U-divergence を損失とするベイズリスクに関する考察

池田　思朗　数理・推論研究系　准教授

はじめに

$p(x)$ と $q(x)$ を確率密度関数とし，Kullback-Leibler (K-L) 情報量は以下の通りである．

$$D_{KL}(p,q)=\int_{\Re}p(x)\log\frac{p(x)}{q(x)}dx.$$

対象の確率分布は θ をパラメータとして $p(x|\theta)$ とかけることがわかっている．ただし θ の値は分らないとする．この分布に対して $q(x)$ という密度関数を用意し $p(x|\theta)$ を正しく推定したい．

パラメーター θ は1次元であるとし，ある累積密度分布 $F(\theta)$ にしたがっているとする．K-L 情報量を損失とするベイズリスクは次の通りである．

$$R(F,q)=\int_{\Re}D_{KL}(p(x|\theta);q(x))dF(\theta). \tag{1}$$

このベイズリスクを最小にする $q(x)$ は $q(x)=p(x;F)=\int_{\Re}p(x|\theta)dF(\theta)$ となる [1]．次の max-min 問題を考える [2, 3]．

$$C=\sup_{F\in\mathcal{F}}\inf_{q\in Q}R(F,q)=\sup_{F}\int_{\Re}D_{KL}(p(x|\theta);p(x;F))dF(\theta). \tag{2}$$

上式は最悪のベイズリスクを考えており，こうして求まる $F(\theta)$ はベイズ統計では Reference 事前分布 [4] と呼ばれる．一方で，この問題は情報理論における通信路容量の問題と密接な関係がある．すなわち θ を通信路への入力とし，x を通信路からの出力とすると，(2) 式は通信路容量の定義であり，最適な F は通信路容量を達成する入力の分布を与える．

通信路容量を考える際には通常入力に関するなんらかの制約を考える．多くの場合，そうした制約は以下のように書ける．

$$\mathcal{F}=\left\{F\ \middle|\ \int_{\Re}m(\theta)dF(\theta)\le A\right\}. \tag{3}$$

例えば $m(\theta)=\theta^2$ とすれば入力の平均電力を制約したことになる．

こうした制約の下で (2) 式の最適な F を考えると，次の Karush Kuhn Tucker (KKT) 条件が得られる [5, 2, 3]．

Corollary 1 (KKT 条件)**.** 最適な事前分布が F^* であるとし，F^* の増加点の集合を E^* とする．このとき，以下の関係が成り立つ．

$$D_{KL}(p(x|\theta);p(x;F^*))\begin{cases}=C+\lambda(m(\theta)-A) & \text{for } \theta\in E^*\\ \le C+\lambda(m(\theta)-A) & \text{for } \theta\notin E^*\end{cases}. \tag{4}$$

情報理論の分野ではこの結果を用い，様々なモデル $p(x|\theta)$ と制約 $m(\theta)$ の下で最適な F を求める研究が続いている [5]．一般に $\mathcal{F}$ は連続分布，離散分布の両方を許容するが，多くの通信路では現実的な制約の下で最適な入力分布が離散分布となる．連続分布が最適となる広く知られている例は，$p(x|\theta)$ が θ を平均とする分散が既知の正規分布で $m(\theta)=\theta^2$ の場合である．

U-Divergence への拡張

U-divergence [6] は以下のように定義される．

$$\begin{aligned}D_U(p,q)&=H_U(p,q)-H_U(p,p),\\ H_U(p,q)&=\int_{\mathcal{Y}}\Big[U(\xi(q(x)))-p(x)\xi(q(x))\Big]dx,\end{aligned} \tag{5}$$

U は実数上で厳密に凸な C^2 級関数とする．u は U の導関数 $u=U'$，ξ は u の逆関数 $\xi=(u)^{-1}$ である．U-divergence を損失とするベイズリスクを考える．

$$R_U(F,q)=\int_{\Re}D_U(p(x|\theta);q(x))dF(\theta). \tag{6}$$

このベイズリスクを最小にする $q(x)$ は $q(x)=p(x;F)=\int_{\Re}p(x|\theta)dF(\theta)$ となる [8]．したがって前と同様に max-min 問題を考えることができる．

$$C_U=\sup_{F\in\mathcal{F}}\inf_{q\in Q}R_U(F,q)=\sup_{F}\int_{\Re}D_U(p(x|\theta);p(x;F))dF(\theta). \tag{7}$$

(3) 式の $\mathcal{F}$ からこの問題の最適値を与える F を求めるとすると，KKT 条件は次のようになる．

Corollary 2 (KKT 条件)**.** 最適な事前分布が F^* であるとし，F^* の増加点の集合を E^* とする．このとき，以下の関係が成り立つ．

$$D_U(p(x|\theta);p(x;F^*))\begin{cases}=C_U+\lambda(m(\theta)-A) & \text{for } \theta\in E^*\\ \le C_U+\lambda(m(\theta)-A) & \text{for } \theta\notin E^*\end{cases}. \tag{8}$$

以下では β-divergence を考える．β-divergence は U-divergence において $U(z)=\frac{1}{\beta+1}(\beta z+1)^{\frac{\beta+1}{\beta}}$ として得られ，Tsallis 統計と関係していることが知られている [7]．

$$D_\beta(p,q)=\frac{1}{\beta+1}\int_{\Re}(q(x)^{\beta+1}-p(x)^{\beta+1})dx-\frac{1}{\beta}\int_{\Re}p(x)(q(x)^{\beta}-p(x)^{\beta})dx.$$

β-divergence に対して (7) 式の最適化問題を考えると，通信路容量の場合と同様に様々なモデルと制約条件の下で最適な F は離散分布となる．一方で連続分布が最適となるような組合せを考えると，Tsallis 統計で重要な q-正規分布が最適となるような制約条件は存在しない．一方，$p(x|\theta)$ を 平均 θ, 分散 σ_N^2 の正規分布, $m(\theta)$, A を次のように定めると，

$$m(\theta)=\sqrt{\frac{\sigma_N^2+\sigma_s^2}{(\beta+1)\sigma_N^2+\sigma_s^2}}\exp\Big(-\frac{\beta\theta^2}{2((\beta+1)\sigma_N^2+\sigma_s^2)}\Big),\quad A=\frac{1}{\sqrt{\beta+1}}.$$

最適な F は平均 0 分散 σ_S^2 の正規分布となり，最適値は次の通りである．

$$C_U=\frac{2\beta+1}{(2\pi)^{\beta/2}\beta(\beta+1)^{3/2}}\left(\frac{1}{(\sigma_N^2)^{\beta/2}}-\frac{1}{(\sigma_N^2+\sigma_S^2)^{\beta/2}}\right).$$

参考文献

[1] J. Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975.

[2] D. Haussler. A general minimax result for relative entropy. *IEEE Trans. Inf. Theory*, 43(4):1276–1280, 1997.

[3] P. D. Grünwald and A. P. Dawid. Game thoery, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

[4] J. M. Bernardo. Reference posterior distributions for Bayesian inference. *J. R. Statistical Society, Series B*, 41(2):113–147, 1979.

[5] J. G. Smith. The information capacity of amplitude- and variance-constrained scalar Gaussian channels. *Information and Control*, 18:203–219, 1971.

[6] M. Minami and S. Eguchi. Robust blind source separation by beta divergence. *Neural Computation*, 14:1859–1886, 2002.

[7] A. Ohara and T. Wada. Information geometry of q-Gaussian densities and behaviors of solutions to related diffusion equations. *Journal of Physics A: Mathematical and Theoretical*, 43(3):035002, 2010.

[8] N. Murata and Y. Fujimoto. Bregman divergence and density integration. *Journal of Math-for-industry*, 1(2009B-3):97–104, 2009.