SVM Multi-Task Learning and Non convex Sparsity Measure

R. Flamary, A. Rakotomamonjy, G. Gasso and S. Canu

February 19, 2009

Abstract

Recently, there has been a lot of interest around multi-task learning (MTL) problem with the constraints that tasks should share common features. Such a problem can be addressed through a regularization framework where the regularizer induces a joint-sparsity pattern between task decision functions. We follow this principled framework but instead we focus on $\ell_p - \ell_2$ (with $p \leq 1$) mixed-norms as sparsity-inducing penalties. After having shown that the $\ell_1 - \ell_2$ MTL problem is a general case of Multiple Kernel Learning (MKL), we adapted the available efficient tools of solving MKL to the sparse MTL problem. Then, for the more general case when $p < 1$, the use of a DC program provides an iterative scheme solving at each iteration a weighted $\ell_1 - \ell_2$ sparse MTL problem.

Work description

Multi-Task Learning (MTL) is a statistical learning framework which seeks at learning different models in a joint manner. The idea behind this paradigm is that, when the tasks to be learned are similar enough or are related in some sense, it may be advantageous to take into account these relations between tasks. Several works have experimentally highlighted the benefit of such a framework (Caruana, 1997).

In this work, we consider that tasks to be learned share a common subset of features or kernel representation. This means that while learning the tasks, we jointly look for features or kernels that are useful for all tasks. A way to address this issue is to use a regularization principle and thus minimizes a regularized empirical risk while the regularization term favors a common sparsity profile in features for all tasks (Argyriou et al., 2008; Obozinski et al., 2007).

This paper also considers this regularization principle for joint feature selection across tasks. Our contribution is two fold. First we consider the multi-task learning problem in a SVM framework with a kernel representation. The proposed algorithms rely on sparsity-inducing ($\ell_p - \ell_2$) mixed-norms regularizers

which encourage sparse kernel selection among a prescribed set of kernels. This set of *basis* kernels can be made large enough at will, gathering information about the different sources of the input samples. From this framework, it comes up the convex case turns into a multiple kernel learning problem. At the second stage, we extend the analysis to a non-convex regularization term in order to gain in sparsity The difficulty raised by this formulation is tackled via a DC programming (Horst & Thoai, 1999).

Framework and algorithms

Suppose we are given T classification tasks to be achieved from T different datasets $\{x_{i,1}, y_{i,1}\}_i^{n_1}, \cdots, \{x_{i,T}, y_{i,T}\}_i^{n_T}$, where any $x_{i,\cdot} \in \mathcal{X}$ and $y_{i,\cdot} \in \{+1, -1\}$ and n_i denotes the i^{th} dataset size. For a given task t, we are looking for a decision function of the form: $f_t(x) = \sum_{k=1}^{M} f_{t,k}(x) + b_t \qquad \forall t \in \{1, \cdots, T\}$ where any function $f_{\cdot,k}$ belongs to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ of kernel K_k, b_t is the bias term and M is the number of *basis* kernels provided. To learn the decision function f_t of each task under the constraints that all these functions share a common sparse profile of their kernel representation, let consider the following optimization problem:

$$\min_{f_1, \cdots, f_T} C \cdot \sum_{t,i} L(f_t(x_{i,t}), y_{i,t}) + \Omega(f_1, \cdots, f_T)$$

where $L(f_t(x), y)$ is a loss function, Ω a sparsity-inducing penalty function involving all f_t and C a trade-off parameter that balances both antagonist objectives. We propose the penalty function $\Omega_{p,q}(f_1, \cdots, f_T) = \sum_{k=1}^{M} \left(\sum_{t=1}^{T} \|f_{t,k}\|_{\mathcal{H}_k}^q \right)^{p/q}$ with typically $p \leq 1$ and $q \geq 1$.

Easy case Let $p = 1$ and $q = 2$ and consider a hinge loss function. Therefore, it can be shown this multi-task SVM problem is strongly related to the Multiple Kernel Learning (MKL) problem. Especially, it can be shown that the problem boils down to solve T SVM tasks over a kernel defined as a convex combination of the M kernels ($k(x) = \sum_{k=1}^{M} d_k k_k(x), d_k \geq 0, \sum d_k = 1$). The ℓ_1-type penalty encourages the vanishing of some coefficients d_k. Hence, an efficient algorithm is derived based on off-the-shelf MKL solvers (Rakotomamonjy et al., 2008).

Non convex case ($p < 1$) To address this case, the ℓ_p penalty is decomposed as a difference of two convex functions that is $g(u) = \|u\|_p = |u| - (|u| - |u|^p)$. Using the DC programming (Horst & Thoai, 1999), we establish after few algebras that the optimisation problem can be solved iteratively where at each iteration, we resolve a weigthed version of the $\ell_1 - \ell_2$ problem. The weigths are proportional to the inverse of the norm $\|f_{\cdot,k}\| = \left(\sum_{t=1}^{T} \|f_{t,k}\|_{\mathcal{H}_k}^2 \right)^{1/2}$.

The extended version of the paper provides empirical evidences that show the benefit of the proposed approaches and algorithms. These experimental results show the improvment of sparsity with compelling classification performances.

References

Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning, to appear*.

Caruana, R. (1997). Multi-task learning. *Machine Learning*, *28*, 41–75.

Horst, R., & Thoai, N. V. (1999). Dc programming: overview. *Journal of Optimization Theory and Applications*, *103*, 1–41.

Obozinski, G., Taskar, B., & Jordan, M. (2007). *Multi-task feature selection* (Technical Report). UC Berkeley Technical Report.

Rakotomamonjy, A., Bach, F., Grandvalet, Y., & Canu, S. (2008). SimpleMKL. *Journal of Machine Learning Research*, *9*, 2491–2521.