Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2011-029
CBCL-299

June 3, 2011

Regularization Predicts While Discovering Taxonomy

Youssef Mroueh, Tomaso Poggio, and Lorenzo Rosasco

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Regularization Predicts While Discovering Taxonomy

Youssef Mroueh[♯,†], [♯], Lorenzo Rosasco[♯,†]

♯ - *CBCL, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA*

† -*IIT@MIT Lab, Istituto Italiano di Tecnologia, Genova, Italy*

{ymroueh, poggio, lrosasco}@mit.edu

June 3, 2011

Abstract

In this work we discuss a regularization framework to solve multi-category when the classes are described by an underlying class taxonomy. In particular we discuss how to learn the class taxonomy while learning a multi-category classifier.

1 Introduction

Multi-category classification problems are central in a variety of fields in applied science: given a training set of input patterns (e.g. images) labeled with one of a finite set of labels identifying different categories/classes, infer – that is *learn* – the class of a new input. Within this class of problems the special case of binary classification has received special attention since multi-category problems are typically reduced to a collection of binary classification problems. A common approach is to learn a classifier to discriminate each class from all the others, the so called *one-vs-all* approach.

The availability of new datasets with thousands (and even millions) of examples and classes, has triggered the interest in novel learning schemes for multi-category classification. The one-vs-all approach does not scale as the number of classes increases, since the computational cost for training and testing is essentially linear in the number of classes. More importantly taxonomies that may lead to a more efficient classification are not exploited.

Several approaches have been proposed to address these shortcomings. A group of methods assume the class taxonomy to be known: this includes approaches such as structured and multi-task learning (see for example [13, 12]). In some applications it is possible that the taxonomy is available as a prior information, for example In image classification taxonomies are often derived from databases such as WordNet [11]. Another group of methods addresses the question of learning the taxonomy itself from data. Taxonomies can be learned separately from the actual classification [8, 4], or, more interestingly, while training a classifier (see next sections).

The approach we develop here is of this second type. It is based on a mathematical framework which unifies most of the previous approaches and allows to develop new efficient solutions. Key to our approach is the idea that multi-category classification can be cast as vector valued regression problem where we can take advantage of the theory of reproducing kernel Hilbert spaces. As we discuss in the following, our framework formalizes the idea of class taxonomies using functional and geometric concepts, related to graphical models, while allowing us to make full use of the tools from regularization theory.

The paper is organized as follows. We start recalling some basic concepts on reproducing kernel Hilbert spaces for vector valued functions in Section 2. In Section 3 we first recall how a known taxonomy can be used to design a suitable vector valued kernel, then we discuss how the taxonomy can be learned from data while training a multi-category classifier. We end with some discussion and remarks in Section (4).

2 Reproducing Kernel Hilbert Spaces for vector valued functions

Let $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^T$, we denote by $\langle \cdot, \cdot \rangle$ the euclidean inner product in the appropriate space.

Definition 1. *Let $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$ be a Hilbert space of functions from $\mathcal{X} \longrightarrow \mathcal{Y}$. A symmetric, positive definite, matrix valued function $\Gamma : \mathcal{X} \times \mathcal{X} \longrightarrow \mathcal{Y} \times \mathcal{Y}$ is a reproducing Kernel for $\mathcal{H}$, if for all $x \in \mathcal{X}, c \in \mathcal{Y}$, $f \in \mathcal{H}$ we have that $\Gamma(x, .)c \in \mathcal{H}$ and the following reproducing property holds*

$$\langle f(x), c \rangle = \langle f, \Gamma(x, \cdot)c \rangle_{\mathcal{H}} .$$

In the rest of the paper we will consider for simplicity kernels of the form

$$\Gamma(x, x') = K(x, x')A,$$

where K is a scalar kernel on $\mathcal{X}$ and A a $T \times T$ positive semi-definite matrix, that we call the the *taxonomy* matrix. In this case, functions in $\mathcal{H}$, can be written using the kernel as $f(x) = \sum_{i=1}^{d} K(x, x_i)Ac_i^f$, where $d \leq \infty$, with inner product $\langle f, g \rangle_{\mathcal{H}} = \sum_{i,j=1}^{d} K(x, x_i) \left\langle c_i^f, Ac_j^g \right\rangle$, and norm $\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{d} K(x, x_i) \left\langle c_i^f, Ac_i^f \right\rangle$.

As we discuss below, the taxonomy, when known, can be encoded designing a suitable taxonomy matrix A. If the taxonomy is not known, it can be learned estimating A from data under suitable constraints.

3 A Regularization Approach

In this section we describe a regularization framework for multi-category classification when the categories have an underlying taxonomy. First, we discuss the case where the taxonomy is known and show how such an information can be incorporated in a learning algorithm. Seemingly different strategies can be shown to be equivalent using our framework. Second, if the taxonomy is not known, we discuss how it can be learned *while* training a multi-category classifier. Connection with other methods and in particular Bayesian approaches are discussed.

3.1 Classification with Given Taxonomy

We assume that A is given

$$f_* = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{1}$$

where the T class labels have been turned into vector codes using the encoding $j \mapsto e_j$, where e_j is the canonical basis in $\mathbb{R}^T$. The final classifier is obtained using the classification rule naturally induced by the coding, that is

$$\arg\max_{j=1,\ldots T} f_j^*(x), \tag{2}$$

where $f_j(x) = \langle f(x), e_j \rangle$. The role played by A and its design can be understood from different perspectives listed below.

Kernels and Regularizers. A first insight is obtained by writing [10, 3]

$$\|f\|_{\mathcal{H}}^2 = \sum_{t=1}^{T} \sum_{t'=1}^{T} A_{t,t'}^{\dagger} \langle f_t, f_{t'} \rangle_K ,$$

where $\langle \cdot, \cdot \rangle_K$ is the inner product induced by K and $A^{\dagger}$ is the pseudo-inverse of A. The entries of A or rather $A^{\dagger}$ encode the coupling of the different components. It is easy to see that the one-vs-all approach corresponds to $A_{tt'}^{\dagger} = \delta_{tt'}$ [3]. Different choice of A induce couplings of the different classifiers ([12]). For example if the taxonomy is described by a graph with adjacency/weight matrix W, one can consider the regularizer

$$\sum_{j,t=1}^{T} W_{j,t} \|f_j - f_t\|_K^2 + \gamma \sum_{t=1}^{T} \|f_t\|_K^2, \quad \gamma > 0 \tag{3}$$

which corresponds to $A^\dagger = L_W + \gamma I$ where L_W is the graph laplacian induced by W. We refer to [10, 3] for further references and discussions.

Kernels and Coding. One could replace the standard coding (the one using the canonical basis) with a different coding strategy ([8],[13])

$$i \mapsto b_i, \quad b_i \in \mathbb{R}^L,$$

where the code $\mathcal{B} = \{b_1, \ldots, b_T\}$ reflects our prior knowledge on the class taxonomy. For example if the classes are organized in a tree we can define $(b_i)_j = 1$ if node j is parent of leaf i and $(b_i)_j = -1$ otherwise. Then L is given by T plus the number of nodes minus the number of leafs. In this case one would choose $A = I$ and the classification rule (2) would be replaced by a different strategy that will depend on $\mathcal{B}$. In fact it is easy to show that the above approach is equivalent to using the standard coding and $A = BB^\top$, where B is the $T \times L$ matrix of code vectors. On the contrary, any matrix A can be written as $A = U\Sigma U^\top$ and the rows of $\Sigma^{1/2}U$ define a taxonomy dependent coding.

Kernels and Metric. One can encode prior information on the class taxonomy by defining a new metric on $\mathbb{R}^T$ by the inner product

$$y^\top B y',$$

and choosing $A = I$, see for example [11]. This is equivalent to taking the canonical inner product in $\mathbb{R}^T$ and $A = B$.

3.2 Learning the Taxonomy while Training a Classifier

Next we consider the situation where A is not known and needs to be learned from data. The idea is to replace (1) with

$$\min_{A \in \mathcal{A}} \min_{f \in \mathcal{H}_A} \left\{ \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_A} \right\}, \tag{4}$$

where we wrote $\mathcal{H}_A$ in place of A to emphasize the dependence on A. Here $\mathcal{A}$ is a set of constraints reflecting our prior beliefs about the class taxonomy we want to learn. We give several examples.

Spectral Regularization We consider $\mathcal{A} = \{A \mid A \geq 0, \|A\|_p \leq 1\}$, where $\|A\|_p$ are different matrix norms encoding different prior and $A \geq 0$ denotes the positive semi-definiteness constraint. For example we may consider the the trace norm $\mathrm{Tr}(A)$, forcing A to be low rank.

Information-Theoretic Regularization. A different set of constraints is given by $\mathcal{A} = \{A \geq 0, D(A, I) = \mathrm{Tr}(A) - \log(\det(A)) \leq 1\}$. The above set of constraints can be derived from a probabilistic/information theoretic argument [6]. From a regularization perspective, one can see that, the constraint use a convex constraint $\mathrm{Tr}(A)$ and a strictly convex term $-\log(\det(A))$.

Sparse Graph. ([9, 4]) A natural constraint while consider the regularizer (3) is that of imposing the underlying graph to be sparse. To do this we consider a family of taxonomy matrices parameterized by the weight matrix W subject to a sparsity constraint, i.e.

$$\mathcal{A} = \{A \mid A \geq 0, A^\dagger = L_W, W \in \mathcal{W}\}, \quad \text{with} \quad \mathcal{W} = \{W \mid \|W\|_1 \leq 1, W_{ii} = 0, W_{ij} \geq 0\}, \tag{5}$$

and $\|W\|_1 = \sum_{i,j} W_{i,j}$.

The approach described so far is related to previous work such as [2, 9] which is in the same spirit. Connection with graphical models estimation are discussed in the next section.

3.3 Connections to Unsupervised Learning of Graphical Models

A (Gaussian) graphical model is described by a T-dimensional multivariate normal distribution with mean μ and covariance Σ; it is well known that any i, j component is conditionally independent if the corresponding entry of the inverse covariance matrix satisfies $(\Sigma^{-1})_{i,j} = 0$. The estimate of the Gaussian graphical model, from N sample with empirical covariance S requires the estimation of a sparse approximate inverse covariance Θ by minimizing the functional

$$\mathrm{Tr}(\Theta S) - \log(\det(\Theta)) + \|\Theta\|_1 . \tag{6}$$

This functional can be derived from a Bayesian perspective, see for example [5] and references therein.

The approach outlined so far can be used to learn a taxonomy (before training a classifier), if for each category a feature description is available. Each feature can then be seen as an observation of a T-dimensional normal distribution, T being the number of categories. This is the point of view taken in [4] (see also [7]).

A feature description of the classes is almost equivalent to the taxonomy matrix A. It is natural to ask what may be done when such a feature description is not available. We propose here that for available supervised data we can leverage the indirect knowledge of the input-output relation to learn the taxonomy. To see the connection to (6) it is useful to rewrite (4) in the following way. For simplicity let us consider the case where K is a linear kernel. Then, given a taxonomy matrix A, the functions in the corresponding vector valued RKHS can be written as $f(x) = W_A x$, where [1] W_A is a $T \times p$ matrix which depends on A and the corresponding norm is $\|f\|^2_{\mathcal{H}_A} = \mathrm{Tr}(W_A^\top W_A)$. In the case of the square loss, the simple change of variable $W_A = A^\dagger W$ leads to (4)

$$\min_{W \in \in \mathbb{R}^{T \times p}} \left\{ \frac{1}{n} \|Y - WX\|_F^2 + \lambda \min_{A \in \mathcal{A}} \mathrm{Tr}(W^\top A^\dagger W) \right\}, \tag{7}$$

where, due to the re-parameterization, the minimization over A is only over the regularization term.

We can see now how the problem of Equation (6) is related to Equation (7)– and hence to Equation (4) when $\mathcal{A} = \{A \geq 0 \mid \mathrm{Tr}(A) - \log(\det(A)) \leq 1\}$. By Solving (7) we use the input output data to estimate a matrix W which capture how the classes are related. At the same time A is learned to give a better estimate of the taxonomy information encoded in W. The two estimation problems depend on each other: the best prediction is sought while estimating the best taxonomy. We end noting that more complex priors can be considered and that they will correspond to more complex regularizers. For example one can show that the hierarchical Bayesian prior considered in [4] corresponds exactly to the regularizer in (5).

4 Discussion

In this paper we describe a framework for solving multi-category classification when there is an underlying taxonomy. In particular, we discuss how to learn simultaneously a classifier and the underlying class taxonomy. The proposed framework has some interesting properties:

- It leads to a unified view of several seemingly different algorithms, clarifying the connection between regularization approaches and other methods such as graphical models.

- A few computational advantages follow from the approach described here. In the case when the matrix A is known (1), we can prove [3] that, for a large class of convex loss functions, the computational cost for training a multi-category classifier is the same as a binary classification problem. In the case of the problem of Equation (4) it is possible to prove that for a large class of loss functions and for broad sets of constraints, the corresponding optimization problem is convex and can be solved efficiently, using ideas from [1].

- Finally, since the problem is cast in a regularization framework, the analysis of the sample complexity of the algorithms can be studied in a unified way.

[1] Note that by definition we have $f(x) = \sum_{i=1}^d \langle x_i, x \rangle Ac_i = \sum_{i=1}^d \sum_{j=1}^p x_i^j x^j Ac_i = \sum_{j=1}^p x^j \sum_{i=1}^d x_i^j Ac_i$ so that W' has columns $\sum_{i=1}^d x_i^j Ac_i$.

Acknowledgments

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-05- C-7262) Additional support was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73, 2008.

[2] A. Argyriou, A. Maurer, and M. Pontil. An algorithm for transfer learning in a heterogeneous environment. In *ECML/PKDD*, pages 71–85, 2008.

[3] L. Baldassarre, A. Barla, B. Gianesin, and M. Marinelli. Vector valued regression for iron overload estimation. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, Dec. 2008.

[4] Brenden Lake and Joshua Tenenbaum. Discovering structure by learning sparse graph. *Proceedings of the 33rd Annual Cognitive Science Conference*, 2010.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008.

[6] Jason V.Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S.Dhillon. Information theoretic metric learning. *ICML*, 2007.

[7] Charles Kemp and Joshua B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, August 2008.

[8] Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. *Journal of machine learning*, 2003.

[9] Laurent Jacob, Francis Bach, and Jean Philippe Vert. Clustered multitask learning: A convex relaxation. *Advances in Neural Information Processing Systems (NIPS 21)*, 2009.

[10] C. A. Micchelli and M. Pontil. Kernels for multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2004.

[11] Rob Fergus, Hector Bernal, Yair Weiss, and Antonio Torralba. Semantic label sharing for learning with many categories. *Proc. of IEEE European Conference on Computer Vision*, 2010.

[12] Massimiliano Pontil Theodoros Evgeniou, Charles A.Michelli. Learning multiple tasks with kernel methods. *Journal of machine learning*, 2006.

[13] Thorsten Joachims, Thomas Hofmann, Yisong Yue, and Chun-Nam Yu. Predicting structured objects with support vector machines. *Communications of the ACM, Research Highlight, 52(11):97-104*, 2009.