A Sober look at Clustering Stability

Shai Ben-David[1] Ulrike von Luxburg[2] Dávid Pál[1]

[1]School of Computer Science
University of Waterloo

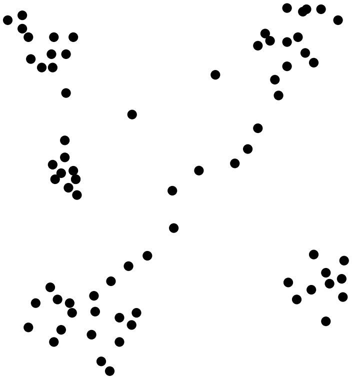
[2]Fraunhofer IPSI, Darmstadt, Germany

COLT 2006

University of Waterloo

What is clustering?

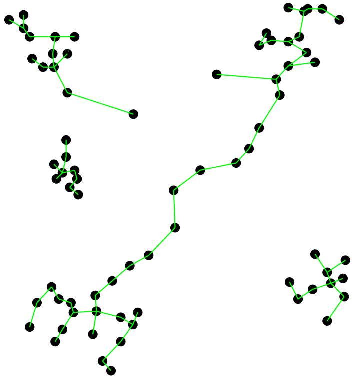
By clustering we mean grouping data according to some distance/similarity measure.



Data

What is clustering?

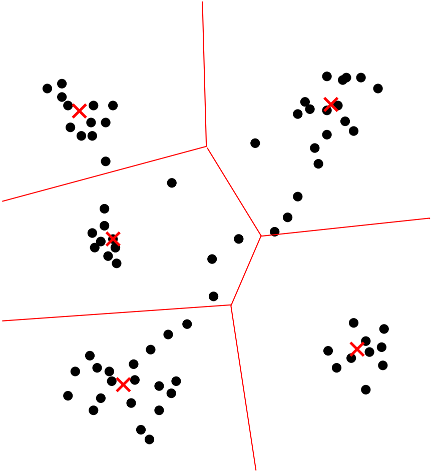
By clustering we mean grouping data according to some distance/similarity measure.



Clusters (Linkage algorithm)

What is clustering?

By clustering we mean grouping data according to some distance/similarity measure.



Clusters (Center-based algorithm)

Correctness of clustering

Q: Clustering is not well defined problem.
How do we know that we cluster correctly?

A: Common solution – Stability.

Correctness of clustering

Q: Clustering is not well defined problem.
How do we know that we cluster correctly?

A: Common solution – Stability.

Stability: Idea of our definition

- Pick your favorite clustering algorithm A.
- Generate two independent samples S_1 and S_2.

Stability

How much will clusterings $A(S_1)$ and $A(S_2)$ differ?

If for large sample sizes clusterings $A(S_1)$ and $A(S_2)$ are almost identical, we say that A is *stable*. Otherwise *unstable*.

Stability: Idea of our definition

- Pick your favorite clustering algorithm A.
- Generate two independent samples S_1 and S_2.

Stability

How much will clusterings $A(S_1)$ and $A(S_2)$ differ?

If for large sample sizes clusterings $A(S_1)$ and $A(S_2)$ are almost identical, we say that A is *stable*. Otherwise *unstable*.

Stability: Idea of our definition

- Pick your favorite clustering algorithm A.
- Generate two independent samples S_1 and S_2.

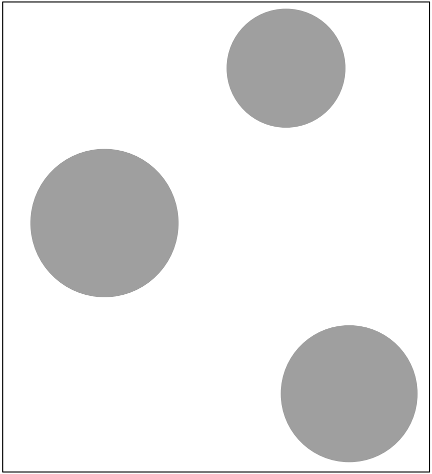
Stability

How much will clusterings $A(S_1)$ and $A(S_2)$ differ?

If for large sample sizes clusterings $A(S_1)$ and $A(S_2)$ are almost identical, we say that A is *stable*. Otherwise *unstable*.

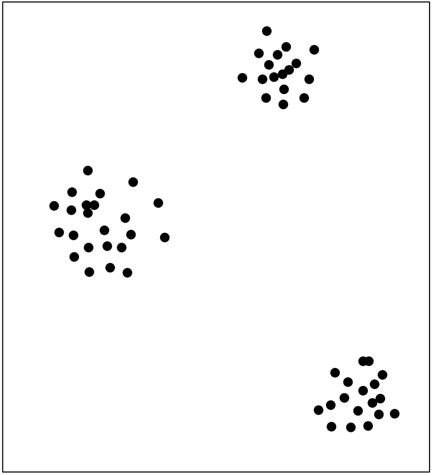
Example of stability

Probability distribution



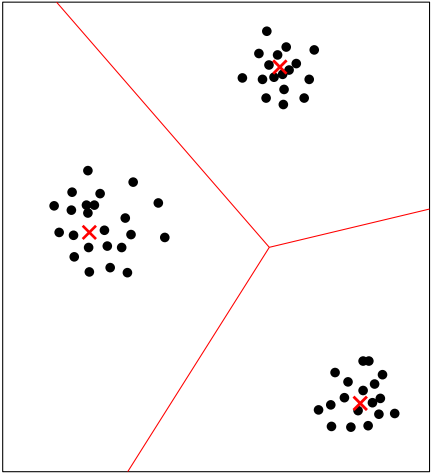
Example of stability

Sample S_1



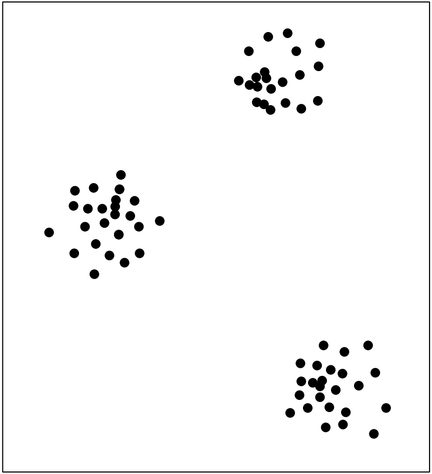
Example of stability

Clustering $A(S_1)$



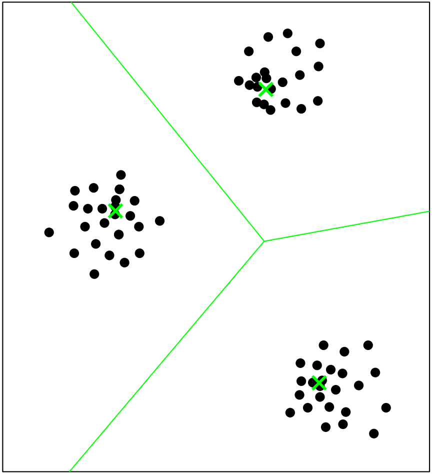
Example of stability

Sample S_2



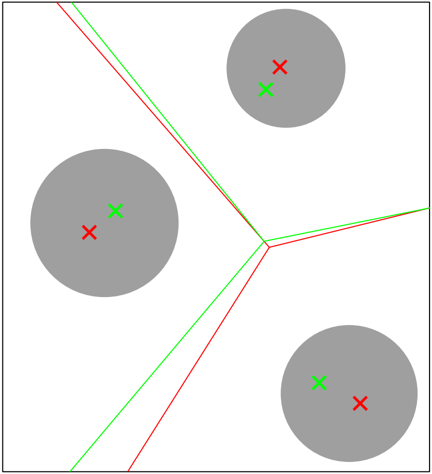
Example of stability

Clustering $A(S_2)$



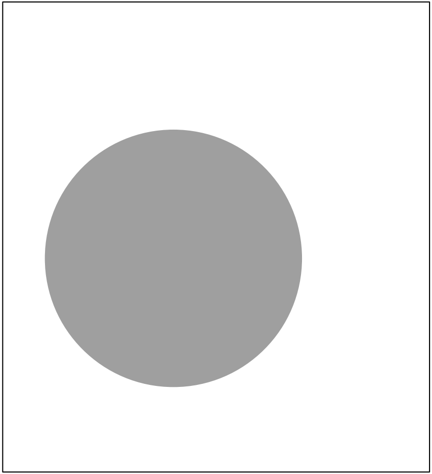
Example of stability

Clusterings $A(S_1)$ and $A(S_2)$ are equivalent.



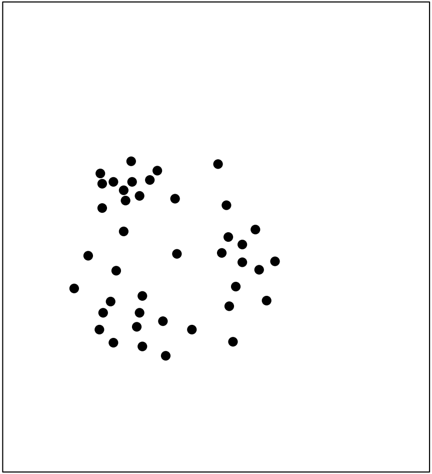
Example of instability

Probability distribution



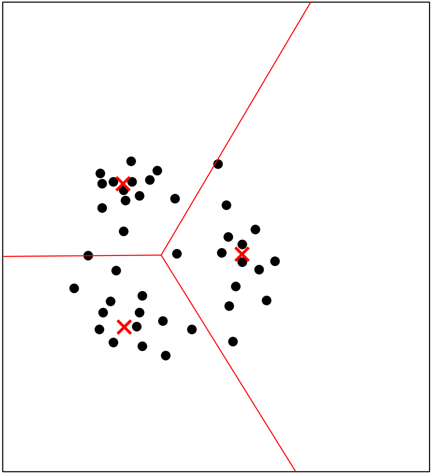
Example of instability

Sample S_1



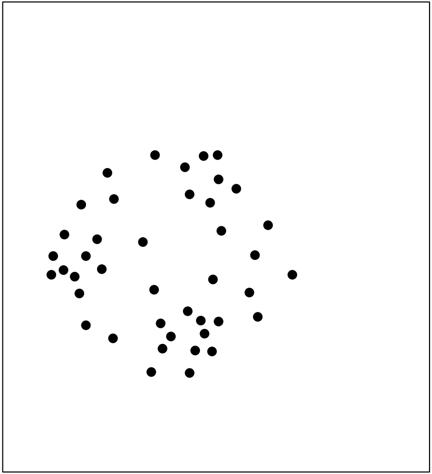
Example of instability

Clustering $A(S_1)$



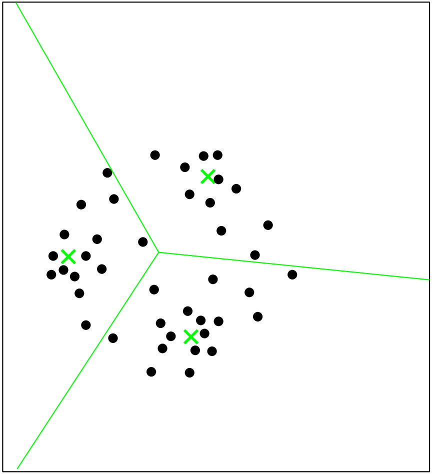
Example of instability

Sample S_2



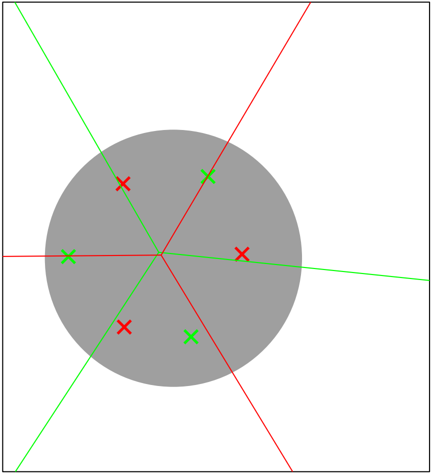
Example of instability

Clustering $A(S_2)$



Example of instability

Clusterings $A(S_1)$ and $A(S_2)$ are different



Motivation

Why do people think stability is important?

- For tuning parameters of clusterings algorithms, such as number of clusters
- To verify meaningfulness of clustering outputted by algorithm.

Motivation

Why do people think stability is important?

- For tuning parameters of clusterings algorithms, such as number of clusters
- To verify meaningfulness of clustering outputted by algorithm.

Motivation

Our intention:

Provide theoretical justification.

We discovered:

The popular belief is false.

Motivation

Our intention:

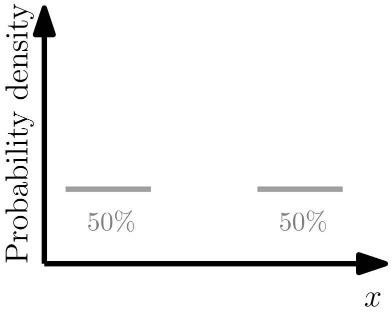
Provide theoretical justification.

We discovered:

The popular belief is false.

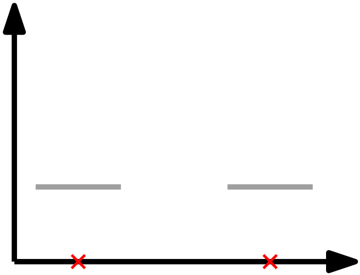
First example

1D probability distribution



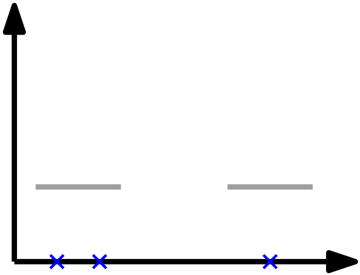
First example

2 centers – stable



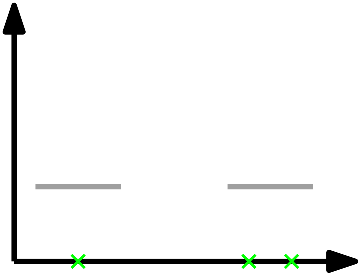
First example

3 centers – solution #1



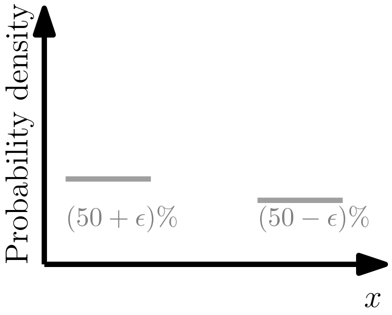
First example

3 centers – solution #2 $\Longrightarrow$ unstable



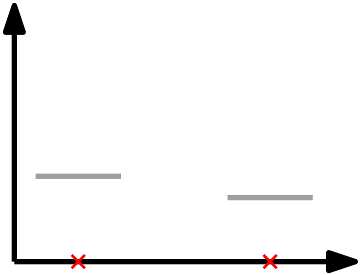
First example

slightly asymmetric distribution



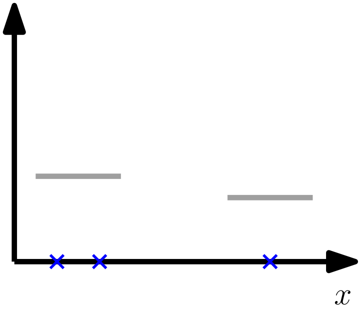
First example

2 centers – stable



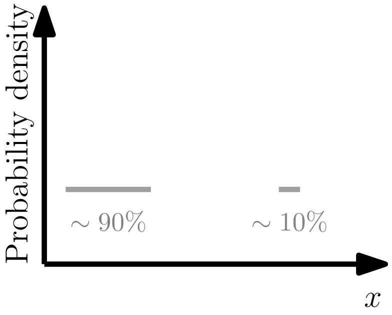
First example

3 centers – stable



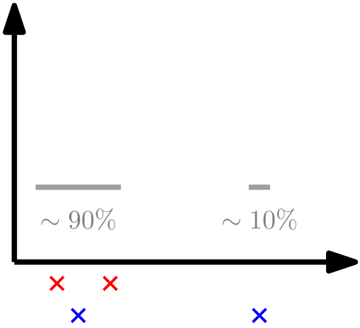
Second example

1D probability distribution



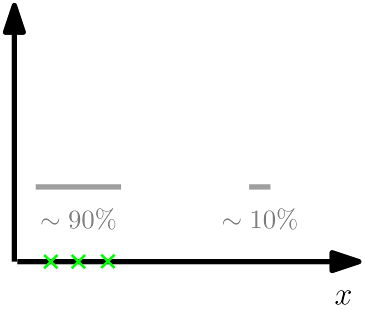
Second example

2 centers – unstable



Second example

3 centers – stable



Our results

Theorem

For a cost based algorithm (e.g. k-means, k-medians):

- *If the optimization problem has unique optimum, then the algorithm is stable.*
- *If the underlying probability distribution is symmetric and the optimization problem has multiple symmetric optima, then the algorithm is unstable.*

Our results

Theorem

For a cost based algorithm (e.g. k-means, k-medians):

- *If the optimization problem has unique optimum, then the algorithm is stable.*
- *If the underlying probability distribution is symmetric and the optimization problem has multiple symmetric optima, then the algorithm is unstable.*

Our results

Theorem

For a cost based algorithm (e.g. k-means, k-medians):

- *If the optimization problem has unique optimum, then the algorithm is stable.*
- *If the underlying probability distribution is symmetric and the optimization problem has multiple symmetric optima, then the algorithm is unstable.*

Conclusion

- Stability, contrary to common belief, does not measure validity of a clustering or meaningfulness of choice of number of clusters.
- Instead, it measures the number of solutions to the clustering optimization problem for the underlying probability distribution.

Open problems

Q: Is symmetry really needed for instability?

A: No!
(Work in progress, together with Shai Ben-David & Hans Ulrich Simon)

Analyze finite sample sizes, and give explicit bounds.

Analyze other types of algorithms e.g. linkage algorithms.

Open problems

Q: Is symmetry really needed for instability?

A: No!
(Work in progress, together with Shai Ben-David & Hans Ulrich Simon)

Analyze finite sample sizes, and give explicit bounds.

Analyze other types of algorithms e.g. linkage algorithms.

Open problems

Q: Is symmetry really needed for instability?

A: No!
(Work in progress, together with Shai Ben-David & Hans Ulrich Simon)

Analyze finite sample sizes, and give explicit bounds.

Analyze other types of algorithms e.g. linkage algorithms.

Open problems

Q: Is symmetry really needed for instability?

A: No!
(Work in progress, together with Shai Ben-David & Hans Ulrich Simon)

Analyze finite sample sizes, and give explicit bounds.

Analyze other types of algorithms e.g. linkage algorithms.

Concrete demonstration of our analysis: k-means

Consider k-means in metric space (X, ℓ).

Given a sample $S = \{x_1, x_2, \ldots, x_m\}$, we search centers $c_1, c_2, \ldots, c_k$. The k-means algorithm minimizes the *empirical cost*

$$cost(S; c_1, c_2, \ldots, c_k) = \frac{1}{m} \sum_{x \in S} \min_{1 \le i \le k} (\ell(c_i, x))^2$$

As $m \to \infty$ this converges to the *true cost* [Ben-David, COLT04]

$$cost(P; c_1, c_2, \ldots, c_k) = \operatorname*{Exp}_{x \in P} \min_{1 \le i \le k} (\ell(c_i, x))^2$$

Minimizing $cost(S; .)$ is for large samples almost the same as minimizing $cost(P; .)$.

Concrete demonstration of our analysis: k-means

Consider k-means in metric space (X, ℓ).

Given a sample $S = \{x_1, x_2, \dots, x_m\}$, we search centers $c_1, c_2, \dots, c_k$. The k-means algorithm minimizes the *empirical cost*

$$cost(S; c_1, c_2, \dots, c_k) = \frac{1}{m} \sum_{x \in S} \min_{1 \le i \le k} (\ell(c_i, x))^2$$

As $m \to \infty$ this converges to the *true cost* [Ben-David, COLT04]

$$cost(P; c_1, c_2, \dots, c_k) = \operatorname*{Exp}_{x \in P} \min_{1 \le i \le k} (\ell(c_i, x))^2$$

Minimizing $cost(S; .)$ is for large samples almost the same as minimizing $cost(P; .)$.

Concrete demonstration of our analysis: k-means

What happens if the function $cost(P; c_1, c_2, \ldots, c_k)$ has more than one k-tuple of centers minimizing it?

Instability !

Concrete demonstration of our analysis: k-means

What happens if the function $cost(P; c_1, c_2, \ldots, c_k)$ has more than one k-tuple of centers minimizing it?

Instability !

Concrete demonstration of our analysis: k-means

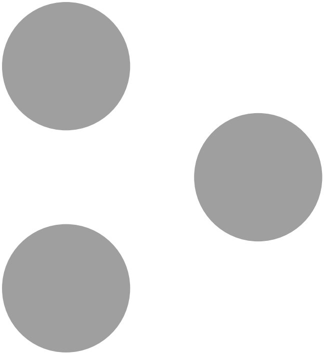
What happens if the function $cost(P; c_1, c_2, \ldots, c_k)$ has more than one k-tuple of centers minimizing it?

Instability !

Example of instability

Searching 2 centers

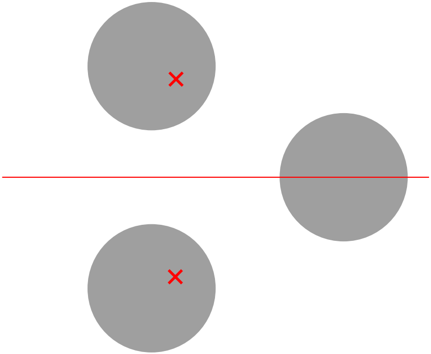
Probability distribution (perfectly symmetric)



Example of instability

Searching 2 centers

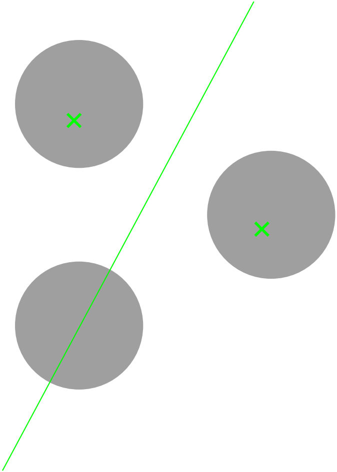
Optimal solution #1



Example of instability

Searching 2 centers

Optimal solution #2



Example of instability

Searching 2 centers

Optimal solution #3