NIPS2007読む会

Catching Up Faster in Bayesian Model Selection and Model Averaging （論文紹介）

2008/1/18

東京大学大学院　情報理工学系研究科
数理情報学専攻 数理第四研究室
博士二年
鈴木　大慈

紹介する論文

- “Catching Up Faster in Bayesian Model Selection and Model Averaging”:

  Tim van Erven, Peter Grunwald, Steven de Rooij,

  In *proceeding of NIPS 2007.*

概要

モデル選択規準の一致性と予測性能の話

	予測性能	モデルの一致性
AIC Mallow's Cp LOOCV	○	×
BIC Bayes Factor (MDL)	×	○
本研究	○	○

前置き

予測誤差について　（ノンパラ）

- AIC,Mallows Cpは漸近的に最良の予測誤差を持つ. [Shibata, 1983]
  - regression: 真が有限次元モデルに含まれていないが次元を上げて行けば任意に近時出来る場合.

- BIC, Bayes Factorは予測誤差の収束が遅い. [Rissanen, Speed, Yu, 1992]
  - ヒストグラムによるdensity estimation, ビンの数を選択.

モデルの一致性について　（パラメトリック）

- AICにはモデルの一致性がない.

- BICにはモデルの一致性がある. (例えば[Barron, Rissanen, Yu,1998]は参考になるかも)

興味深い事実

Regression：パラメトリック，真を含む

$$Y_i = \beta_k^{\mathrm{T}} \varphi_k(X_i) + \epsilon_i$$

正規分布

(ある条件の下)
最尤推定する場合，
全ての一致性を持つモデル選択規準は
mini-max rateを達成しない．[Yang, 2005]

モデル選択をする場合，AICとBICの性質は両立しない．
(BICは縮小し過ぎている)

リスク

$X^n = (X_1, X_2, \ldots, X_n)$:nサンプル

$R_n(P^*, \hat{P}) = \mathrm{E}_{X^{n-1} \sim P^*}[D(P(X_n|X^{n-1})||\hat{P}(X_n|X^{n-1}))]$ **:真とのKL-Div**

$\hat{P}(X_n|X^{n-1})$ **:データ X^{n-1} から構成した予測分布**

Accumulated risk

$$\sum_{i=1}^{n} R_i(P^*, \hat{P}) = \mathrm{E}_{X^n \sim P^*}[-\log \hat{p}(X^n) + \log p^*(X^n)]$$

Stochastic Complexity

Catch-Up Phenomenon

$\mathcal{M}_k$: モデルk (k=1, 2), M1の方が単純なモデル

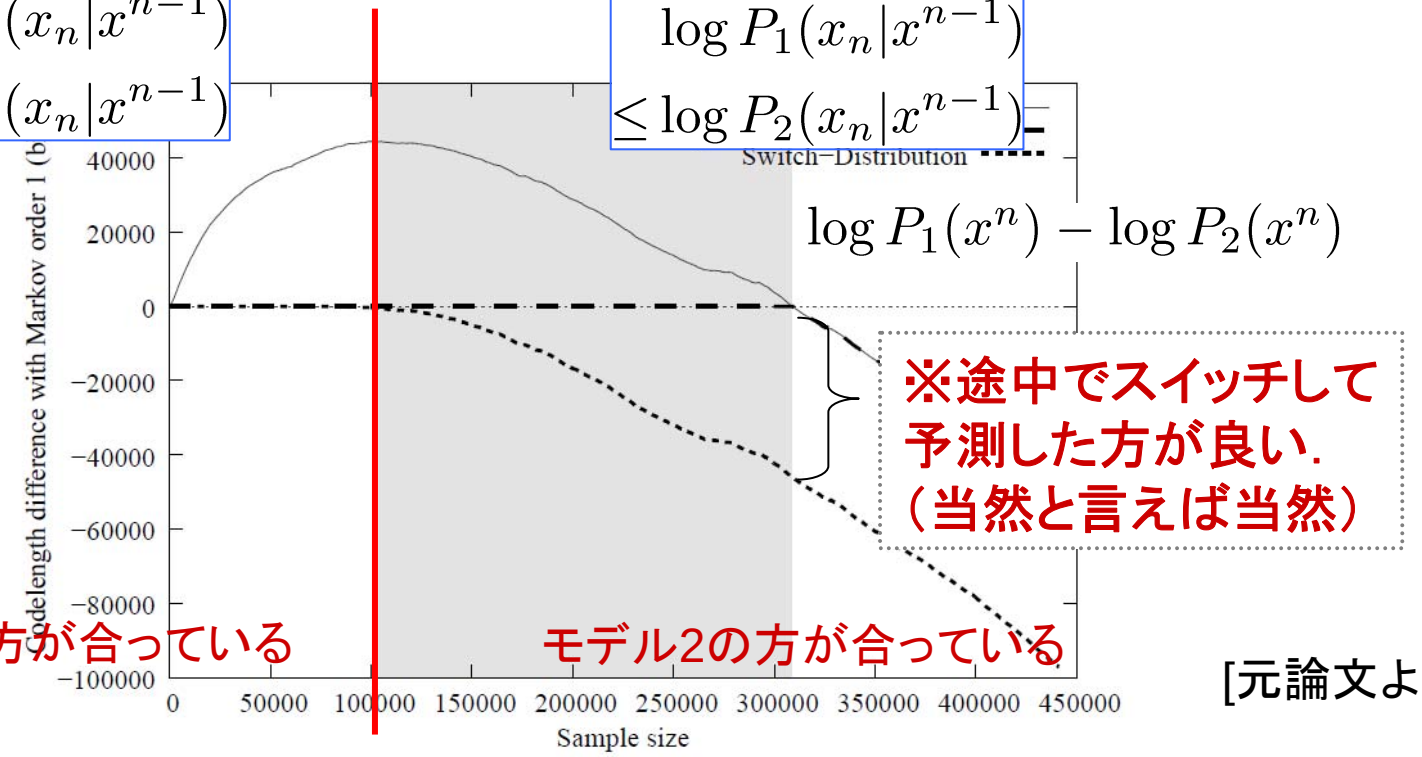
$$\log P_k(x^n) = \log P_k(x_1) + \log P_k(x_2|x^1) + \cdots + \log P_k(x_n|x^{n-1})$$

$\log \sum_k w(k) P_k(x^n)$: Bayes Model Averaging (普通のベイズ)

$$\log P(x^n|s) = \sum_{i=1}^{100000} \log P_1(x_i|x^{i-1}) + \sum_{i=100001}^{n} \log P_2(x_i|x^{i-1})$$

$\log P_1(x_n|x^{n-1}) \geq \log P_2(x_n|x^{n-1})$

$\log P_1(x_n|x^{n-1}) \leq \log P_2(x_n|x^{n-1})$



[元論文より]

モデル –Switch Distribution–

どこでスイッチすべきか分からない → モデルの列に事前分布

$\mathcal{M}_1, \mathcal{M}_2, \ldots$: 可算個のモデルの列

$p_k(x_n|x^{n-1})$: k番目のモデルの予測分布

$s = (k_1, k_2, k_3, \ldots)$:モデルの番号の列

$\pi(s)$:モデルの列の事前分布

$$q_s(x^n) = \prod_{i=1}^{n} p_{k_i}(x_i|x^{i-1})$$

$$p_{\mathrm{sw}}(x^n) = \sum_s q_s(x^n)\pi(s) \quad \text{(Switch Distribution)}$$

※ポイント：モデルにではなく，モデルの列に事前分布を入れる．
（サンプル数に合わせてモデルの事前分布が変わっている）

定理1 – Consistency

$$\pi(K_{n+1}=k|x^n)=\frac{\sum_{s:s_{n+1}=k}\pi(s)q_s(x^n)}{p_{\mathrm{sw}}(x^n)}$$

次にどのモデルを用いるかの事後分布（動的計画法で求められる）

$P_{\theta^*}\in\mathcal{M}_{k^*}$ ：真の分布

どのモデルも異なっていて，あまりスイッチしないような事前分布なら

$$\pi(K_{n+1}=k^*|X^n)\overset{n\to\infty}{\longrightarrow}1 \qquad P_{\theta^*}\text{-a.s.}$$

モデルの事後分布には一致性がある.

（パラメトリックな話：真が我々の用意するモデルに入っている）

定理2 – minimax rate

$$\mathcal{M}^* = \overline{\bigcup_k \mathcal{M}_k}$$ (KL-Divergenceによる閉包，詳細略)

$$h(n) = \inf_{\delta:\text{model selection}} \sup_{P^* \in \mathcal{M}^*} \sup_{n' \geq n} R_{n'}(P^*, P_\delta)$$

全てのモデル選択規準の中でのmini-max risk

事前分布が良い性質を持っていて $nh(n)/(\log n)^2 \to \infty$ なら，

$$\limsup_{n \to \infty} \frac{\sup_{P^* \in \mathcal{M}^*} \sum_{i=1} R_i(P^*, P_{\mathrm{sw}})}{\sum_{i=1}^n h(i)} \leq 1$$

$nh(n)/(\log n)^2 \to \infty$: どのモデルにも含まれない $P^* \in \mathcal{M}^*$ が存在すれば成り立つ．

（ノンパラメトリックな話）

実は

- 真がある M_k に含まれている場合，前述の定理が成り立つか分からない.

  （問題が簡単すぎる）

- Accumulated Riskではなくて一期先予測の誤差についてはまだ分からない.

まとめ

- モデルの事前分布の代わりにモデルの列に事前分布を考えた.

  サンプル数に応じてモデルの事前分布が変わっていることがポイント.

→　モデルの一致性を保ちつつ,

（ノンパラの場合の）mini-max rateを達成.

関連する参考文献

- Y. Yang. Can the strengths of AIC and BIC be shared? *Biometrica*, 92(4):937–950, 2005.
- A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE T. Inform. Theory*, 44(6):2743–2760, 1998
- R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *Ann. I. Stat. Math.*, 35: 415–423, 1983.
- H. Akaike. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, 66(2): 237–242, 1979.