振り返りと導入

前回は最尤推定量と KL ダイバージェンスを定義した。本稿では次のことを行う:

- KL ダイバージェンスの性質を調べる。
- [TODO] 一般のダイバージェンスの定義
- [TODO] ダイバージェンスから誘導される構造

1 Kullback-Leibler ダイバージェンス

定義 1.1 (Kullback-Leibler ダイバージェンス)**.** 関数 $D\colon \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X})\to[0,\infty]$,

$$D(p\|q) \coloneqq \begin{cases} E_q\left[\frac{dp}{dq}\log\frac{dp}{dq}\right] = E_p\left[\log\frac{dp}{dq}\right] & (p \ll q) \\ \infty & (p \not\ll q) \end{cases} \tag{1.1}$$

を $\mathcal{P}(\mathcal{X})$ 上の **Kullback-Leibler ダイバージェンス** と呼ぶ。

命題 1.2. $\mathcal{P}(\mathcal{X})$ に全変動で定まる位相を入れると、KL ダイバージェンスは連続とは限らない。

証明 $\mathcal{X} \coloneqq \{0,1\}$ として $p_n \coloneqq \frac{1}{n}\delta^0 + \left(1-\frac{1}{n}\right)\delta^1$, $q_n \coloneqq e^{-n}\delta^0 + (1-e^{-n})\delta^1$ が反例のひとつ。 □

命題 1.3 (指数型分布族と KL ダイバージェンス)**.** $\mathcal{P}$ を指数型分布族とする。最小次元実現 (V,T,μ) に対し対数分配関数 ψ、自然パラメータ座標 θ、期待値パラメータ座標 η を考える。このとき

$$D(p\|q) = \psi(\theta_q) + \psi^\vee(\eta_p) - \langle\theta_q,\eta_p\rangle \quad (\forall p,q\in\mathcal{P}) \tag{1.2}$$

が成り立つ。ただし $\psi^\vee$ は ψ の Legendre 変換である。

証明 Legendre 変換の定義より $\psi(\theta_p)+\psi^\vee(\eta_p) = \langle\theta_p,\eta_p\rangle$ ゆえに

$$\begin{aligned} \psi(\theta_q)+\psi^\vee(\eta_p) - \langle\theta_q,\eta_p\rangle &= \psi(\theta_q)-\psi(\theta_p)+\langle\theta_p,\eta_p\rangle - \langle\theta_q,\eta_p\rangle && (1.3)\\ &= E_p\left[\psi(\theta_q)-\psi(\theta_p)+\langle\theta_p,T\rangle - \langle\theta_q,T\rangle\right] && (1.4)\\ &= E_p\left[\log\frac{dp}{dq}\right] && (1.5)\\ &= D(p\|q) && (1.6) \end{aligned}$$

□

$\mathcal{X}=\{1,\dots,n\}, n\in\mathbb{N}$ (カテゴリカル分布) の場合に最尤推定量と KL ダイバージェンスの関係を考える。

[TODO] カテゴリカル分布でない場合は?

定義 1.4 (経験分布). $x = (x_1, \ldots, x_k) \in \mathcal{X}^k$ に対し

$$\hat{p}_x := \frac{1}{k} \sum_{i=1}^{k} \delta^{x_i} \tag{1.7}$$

を x により定まる**経験分布 (empirical distribution)** という。

命題 1.5 (最尤推定量と KL ダイバージェンス). $(\Theta, \mathbf{p})$ を $\mathcal{X}$ 上の統計モデルとし、k 個の i.i.d. 拡張 $(\Theta, \mathbf{p}^k)$ を考える。$x = (x_1, \ldots, x_k) \in \mathcal{X}^k$ とし、$\hat{p}_x$ を x により定まる経験分布とする。このとき、$\mathbf{p}^k(\Theta)$ が $\hat{p}_x$ を支配する確率測度を少なくともひとつ含むならば、次が成り立つ:

$$\operatorname{argmin}_{\theta \in \Theta} D(\hat{p}_x \| \mathbf{p}^k(\theta)) = \operatorname{argmax}_{\theta \in \Theta} p_\theta^k(x) \tag{1.8}$$

証明 [TODO] もう少し丁寧に $\forall \theta \in \Theta$ に対し

$$D(\hat{p}_x \| \mathbf{p}^k(\theta)) = E_{\hat{p}_x}\left[\log \frac{d\hat{p}_x}{d(\mathbf{p}^k(\theta))}\right] \tag{1.9}$$

$$= E_{\hat{p}_x}\left[\log \frac{d\hat{p}_x}{di}\right] - E_{\hat{p}_x}\left[\log \frac{d(\mathbf{p}^k(\theta))}{di}\right] \tag{1.10}$$

$$= (\theta \text{ によらない項}) - \frac{1}{k} \log p_\theta^k(x) \tag{1.11}$$

ゆえに命題の主張が従う。 □

今後の予定

- m-射影と最尤推定
- 混合型分布族と識別不能性

参考文献

[AJLS17] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer, **Information Geometry**, Ergebnisse der Mathematik und ihrer Grenzgebiete 34, vol. 64, Springer International Publishing, Cham, 2017.

[Ama16] Shun-ichi Amari, **Information Geometry and Its Applications**, Applied Mathematical Sciences, vol. 194, Springer Japan, Tokyo, 2016 (en).

[AN07] Shun-ichi Amari and Hiroshi Nagaoka, **Methods of Information Geometry**, Translations of Mathematical Monographs, vol. 191, American Mathematical Society, Providence, Rhode Island, April 2007 (en).