 予測誤差と対数周辺尤度はどちらもサンプル X^n の関数であり、確率変数です。統計学としての重要な課題に、(a) 予測誤差と対数周辺尤度がどのような挙動を持つのかを解明すること、および (b) 現実の問題においてそれらの値を算出するための方法を与えること、があります。このPDFでは、この二つの課題を解決するために必要な定理が記載されています。

指標1　テスト時の誤差

予測分布の定義　$p^*(x) = E_w[\, p(x|w) \,]$

予測分布は、ベイズ法において真の分布を推測した分布（ベイズ学習の結果）を表しています。

予測分布 $p^*(x)$ は真の分布 $q(x)$ をどのくらい正確に推測しているでしょうか。

学習誤差の定義　$T = -\frac{1}{n} \sum_{i=1}^{n} \log p^*(X_i)$

予測誤差の定義　$G = -E_X[\, \log p^*(X) \,]$

G が小さいほど真の分布 q(x) と予測分布 p*(x) のカルバック・ライブラ距離が小さいので、G がわかれば学習モデルの観測や設計に役立ちます。しかしながら、真の分布が不明なので G の値を直接に知ることは出来ません。どうしたら G を知ることが出来るでしょうか。学習誤差 T は G の近似値を与えますが、T は確率モデルが複雑であるほど小さな値になるため T を最小化することで確率モデルや事前分布を最適化することはできません。

指標2　モデルと事前分布の尤度

モデルと事前分布が与えられたときデータが得られる確率（周辺尤度）は

$$\mathrm{Prob}(X^n|p,\varphi) = Z = \int \varphi(w) \prod_{i=1}^{n} p(X_i|w)\ dw$$

その対数値にマイナスをつけたものは

$$F = -\log \int \varphi(w) \prod_{i=1}^{n} p(X_i|w)\ dw$$

F を対数周辺尤度、ベイズ符号長、確率的複雑さ、自由エネルギーといいます。統計学、情報理論、統計力学において重要な役割を果たすことが知られています。
F が小さいほどデータが与えられたときの(p,φ)の尤度が大きいので、F がわかれば学習モデルの観測や設計に役立ちます。しかしながら、F の値は容易には計算できません。
どうしたら F の値を求めることができるでしょうか。

G と F の漸近挙動および計算法

目的: 予測誤差 G と対数周辺尤度 F を求めることが目的です。

(1) 漸近挙動: (q,p,φ)が与えられたとき、サンプル数 n が大きいときの
F と G の漸近挙動を導出してください。

(2) 推定法: (p,φ,X^n)が与えられたとき、F と G の計算法を作ってください。

(3) 条件の相違:尤度関数がガウス近似できる場合と、できるとは限らない場合の
二つの条件のもとで(後者は前者を含みます)理論と計算法を望みます。

すなわち （F,G)×(漸近挙動,推定法)×(ガウス近似, 一般） の8通りの定理があります。

このPDFでは、その8通りの定理を述べます。

予測誤差の漸近挙動(ガウス近似)

経験対数損失関数の定義　$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|w)$

平均対数損失関数の定義　$L(w) = -E_X[\log p(X|w)]$

最適パラメータ集合の定義　$W_0 = \{ w ; L(w) \text{ は最小値} \}$,　$w_0 \in W_0$

予測誤差:漸近挙動:ガウス近似

定理①. 尤度関数がガウス近似できるとき dim(w) = d とし、E[ ]をサンプルセットの出方についての平均値とすると

$$E[G] = L(w_0) + d/(2n) + o(1/n)$$

(注)この定理はベイズ予測誤差について述べたものです。ガウス近似できれば定理①は真の分布がモデルで実現可能でなくても成立します。実現可能でないとき最尤とベイズで G は異なります。なお、「実現可能」とは $q(x) = p(x|w_0)$ が成り立つことです。

予測誤差の推測(ガウス近似)

◎ 赤池情報量規準 の定義(赤池, 1974) w* を最尤推定量として

$$AIC = L_n(w^*) + d/n$$

◎ 偏差情報量規準の定義 (Spiegelhalter 他, 2002)

$$DIC = L_n(E_w[w]) + 2\{ E_w[L_n(w)] - L_n(E_w[w]) \}$$

予測誤差:推定法:ガウス近似

定理②. 尤度関数がガウス近似でき、真の分布がモデルで実現可能であれば

$$E[G] = E[AIC] + o(1/n)$$
$$E[G] = E[DIC] + o(1/n)$$

(注)ガウス近似できても真の分布がモデルで実現可能でないときは定理②の式は両方とも成立しません。

実対数閾値(RLCT)

定義. $K(w) = L(p_{w_0}||p_w)$: $p(x|w_0)$ と $p(x|w)$ のカルバック距離で w の関数です。

$$\zeta(z) = \int K(w)^z \varphi(w)\, dw$$

は複素平面全体に一意に解析接続される有理型関数です。その極は全て実数であり負の有理数です。最大の極を $(-\lambda)$ とし、その位数を m とします。どちらも (q,p,φ) により定まります。定数 λ のことを実対数閾値と呼びます。

（余談） このページが理解できなくても次ページ以下は理解できます（以下は数学に関心があるかたのための注です。関心がなければお読みになる必要はありません）。$\zeta(z)$はゼータ関数の一種です。純粋数学が現実の問題解決に決定的であるという典型的な例です。この概念は次の数学者のかたにより発見され研究されてきました。Gel'fand, Shilov, 広中, Atiyah, Igusa, 佐藤, 新谷, 柏原, Arnold, Varchenko, 高山, 大阿久, 斎藤, Kollar, Mustata, … 代数幾何、代数解析、計算代数において現代も研究されている重要な概念です。この概念がなければ「GやFが漸近展開できること」自体を導出することができません。

予測誤差の漸近挙動(一般)

◎ 前ページを要するに (q,p,φ) から固有の定数 λ と m が定まります。

◎ 尤度関数がガウス近似できるときは $\lambda = d/2,\ m=1.$

◎ 簡単ではないですが，一般の場合にも数学的に値を求めることができます。

(山崎，青柳，Rusakov, Geiger, 永田, Lin, Drton, Zwiernik, Uhler)

予測誤差:漸近挙動:一般

定理③. (W., 2001) 尤度関数が正規分布で近似できてもできなくても、
真の分布がモデルで実現できてもできなくても

$$E[G] = L(w_0) + \lambda/n + o(1/n)$$

S. Watanabe, Algebraic geometry and statistical learning theory. Cambridge University Press, 2009.

予測誤差の推測(一般)

定理. 広く使える情報量規準の定義

$$\mathrm{WAIC} = T + (1/n) \sum_{i=1}^{n} V_w[\log p(X_i|w)]$$

(p,φ,X^n) が与えられれば WAIC は簡単に計算できます。

予測誤差:推定法:一般

定理④. (W.,2010) 尤度関数がガウス近似できてもできなくても, 真の分布がモデルで実現できてもできなくても

$$E[G] = E[\mathrm{WAIC}] + O(1/n^2).$$

S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, Journal of Machine Learning Research, 11, 3571-3591, 2010.

対数周辺尤度の漸近挙動(ガウス近似)

ラプラス近似法　　n が大きくてヘッセ行列 $f''(w_0)$ が正定値ならば

$$\int \exp(-nf(w))\, dw = \frac{(2\pi)^{d/2} \exp(-nf(w_0))}{n^{d/2} \det(f''(w_0))^{1/2}}$$

対数周辺尤度:漸近挙動:ガウス近似

定理⑤. 尤度関数がガウス近似できるとき　dim(w)=d として

$$F = n L_n(w_0) + (d/2) \log n + O_p(1)$$

(注)定理⑤は真の分布がモデルで実現できなくても成立します。

対数周辺尤度の漸近挙動(ガウス近似)

定義. ベイズ情報量規準 (Schwarz, 1978)　w*を最尤推定量として

$$BIC = n\, L_n(w^*) + (d/2) \log n$$

対数周辺尤度:推定法:ガウス近似

定理⑥. 尤度関数がガウス近似できるとき

$$F = BIC + O_p(1)$$

(注)定理⑥は、真の分布がモデルで実現可能でなくても成立します。

対数周辺尤度の漸近挙動(一般)

特異積分の漸近挙動　　n が大きいとき

$$\int \exp(-nf(w))\, dw = \frac{C\, (-\log n)^{m-1} \exp(-nf(w_0))}{n^{\lambda}}$$

対数周辺尤度:漸近挙動:一般

定理⑦. (W.,2001) 尤度関数がガウス近似できてもできなくても，真の分布がモデルで実現できてもできなくても

$$F = nL_n(w_0) + \lambda \log n - (m-1)\log\log n + O_p(1)$$

S. Watanabe, Algebraic analysis for nonidentifiable learning machines. Neural Computation, 13(4), 899-933,2001

一般逆温度の事後分布

一般逆温度 β の事後分布

$$E_w{}^{\beta}[\quad] = \frac{\int [\quad] \prod_{i=1}^{n} p(X_i|w)^{\beta} \varphi(w)\, dw}{\int \prod_{i=1}^{n} p(X_i|w)^{\beta} \varphi(w)\, dw}$$

（注意）マルコフ連鎖モンテカルロ法(MCMC)を用いる場合、F は一回のMCMCでは算出できません。Fの計算を行うためには、逆温度の区間[0,1]を細かく刻んで各小区間における F の増分を算出して総和する必要があるため、非常に多くの個数の事後分布を実現する必要があります。一方，一般温度による平均は1回のMCMCで算出できます。

対数周辺尤度の推測（一般）

定義．広く使えるベイズ情報量規準を

$$\mathrm{WBIC} = E_w^{1/\log(n)}[\, nL_n(w)\,].$$

(p,φ,X^n) が与えられれば WBIC は簡単に計算できます。

対数周辺尤度：推定法：一般

定理⑧．(W.,2013) 尤度関数がガウス近似できてもできなくても，真の分布がモデルで実現できてもできなくても

$$F = \mathrm{WBIC} + O_p((\log n)^{1/2}).$$

尤度関数がガウス近似できるときには

$$\mathrm{WBIC} = \mathrm{BIC} + o_p(1).$$

S. Watanabe, A widely applicable Bayesian information criterion, Journal of Machine Learning Research, 14, 867-897, 2013.

使用上の注意

(1) AIC, DIC, WAIC は予測誤差を推測する規準であり，真の分布がモデル候補に含まれていてサンプル数が増えても真の分布を選ぶ確率は1には近づきません（モデル選択における一致性はありません）。問題にもよりますが10%から30%くらいの確率で真の分布でないモデルが選ばれます。つまり、「真のモデルを必ず当てること」はできません。

(2)ガウス近似できるときでも WAIC は AIC や DIC よりも高精度であり，階層ベイズ法の評価に有用です（Gelman et.al., Bayesian data analysis, 3rd edition, CRC press, 2013)。

(3) 尤度関数がガウス近似できるときには、AIC,DIC,WAICはクロスバリデーションと漸近等価です。ガウス近似できないときでもクロスバリデーションとWAICは漸近等価です。WAICの理論的挙動がわかっているのでクロスバリデーションの理論的挙動も導出されます。

(4) ベイズ法でクロスバリデーションを計算するとき、事後分布を参照分布とした重点サンプリングを用いる方法があります（重点サンプリングクロスバリデーション）。この方法では事後分布による平均計算の分散が発散することがあることが知られています。WAICでは発散は起こりません。

(5) BIC, WBIC はモデル選択における一致性を持っていますが、予測損失を推測することはできません。BIC, WBIC はクロスバリデーションとは違う値になります。

(6) 実際の計算において，AIC, BIC を計算するためには最尤推定量を見つける必要があります。一方，DIC, WAIC, WBIC の計算では事後分布による平均を求める必要があります。