Stability of Clustering Methods

Sasha Rakhlin

rakhlin@mit.edu

Massachusetts Institute of Technology

Abstract

Stability of clustering has been proposed as a measure of quality for clustering algorithms (see [9, 7]). Since stability is a desireable property, the natural next step is to examine stability of some known clustering methods. We propose to use the recent result of Caponnetto and Rakhlin [4] for this purpose. In this talk, we would like to outline directions of on-going research, which aim to answer some fundamental theoretical questions about the stability of clustering.

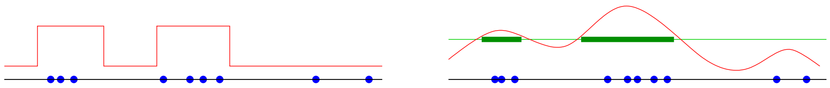
The main stability result of [4] states that the approximate Empirical Risk Minimization algorithm over Donsker classes is stable in the $L_1(\mu)$ sense, where μ is the measure generating the data. Donsker classes are classes of functions satisfying a version of the Central Limit Theorem, and are somewhat more restrictive than Glivenko-Cantelli classes, which only satisfy a version of the Law of Large Numbers. In the binary case, the Donsker property coincides with the finiteness of the VC dimension. More precisely, the stability result says that

$$\|f_S - f_T\|_{L_1(\mu)} \xrightarrow{P} 0,$$

where f_S is an approximate empirical minimizer over the set S of n points and f_T is an approximate empirical minimizer over the set T, such that S and T differ on $o(\sqrt{n})$ points. Specific rates of the above convergence in probability are available under certain entropy assumptions (see [4]). In plain words, as n increases, it becomes less and less likely that changing fewer than $\sqrt{n}$ points of the dataset results in a jump to a far (in the L_1 sense) hypothesis. This result is non-trivial and relies on control of fluctuations of empirical means over the whole function class by studying the limiting Gaussian process.

There are several applications of the above stability of approximate ERM to clustering. Let us briefly sketch them.

First, we apply the L_1 stability to methods of clustering which identify high-density regions (e.g. [5]). We illustrate the idea by exhibiting two situations where high-density regions are stable with respect to perturbations of $o(\sqrt{n})$ samples. Roughly speaking, it takes at least $\sqrt{n}$ additional points to shift the high-density region by any constant amount. One method of identifying these regions is Empirical Risk Minimization over a binary-valued function class. In this setting, one searches for a binary function which best fits the given data by placing "steps" over the high-density regions, as illustrated in the figure below (left). If the function class is VC, our stability result implies that small perturbations of the data are likely to result in only small perturbations of the clusters. Note that even this seemingly simple result is non-trivial: it is not clear a-priori whether ERM over a VC class is L_1-stable. Instead of searching a binary function class, one can find the best density via Maximum Likelihood



Estimation and consider the level sets to cluster the data. Since MLE is minimizing empirical log-loss, the result of [4] guarantees that the L_1 stability holds if the class of densities is Donsker. Closeness in the L_1 distance implies (under some smoothness and boundedness assumptions) the stability of respective clusters, extended to the full domain.

The drawback of the methods mentioned above is that ERM is often hard (see [2]). Some approximation methods, like the greedy procedure of A. Barron [6, 8], can be employed to alleviate this problem. Regardless of the computational issues for these methods, we believe that stability properties of these clustering procedures are important for the theoretical foundations of clustering.

The second application of the L_1 stability is to k-means/medians clustering [1]. In the particular case of k-means clustering with square loss, the problem can be written as an Empirical Risk Minimization over the class

$$\mathcal{H} = \{\|x - y_{m(x)}\|^2 : m(x) \text{ is a } k \text{ partition and } y_{m(x)} \text{ are centroids}\}$$

(see [3]). Again, the stability result of [4] can be applied if this class is shown to be Donsker, which is possible under some regularity assumptions. We believe that the above reasoning can be extended to other clustering methods which minimize some objective function, and we expect to obtain general results in this direction.

To summarize, we sketched how the L_1 stability of approximate ERM facilitates the theoretical study of clustering stability, which has been delineated as a quality measure for clustering.

References

[1] Shai Ben-David. A framework for statistical clustering with a constant time approximation algorithms for k-median clustering. In *COLT*, pages 415–426, 2004.

[2] Shai Ben-David, Nadav Eiron, and Philip M. Long. On the difficulty of approximately maximizing agreements. *J. Comput. Syst. Sci.*, 66(3):496–514, 2003.

[3] Joachim M. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical Report IAI-TR-98-3, 3, 1998.

[4] A. Caponnetto and A. Rakhlin. Some properties of empirical risk minimization over Donsker classes. Available at http://cbcl.mit.edu/people/rakhlin/erm.pdf. Submitted to JMLR, 2005.

[5] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Cluster analysis: a further approach based on density estimation. *Comput. Stat. Data Anal.*, 36(4):441–459, 2001.

[6] J. Li and A. Barron. Mixture density estimation. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural information processings systems 12*, San Mateo, CA, 1999. Morgan Kaufmann Publishers.

[7] Vijay V. Raghavan. Approaches for measuring the stability of clustering methods. *SIGIR Forum*, 17(1):6–20, 1982.

[8] A. Rakhlin, D. Panchenko, and S. Mukherjee. Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, 9:220–229, 2005.

[9] Ulrike von Luxburg and Shai Ben-David. Towards a statistical theory of clustering. PASCAL Workshop on Statistics and Optimization of Clustering, 2005.