Stability of Clustering

under the ERM scheme with infinite large dataset(Ben-David and Luxburg, 2006)

Zheng Liu

May 10, 2017

Nuclear, Plasma, and Radiological Engineering, UIUC

Table of contents

ERM framework for clustering

ERM framework for clustering

Basic definitions:

- Data space $X \in \mathbb{R}^n$ endowed with probability measure $P \in \mathbb{P}$
- A Sample $S = \{x_1, x_2, ..., x_m\}$ is drawn i.i.d from (X^m, P^m)
- A clustering C of set 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 $\mathcal{C}$.
- A clustering distance is a mapping $d : \mathbb{P} \times \mathcal{C} \times \mathcal{C} \to [0, 1]$. It has the following properties: $\forall P \in \mathbb{P}$ and $\forall C_a, C_b, C_c \in \mathcal{C}$
    1. $d_P(C_a, C_a) = 0$
    2. $d_P(C_a, C_b) = d_P(C_b, C_a)$
    3. $d_P(C_a, C_b) \leq d_P(C_a, C_c) + d_P(C_c, C_b)$

ERM framework for clustering

Basic definitions (continued):

- A clustering algorithm A is a mapping $A : X^m \to \mathcal{C},\ m \in \mathbb{N}$
- The *instability* of algorithm A for sample size m with respect to the probability distribution P is :

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

  Further, the instability under infinite large dataset is defined as

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

- The risk of clustering C is $R(P, C) : \mathbb{P} \times \mathcal{C} \to \mathbb{R}_0^+$.
- The optimized risk is $R_P^* := inf_{C \in \mathcal{C}} R(P, C)$

ERM framework for clustering

Why do we need to measure the stability of a clustering algorithm?

- Model selection: Find the optimized cluster number k.
- Connection between stability and generalization(Shamir and Tishby, 2008).

ERM framework for clustering

Emperical Risk Minimization Scheme

Given $(X, \mathcal{C}, \mathbb{P}, R)$, the clustering algorithm aims to find $C \in \mathcal{C}$ that minimize the risk R(P,C).

- Example: k-means algorithm. It aims to minimize the following risk:

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

where $(a_1, ..., a_k)$ are the centers of k clusters.

ERM framework for clustering

Emperical risk

For a sample $S \subseteq X$, the empirical risk of clustering C is $R(P_S, C)$. Here the P_S is the empirical probability distribution of S.

R-minimizing

A clustering algorithm is called *R-minimizing* if $R(P_S, A(S)) = R^*_{P_S}$

Risk Convergence

The R-minimizing algorithm is called risk converging if $\forall \epsilon > 0$ and $\forall \delta \in (0, 1)$, $\exists m_0 \in \mathbb{N}$ such that $\forall m \geq m_0$, $S = \{x_1, x_2, ..., x_m\} \in P^m$

$$Pr\left\{R(P, A(S)) < R^*_P + \epsilon\right\} > 1 - \delta \tag{3}$$

ERM framework for clustering

Unique minimizer

A probability distribution P is said to have a unique minimizer C^* if $\forall \eta > 0, \exists \epsilon > 0$ such that

$$R(P, C) < R_P^* + \epsilon \Rightarrow d_p(C^*, C) < \eta \tag{4}$$

Several distinct minimizers

A probability distribution P is said to have n distinct minimizers $\{C_1^*, ...C_n^*\}$ if $\forall \eta > 0, \exists \epsilon > 0$, and $\exists 1 \leq i \leq n$ such that

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

$$and\ \forall i \neq j, d_P(C_i^*, C_j^*) > 0 \tag{6}$$

Stability of clustering under infinite large dataset

Unique minimizer under infinite large dataset

Theorem 1

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

Proof of Thm 1 (Ben-David and Luxburg, 2006)

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

$$instab(A, P, m) < \zeta, \forall \zeta > 0 \tag{7}$$

Unique minimizer under infinite large dataset

Proof of Thm 1 (continued)

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

$$R(P, C) < R_P^* + \epsilon \Rightarrow d_p(C^*, C) < \eta \tag{8}$$

Because A is risk converging, there $\exists m_0$ such that $\forall m > m_0$,

$$Pr\left\{R(P, A(S)) \geq R_P^* + \epsilon\right\} < \delta \tag{9}$$

Combining (8) and (9), we have:

$$Pr\left\{d_p(C^*, A(S)) \geq \eta\right\} \leq Pr\left\{R(P, A(S)) \geq R_P^* + \epsilon\right\} < \delta \tag{10}$$

Unique minimizer under infinite large dataset

Proof of Thm 1 (continued)

Finally we have

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

n distinct minimizers under infinite large dataset

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 and Luxburg, 2006). Intuitively, if P has multiple minimizers, the ERM scheme will ”randomly” converge to one of the minimizers.

Examples for clustering stability

Empirical estimator

Empirical estimator of instab(A,P)

Given $(X, \mathcal{C}, \mathbb{P}, R)$, we i.i.d sampled (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 were used to train clusterings $\{C_1, ..., C_{2m}\}$,and the last dataset was used to calculate the distance between clusterings:

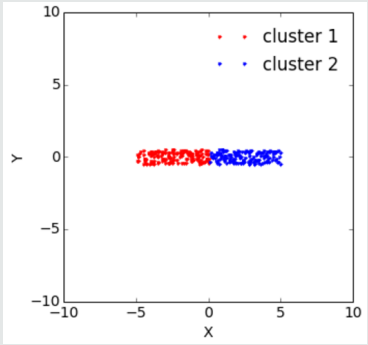
$$d(C, C') = \min_{\pi} \frac{1}{n} \sum_{i: x_i \in S^{2m+1}} 1_{\{C(x_i) \neq \pi(C'(x_i))\}} \tag{11}$$

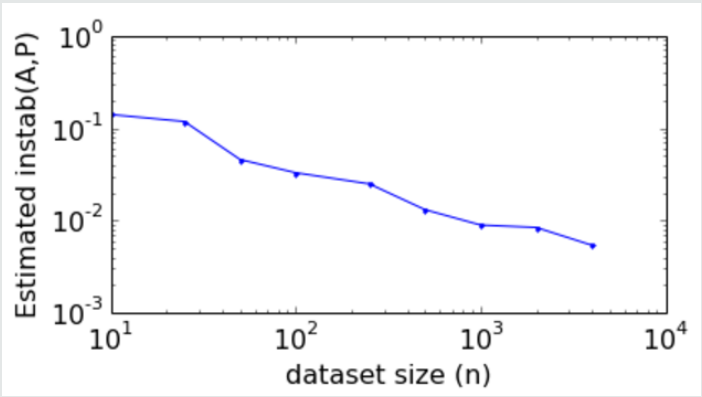
The minimum above is taken over all possible permutations π of the clusters. The empirical estimator of $instab(A, P)$ is defined as

$$\widehat{instab(A, P)}_m = \frac{1}{m} \sum_{i=1}^{m} d(A(S_{2i-1}), A(S_{2i})) \tag{12}$$

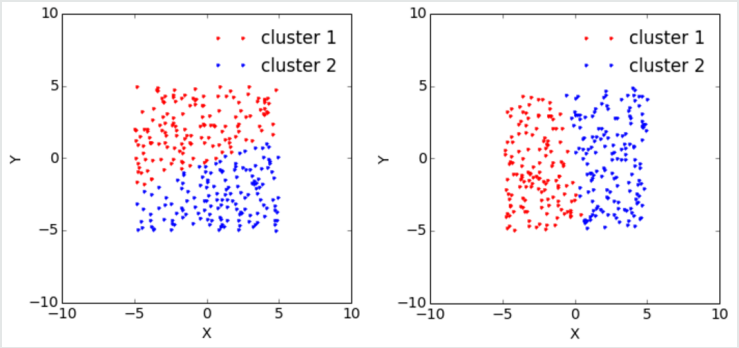
Examples: K-means algorithm, k=2

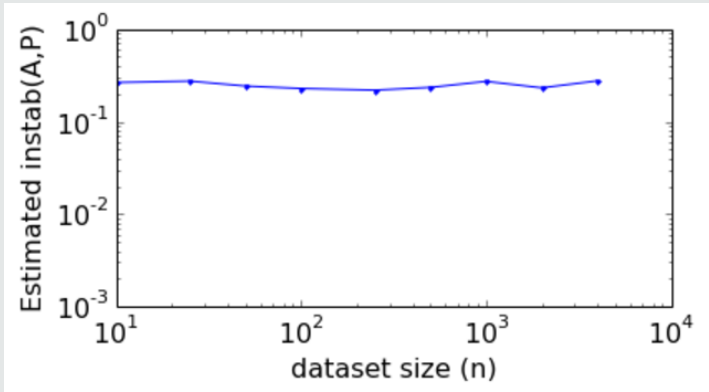
Example one: stable





Example two: instable





Conclusion and further discussions

Conclusion and further discussions

Conclusion

- Represented clustering algorithms in an ERM scheme.
- The idealized risk convergence algorithm is stable if there has unique global minimizer.
- The idealized risk convergence algorithm is not stable if there are multiple global minimizers.
- The stability of an algorithm should not be used to select models.

Further discussions

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).

Reference

1. S. Ben-David, U. von Luxburg, and D. P´al, “A sober look on clustering stability,” COLT, 2006.
2. O. Shamir and N. Tishby, “Cluster stability for finite samples” NIPS, 21, 2008.
3. O. Shamir and N. Tishby, “On the reliability of clustering stability in the large sample regime,” NIPS, 2009.