“Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization” のまとめ

冨岡 亮太
ryotat@sat.t.u-tokyo.ac.jp
T-PRIMAL NIPS 2007 勉強会

2008/01/18

紹介する文献

“Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization” by XuanLong Nguyen, Martin Wainwright, Michael Jordan

概要

- 2 組のサンプル点の集合 $\{X_i\}_{i=1}^n \sim Q$，$\{Y_j\}_{j=1}^n \sim P$[1]のみが与えられたときに P と Q の間の ϕ-ダイバージェンスを直接（分布を推定しないで）推定する．
- ダイバージェンスの中で分数になっている部分 q/p をサンプルから推定するのが難しい．⇒ 変分表現（variational representation）を使う．
- 関数空間上の最適化 ⇒ 表現定理（representer theorem）を使うと双対問題が有限 (サンプル数）次元の問題で書ける．
- KL ダイバージェンスの場合， 2 つのパラメトリゼーション (M1) $g = p/q$，(M2) $g = \log p/q$ が考えられる．
- M1 は理論的な性能保証あり．M2 の方が実験的には良い結果．

ダイバージェンスとその変分表現

[**定義**] ϕ-ダイバージェンス

$$D_\phi(P, Q) = \int p(x)\phi\left(\frac{q(x)}{p(x)}\right)\mathrm{d}\mu(x),$$

ただし ϕ は凸関数．

[**例**] KL-ダイバージェンス

$$D_K(P, Q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x).$$

ここで $\phi(u) = -\log(u)$.

[1]論文でサンプルの数が同数になっているのは本質的ではない．

ϕ は凸関数なので双対関数 ϕ^* を使って以下のように書ける．

$$\phi(u) = \sup_v (uv - \phi^*(v)).$$

これを ϕ-ダイバージェンスの定義に代入すると，以下を得る．

$$\begin{aligned} D_\phi(P,Q) &= \int p(x) \sup_{f(x)\in\mathbb{R}} \left(\frac{q(x)}{p(x)} f(x) - \phi^*(f(x)) \right) \mathrm{d}\mu(x), \\ &\geq \sup_{f\in\mathcal{F}} \left(\int f(x) \mathrm{d}Q(x) - \int \phi^*(f(x)) \mathrm{d}P(x) \right). \end{aligned}$$

ϕ の劣微分（sub gradient）を $\partial\phi$ と書くと等号成立は $\mathcal{F}$ と q/p での ϕ の劣微分が交わりを持つ，$\mathcal{F} \cap \partial\phi(q/p) \neq \emptyset$ となる[2]．

[**例**] KL-ダイバージェンスの場合

$$\phi(u) = -\log(u), \qquad \phi^*(v) = -1 - \log(-v) \Big(= \sup\left(uv + \log(u)\right) \Big).$$

$$D_K(P,Q) \geq \sup_{g\geq 0} \left(-\int g(x) \mathrm{d}Q(x) + \int \left(1 + \log g(x)\right) \mathrm{d}P(x) \right)$$

ただし $(u,v) = (q/p, -g)$．等号成立条件は $\exists g = p/q$（なぜなら $\partial\phi/\partial u = -1/u$）．ちなみに $g = -f$ は単に符号を反転しただけ．

定式化 M1: $g = p/q$

$\mathcal{H}$ を RKHS とし，$w \in \mathcal{H}$，RKHS の再生性より $g(x) = \langle w, \Phi(x) \rangle$ と定義する．
主問題:

$$\underset{w}{\text{minimize}} \qquad \frac{\lambda_n}{2} \|w\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n \langle w, \Phi(x_i) \rangle - \frac{1}{n} \sum_{j=1}^n \log \langle w, \Phi(y_j) \rangle .$$

補遺の公式 (5) にあてはめて考えると

$$\begin{aligned} f_0(w) &= \frac{\lambda_n}{2} \|w\|_{\mathcal{H}}^2 + \left\langle w, \tfrac{1}{n} \textstyle\sum_{i=1}^n \Phi(x_i) \right\rangle, & f_0^*(\beta) &= \frac{1}{2\lambda_n} \left\| \beta - \tfrac{1}{n} \textstyle\sum_{i=1}^n \Phi(x_i) \right\|_{\mathcal{H}}^2, \\ f_i(z) &= -\frac{1}{n} \log(z), & f_i^*(\alpha) &= -\frac{1}{n}(1 + \log(-n\alpha)), \\ a_i &= \Phi(y_i). \end{aligned}$$

ここで，双対関数の導出に補遺の定数倍の公式 (1) および線形関数の加算の公式 (2) を使った．従って，以下のように双対問題が得られる．

$$\underset{\alpha>0}{\text{minimize}} \qquad \frac{1}{2\lambda_n} \left\| \textstyle\sum_{j=1}^n \alpha_j \Phi(y_j) - \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \right\|_{\mathcal{H}}^2 . - \frac{1}{n} \sum_{i=1}^n \left(1 + \log(n\alpha_i)\right)$$

カーネル $K(\cdot,\cdot)$ を使ってノルムの項を書き直すと論文の Lemma 3 を得る．さらに主変数 w は双対変数 α を用いて以下のように書ける．

$$w = \frac{1}{\lambda_n} \left(\sum_{j=1}^n \alpha_j \Phi(y_j) - \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \right).$$

[2] ほとんどいたるところで $(u,v) = (q(x)/p(x), f(x))$ が双対ペア条件 $u \in \partial\phi^*(v)$ および $v \in \partial\phi(u)$ をみたさなくてはいけないから．

すなわち

$$g(x) = \frac{1}{\lambda_n}\left(\sum_{j=1}^{n}\alpha_j K(y_j, x) - \frac{1}{n}\sum_{i=1}^{n} K(x_i, x)\right).$$

さらに主・双対問題の $\|\cdot\|_{\mathcal{H}}^2$ の項が等しいことを用いると

$$\begin{aligned}
\hat{D}_K &= 1 - (\text{主問題の最適値}) + \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2 \\
&= 1 + (\text{双対問題の最適値}) + \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2, \\
&= 1 - \frac{1}{n}\sum_{i=1}^{n}(1 + \log(n\alpha_i)) + \lambda_n\|w\|_{\mathcal{H}}^2, \\
&= -\frac{1}{n}\sum_{i=1}^{n}\log(n\alpha_i) + \lambda_n\|w\|_{\mathcal{H}}^2.
\end{aligned}$$

最後の項は $\lambda_n \to 0$ $(n \to \infty)$ で小さくなるとしている．

定式化 M2: $g = \log p/q$

主問題:

$$\underset{w}{\text{minimize}} \qquad \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^{n} e^{\langle w, \Phi(x_i)\rangle} - \frac{1}{n}\sum_{j=1}^{n}\langle w, \Phi(y_j)\rangle.$$

上と同様に補遺の方式 (5) にあてはめて考えると

$$\begin{aligned}
f_0(w) &= \frac{\lambda_n}{2}\|w\|_{\mathcal{H}}^2 - \left\langle w, \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\Phi(y_j)\right\rangle, & f_0^*(\beta) &= \frac{1}{2\lambda_n}\|\beta + \tfrac{1}{n}\textstyle\sum_{j=1}^{n}\Phi(y_j)\|_{\mathcal{H}}^2, \\
f_i(z) &= \frac{1}{n}e^z, & f_i^*(\alpha) &= \alpha\log(n\alpha) - \alpha, \\
a_i &= \Phi(x_i). & &
\end{aligned}$$

従って，双対問題は以下のように得られる．

$$\underset{\alpha<0}{\text{minimize}} \qquad \frac{1}{2\lambda_n}\|\textstyle\sum_{i=1}^{n}\alpha_i\Phi(x_i) + \frac{1}{n}\sum_{j=1}^{n}\Phi(y_j)\|_{\mathcal{H}}^2 + \displaystyle\sum_{i=1}^{n}(-\alpha_i\log(-n\alpha_i) + \alpha_i).$$

α の符号を反転すると以下の式を得る．(論文の p5 の最初の式)

$$\underset{\alpha>0}{\text{minimize}} \qquad \frac{1}{2\lambda_n}\|-\textstyle\sum_{i=1}^{n}\alpha_i\Phi(x_i) + \frac{1}{n}\sum_{j=1}^{n}\Phi(y_j)\|_{\mathcal{H}}^2 + \displaystyle\sum_{i=1}^{n}(\alpha_i\log(n\alpha_i) - \alpha_i).$$

このとき主変数 w は以下のように書ける．

$$w = \frac{1}{\lambda_n}\left(-\sum_{i=1}^{n}\alpha_i\Phi(x_i) + \frac{1}{n}\sum_{j=1}^{n}\Phi(y_j)\right).$$

また定式化 M1 と同様に

$$\hat{D}_K = 1 + \sum_{i=1}^{n}(\alpha_i\log(n\alpha_i) - \alpha_i) + \lambda_n\|w\|_{\mathcal{H}}^2.$$

理論的な結果

[**定理 2**] 以下の（上の議論より一般的な推定量 $\hat{g}_n$ を考える．

$$\hat{g}_n = \operatorname*{argmin}_g \int g(x)\mathrm{d}Q(x) - \int \log g(x)\mathrm{d}P(x) + \frac{\lambda_n}{2} I^2(g).$$

ここで $I(g)$ は非負の複雑さの指標とし，$\mathcal{G}_M := \{g | I(g) \leq M\}$ とする．以下の 3 つを仮定．

1. 真の $g_0 = p_0/q_0$ が上下からバウンドされている，$0 < \eta_0 \leq g_0 \leq \eta_1$ $(\exists \eta_0, \eta_1)$.
2. $I(g)$ のリプシッツ性．$\sup_{g \in \mathcal{G}_M} |g|_\infty \leq cM$ $(\forall M \geq 1)$.
3. ブラケットエントロピー $\mathcal{H}_\delta^{\mathcal{B}}(\mathcal{G}_\mathcal{M}, \mathcal{L}_\in(\mathcal{Q})) = O(M/\delta)^\gamma$.

さらに，$n \to \infty$ で λ_n を以下のオーダーで小さくする．

$$\lambda_n^{-1} = O_P(n^{2/(2+\gamma)})(1 + I(g_0)).$$

すると，g_0 と $\hat{g}_n$ のヘリンジャー距離が以下のオーダーで小さくなる．

$$h_Q(g_0, \hat{g}_n) = O_P(\lambda_n^{1/2})(1 + I(g_0)), \quad \text{また,} \quad I(\hat{g}_n) = O_P(1 + I(g_0)).$$

さらに $\inf_g g(x) \geq \eta_0$ $(\forall x)$ が成り立つとき

$$|\hat{D}_K - D_K(P, Q)| = O_P(\lambda_n^{1/2})(1 + I(g_0)).$$

補足

- λ_n を小さくしていくことと $I(\hat{g}_n)$ が定数オーダーであることを $\hat{D}_K$ を計算する時に正則化項を無視するために使っている．
- 定式化 M1 では $|g(x)| = |\langle w, \Phi(x)\rangle| \leq \|w\|_\mathcal{H} \|\Phi(x)\|_\mathcal{H} \leq I(g)\sqrt{K(x,x)}$ より，例えばガウシアンカーネルなどの場合リプシッツ条件をパスする．一方定式化 M2 では $I(g) = \|\log g\|$ のため，M に対して $\sup_{g \in \mathcal{G}_M} |g|_\infty$ は指数的に増えるため，リプシッツ条件を満たさない．
- 実験では上のオーダーより速く $\lambda_n = 1/n$ で小さくしている．

補遺: Lagrange-Fenchel 双対性公式集

[**定義**]: 双対関数

$$f^*(\alpha) = \sup_x \left(\langle \alpha, x \rangle - f(x) \right).$$

定数倍:

$$(c \cdot f)^*(\alpha) = c f^*(\alpha/c). \tag{1}$$

線形関数を加える:

$$(f + \langle w, x \rangle)^*(\alpha) = f^*(\alpha - w). \tag{2}$$

$$(\mathrm{P}) \operatorname*{minimize}_x f(x) + g(x), \qquad (\mathrm{D}) \operatorname*{minimize}_\alpha f^*(\alpha) + g^*(-\alpha), \tag{3}$$

$$(\mathrm{P}) \operatorname{minimize} f_0(x) + \sum_{i=1}^n f_i(x), \qquad (\mathrm{D}) \operatorname*{minimize}_\alpha f_0^*(\sum_{i=1}^n \alpha_i) + \sum_{i=1}^n f_i^*(-\alpha_i), \tag{4}$$

$$(\mathrm{P}) \operatorname{minimize} f_0(x) + \sum_{i=1}^n f_i(\langle x, a_i \rangle), \qquad (\mathrm{D}) \operatorname*{minimize}_\alpha f_0^*(\sum_{i=1}^n \alpha_i a_i) + \sum_{i=1}^n f_i^*(-\alpha_i). \tag{5}$$