変数変換に基づく転移学習

南　俊匠　　総合研究大学院大学　統計科学専攻　博士課程(5年一貫制)4年

始めに

- 転移学習(transfer learning)とは，元タスクで学習した知識を別のタスクに再利用する手法であり，機械学習のフレームワークとして広く普及している.
- 特に，訓練データが限られており，ゼロからの学習が効果的ではないときに，転移学習により予測性能を大幅に向上できる可能性がある.
- 転移学習では多くの場合，元タスクの情報として元タスクのサンプルやモデルのパラメータ等の情報を使用する.
- 一方で，データサイズやプライバシーの問題から，それらの情報が使用できない場合も多い.
- 本発表では，元タスクのモデルからの出力のみを使用した転移学習を考え，そのための変数変換法の選択基準を模索する.

問題設定と学習の方針

元タスク（ソースタスク）の事前学習モデル $f_s(x)$ と，目標タスク（ターゲットタスク）の n個のサンプル $\{(x_i, y_i)\}_{i=1}^n$ を用いて，ターゲットモデル $\hat{f}_t(x)$ を学習する.

変数変換を用いた転移学習の方針

- ソースモデルの出力 $f_s(x)$ とターゲットタスクの説明変数 y を使って，新しい変数 $z = \phi(y, f_s(x))$ を作成し，それに対するモデル $\hat{f}_w(x)$ を訓練する.
- 訓練されたモデル $\hat{f}_w(x)$ とソースモデル $f_s(x)$ を使って，ターゲットモデル $\hat{f}_t(x) = \psi(\hat{f}_w(x), f_s(x))$ を構築する.

訓練データ

$$S = \{(x_i, y_i, f_s(x_i))\}_{i=1}^n$$

→ **変数変換**

$$z_i = \phi(y_i, f_s(x_i))$$

→ **学習**

$$\hat{f}_w(x) = \arg\min_{f_w} \frac{1}{n} \sum_{i=1}^n L(f_w(x_i), z_i)$$

→ **変数変換**

$$\hat{f}_t(x) = \psi(\hat{f}_w(x), f_s(x))$$

具体例

例①: ソースモデルとの差を学習する [Kuzborskij+ (2013)]

$$\phi(y_i, f_s(x_i)) = y_i - f_s(x_i)$$
$$\psi(\hat{f}_w(x), f_s(x)) = \hat{f}_w(x) + f_s(x)$$

例②: ソースモデルとの比を学習する

$$\phi(y_i, f_s(x_i)) = \frac{y_i}{f_s(x_i)}$$
$$\psi(\hat{f}_w(x), f_s(x)) = \hat{f}_w(x) f_s(x)$$

例③: 内分点（外分点）を学習する [Minami+ (2021)]

$$\phi(y_i, f_s(x_i)) = \frac{y_i - \tau f_s(x_i)}{1 - \tau}$$
$$\psi(\hat{f}_w(x), f_s(x)) = (1 - \rho)\hat{f}_w(x) + \rho f_s(x)$$

誤差の上界とそれに基づく変数変換法の選択基準

$f_w(x)$ の推定にカーネルリッジ回帰を使用すると，いくつかの仮定の下で，汎化誤差についての以下の不等式が確率$1-\delta$で成立する.

$$R(\hat{f}_t) \le \underbrace{3\mu_\psi^2 \mathbb{E}_{x,y}\left[\psi^{-1}(f_t(x), f_s(x)) - \phi(f_t(x), f_s(x))\right]}_{\psi^{-1}\text{と}\phi\text{の違いに関する項}} + \underbrace{(3\mu_\psi^2\mu_\phi^2 + 1)\sigma^2}_{\text{ノイズの分散に関する項}} + \underbrace{3\mu_\psi^2 \left(\hat{R}_S(\hat{f}_w) + \frac{2Mr}{n\sqrt{\lambda}}\sqrt{\sum_{i=1}^n \phi(y_i, f_s(x_i))^2} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right)}_{\hat{f}_w(x)\text{の予測誤差に関する項}}$$

ただし，

$$R(\hat{f}_t) = \mathbb{E}_{x,y}\left[(y - \hat{f}_t(x))^2\right], \quad \hat{R}_S(\hat{f}_w) = \frac{1}{n}\sum_{i=1}^n \left[(z_i - \hat{f}_w(x_i))^2\right]$$

μ_ψ : Lipschitz constant for $\psi(\cdot,\cdot)$,　μ_ϕ : Lipschitz constant for $\phi(\cdot,\cdot)$

$(z - f_w(x))^2 < M,\ k(x, x') < r$ for all x, x', z, f_w

→ $\psi^{-1} = \phi, \mu_\psi = \frac{1}{\mu_\phi}$ を仮定すると，

$$\frac{\sqrt{\sum_{i=1}^n \phi(y_i, f_s(x_i))^2}}{\mu_\phi^2}$$

を最小にするように変数変換ϕを選択すればよい

実験

ターゲットタスクの真の関数が，ソースモデルからの出力の変数変換とターゲットタスク固有の項の和である，という状況を考える.

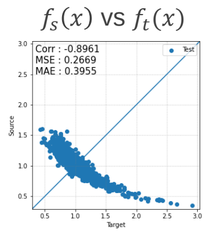
$$f_s(x) = x^{\mathrm{T}}\theta_s + 1, \quad f_t(x) = x^{\mathrm{T}}\theta_w + g(f_s(x))$$

$\theta_s \sim \mathcal{N}(0, 1/40)$, $\theta_s \sim \mathcal{N}(0, 1/100)$, $\epsilon \sim \mathcal{N}(0, 1/100)$, $n = 10$, $x \in \mathbb{R}^{100}$ の下でデータを生成し，下記の関数の候補から，上記の基準に基づいて最適な関数を選択する.

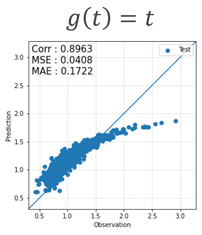
$$\phi(y, f_s(x)) = y - g(f_s(x)), \quad g(t) = 0,\ t,\ 1/t,\ \sqrt{t},\ \log t,\ \exp t$$

Case1: $g_{true}(t) = 1/t$ （関数の候補の中に真の関数が含まれている場合）

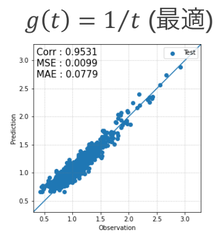
$f_s(x)$ vs $f_t(x)$



$g(t) = t$



$g(t) = 1/t$ (最適)

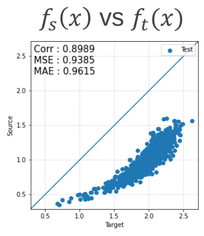


$g(t)$	選択基準の値
0	0.44658379
t	0.76374776
t^2	1.07563255
$1/t$	0.08133291
$\sqrt{t}$	0.61114573
$\log t$	0.79999025
$\exp t$	1.3344229

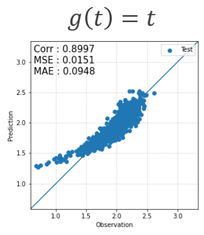
→ 真の関数を同定

Case2: $g_{true}(t) = -2/\sqrt{t}$ （関数の候補の中に真の関数が含まれていない場合）

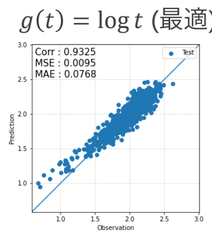
$f_s(x)$ vs $f_t(x)$



$g(t) = t$



$g(t) = \log t$ (最適)



$g(t)$	選択基準の値
0	0.39408597
t	0.14439888
t^2	0.35011503
$1/t$	0.82661393
$\sqrt{t}$	0.23785173
$\log t$	0.09882886
$\exp t$	0.58820885

→ 真の関数に類似した関数を選択

まとめ・今後の展望

- 元タスクの出力のみを使用した転移学習のための変数変換法の選択基準を，汎化誤差の理論的上界に基づいて構築した.
- 数値実験を用いて，関数の候補の中から適切に関数を選択できていることを確認した.
- 以下のケースについてより詳細な方法論の構築が必要である.
  - 変換に使用する関数 ϕ と ψ が逆関数の関係にないケース
  - 変換に使用する関数 ϕ や ψ を機械学習のモデルとして推定するケース

1. Kuzborskij, I., Orabona, F. Stability and hypothesis transfer learning. In International Conference on Machine Learning, 942–950 (2013).
2. Minami, S., Liu, S., Wu, S., Fukumizu, K., Yoshida, R. A general class of transfer learning regression without implementation cost. Proceedings of the 35th AAAI Conference on Artificial Intelligence (2021), in press.