Inductive Bias and Performance Bounds: For the Case of Multi-Task Learning

Jin Kim

I. Introduction

Machine learning typically relies on large number of data. Traditional statistics used to be studied from decades ago, but availability of large number of dataset and increasing computational power surely affected recent rush on machine learning. However on the other hand, if the number of samples is limited, the performance is not guaranteed to meet the goal. Moreover, theories about performance lower bound show that with small number of samples, there is certain amount of regret for worst-case probability distribution[1]. To overcome this limitation, everything that can be utilized to improve convergence rate and generalization shall be considered. In general, any factors making some hypotheses more admissible than others called inductive bias. For example, setting up the model for the problem and choosing appropriate algorithm by human expert is one of inductive bias. Also, knowledge on domain or prior over feature space and preprocessing on data—e.g. clustering, hierarchical structuring, adjustments—can be considered as inductive biases.

Multi-task learning(MTL) was first proposed by Caruana[3], and it deals with concurrent learning of several related tasks. According to the author, each task can behave like if they are the mutually inductive bias. Originally MTL designed for neural network and decision tree method and showed its benefit numerically, but numerous studies have conducted on both theory and applications.

In this report, Liu's paper[2] will be mainly discussed. Indeed, Improvement on generalization performance has been studied for a long time from Caruana's original paper[3]. Baxter[4] also proposed firm theoretical framework on inductive biases including MTL by considering a hypotheses family, set of hypothesis sets. [2] works on explicit formula for performance bounds for specific task of interest. For the earlier part of this report, the problem definition will be summarized. Then the major contribution of the paper will be followed. Finally, some remarks from the paper and my own discussion will be presented.

II. Preliminaries

A. Multi-Task Learning

The multi-task learning problem discussed in this paper considers some Hilbert space for feature vector and T binary classification tasks. Linear hypothesis, i.e. defined by RKHS will be used. Notations are:

- $\mathscr{H}$: Feature Hilbert space with $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$
- $z=(x,y)\in\mathscr{H}\times\{-1,+1\}$: A training sample
- $S_t=\{z_{t,1},\cdots,z_{t,n_t}\}$: Sample sets for t^{th} task
- $H=\{h(x)=\langle h,x\rangle : h\in\mathscr{H}$ is a linear class
- $l(z;h)=l(y,h(x))$ is a loss function which is differentiable and σ-strongly convex over h and L Lipschitz for any given sample $z=(x,y)$

- The algorithm is to find:

$$\min_{w_1,\cdots,w_T,\theta} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n_t} l(y_{t,i},\langle w_t+\theta, x_{t,i}\rangle) \tag{1}$$

Note that each task may have individual underlying probability distribution, and each task have its own sample set S_t. However, the learning is performed over all samples, $\{x_{1,1},\cdots,x_{1,n_1},\cdots,x_{T,n_T}\}$. Also, since hypothesis h_t is linear sum of w_t and θ, the minimizer is not unique. According to the Authors[2], more constraints on w_t and θ are required to make the problem well posed, such as norm regulation or sparsity of w_t or maximum condition of $\|\theta\|$. However, the author does not provide an specific regulation on them in the paper.

Assumption 1 (Reconstruction): There exists a subset $B \subseteq \{x_{1,1},\cdots,x_{T,n_T}\} - \{x_{t,1},\cdots,x_{t,n_t}\}$ so that any feature x drawn from the probability measure of t^{th} task, $x = \sum_{j=1}^{N}\alpha_j b_j + \eta$ for some real valued vector $\|\alpha\| \le r$ and small error $\|\eta\| \le \varepsilon$

This assumption provides the least amount of relationship between each tasks. If task of our interest does not share a support on its feature distribution with others, then it is impossible to achieve with small ε. If sample sizes of 'other' tasks are too small, the assumption also becomes hard to meet.

B. Stability and Generalization bound

To evaluate stability, let $S^{(i)}$ denotes S with single i^{th} sample replaced by fresh sample z^i for the task. The hypothesis predicted by set S will be denoted as h_S.

Definition 1 (Uniform stability): An algorithm is β uniformly stable on $l(z;h)$ if

$$\forall S \in \mathscr{Z}^n, \forall i \in [n], \forall z, z^i \in \mathscr{Z}, |l(y, h_S^{(i)}) - l(y, h_S) \le \beta$$

Note that if upper bound β is small, then the algorithm generalizes well. Suppose β is a function of n, and $\beta \to 0$ for $n \to \infty$, then $\beta(n)$ is a generalization bound.

III. Major Contributions

In this chapter, few theorems related to my focus will be introduced. However, detailed proof and mathematical tools are omitted in this report.

A. Generalization bound for θ

Suppose we are interested in a specific task t, then perturbation on S_t result in θ these theorems.

Theorem 1: For any $z'_{t,i}$ distributed from t^{th} task, given w_t, $\theta_{S_t^{(i)}}$ and θ_{S_t} meet following inequality

$$\begin{aligned}
&|l(y_t,\langle w_t+\theta_{S_t^{(i)}}, x_t\rangle) - l(y_t,\langle w_t+\theta_{S_t}, x_t\rangle| \\
&\le \max_{z'_{t,i}\in\mathscr{Z}_t} L|\langle \theta_{S_t^{(i)}} - \theta_{S_t}, x_t\rangle| \\
&\le \frac{Lr\max\{n_\tau : \tau \ne t\}}{2\sigma}\left(\sqrt{\left(\frac{2Lr}{n_t}\right)^2 + \frac{4\sigma O(\varepsilon)}{n_t \max\{n_\tau : \tau \ne t\}}} + \frac{2Lr}{n_t}\right)
\end{aligned}$$

1 show that θ is uniformly stable with respect to the domain of the t^{th} task. For the second theorem, $\varepsilon = 0$ and $n_1 = n_2 = \cdots = n_T = n$ are assumed for simplification.

Theorem 2: If $\varepsilon = 0$ and $n_1 = n_2 = \cdots = n_T = n \geq 2$, and $\mathscr{Z} = \bigcup_t \mathscr{Z}_t$

$$\max_{z'_{t,i} \in \mathscr{Z}} L|\langle \theta_{S_t^{(i)}} - \theta_{S_t}, x_t \rangle| \leq \frac{2Lr^2}{\sigma T}$$

The second theorem claims that increasing number of multiple tasks, $T \gg 1$, provides non-vanishing inductive bias and make generalization bound decrease.

B. Generalization bound for h

[2] further claim stability over h

Theorem 3: For any $z'_{t,i}$ distributed from t^{th} task, for any $h_{t,S_t^{(i)}}$ and h_{t,S_t} by ERM of multi-task problem, the inequality holds.

$$\begin{aligned}
&|l(y_t, h_{t,S_t^{(i)}}(x_t)) - l(y_t, h_{t,S_t}(x_t))| \\
&\leq \max_{z'_{t,i} \in \mathscr{Z}_t} L|\langle h_{t,S_t^{(i)}} - h_{t,S_t}, x_t \rangle| \\
&\leq \frac{Lr \max\{n_\tau : \tau \neq t\}}{2\sigma} \left(\sqrt{\left(\frac{2Lr}{n_t}\right)^2 + \frac{4\sigma O(\varepsilon)}{n_t \max\{n_\tau : \tau \neq t\}}} + \frac{2Lr}{n_t} \right)
\end{aligned}$$

For simplicity, suppose $\varepsilon = 0$ then,

$$|l(y_t, h_{t,S_t^{(i)}}(x_t)) - l(y_t, h_{t,S_t}(x_t))| = \frac{2L^2 r^2 \max\{n_\tau : \tau \neq t\}}{2n_t \sigma}$$

This theorem implies that if the set $\{n_\tau : \tau \neq t\}$ is fixed, generalization bounds of specific task is of order $O(1/n_t)$, which is faster than $O(1/\sqrt{n})$ for single task learning problem.

IV. Discussions

One of the main result of [2] is that; for linear MTL problem posed on (II-A), a particular task has generalization bound with a convergence rate of order $O(1/n)$ and $O(1/T)$ under some assumptions: restriction on l, distribution on samples and how close each tasks are. With the rate faster than $O(\sqrt{\log n/n})$ or $O(\sqrt{1/n})$, MTL likely to generalize better with relatively few samples than single task learning. The author also states that his approach can be applied to other algorithms as well, rather than empirical risk minimizer.

Still there are some arguable points. First, single task learning problem can achieve of order $O(1/n)$ generalization bound[1]. The author briefly mention about this, and claim that their bounds too much rely on regularization on the hypothesis set. However, considering multiple learning tasks on the same feature space may be much more tough regulation. Also, an implicit regularization on w_t or θ is mentioned in MTL case as well.

Another thing to be discussed is trade-off between number of samples for additional tasks and ε. In order to achieve small r and ε, $B \subseteq \{x_{1,1}, \cdots, x_{T,n_T}\} - \{x_{t,1}, \cdots, x_{t,n_t}\}$ span and cover feature space. To

insure this diversity, large number of samples on $S_\tau : \tau \neq t$ is necessary. Meanwhile, the upper bound convergence rate is proportional to $\max\{n_\tau : \tau \neq t\}$, which penalized by number of samples for other tasks.

On the other hand, interesting questions arise from the paper. One is relationship with multi dimensional label. The key difference is that MTL draw samples for each task separately, while later case considers the situation that each sample has multiple label. They are essentially different, but as Caruana[3]'s original formulation is more close to neural net with multiple output, The approach in [2] might be able to multi dimensional label problem. If it is possible, considering augmented label would be interesting as well. Sometimes estimation problem becomes easier if more parameters are augmented on the original problem, for instance, EM algorithm. MTL can be applied to existing learning problem by augmenting labels.

REFERENCES

[1] M. Raginsky, "Maxim raginsky: Statistical learning theory."

[2] T. Liu, D. Tao, M. Song, and S. J. Maybank, "Algorithm-dependent generalization bounds for multi-task learning," vol. 39, no. 2, pp. 227–241.

[3] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *Proceedings of the Tenth International Conference on Machine Learning*, pp. 41–48, Morgan Kaufmann.

[4] J. Baxter, "A model of inductive bias learning,"