On the Value of Target Data in Transfer Learning

Steve Hanneke  
steve.hanneke@gmail.com  
TTIC

Samory Kpotufe  
skk2175@columbia.edu  
Columbia University

Transfer Learning:
- In many learning problems, we have tons of data from related problems.
- It might be helpful to use it.
- But how helpful? And what is the “right” way to use it?

Definitions: Hypothesis class: $\mathcal{H}$, VC dimension $d$
Target distribution: $Q$ over $X \times \{-1, 1\}$
Source distribution: $P$ over $X \times \{-1, 1\}$
Source and target loss: $R_Q(h) = E_Q(h(X) \neq Y)$, $h^*_Q = \arg \min_R R_Q(h), \mathcal{E}_Q(h) = R_Q(h) - R_Q(h^*_Q)$
Source and target Bayes risk: $R_P(h) = P_P(h(X) \neq Y), h^*_P = \arg \min_R R_P(h), \mathcal{E}_P(h) = R_P(h) - R_P(h^*_P)$

Receive $n_P$ samples $S_P$ iid $P$, $n_Q$ samples $S_Q$ iid $Q$
Learner $h$ can use both $S_Q$ and $S_P$.

Main question: What is the optimal rate of convergence of $\mathcal{E}_Q(h)$?
(in terms of $n_P$ and $n_Q$)

Transfer exponent: $\rho$
For some $C, \rho,$
\[ \forall h \in \mathcal{H}, \quad \mathcal{E}_Q^\rho(h) \leq C \mathcal{E}_P(h). \]

Bernstein class condition:
For some $C, \rho, \epsilon Q, \forall h \in \mathcal{H},$
\[ P_X(h \neq h^*_P) \leq C \mathcal{E}_P(h) \quad \text{and} \quad Q_X(h \neq h^*_Q) \leq C \mathcal{E}_Q(h). \]

Benefits of $\rho$: 1. Unlike KL, $\rho$ handles non-overlapping supports: Figure 1 2. Unlike $d_\Delta, d_\gamma$ discrepancies, $\rho$ is asymmetric: Figures 2,3 3. and $\rho$ is inherently localized (focused on small $\mathcal{E}_P(h)$): Figure 3

Main Theorem:
Optimal $\mathcal{E}_Q(h) = \Theta \left( \min \left\{ \left( \frac{d}{n_P} \right)^{\frac{1}{2-\rho/P}}, \left( \frac{d}{n_Q} \right)^{\frac{1}{2-\rho/Q}} \right\} \right)$

Algorithm 1:
Minimize $\hat{R}_{Q|P}(h)$
subject to $\hat{R}_{Q|P}(h) - R_{Q|P}(h^*_Q) \leq \epsilon \sqrt{R_{Q}(h \neq h^*_Q) \frac{d}{n_Q}} + \epsilon \frac{d}{n_Q}$
$h \in \mathcal{H}$
where $h^*_Q = \arg \min_R \hat{R}_{Q|P}(h)$

Remark: $\text{ERM}(S_P \cup S_Q) = \arg \min_R \hat{R}_{SP|U|S_Q}(h)$ does not achieve optimal rate $h \in \mathcal{H}$

Marginal transfer exponent: $\gamma$
For some $C, \gamma,$
\[ \forall h \in \mathcal{H}, \quad \mathcal{E}_E^\gamma(h) \leq C \mathcal{E}_E(h) \]

Applications:
1. Adapting to sampling costs
2. Reweighting the source data
3. Adaptive transfer from multiple sources

Application: Adapting to sampling costs using unlabeled data
$\epsilon_P(n), \epsilon_Q(n) = \text{cost to obtain } n \text{ samples from } P, Q$ resp
Assume $\epsilon_P, \epsilon_Q$ increasing, concave, unbounded

Suppose $U_Q$ is a large unlabeled $Q_X$ sample
Define $\hat{\delta}(S, U_Q) = \max \left\{ \hat{R}_{Q|P}(h \neq h^*_S) : h \in \mathcal{H}, \right.$
\[ \hat{R}_{S}(h) - \hat{R}_{S}(h^*_S) \leq \epsilon \sqrt{\hat{R}_{S}(h \neq h^*_S) \frac{d}{|S|}} + \epsilon \frac{d}{|S|} \left. \right\}, \]
where $h^*_S = \arg \min_R \hat{R}_{S}(h)$

Algorithm 2:
0. $S_P \leftarrow \{\}; \quad S_Q \leftarrow \{\}$
1. For $t = 1, 2, \ldots$
2. Let $n_{t,P}$ be minimal such that $\epsilon_P(n_{t,P}) \geq 2^{-t}$
3. Sample $n_{t,P}$ samples from $P$ and add them to $S_P$
4. Let $n_{t,Q}$ be minimal such that $\epsilon_Q(n_{t,Q}) \geq 2^{-t}$
5. Sample $n_{t,Q}$ samples from $Q$ and add them to $S_Q$
6. If $\delta(S_Q, U_Q) \leq \epsilon / 4$, return $h^*_Q$
7. If $\delta(S_P, U_Q) \leq \epsilon / 4$, return $h^*_P$

Theorem:
Let $n^*_Q = \frac{d}{\epsilon^2 Q}, n^*_P = \frac{d}{\epsilon^2 P}$. Algorithm 2 has $\mathcal{E}_Q(h) \leq \epsilon$
with total sampling cost $\tilde{O}(\min \{\epsilon_Q(n^*_Q), \epsilon_P(n^*_P)\})$
and this is optimal (up to logs).