Hypothesis Transfer Learning via Transformation Functions

Simon S. Du, Jayanth Koushik, Aarti Singh, Barnabás Póczos
Machine Learning Department, Carnegie Mellon University

Hypothesis Transfer Learning

1.1 Set Up

1. Model: $Y^{so} = f^{so}(X) + \epsilon^{so}$ and $Y^{ta} = f^{ta}(X) + \epsilon^{ta}$
   $f^{so} \neq f^{ta}$ but related.
2. Data from source domain and target domain :
   $$\{X_i^{so}, Y_i^{so}\}_{i=1}^{n_{so}} \text{ and } \{X_i^{ta}, Y_i^{ta}\}_{i=1}^{n_{ta}}, \quad n_{so} >> n_{ta}.$$
3. Goal: estimate f^{ta}.

1.1 Motivating Examples

1. Reaction time prediction using fMRI images.
2. Distance to object prediction using attributes of robotic arm.

Transformation Functions

2.1 Definitions

Transformation Function: $G(a,b) : \mathbb{R}^2 \to \mathbb{R}$ s.t. given $a \in \mathbb{R}, G(a,\cdot)$ invertible.

Auxiliary Function: $w_G(x) = G^{-1}_{f^{so}(x)}(f^{ta}(x)) \Leftrightarrow G(f^{so}(x), w_G(x)) = f^{ta}(x)$.

Idea: w_G may be easier to learn than f^{ta}.

2.2 Examples

1. Offset-Transfer: $G(a,b) = a + b$, $w_G(x) = f^{ta}(x) - f^{so}(x)$.
2. Scale-Transfer: $G(a,b) = ab$, $w_G(x) = f^{ta}(x)/f^{so}(x)$.
3. Non-Transfer $G(a,b) = b$, $w_G(x) = f^{ta}(x)$ (f^{so} is irrelevant).

Algorithm: Transformation Function based Transfer Learning

Inputs: Source domain data: $\mathcal{T}^{so} = \{(X_i^{so}, Y_i^{so})\}_{i=1}^{n_{so}}$, target domain data: $\mathcal{T}^{ta} = \{(X_i^{ta}, Y_i^{ta})\}_{i=1}^{n_{ta}}$, transformation function: G, algorithm to train f^{so}: $\mathcal{A}_{so}$ and an algorithm to train w_G: $\mathcal{A}_{w_G}$.

Outputs: Regression function for the target domain: $\hat{f}^{ta}$.

1: Train the source domain regression function $\hat{f}^{so} = \mathcal{A}_{so}(\mathcal{T}^{so})$.
2: Construct new data to train w_G with $\hat{f}^{so}$ and $\mathcal{T}^{ta}$: $\mathcal{T}^{w_G} = \{(X_i^{ta}, W_i)\}_{i=1}^{n_{ta}}$, where $W_i = G^{-1}_{(f^{so}(X_i)}(Y_i^{ta})$.
3: Train the auxiliary function: $\hat{w}_G = \mathcal{A}_{W_G}(\mathcal{T}^{w_G})$.
4: $\hat{f}^{ta}(X) = G\left(\hat{f}^{so}(X), \hat{w}_G(X)\right)$.

Theoretical Analysis

3.1 Excess Risk Analysis for Kernel Smoothing

Theorem 1 *Assume f^{so} is $(\lambda_{so}, \alpha_{so})$ Hölder and w_G is $(\lambda_{w_G}, \alpha_{w_G})$ Hölder . If A_{so} and A_{w_G}kernel smoothing estimation algorithms for f^{so} and w_G with bandwidth $h_{so} \asymp n_{so}^{-1/(2\alpha_{so}+d)}$ and $h_{w_G} \asymp n_{ta}^{-1/(2\alpha_{w_G}+d)}$, with probability at least $1-\delta$ the risk satisfies:*

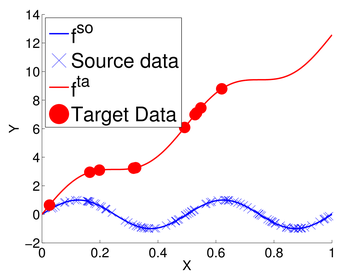
$$R\left(\hat{f}^{ta}\right) - R\left(f^{ta}\right) = O\left(\left(n_{so}^{-2\alpha_{so}/(2\alpha_{so}+d)} + n_{ta}^{-2\alpha_{w_G}/(2\alpha_{w_G}+d)}\right)\log(1/\delta)\right).$$

3.2 Other Extensions

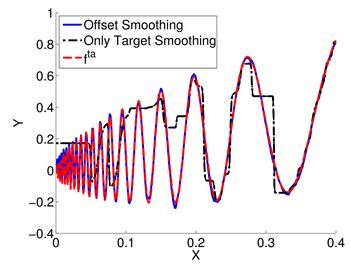
1. Excess risk analysis for kernel ridge regression.
2. A general theorem for this procedure based on algorithmic stability.
3. A theorem for finding the best transformation function among a function class using cross-validation.

Experiments

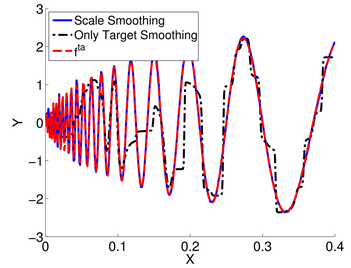
4.1 Synthetic Data



Model shift: more data from the source domain than the target domain.



$f^{so}(x)$: Doppler Function. $f^{ta}(x) = f^{so}(x) + x$. $n_{so} = 10000$ and $n_{ta} = 100$



$f^{so}(x)$: Doppler Function. $f^{ta}(x) = f^{so}(x) \cdot x$. $n_{so} = 10000$ and $n_{ta} = 100$

4.2 Robotic Data

Transferring from linear low noise domain to non-linear high noise domain, $n_{so} = 320$.

	$n_{ta} = 10$	$n_{ta} = 20$	$n_{ta} = 40$	$n_{ta} = 80$	$n_{ta} = 160$	$n_{ta} = 320$
Only Target Smoothing	0.086 ± 0.022	0.076 ± 0.010	0.066 ± 0.008	0.064 ± 0.007	0.065 ± 0.006	0.063 ± 0.005
Only Target Ridge	0.080 ± 0.017	0.078 ± 0.022	0.063 ± 0.013	0.050 ± 0.007	0.048 ± 0.006	$\mathbf{0.040 \pm 0.005}$
Only Source Ridge	0.098 ± 0.017	0.098 ± 0.017	0.098 ± 0.017	0.098 ± 0.017	0.098 ± 0.017	0.098 ± 0.017
Combined Smoothing	0.092 ± 0.011	0.084 ± 0.008	0.077 ± 0.009	0.075 ± 0.006	0.074 ± 0.006	0.067 ± 0.006
Combined Ridge	0.087 ± 0.025	0.077 ± 0.015	0.062 ± 0.009	0.061 ± 0.005	0.047 ± 0.003	0.041 ± 0.004
Offset Smoothing	0.080 ± 0.026	0.066 ± 0.023	$\mathbf{0.052 \pm 0.006}$	0.054 ± 0.006	0.050 ± 0.003	0.052 ± 0.004
Offset Ridge	0.146 ± 0.112	0.066 ± 0.017	0.053 ± 0.007	$\mathbf{0.048 \pm 0.006}$	$\mathbf{0.043 \pm 0.004}$	0.041 ± 0.003
Scale Smoothing	$\mathbf{0.078 \pm 0.022}$	$\mathbf{0.065 \pm 0.013}$	0.056 ± 0.009	0.056 ± 0.005	0.054 ± 0.008	0.055 ± 0.004
Scale Ridge	0.102 ± 0.033	0.095 ± 0.100	0.057 ± 0.014	0.052 ± 0.010	0.044 ± 0.004	0.042 ± 0.002