カーネル求積に基づく浅い学習モデルの学習法

理化学研究所・革新知能統合研究センター　園田翔

特徴量写像 φ の重み付き積分 $\int \varphi \mathrm{d}\mu$ で表される学習モデル（浅い学習モデル）に対して，重みの推定と離散化を同時に行う方法を提案する．$\mathcal{X} \subset \mathbb{R}^m, \mathcal{Y} \subset \mathbb{R}$ を入出力データの空間とし，$\mathcal{A} \subset \mathbb{R}^d$ をパラメータ空間とする．$\mathcal{M}$ を $\mathcal{A}$ 上の複素数値 Radon 測度の全体とし，全変動ノルムを入れる．特徴量写像 $\varphi : \mathcal{A} \to (\mathcal{X} \to \mathcal{Y})$ のパラメータ分布 $\mu \in \mathcal{M}$ による線形結合

$$S[\mu](x) := \int_{\mathcal{A}} \varphi(x; a) \mathrm{d}\mu(a), \quad x \in \mathcal{X}$$

で表される学習モデルを浅い学習モデルという．Radon 測度は Dirac 測度を含むので，パラメータ分布として質点の重み付き和 $\mu_p = \sum_{j=1}^p w_j \delta_{a_j} (w_j \in \mathbb{C}, a_j \in \mathcal{A})$ をとることで，離散的なモデルも表現できる:

$$S[\mu_p](x) = \sum_{j=1}^p w_j \varphi(x; a_j), \quad x \in \mathcal{X}.$$

従って，基底関数（特徴量写像）φ を適切に選ぶことで，$S[\mu]$ は Fourier 基底展開（$\varphi(x; \omega) = \exp(-ix \cdot \omega)$）や，決定木（$\varphi(x; A) = 1_A(x)$），浅いニューラルネット（$\varphi(x; a, b) = \sigma(a \cdot x - b)$）などを包括的に表現できる．ただし，特徴量写像の合成写像を含むような，深層ニューラルネットは含めない．パラメータを $\boldsymbol{w}_p = (w_1, \ldots, w_p), \boldsymbol{a}_p = (a_1, \ldots, a_p)$ というベクトル形式で表すと，特徴量写像 $a \mapsto \varphi(\cdot; a)$ の非線形性によって学習問題は一般に非凸最適化問題になるが，μ_p という分布形式で表現することにより，積分作用素 $\mu \mapsto S[\mu]$ の線形性によって学習問題は（Banach 空間上の）凸最適化問題になるというメリットがある．また，パラメータの次元 p が異なるモデルも区別なく同時に扱えるというメリットもある．

このように積分表示された学習モデルに対し，準 Monte Carlo 法の一種であるカーネル求積法を応用して，有限データ $D = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ から離散モデル $S[\mu_p]$ を推定する．勾配法ではなく求積法を用いることのメリットは，逐次近似によってパラメータの次元が自動的に決定できること，元のモデルが解析的な逆変換（e.g., Fourier 変換，リッジレット変換）を持つ場合には学習済パラメータの解釈ができること，勾配法では非凸な学習問題が求積法では凸になることなどがある．カーネル求積の基本方針は，$\mathcal{A}$ 上の適当な再生核 $K : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ を用いて μ を再生核 Hilbert 空間 $\mathcal{H}_K$ に埋め込み，埋め込んだ元 $K[\mu] \in \mathcal{H}_K$ を有限和 $K[\mu_p] \in \mathcal{H}_K$ で逐次近似する．近似は RKHS ノルム $\|K[\mu_p] - K[\mu]\|_K$ を最小化する意味でとるが，これは分布間の maximum mean discrepancy (MMD) に等しいことが知られており，MMD を介して統計的な再解釈も可能である．本研究では，ユニタリ・カーネルと呼ばれるクラスのカーネルを用いることで，複素数値測度に対するカーネル求積を構成し，勾配法による大域最適解と同等な解に収束することを示した．また，二乗誤差 $L[\mu_p] := \mathbb{E}_X |S[\mu_p](X) - S[\mu](X)|^2$ に対する誤差解析の結果，汎化誤差は $O(1/p + 1/\sqrt{n})$ で押さえられることを示した．2019 年現在，カーネル求積が Monte Carlo 積分よりも速いことは極めて限定的な条件の下でしか証明されていないが，多くの数値実験では条件を満たさない場合でも $O(1/p^2)$ ないし $O(\exp(-p))$ で収束することが知られており，本研究で実施した数値実験でも $O(1/p^2)$ 相当の収束レートが得られている．

謝辞. 本研究は早稲田大学の村田昇教授，同修士課程の松原拓夫氏，東京大学・理研 AIP の鈴木大慈准教授および二反田篤史助教との議論を経て，現在の形に収束しました．皆様の建設的なご意見に感謝致します．

参考. S. Sonoda, “Unitary Kernel Quadrature for Training Parameter Distributions,” arXiv:1902.00648v2.