On Asymptotic Generalization Error of Asymmetric Multitask Learning

Keisuke Yamazaki
Precision and Intelligence Laboratory
Tokyo Institute of Technology
R2-5, 4259 Nagatsuta, Midori-ku,
Yokohama, 226-8503 Japan
k-yam@pi.titech.ac.jp

Samuel Kaski
Department of Information and Computer Science,
Helsinki University of Technology
P.O. Box 5400, 02015 TKK, Finland
samuel.kaski@tkk.fi

Abstract

A recent variant of multi-task learning uses the other tasks to help in learning a task-of-interest, for which there is too little training data. The task can be classification, prediction, or density estimation. The problem is that only some of the data of the other tasks are relevant or representative for the task-of-interest. It has been experimentally demonstrated that a generative model works well in this *relevant subtask learning* task. In this paper we analyze the generalization error of the model, to show that it is smaller than in standard alternatives, and to point out connections to semi-supervised learning, multi-task learning, and active learning or covariate shift.

1 Introduction

Lack of a sufficient amount of training data is a recurring problem in practical data analysis and machine learning studies. In bioinformatics, for instance in microarray data analysis, this has been referred to as the large p, small n problem: It is hard to learn classifier or regression models when the dimensionality of the data samples p is large and the number of samples n is small. Recommender systems and user modeling share an analogous problem: it would be useful to give good predictions or recommendations already when only a few observations of the user's behavior have been made.

The problem can be alleviated by regularizing the predictors, and by including more prior knowledge to the model structure or in the prior distribution of the parameters. Collecting new data helps too. Several machine learning scenarios have been developed to help in cases where these straightforward alternatives are not applicable or available. For instance, in semisupervised learning a classification task where labeled data is the scarce resource, can be aided by including non-labeled data from the same distribution. In multi-task learning several tasks are learned together, in the hope that the tasks share properties which help in learning each task. Common to these scenarios is that they try to incorporate more data into the training data set.

A central practical problem in adding more data is that most models assume all training data to be "relevant"; typically the implicit assumption is that all training data come from the same distribution, or at least that adding the data to the learning set improves the performance. Requiring all data to come from the same distribution is a strong assumption, and if it could be relaxed, it would be possible to include data sets or tasks containing only partly relevant data. Useful sets abound in genomic databanks and measurement databases in bioinformatics, for instance, or data about other users or products in recommender systems.

Relevant subtask learning is a recent variant of multi-task learning, where the assumption of representative data is relaxed by assuming that the learning data is a mixture of relevant and irrelevant samples. The setup is that there is one task-of-interest, which is special in that the test data is known

to come from its distribution. In the other tasks, some samples come from the same distribution and some not, and it is naturally not known which. In other words, the other tasks are contaminated by data from irrelevant subtasks, and we would only like to use the data from the relevant subtasks. The relevant subtask model (RSM) (Kaski & Peltonen, 2007) builds a classifier for the task-of-interest under these assumptions; we will generalize this setup to unsupervised learning, which can naturally model classes as well. RSM has been empirically demonstrated to outperform standard multi-task learning and the straightforward alternatives of learning the task-of-interest separately and pooling all data together.

From the statistical point of view good performance of RSM is not trivial since the increased complexity of the learning model increases the generalization error. In this paper we derive the asymptotic generalization error for maximum likelihood estimates. Comparing the generalization error of alternative models we prove that RSM is still better than the others.

2 Relevant Subtask Learning

Relevant subtask learning is a variant of multi-task learning. In multi-task learning (Caruana, 1997; Marx et al., 2005; Raina et al., 2005) there are several classification tasks, and the question is whether solving the problems together improves performance compared to solving each separately. In practice, the tasks are different data sets. More generally, instead of classification problems the tasks could be other statistical modeling tasks such as regression, clustering or density estimation; in all cases the research problem is to learn a good estimator for each task, transferring information between the tasks as needed.

In relevant subtask learning the setup is asymmetric. One of the tasks is a target task, the "*task-of-interest*", and the research problem is whether the other tasks can be used to help in better solving the target task. In practice the tasks are data sets, and the goal is to find more data to complement the scarce learning data in the task-of-interest. The obvious problem is that not all data in the other tasks are relevant in the sense of coming from the same distribution or at least helping in learning the task-of-interest. Relevant subtask learning makes the assumption that each task is a mixture of relevant and irrelevant data, that is, each task is a combination of a relevant subtask and an irrelevant subtask.

In this paper we will consider the unsupervised setting of density estimation, and two tasks without loss of generality: the task of interest (task number 1) and a supplementary task (task number 2). Let us denote the distribution of data in the task-of-interest by $q_1(x) = q_{01}(x)$, where the datum $x \in R^M$. The data of the irrelevant subtask within task 2 follow the distribution $q_{02}(x)$.

For learning we are given two training data sets D_1 and D_2, one for each task, where $\#D_1 = \alpha n$ and $\#D_2 = (1-\alpha)n$ for $0 < \alpha < 1$. Here αn of course needs to be a natural number. D_1 is known to be sampled from $q_{01}(x)$, but generally $\#D_1$ is too small for learning an adequate model for q_{01}.

The supplementary data set D_2 contains samples from both the same distribution as D_1, and from the irrelevant subtask. Hence the density is a mixture,

$$q_2(x) = c^* q_{01}(x) + (1 - c^*) q_{02}(x),$$

where $0 < c^* < 1$ is a constant. All quantities above, the q and c^*, are unknown to us. Then, we formally rewrite $D_2 = D_{21} + D_{22}$, where D_{21} and D_{22} have been sampled from q_{01} and q_{02}, respectively. Obviously, the data in D_2 do not have a label showing which distribution they come from, q_{01} or q_{02}. The data in D_1 have the label since they are all from q_{01}. This situation is analogous to semi-supervised learning (Zhu, 2007), in which a small labeled training data set of a classifier is complemented with additional unlabeled data. The difference is that in RSL the goal is not to classify samples according to the labels but instead to build a good model for $q_1(x)$ based on D_1 and D_{21}. In this sense relevant subtask learning even resembles "one-class classification" (Tax, 2001; Tax & Duin, 2001).

The test data set D_t is known to come from $q_{01}(x)$. This fundamentally separates RSL from standard multitask learning; we are only interested in the target task. As a result, since we use both data sets D_1 and D_2 for training, the test distribution is different from the training distribution. Moreover, we can change the training distribution by selecting the data for training. This scenario is similar to active learning (Fedorov, 1972), which focuses on doing the active selection, and the covariate

Table 1: Data structures and learning models

method	data	training model	prediction model with MLE	true parameter
M1	D_1	p_1 (single dist.)	$p_1(x\|\hat{a})$	a^*
M2	D_2	p_2 (mixture)	$p_{01}(x\|s(\hat{w}_2))$	$w_2^* = (a^*, b^*, c^*)$
M3	$D_1 \oplus D_2$	$p_1 \oplus p_2$ (RSM)	$p_{01}(x\|s(\hat{w}_3))$	$w_3^* = (a^*, b^*, c^*)$
M4	$D_1 + D_2$	p_2 (mixture)	$p_{01}(x\|s(\hat{w}_4))$	$w_4^* = (a^*, b^*, \alpha + (1-\alpha)c^*)$

shift (Shimodaira, 2000), where the research interest is in studying the effects of the difference in the training and testing distributions.

Let us assume that our learning model can attain the true distributions which generate the training data. In other words, we prepare models $p_{01}(x|a) = p_1(x|a)$ and $p_{02}(x|b)$, where $a \in R^d$ and $b \in R^d$ are the parameters, respectively. Then the assumption translates to the following: there exist true parameters a^* and b^* such that

$$q_{01}(x) = p_{01}(x|a^*), \quad q_{02}(x) = p_{02}(x|b^*).$$

Let us define a mixture model for modeling D_2:

$$p_2(x|w) = cp_{01}(x|a) + (1-c)p_{02}(x|b),$$

where $w = \{a, b, c\}$. More precisely, $a = (w_1, \ldots, w_d)$, $b = (w_{d+1}, \ldots, w_{2d})$, and $c = w_{2d+1}$.

With these definitions we can formulate the possible solutions to the learning problem more precisely. Table 1 summarizes the methods. We will use maximum likelihood estimator (MLE) $\hat{w}_i$ of Mi for $i = 1, \ldots, 4$. Each Mi is defined as follows:

(M1) Single-task learning. The first and the simplest approach is to use only data from the task of interest, discarding the supplementary data altogether.

(M2) Learning only from the supplementary task. A slightly artificial choice is to discard D_1 altogether and learn only from D_2. The model M2 is expected to be better than M1 when $\#D_1$ is too small and the estimation using D_1 is not reliable.

(M3) Relevant subtask model. M3 corresponds to the RSM of (Kaski & Peltonen, 2007). The likelihood is defined by

$$\hat{w}_3 = \arg\max_w \left\{ \sum_{x_i \in D_1} \ln p_1(x_i|a) + \sum_{x_i \in D_2} \ln p_2(x|w) \right\}.$$

Note that the complexity of the model is more than M1. You can see the dimension of the parameters increases from d to $2d+1$, which could cause worse generalization.

(M4) Pooled data model. The simplest approach, optimal when data of all tasks come from the same distribution, is to pool all data and estimate a single model. Since we know the data may be a mixture of relevant and irrelevant data, we will learn the mixture model.

In Table 1, $D_1 \oplus D_2$ is the combined data set with the task label and $D_1 + D_2$ is the merged data set, where we cannot distinguish the difference between the original data sets any more. The $s(\cdot)$ is a function that chooses the parameters corresponding to the target task from w, namely $s(w) = a$.

3 Analysis of the Generalization Error

Let us define the generalization error of the learning methods by

$$G_i(n) = E_{D_1 \oplus D_2}\left[\int q_1(x) \ln \frac{q_1(x)}{p_{01}(x|s(\hat{w}_i))} \right],$$

where $E_{D_1 \oplus D_2}[\cdot]$ denotes the expectation over the training data D_1 and D_2, the suffix i stands for the error of the learning method Mi, and we define $s(\hat{w}_1) = \hat{a}_1$ to simplify the notation. The following theorem is the main contribution of this paper:

Theorem 1 *The generalization error has the asymptotic form*

$$G_1(n) = \frac{d}{2\alpha n} + O\left(\frac{1}{n^2}\right),$$
$$G_2(n) = \frac{1}{2(1-\alpha)n}\mathrm{Tr}[I(a^*)J(w_2^*)^{-1}] + O\left(\frac{1}{n^2}\right),$$
$$G_3(n) = \frac{1}{2n}\mathrm{Tr}[I(a^*)K(w_3^*)^{-1}] + O\left(\frac{1}{n^2}\right),$$
$$G_4(n) = \frac{1}{2n}\mathrm{Tr}[I(a^*)J(w_4^*)^{-1}] + O\left(\frac{1}{n^2}\right),$$

where the I, J, K are $(2d+1) \times (2d+1)$-dimensional matrices where

$$I(a^*)_{ij} = \int \frac{\partial \ln p_1(x|a^*)}{\partial w_i}\frac{\partial \ln p_1(x|a^*)}{\partial w_j} q_1(x)dx,$$
$$J(w^*)_{ij} = \int \frac{\partial \ln p_2(x|w^*)}{\partial w_i}\frac{\partial \ln p_2(x|w^*)}{\partial w_j} q_2(x)dx,$$
$$K(w^*) = \alpha I(a^*) + (1-\alpha)J(w^*)$$

for $a^* = s(w^*) \subset w^*$.

As a reminder, d is the number of parameters in p_{01}, n is the total number of data samples, and α is the proportion of samples in the task-of-interest data $\#D_1$. Note that $I(a^*)$ has non-zero elements in the top-left $d \times d$ submatrix. Comparing to the coefficients of the generalization errors, we can find the following:

Corollary 1 *Generalization error of M3 is smaller than that of M1 or M2.*

This corollary implies that the advantage resulting from increasing the number of training data is stronger than the disadvantage caused by the cost for parameter tuning.

4 Discussion

In this paper we have derived the asymptotic generalization error of relevant subtask learning models, estimated using maximum likelihood. The results gave an interesting insight that RSM is the best alternative when the model can attain the true distribution. This means that the advantage to get more data is larger than the disadvantage to increase the complexity of the model. Since our model setting is general, applications to practical data such as bioinformatics are expected in future studies.

References

Caruana, R. (1997). Multitask learning. *Machine Learning, 28*, 41–75.

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Kaski, S., & Peltonen, J. (2007). Learning from relevant tasks only. *ECML* (pp. 608–615).

Marx, Z., Rosenstein, M. T., Kaelbling, L. P., & Dietterich, T. G. (2005). Transfer learning with an ensemble of background tasks. *NIPS workshop on inductive transfer*.

Raina, R., Ng, A. Y., & Koller, D. (2005). Transfer learning by constructing informative priors. *NIPS workshop on inductive transfer*.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference, 90*, 227–244.

Tax, D. M. J. (2001). *One-class classification*. Doctoral dissertation, Delft University of Technology.

Tax, D. M. J., & Duin, R. P. W. (2001). Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research, 2*, 155–173.

Zhu, X. (2007). *Semi-supervised learning literature survey* (Technical Report TR1530). Computer Science, University of Wisconsin Madison.