LOSS FUNCTIONS FOR MULTICLASS LEARNING AND THE GEOMETRY OF THE LABEL SPACE

YOUSSEF MROUEH[♯,‡], TOMASO POGGIO[♯], LORENZO ROSASCO[♯,‡] JEAN-JACQUES E. SLOTINE[†]
♯ - *CBCL, MCGOVERN INSTITUTE, BCS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY*
† - ISTITUTO ITALIANO DI TECNOLOGIA
† - *ME, BCS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY*
{YMROUEH,TP, LROSASCO,JJS}@MIT.EDU

As bigger and more complex datasets are available, multiclass learning is becoming increasingly important. While theory and algorithms for solving binary classification problems are available, the problem of multicategory classification is much less understood. Indeed, practical multiclass algorithms often reduce the problem to a collection of binary classification problems. Binary classification algorithms typically starts from a *relaxation approach*: classification is posed as a non-convex minimization problem and hence relaxed to convex one, defined by suitable convex loss functions. In this context, results in statistical learning theory quantify the error incurred by relaxation (to further derive consistency and sample complexity results), see for example [1].

Generalizing the above approach to more than two classes is not straightforward. Previous results show that naive extension of binary classification approaches might lack basic properties such as statistical consistency [5]. At the root of these difficulties there is the more complex label structure when dealing with multiple categories: suitable coding strategies [2] and loss functions [6] need to be defined to avoid unnatural ordering among the classes, and constraints on the function class needs to be imposed, see[1, 6]. In particular, the latter constraints often lead to complications from a computational point of view, and is usually ignored in practice at the price of a loss on the consistency guarantees.

Here we study a suitable coding/decoding strategy, namely the simplex coding [3]. This strategy is the natural generalization of the ± 1 coding for binary classification and allows to cast multicategory classification as a vector valued regression problem. Our analysis shows that the simplex coding allows to generalize concepts, results and proof techniques from the binary case, which, interestingly, is recovered as a special case. In particular we prove new explicit results for the relaxation error, when considering convex loss functions, e.g. the least square and the hinge loss functions. We prove that these loss functions, coupled with the simplex coding/decoding strategies leads to fisher consistent algorithms and, especially, we derive explicit comparison theorems relating the excess misclassification error with the excess expected loss. We further describe the practical implementation of the corresponding regularization schemes in the context of vector valued reproducing kernel Hilbert spaces [4]. In particular, we discuss: 1) the complexity of the obtained algorithms (batch and online), and how the possible structure among the classes can be incorporated or learned using a constrained optimization approach.

REFERENCES

[1] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, March 2005.
[2] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.
[3] Simon I. Hill and Arnaud Doucet. A framework for kernel-based multi-category classification. *J. Artif. Int. Res.*, 30:525–564, December 2007.
[4] C.A. Micchelli and M. Pontil. On learning vector–valued functions. *Neural Computation*, 17:177–204, 2005.
[5] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. In *Proceedings of the 18th Annual Conference on Learning Theory*, volume 3559, pages 143–157. Springer, 2005.
[6] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

Topic: learning theory; Learning algorithms;
Preference: oral