Stability of Clustering Algorithms Under the ERM Scheme with Infinite Large Datasets

Zheng Liu ZLIU86@ILLINOIS.EDU
Department of Nuclear, Plasma, and Radiological Engineering
University of Illinois
Urbana, IL 61801, USA

Abstract

This ECE543 final report studies the stability of clustering algorithms under infinite large dataset. An empirical risk minimization(ERM) scheme is introduced to represent the center based clustering algorithms, such as k-means algorithms and spectral clustering algorithms. Within the ERM scheme, two theorems are stated regarding to the stability of clustering algorithms for different data distributions. If the dataset has unique minimizer, the clustering algorithm tends to be stable. If the dataset has multiple distinct minimizers, the clustering algorithm tends to be unstable. Because the number of minimizers does not related to the correctness of models, thus the stability of clustering algorithm should not be used in model selection under infinite large dataset. Two k-means clustering examples are also presented to demonstrate those two theorems.

Keywords: Clustering, Stability, Empirical Risk Minimization

1. Introduction

There are two different groups of algorithms in machine learning: supervised learning, and unsupervised learning. Supervised learning algorithms aim to learn the relationship between input and output using labeled datasets, while unsupervised learning algorithms aim to learn the intrinsic structure of a dataset. In the class of ECE543, we have discussed the learnability of supervised learning algorithms under different conditions, such as the uniform convergence property and the strong convexity of loss functions. We represented the supervised learning algorithms in the empirical risk minimization (ERM) scheme and analyzed those algorithms' stability, generality, and consistency. In this report, we represent one class of unsupervised learning algorithms in the ERM scheme and analyzed their stability under infinite large dataset.

In clustering algorithms, users need to define several model properties before running the algorithm. Different model properties represent different models. People want to select the model that represents data the best. In application, the stability of clustering algorithms is widely used in model selection. However, the theoretical foundation of model selection via clustering stability was not very clear until recently. Ben-David and Luxburg proposed a way to represent the center based clustering algorithms within the ERM scheme and analyzed the stability of such algorithms with infinite large dataset(Ben-David et al. (2006)). Shamir and Tishby analyzed the stability of clustering algorithms with finite large

dataset, and theoretically explained the connection between stability and model selection (Shamir and Tishby (2008a,b, 2009)).

This report basically adopts the setup of Ben-David and Luxburg's work. In section two, we introduced the ERM framework for clustering. In section three, we presented and proved two important theorems for clustering stability under infinite large dataset. We also demonstrated those theorems with intuitive samples through the k-means algorithm. In section four, we summarized our work and expanded our discussion for finite large dataset.

2. ERM framework for clustering

Here we setup the standard notation for this report. There is a data space $X \in \mathbb{R}^n$ endowed with probability measure $P \in \mathbb{P}$. With a specific probability P, we i.i.d. sample each element of the dataset $S = \{x_1, x_2, ..., x_m\} \in (X^m, P^m)$.

Definition 1 (Clustering) A clustering C of data space X is a finite partition $C : X \to \mathbb{N}$, and a specific data cluster is defined as $c_i := \{x \in X; C(x) = i\}$. C belongs to a family of clusterings $\mathcal{C}$.

Definition 2 (Clustering distance) A clustering distance is a mapping $d : \mathbb{P} \times \mathcal{C} \times \mathcal{C} \to [0, 1]$. For a specific probability distribution P, it maps two clusterings to a real number between zero and one. For all $P \in \mathbb{P}$, and for all possible clusterings $C_a, C_b, C_c \in \mathbb{C}$, it has the following properties:

- $d_P(C_a, C_a) = 0$
- $d_P(C_a, C_b) = d_P(C_b, C_a)$ (Symmetry)
- $d_P(C_a, C_b) \leq d_P(C_a, C_c) + d_P(C_c, C_b)$ (Triangle inequality)

Definition 3 (Clustering algorithm) A clustering algorithm A is a function that maps a dataset with m examples to a clustering. $A : X^m \to \mathcal{C},\ m \in \mathbb{N}$

Definition 4 (Instability of algorithm A) The *instability* of algorithm A for sample size m with respect to the probability distribution P is defined as the expected distance between two clusterings obtained from A using two independent datasets:

$$instab(A, P, m) := \mathbb{E}_{S_1, S_2 \in P^m} d_P(A(S_1), A(S_2)) \tag{1}$$

Further, the instability under infinite large dataset is defined as

$$instab(A, P) := \lim_{m \to \infty} instab(A, P, m). \tag{2}$$

Definition 5 (Risk of clustering) The risk of clustering C with respect to a specific probability distribution P is a function that maps a clustering into a nonnegative real number. $R(P, C) : \mathbb{P} \times \mathcal{C} \to \mathbb{R}_0^+$.

Definition 6 (Optimized risk) The optimized risk R_P^* is the infimum of the clustering risk. $R_P^* := inf_{C \in \mathcal{C}} R(P, C)$

With those definitions, we are able to define the ERM scheme.

Definition 7 (Empirical risk minimization scheme) Given $(X, \mathcal{C}, \mathbb{P}, R)$, the ERM scheme is that a clustering algorithm aims to find $C \in \mathcal{C}$ that minimize the risk R(P,C).

All the center based clustering algorithms can be represented within the ERM scheme(Ben-David et al. (2006)). For example, the k-means algorithm tries to minimize the following risk:

$$R(P, C) = \mathbb{E}_P \min_{1 \leq i \leq k} ||x - a_i||_2^2 \tag{3}$$

where $(a_1, ..., a_k)$ are the centers of k clusters. For different k value, the k-means algorithm has different risk functions.

Definition 8 (Empirical risk) Define the empirical probability distribution of a dataset S to be P_S. The empirical risk of clustering C for dataset S is $R(P_S, C)$.

Definition 9 (R-minimizing) Clustering algorithm A is called R-minimizing if it returns the optimal empirical risk, that is $R(P_S, A(S)) = R_{P_S}^*$

Definition 10 (Risk convergence) The R-minimizing algorithm is called risk converging if for any $\epsilon > 0$ and for any $\delta \in (0, 1)$, there exists $m_0 \in \mathbb{N}$ such that for all $m \geq m_0$, $S = \{x_1, x_2, ..., x_m\} \in P^m$, the probability is lower than δ for the event that the risk of algorithm A on dataset S exceeds the optimized risk by ϵ.

$$Pr\ \{R(P, A(S)) < R_P^* + \epsilon\} > 1 - \delta \tag{4}$$

Definition 11 (Unique minimizer) A probability distribution P is said to have a unique minimizer C^* if for any $\eta > 0$, there exists $\epsilon > 0$ such that

$$R(P, C) < R_P^* + \epsilon \Rightarrow d_p(C^*, C) < \eta \tag{5}$$

Definition 12 (Multiple distinct minimizers) A probability distribution P is said to have n distinct minimizers $\{C_1^*, ...C_n^*\}$ if for any $\eta > 0$, there exists $\epsilon > 0$, and $1 \leq i \leq n$ such that

$$R(P, C) < R^* + \epsilon \Rightarrow d_p(C_i^*, C) < \eta \tag{6}$$

$$and\ \forall i \neq j, d_P(C_i^*, C_j^*) > 0 \tag{7}$$

3. Stability of clustering under infinite large dataset

Based on previous definitions, we analyze the stability of risk convergence clustering algorithms under different data distribution. Theorem one states the situation where P has a unique minimizer. Theorem two states the situation where P has multiple minimizers.

The statements and proofs are based on Ben-David and Luxburg's paper (Ben-David et al. (2006)).

Theorem 1 If P has unique minimizer C^*, any risk convergence algorithm is stable on P under infinite large dataset.

Proof: Given $(X, \mathcal{C}, \mathbb{P}, R)$, a risk convergence algorithm A, and a large enough dataset with size m, the goal is to show

$$instab(A, P, m) < \zeta, \forall \zeta > 0 \tag{8}$$

Firstly pick $\delta \in (0, 1)$ and $\eta > 0$ such that $2(\delta + \eta) < \zeta$. Because A has unique minimizer C^*, there exists $\epsilon > 0$ such that

$$R(P, C) < R_P^* + \epsilon \Rightarrow d_p(C^*, C) < \eta \tag{9}$$

Because A is risk converging, there exists m_0 such that for any $m > m_0$,

$$Pr\ \{R(P, A(S)) \geq R_P^* + \epsilon\} < \delta \tag{10}$$

Combining equation (9) and (10), we have:

$$Pr\ \{d_p(C^*, A(S)) \geq \eta\} \leq Pr\ \{R(P, A(S)) \geq R_P^* + \epsilon\} < \delta \tag{11}$$

Then we write the instability of algorithm A as follows:

$$\begin{aligned}
& instab(A, P, m) && (12)\\
= & \mathop{\mathbb{E}}_{S_1, S_2 \sim P^m} d_P(A(S_1), A(S_2)) && (13)\\
\leq & \mathop{\mathbb{E}}_{S_1, S_2 \sim P^m} [d_P(A(S_1), C^*) + d_P(A(S_2), C^*)] && (14)\\
= & 2 \mathop{\mathbb{E}}_{S \sim P^m} d_P(A(S), C^*) && (15)\\
\leq & 2(\eta \cdot \mathop{Pr}_{S \sim P^m}(d_P(A(S), C^*) < \eta) + 1 \cdot \mathop{Pr}_{S \sim P^m}(d_P(A(S), C^*) > \eta) && (16)\\
\leq & 2(\eta + \mathop{Pr}_{S \sim P^m}(d_P(A(S), C^*) > \eta)) && (17)\\
\leq & 2(\eta + \delta) && (18)\\
< & \zeta && (19)
\end{aligned}$$

Line 14 use the triangle inequality of clustering distance. Line 16 uses the property of expectation. Line 17 uses the result of equation 11. ■

Theorem 2 If P has n distinct minimizers (for example, due to the symmetry of P), any risk convergence algorithm is instable on P under infinite large dataset.

The detailed proof is in reference(Ben-David et al. (2006)). Intuitively, if P has multiple minimizers, the ERM scheme will "randomly" converge to one of the minimizers.

In the rest of this section, we'll show two examples demonstrating those two theorems. Before showing the examples, we need to find a way to empirically estimate the instability of algorithm A under (X^m, P). The empirical instability is estimated using the following procedure. Given $(X, \mathcal{C}, \mathbb{P}, R)$, we i.i.d sample (2m+1)n data points from distribution P and split them equally into 2m+1 sets $\{S_1, S_2, ..., S_{2m+1}\}$ with n points in each set. The first 2m datasets are used to train clusterings $\{C_1, ..., C_{2m}\}$,and the last dataset is used to calculate the distance between clusterings:

$$d(C, C') = \min_{\pi} \frac{1}{n} \sum_{i: x_i \in S^{2m+1}} 1_{\{C(x_i) \neq \pi(C'(x_i))\}} \tag{20}$$

The minimum above is taken over all possible permutations π of the clusters. The empirical estimator of $instab(A, P)$ is defined as

$$\widehat{instab(A, P)}_m = \frac{1}{m} \sum_{i=1}^{m} d(A(S_{2i-1}), A(S_{2i})) \tag{21}$$

The following two examples use the k-means algorithm to cluster datasets. In both of the two examples, the dataset has only one cluster, but we set the k equals to two in the k-means algorithm. In other words, both of the two examples use wrong model. Figure one shows the experiment result for example one. When the data is uniformly distributed in a slim rectangular area, the k-means algorithm is stable and always converge to the clustering that split the dataset into left part and right part. The estimated instability converges to zero as dataset size goes to infinity. This situation is described by theorem one.

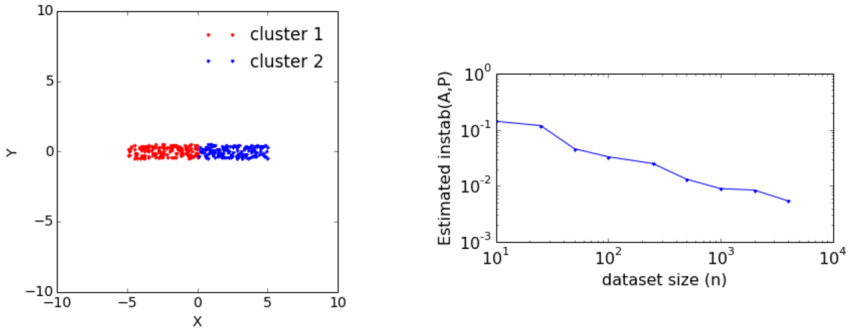

Figure 1: Example one. The data is uniformly distributed in a slim rectangular area. The k is set to two in k-means algorithm. The left plot shows the data distribution with color denoting two clusters obtained from k-means algorithm. The right plot shows that the estimated instability converges to zero as the dataset size increases.

Figure two shows the experiment result for example two, where the data is uniformly distributed in a square area. In this distribution, the dataset has rotation symmetry for every 90 degrees. The k-means algorithm gives two different possible clusterings for this

situation: split the dataset as upper part and lower part, or split the dataset as left part and right part. Under this data distribution, the empirical instability does not converge to zero as dataset size increases. This situation is described by theorem two.

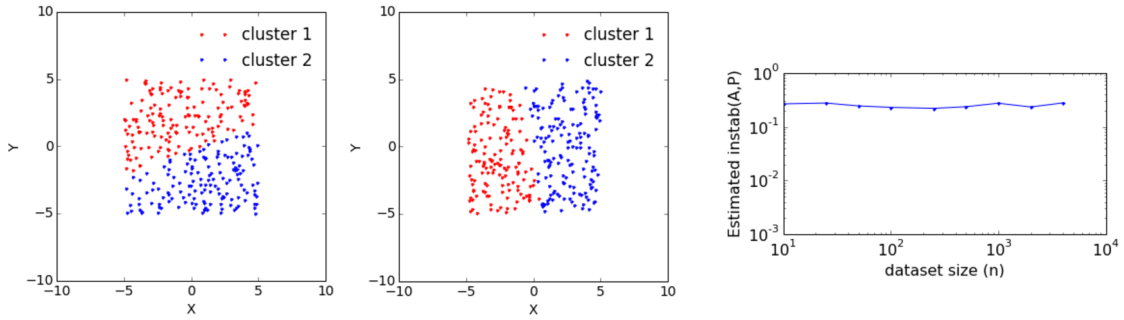

Figure 2: Example two. The data is uniformly distributed in a square area. The k is set to two in k-means algorithm. The left two plots show the data distribution, and also two different clustering results obtained from k-means algorithm. The right plot shows that the estimated instability does not converge.

Both of the two examples use wrong models. However, example one provided stable clustering result, while example two is unstable. This demonstrates that under infinite large dataset, stability should not be used to select models.

4. Conclusion and further discussions

In this report, an empirical risk minimization scheme is introduced to represent the center based clustering algorithms. Based on the scheme, two theorems about clustering stability under infinite large dataset are analyzed. Theorem one states that the idealized risk convergence algorithm is stable if there has unique global minimizer. Theorem two states that the idealized risk convergence algorithm is not stable if there are multiple global minimizers. Thus the stability of an algorithm should not be used to select models. However, the stability is still widely used in model selection in real application. In order to explain this, further study needs to be done to analyze the clustering stability under finite large dataset. Shamir and Tishby showed that for idealized risk convergence algorithm with finite samples, the convergence speed of the empirical instability can be used to select models (Shamir and Tishby (2009)).

References

Shai Ben-David, Ulrike Von Luxburg, and Dávid Pál. A sober look at clustering stability. In *International Conference on Computational Learning Theory*, pages 5–19. Springer, 2006.

Ohad Shamir and Naftali Tishby. Cluster stability for finite samples. In *Advances in neural information processing systems*, pages 1297–1304, 2008a.

Ohad Shamir and Naftali Tishby. Model selection and stability in k-means clustering. In *COLT*, pages 367–378. Citeseer, 2008b.

Ohad Shamir and Naftali Tishby. On the reliability of clustering stability in the large sample regime. In *Advances in neural information processing systems*, pages 1465–1472, 2009.