

# 共変量シフトにおける補助変数を用いた予測と情報量規準

大阪大学 大学院基礎工学研究科 井戸 貴大

大阪大学 大学院基礎工学研究科, 理化学研究所 革新知能統合研究センター 伊森 晋平

京都大学 情報学研究科, 理化学研究所 革新知能統合研究センター 下平 英寿

近年では, 本来興味のあるデータに加え, メタデータと呼ばれる様々な情報が観測されている. 例えば, 画像共有コミュニティにおいて画像に関連付けられるタグやグループ情報等がある. このようなメタデータをモデルに取り入れることでモデルの精度が上昇する研究が McAuley & Leskovec (2012) らによって行われている. また, 元々の解析対象である変数 (主要変数) の推測に補助変数 (Mercatanti et. al., 2015) と呼ばれる, 本来興味のない変数を適切にモデリングすることで混合分布のパラメータの推定精度が上昇することが知られている.

これら副次的に得られる情報を取り入れ構築したモデルを, 変数の分布が学習時とモデルの適用先のデータで異なる状況にも適用していくということを考える. 特に, 教師付き学習の枠組みにおいて共変量の分布が推定時と評価時で変化する状況は, 共変量シフト (Shimodaira, 2000) として知られている. モデルが誤特定され, かつ, 共変量シフトの設定においては, 最尤法の良さが保証されないが, 重み付き最尤法 (Maximum Weighted Log-likelihood Estimate, MWLE) によりこの問題は解消される.

本講演では, 共変量  $x$  を与えたときの目的変数  $y$  の条件付き分布  $q(y|x)$  をモデリングするために, この MWLE を応用した手法を提案する. いま, 補助変数  $a$  をモデルに取り入れているため,

$$q(y|x) = \int q(y|x, a)q(a|x)da$$

が成り立ち,  $q(y|x, a), q(a|x)$  をそれぞれパラメトリックモデル  $p(y|x, a; \alpha), p(a|x; \beta)$  で推定する. ただし,  $\alpha \in \mathbb{R}^{p_1}, \beta \in \mathbb{R}^{p_2}$  は未知パラメータであり,  $\hat{\alpha}_{w_1}, \hat{\beta}_{w_2}$  をこれらの MWLE とする. このとき, 補助変数に関して周辺化したモデル

$$p(y|x; \hat{\alpha}_{w_1}, \hat{\beta}_{w_2}) := \int p(y|x, a; \hat{\alpha}_{w_1})p(a|x; \hat{\beta}_{w_2})da$$

を考えることができる.

最尤法によって推定したモデルの評価基準である AIC は推定時と評価時でデータを生成する真の分布が不変であることを仮定しているため, 共変量シフト下ではその良さが保証されない. Shimodaira (2000) ではこの問題を克服した規準が提案されているが, 補助変数を取り込んだときのように, 推定時と評価時で観測される変数の種類が異なる状況は考慮されていない. そこで本講演では, これらの問題点を解決した新たなモデル選択規準を提案し, 数値例も交え提案手法が有用であることを示す.

## 参考文献

- [1] McAuley, J., & Leskovec, J. Image labeling on a network: using social-network metadata for image classification. Computer Vision-ECCV 2012. 2012, p. 828-841.
- [2] Mercatanti, A., Li, F., & Mealli, F. Improving inference of Gaussian mixtures using auxiliary variables. Statistical Analysis and Data Mining. 2015, Vol. 8, p. 34-48.
- [3] Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference. 2000, Vol. 90, p. 227-244.