A Rate-Distortion One-Class Model and its Applications to Clustering[*]

Koby Crammer Partha Pratim Talukdar Fernando Pereira[†]

CIS Department, University of Pennsylvania
Philadelphia, PA 19104
Email: {crammer,partha,pereira}@cis.upenn.edu

We study the problem of one-class classification, in which we seek a rule to separate a coherent subset of instances similar to a few positive examples from a large pool of instances. For instance, in document retrieval we might want to retrieve a small set of documents similar to a few seed examples. For pathway analysis in genomics, it is useful to find other genes that are co-expressed with a few genes of interest. In both cases, we prefer high-precision answers over high-recall ones.

A popular intuition for this *one-class classification* problem is that of finding a small ball (under some appropriate norm) that contains as many of the seed elements as possible [8]. Most previous approaches to the problem take the point of view of outlier and novelty detection, in which most of the examples are identified as relevant. Rather than keeping all but a few outliers, a small subset of relevant examples is identified in [4].

Current approaches to one-class classification use convex cost functions that focus on large-scale structures in the data. Those functions grow linearly outside class and and are constant inside it [6, 8, 2]. In a related study, [7] seek to separate most of the examples from the origin using a single hyperplane. More recently, [5] generalized that approach to the general case of Bregman divergences. In all of those methods, the convexity of the cost function forces the the solution to shrink to the center of mass as the radius of the ball goes to zero, thus ignoring any local substructure.

Using ideas from rate-distortion theory [3], we propose to express the one-class problem in terms of lossy coding of each instance into a few possible instance-dependent codewords. Unlike previous methods that use just two [4] or a small number [1] of possible codewords for all instances, the total number of codewords in our method is greater than the number of instances. To preclude trivial codings, we force each instance to associate only with a few possible codewords. Finding the best coding function is an optimization problem for which we provide an efficient algorithm. The optimization has an "inverse temperature" parameter that represents the tradeoff between compression and distortion. As temperature decreases, the solution passes through a series of phase transitions associated with different sizes for the one class. This model outperforms two previous algorithms proposed for the problem, which are effective only in more restricted situations.

[*] Category: Learning Algorithms

[†] now at Google, California.

Our one-class model is also effective on the task of clustering a set of instances into multiple classes when some of the instances are clutter that should not be included in any cluster. This task can be reduced to an alternation between applications of the one-class algorithm and hard clustering. Initial experiments with synthetic and real world data show that by leaving some instances out of the clusters, the quality of the clustering improves.

References

[1] R. Bekkerman and A. McCallum. Disambiguating web appearances of people in a social network. In *WWW*, 2005.

[2] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. Support vector clustering. *JMLR*, 2:125–137, 2001.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[4] K. Crammer and G. Chechik. A needle in a haystack: Local one-class optimization. In *ICML 23*, 2004.

[5] K. Crammer and Y. Singer. Learning algorithms for enclosing points in bregmanian spheres. In *COLT 16*, 2003.

[6] B. Schölkopf, C. Burges, and V.N. Vapnik. Extracting support data for a given task. In *KDD 1*, 1995.

[7] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1472, 2001.

[8] D.M.J. Tax and R.P.W. Duin. Data domain description using support vectors. In *ESANN*, pages 251–256, April 1999.