Exploring Regularization in Learning to Predict Structured Objects

Yasemin Altun
TTI-C, Chicago IL 60632, USA
altun@tti-c.org
http://ttic.uchicago.edu/ altun

Discriminative learning framework is one of the very successful fields of machine learning. The methods of this paradigm, such as Boosting, and Support Vector Machines have significantly advanced the state-of-the-art for classification by improving the accuracy and by increasing the applicability of machine learning methods. However, traditionally these methods do not exploit dependencies between class labels where more than one label is predicted. Many real-world classification problems, on the other hand, involve sequential, temporal or structural dependencies between multiple labels. Recently there has been a growing interest to generalize discriminative learning methods to handle structured labels, such as label sequences, parse trees, paths over forests. A variety of learning methods have been generalized to the structured case including logistic regression, perceptron (voted and dual), boosting, SVMs and kernel logistic regression (See [1] for a review on this line of research). These techniques combine the efficiency of dynamic programming methods with the advantages of the state-of-the-art learning methods, in particular the ability to learn efficiently in high dimensional feature spaces, either by the use of implicit data representation via kernels or by explicit feature induction.

The studies mentioned above can be summarized as exploring the optimization of different loss functions $\mathcal{L}$ over a sample S using various algorithms. Another important component of learning is regularization and it has not been explored in structured output prediction problems till now. In order to overcome overfitting problem, a regularization term is commonly added to the objective

$$f^* = \operatorname*{argmin}_f \mathcal{L}(S, f) + \epsilon \|f\|, \tag{1}$$

and the regularization coefficient ϵ is determined by cross validation techniques. Convex duality theory provides a unified treatment for this large set of optimization problems and a natural interpretation of the regularization term. For many common loss functions $\mathcal{L}$ (such as log-loss, exp-loss, hinge loss), it has been shown that (1) is the convex dual of minimizing the divergence of a target distribution p from a reference distribution subject to constraints to fit the data (c. f. [2] and references therein). More formally, the data constraints enforce the

expected value of some *features* (measurements) ϕ with respect to the target distribution p should approximately match the expected value of the features wrt to the sample S,

$$\|E_{z\sim p}[\phi(z)] - E_{z\sim S}[\phi(z)]\| \leq \epsilon, \tag{2}$$

where $\|.\|$ denotes the norm of the Banach space that ϕ range over. An important point is that, ϵ in (1) is exactly the relaxation parameter in (2). Moreover, if one has some prior knowledge on the properties of ϕ, the Banach space can be constructed accordingly leading to more informative regularization, which in turn leads to better generalization guarantees as studied in [3].

Learning to predict structured outputs yields a highly structured set of features. One can leverage the knowledge of the structure of the problem for defining the approximation of the empirical and expected values of the features, which in turn leads to different regularizations. We investigate regularization in learning to predict structured outputs and explore two prominent examples of structured output prediction, namely label sequence learning and hierarchical classification.

In hierarchical classification, we study a max-margin formulation, where a linear discriminant function is trained for all of the nodes in the hierarchy jointly and the prediction is made by comparing the energy or the compatibility score of all possible paths in the hierarchy. This formulation allows to accumulate data in higher nodes, therefore can lead to tighter relaxations in the data constraints. We optimize the convex dual of this max-margin formulation with hierarchically structured constraints in term of relaxations. In label sequence learning, two sets of features are defined, label-label interactions and observation-label interactions. These features present very different properties. In particular, the features in the first set are observed much more frequently than the ones in the second set. Thus, one would like to control complexity of the first set and the sparsity on the second. This can be achieved by imposing ℓ_1 norm on observation-label features and ℓ_2 norm on inter-label interaction features.

In this paper, we study the connection between relaxed moment matching constraints and the regularization in the convex dual for structured output prediction tasks. Some preliminary experiments show the advantage of using carefully constructed regularization, in terms of sparsity and improvement in accuracy.

References

[1] Y. Altun. *Discriminative Methods for Label Sequence Learning*. PhD thesis, Department of Computer Science, Brown University, 2005.

[2] Y. Altun and A.J. Smola. Divergence minimization and convex duality. In *Computational Learning Theory*, 2006.

[3] M. Dudik and R.E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *COLT*, 2006.