Catoni 流の帰納的 PAC-Bayesian 学習に関する一考察

綾野 孝則、鈴木譲

大阪大学 大学院理学研究科 数学専攻

統計数理研究所
2010 年 3 月 29 日 (月)

PAC-Bayesian Supervised Classification - The Thermodynamics of Statistical Learning - (Olivier Catoni, 2007)

$\omega := \{(X_i, Y_i)\}_{i=1}^N$: (パターン, ラベル) の独立に発生した N 個の例
ω の確率測度 $\mathbb{P}$: 未知
f_θ: パターン $\mapsto$ ラベル

$r(\theta, \omega)$: 例から計算した f_θ の誤り率
$R(\theta)$: f_θ の平均の誤り率

中心的な課題

- ω から $\inf_\theta R(\theta)$ に到達する θ を推定
- $\inf_\theta r(\omega, \theta)$ から $\inf_\theta R(\theta)$ を評価

Catoni(2007) のオリジナル的視点

- 推定 $\hat{\theta}(\omega)$ というよりは、事後確率 $\rho(\cdot|\omega)$ の推定
- ρ を $\mathbb{P}$ で平均しても、 事前確率 π に一致しない
- 自由エネルギー

$$-\frac{1}{\lambda}\log\{\int\int \exp[-\lambda r(\theta,\omega)]\pi(d\theta)\mathbb{P}(d\omega)\}$$

λ: 絶対温度の逆数
を用いて、$\inf_\theta r(\omega,\theta)$ を $\inf_\theta R(\theta)$ と関連付ける

- π を Gibbs 分布に一般化 (Localization) して、誤り率の上界を tight に

Radon-Nikodym の定理

$(\Omega, \mathcal{F})$：可測空間
μ, ν: $(\Omega, \mathcal{F})$ の測度

μ が ν に対して絶対連続 $\mu \ll \nu$

$\nu(A) = 0 \Longrightarrow \mu(A) = 0, \quad \forall A \in \mathcal{F}$

Radon-Nikodym の定理

$\mu \ll \nu$ のとき、$\mu(A) = \int_A f d\nu \ \ \forall A \in \mathcal{F}$ なる可測な $f : \Omega \to \mathbb{R}$ が存在

$\dfrac{d\mu}{d\nu} := f$ は ν 測度 0 の点を除いて一意的 (Radon-Nikodym 微分)

Radon-Nikodym の定理 (続)

離散

Ω: 可算集合
$\mathcal{F}$: べき集合からなる
$P(x) := \mu(\{x\})$, $Q(x) := \nu(\{x\})$ (確率)

$$0 \neq Q(x) \Longrightarrow \frac{d\mu}{d\nu} = \frac{P(x)}{Q(x)}$$

連続で、確率密度関数が存在

$f := \frac{d\mu}{d\lambda}$, $g := \frac{d\nu}{d\lambda}$ (確率密度関数)

$$0 \neq g(x) \Longrightarrow \frac{d\mu}{d\nu} = \frac{f(x)}{g(x)}$$

Kullback-Leibler 情報量

Kullback-Leibler 情報量 $K(\mu, \nu)$

$$\begin{cases} \int_\Omega d\mu \log \frac{d\mu}{d\nu}, & \mu << \nu \\ \infty, & \text{otherwise} \end{cases}$$

- $K(\mu, \nu) \geq 0$
- $\mu = \nu \Longrightarrow K(\nu, \mu) = 0$

離散 : $\sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)}$

連続で、確率密度関数が存在 : $\int_\Omega dx f(x) \log \frac{f(x)}{g(x)}$

事前確率と事後確率

独立に生起する N 個の例 $\omega := \{(X_i, Y_i)\}_{i=1}^N \in \Omega$ の分布 $\mathbb{P}$

可測空間 $(\Omega, \mathcal{F})$ の測度

パラメータの事前確率 π

可測空間 $(\Theta, \mathcal{T})$ の測度

$\omega \in \Omega$ のもとでの事後確率 $\rho(\cdot|\omega)$

可測空間 $(\Theta, \mathcal{T})$ の測度 $\rho(\cdot|\omega)$

- $\rho(\cdot|\omega) \ll \pi$
- $\omega \in \Omega$ について、可測

事前確率と事後確率 (続)

記法: 可測関数 h の測度 μ による平均

$$\mu(h) := \int_{\Omega} f d\mu$$

$\mathbb{P}$ 未知ゆえ、$\pi \neq \mathbb{P}(\rho)$

$$K(\rho, \pi) := \rho[\log \frac{d\rho}{d\pi}]$$

$$\begin{aligned} \mathbb{P}[K(\rho, \pi)] &= \mathbb{P}[K(\rho, \mathbb{P}(\rho))] + K(\mathbb{P}(\rho), \pi) \\ &\geq \mathbb{P}[K(\rho, \mathbb{P}(\rho))] \\ &= I(\Omega, \Theta) \end{aligned}$$

誤り率と目標

設計時の誤り率 $r(\theta, \omega) := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[f_\theta(X_i) \neq Y_i]$

運用時の誤り率 $R(\theta) := \mathbb{P}[r(\theta, \omega)]$

目標

- $\rho(R)$ の上界 (なるべく $\mathbb{P}$ を用いないで)
- $\rho(R)$ を最小にする ρ

(最適な θ のみに確率測度 1 をおくような ρ も含む)

Kullback-Leibler 情報量と Gibbs 分布

$h : \Theta \to \mathbb{R}$ (可測関数)

$$\frac{d\pi_{exp(h)}}{d\pi} := \frac{exp(h)}{\pi[exp(h)]}$$

$(\Longleftrightarrow \pi_{exp(h)}(A) = \int_A \frac{exp(h)}{\pi[exp(h)]} d\pi,\ A \in \mathcal{T})$ とおくと、

$$\begin{aligned} log\{\pi[exp(h)]\} &= \rho(h) - K(\rho, \pi) + K(\rho, \pi_{exp(h)}) \\ &= \sup_{\rho} \rho(h) - K(\rho, \pi) \end{aligned} \tag{1}$$

Gibbs 分布

$$\frac{d\pi_{exp(-\lambda r)}}{d\pi} := \frac{exp(-\lambda r)}{\pi[exp(-\lambda r)]}$$

基本不等式: すべての導出はここから

$$\Phi_a(p) := -a^{-1} log\{1 - [1 - exp(-a)]p\} \ , \ a \in \mathbb{R}^* \ , \ p \in (0,1)$$

$$U_\lambda(\theta, \omega) := \lambda\{\Phi_{\lambda/N}[R(\theta)] - r(\theta, \omega)\}$$

(1) に代入して、

$$\mathbb{P}\{\exp[\sup_\rho \rho(U_\lambda(\theta)) - K(\rho, \pi)]\} \leq 1$$

$\lambda\Phi_{\lambda/N}$ の凸性から、

$$\mathbb{P}\{\exp[\sup_\rho \lambda(\Phi_{\lambda/N}(\rho(R)) - \rho(r)) - K(\rho, \pi)]\} \leq 1 \qquad (2)$$

Random Bound

任意の $\lambda > 0$ について

$$\begin{aligned}\mathbb{P}[\rho(R)] &\leq \Psi^{-1}_{\lambda/N}\{\mathbb{P}[\rho(r) + \frac{K(\rho,\pi)}{\lambda}]\} \\ &\leq \mathbb{P}\{\frac{\lambda}{N[1-\exp(-\frac{\lambda}{N})]}[\rho(r) + \frac{K(\rho,\pi)}{\lambda}]\}\end{aligned}$$

$N, \lambda \to \infty,\ \lambda/N \to 0$ で両辺が一致

最適な λ:

$$\lambda = \sqrt{\frac{2N\mathbb{P}[K(\rho,\pi)]}{\mathbb{P}[\rho(r)]\{1-\mathbb{P}[\rho(r)]\}}}$$

問題点

両辺に未知の $\mathbb{P}$ があるため、設計時には使えない

Non-Random Bound

任意の $\lambda > 0$ について

$$\mathbb{P}[\rho(R)] \leq \frac{1}{N[1-\exp(-\frac{\lambda}{n})]}\{\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta + \mathbb{R}[K(\rho, \pi_{\exp(-\lambda r)})]\}$$

特に、$\rho := \pi_{\exp(-\lambda r)}$ とおくと

$$\mathbb{P}[\rho(R)] \leq \frac{1}{N[1-\exp(-\frac{\lambda}{n})]}\{\int_0^\lambda \pi_{\exp(-\beta R)}(R)d\beta\}$$

$$\begin{array}{lll} & N, \lambda \to \infty,\ \lambda/N \to 0 & \\ \Longrightarrow & \dfrac{1}{N[1-\exp(-\frac{\lambda}{n})]} \to \dfrac{1}{\lambda},\ \inf_{\theta\in\Theta} \pi_{-\lambda R}(\theta) \text{ の比率} \to \infty & \\ \Longrightarrow & \text{右辺が } \inf_{\theta\in\Theta} R(\theta) \text{ に到達} & \end{array}$$

Deviation Bound: Probably Approximately Correct 的な評価

Markov の不等式:

$$\mathbb{P}[\exp(h) \geq 1] \leq \mathbb{P}[\exp(h)]$$

と基本不等式から、任意の $\lambda > 0$ について、確率少なくとも $1 - \epsilon$ で

$$\begin{aligned} \rho(R) &\leq \Psi^{-1}_{\lambda/N}[\rho(r) + \frac{K(\rho,\pi) - \log\epsilon}{\lambda}] \\ &\leq \frac{\lambda}{N[1 - \exp(-\frac{\lambda}{N})]}[\rho(r) + \frac{K(\rho,\pi) - \log\epsilon}{\lambda}] \end{aligned}$$

上界を tight にするには、(1) より、$\rho := \pi_{\exp(-\lambda r)}$

Local Bound

π を $\pi_{\exp(-\beta r)}$ $(0 \leq \beta < \lambda)$ に一般化

- $\beta = 0$ なら $\pi_{\exp(-\beta r)} = \pi$
- β の選び方によって、上界が tight になる

$$\mathbb{P}[\rho(R)] \leq \mathbb{P}\{\frac{\lambda - \beta}{N[1 - \exp(-\frac{\lambda}{N})] - \beta}[\rho(r) + \frac{K(\rho, \pi_{-\beta r})}{\lambda - \beta}]\}$$

$$\mathbb{P}[\rho(R)] \leq \frac{1}{N[1 - \exp(-\frac{\lambda}{n})] - \beta}\{\int_{\beta}^{\lambda} \pi_{\exp(-\gamma R)}(R) d\gamma + \mathbb{R}[K(\rho, \pi_{\exp(-\lambda r)})]\}$$

$$\rho(R) \quad \leq \quad \frac{K(\rho, \pi_{-\lambda r}) + \int_{\beta}^{\lambda} \pi_{\exp(-\gamma r)}(r) d\gamma - 2\log(\epsilon))}{N(2 - \exp(-\frac{\lambda}{N}) - \exp(\frac{\beta}{N})}$$

モデル選択を考慮する場合

M: 可算集合
$\Theta = \bigsqcup_{m \in M} \Theta_m$, $\Theta_m \in \mathcal{T}$,

$$\Theta_1 \subseteq \Theta_2 \subseteq \cdots$$

$$\inf_{\theta \in \Theta_1} r(\theta, \omega) \geq \inf_{\theta \in \Theta_2} r(\theta, \omega) \geq \cdots$$

$\pi(\cdot|m)$: $m \in M$ を前提とした (一般化された) 事前確率
$\rho(\cdot|m, \omega)$: $m \in M$ を前提とした (一般化された) 事後確率

$\mu(m)$: $m \in M$ の事前確率
$\nu(m|\omega)$: $\omega \in \Omega$ のもとでの $m \in M$ の事後確率

モデルの事前分布でも Localization を行う

$\eta, \xi > 0$

$$\gamma := \frac{\eta[1 - \exp(-\frac{\lambda}{N})]}{\exp(\frac{\eta}{N}) - 1}$$

$$h(m) := -\xi \int_{\beta}^{\gamma} \pi_{\exp(-\alpha\Phi_{-\eta/N} \circ R)}[\Phi_{-\eta/N} \circ R(m, \cdot)] d\alpha$$

$$\frac{d\bar{\mu}}{d\mu}(m) := \frac{\exp[-h(m)]}{\mu[\exp(-h)]}$$

$$\frac{d\bar{\pi}(\cdot|m)}{d\pi(\cdot|m)} := \frac{\exp[-\beta\Psi_{-\eta/N} \circ R(\theta)]}{\pi\{\exp[-\beta\Psi_{-\eta/N} \circ R(\theta)]|m\}}$$

2 段階の Localization

- μ を $\bar{\mu}$ に一般化 ($\xi := 0$ で $\bar{\mu} = \mu$)
- π を $\bar{\pi}$ に一般化 ($\beta := 0$ で $\bar{\pi} = \pi$)

モデルの事前分布でも Localization を行う

- $\nu := \mu_{(\frac{\exp(-\lambda r)}{\exp(-\beta r)})^{1/2}}$
- $\rho(\cdot|m) := \pi_{\exp(-\lambda r)}$

とおくと、$\nu\rho(R)$ の tight な上界が得られる

まとめ

PAC-Bayesian Supervised Classification
- The Thermodynamics of Statistical Learning -
(Olivier Catoni, 2007)

1. Inductive PAC-Bayesian learning
2. Comparing posterior distributions to Gibbs priors
3. Transductive PAC-Bayesian learning
4. Support Vector Machines

特に基本概念 (1 章) を紹介した。

手法:

- 情報理論
- 統計力学
- 強化学習