Clustering stability: an overview

Ulrike von Luxburg

Докладчик: Токмакова Лада

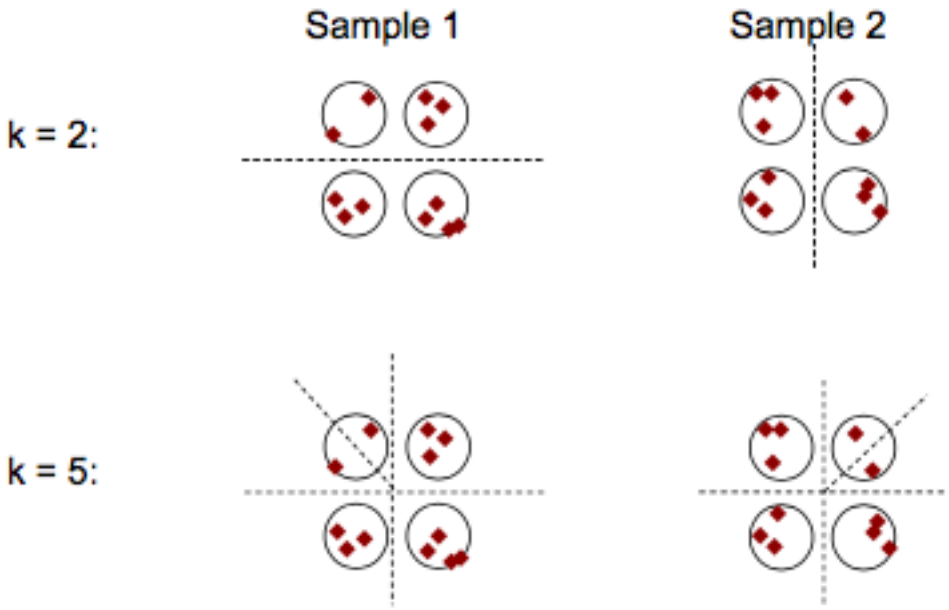
План

- Определение стабильности кластеризации
- The idealized K-means algorithm
- The actual K-means algorithm
- General clustering algorithms

Введение

Основной вопрос:

«Как определить количество кластеров?»



Определение стабильности кластеризации - 1

- Множество:
$$S = (X_1, \dots, X_n)$$
- Функция кластеризации:
$$C_K : S \to \{1, \dots, K\}$$
- Дополнительный параметр на вход алгоритму кластеризации:
$$K$$
- Расстояние между кластеризациями:
$$d(C, C')$$
- ***Нестабильность*** алгоритма кластеризации для фиксированных P, K и n :
$$Instab(K, n) = E\left(d(C_K(S_n), C_K(S_n'))\right)$$

Определение стабильности кластеризации - 2

- $\rightarrow S$, A
- Для $k = 2, \dots, k_{max}$
  - Генерируем зашумленные множества:
    $$S_b \ (b = 1, .., b_{max})$$
  - Для $b = 1, \dots, b_{max}$ вычислим:
    $$C_k(S_b)$$
  - Для $b, b' = 1, \dots, b_{max}$ вычислим:
    $$d(C_k(S_b), C_k(S_{b'}))$$
  - Вычислим нестабильность кластеризации (среднее расстояние между кластеризациями):
    $$\widehat{Instab}(k, n) = \frac{1}{b_{max}^2} \sum_{b,b'=1}^{b_{max}} d(C_k(S_b), C_k(S_{b'}))$$
- Выбираем параметр K, который дает наилучшую стабильность кластеризации:
  $$K := \arg\min_k \widehat{Instab}(k, n)$$

Алгоритм K-means

- $\rightarrow X_1, \ldots, X_n \in \mathbb{R}^d$, фиксированное K.
- Хотим минимизировать целевую функцию:

$$\min Q_K^{(n)}(c_1, .., c_K) = \frac{1}{n} \sum_{i=1}^{n} \min_{k=1,..,K} \|X_i - c_k\|^2$$

Если $n \rightarrow \infty$, то

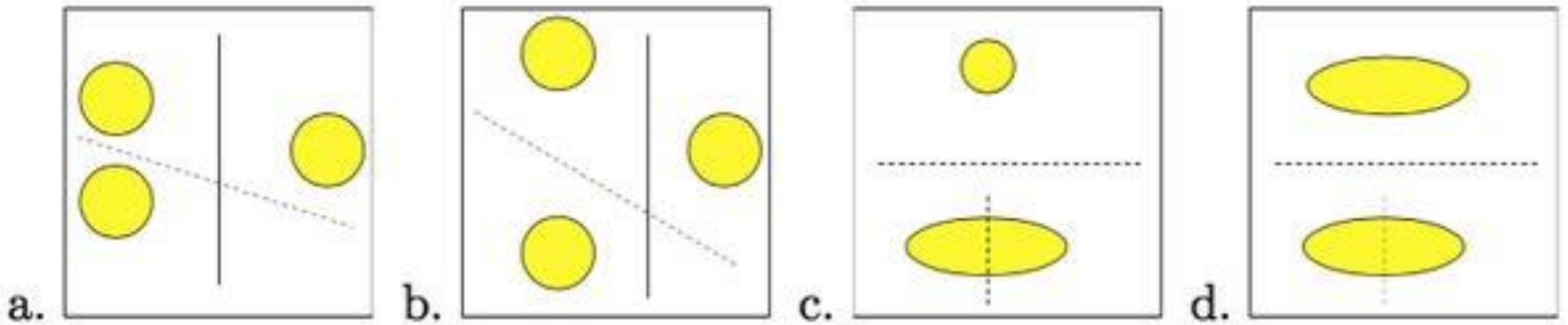
$$\min Q_K^{(\infty)}(c_1, .., c_K) = \int \min_{k=1,..,K} \|X_i - c_k\|^2 \, dP(x)$$

- Определяем $c^{<0>} = \{c_1^{<0>}, \ldots, c_K^{<0>}\}$
- Итерация:
  - $\forall i = 1, .., n: C^{<t>}(X_i) \coloneqq \arg\min_{k=1,\ldots,K} \left\| X_i - c_k^{<t>} \right\|$
  - $\forall k = 1, \ldots, K: c_k^{<t+1>} \coloneqq \frac{1}{N_k} \sum_{\{i | C^{<t>}(X_i) = k\}} X_i$

The idealized K-means algorithm - 1

- Сходится к глобальному минимуму, т.е.

$$c^{(n)} := \left(c_1^{(n)}, \dots, c_K^{(n)}\right) := \underset{c}{\operatorname{argmin}}\, Q_K^{(n)}(c)$$



The idealized K-means algorithm - 2

Лемма 1 (Стабильность и глобальный оптимум)

Пусть есть распределение вероятностей P на $\mathbb{R}^d$, предельная целевая функция $Q_K^{(\infty)}$, фиксированное $K > 1$. Тогда:

- Если $Q_K^{(\infty)}$ имеет единственный глобальный минимум, то the idealized K-means algorithm стабильный, когда $n \to \infty$:

$$\lim_{n\to\infty} Instab(K, n) = 0$$

- Если $Q_K^{(\infty)}$ имеет несколько глобальных минимумов, то the idealized K-means algorithm нестабильный:

$$\lim_{n\to\infty} Instab(K, n) > 0$$

The idealized K-means algorithm - 3

Лемма 2 (сходимость rescaled stability)

Пусть распределение вероятностей P имеет плотность p, для фиксированного параметра K $Q_K^{(\infty)}$ имеет единственный глобальный минимум $c^{(*)} = \left(c_1^{(*)}, \dots, c_K^{(*)}\right)$. Обозначим границу между кластерами i и j за B_{ij}. Пусть $m \in \mathbb{N}$, и $S_{n,1}, \dots, S_{n,2m}$ выборки размера n из P. Пусть $C_K\left(S_{n,i}\right)$ - результат the idealized K-means algorithm для $S_{n,i}$.

$$\widehat{Instab}(K,n) := \frac{1}{m}\sum_{i=1}^{m} d_{MM}\left(C_K\left(S_{n,2i-1}\right), C_K\left(S_{n,2i}\right)\right),$$

где d_{MM} - minimal matching distance:

$$d_{MM}(C,C') := \min_{\pi}\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{C(X_i) \neq \pi(C'(X_i))\}$$

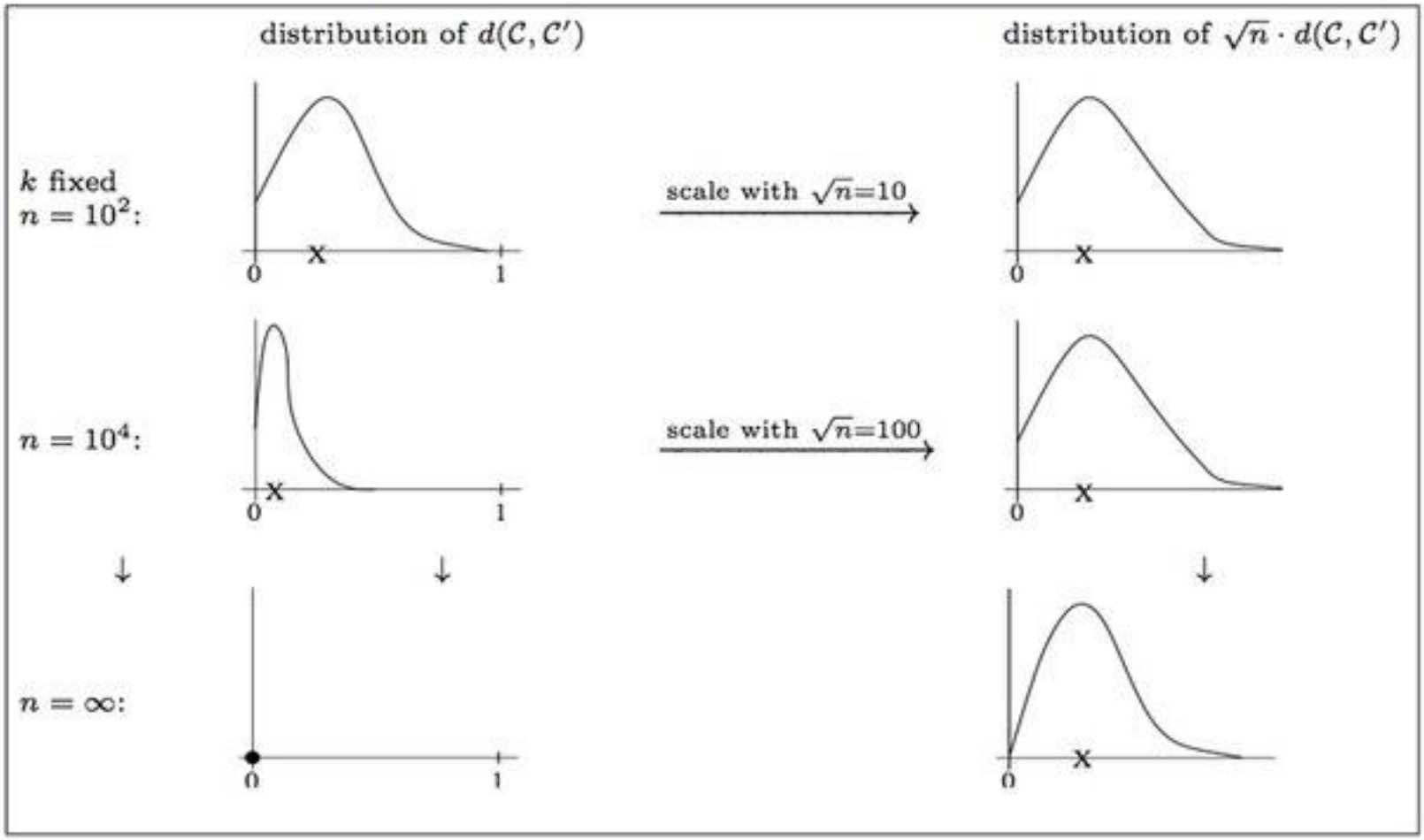
The idealized K-means algorithm - 4

Тогда при $n \to \infty$ и $m \to \infty$ имеет место сходимость по вероятности:

$$\sqrt{n}\widehat{Instab}(K,n) \to \sum_{1 \le i < j \le K} \int_{B_{ij}} \frac{V_{ij}}{\left\| c_i^{(*)} - c_j^{(*)} \right\|} p(x)dx$$

- $\sqrt{n}\widehat{Instab}(K,n)$ - rescaled instability
- V_{ij} - описывает асимптотику случайных колебаний границы между кластером i и кластером j.

The idealized K-means algorithm - 5



The idealized K-means algorithm - 6

Утверждение 1 (нестабильная кластеризация)

Пусть $Q_K^{(\infty)}$ имеет единственный глобальный минимум. Если значение $Instab(K, n)$ большое, то, как правило, the idealized K-means кластеризация имеет границы в регионах с высокой плотностью.

Утверждение 2 (стабильная кластеризация)

Пусть $Q_K^{(\infty)}$ имеет единственный глобальный минимум. Если значение $Instab(K, n)$ маленькое, то, как правило, the idealized K-means кластеризация имеет границы в регионах с низкой плотностью.

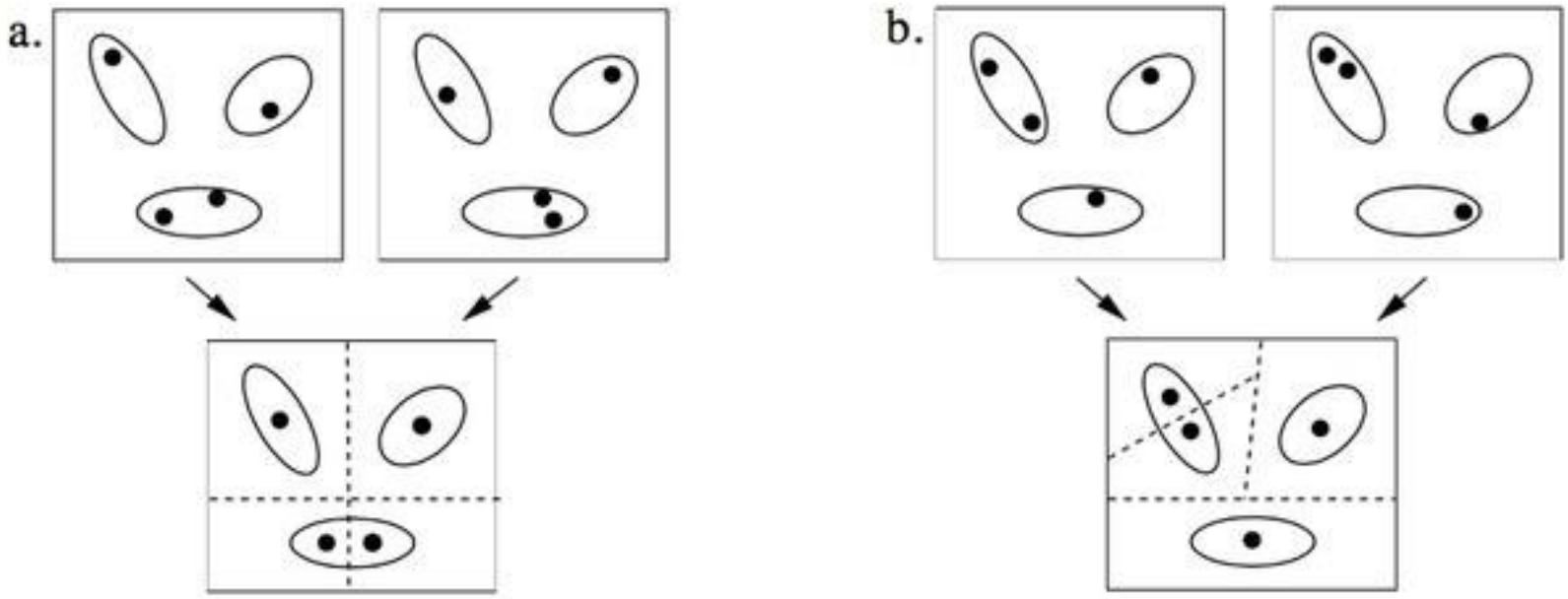
The idealized K-means algorithm - 7

Утверждение 3

Пусть исходное распределение P имеет K хорошо разделенных кластеров. Тогда для the idealized K-means algorithm:

- Если K слишком большое, то кластеризации, полученные с помощью этого алгоритма, как правило, неустойчивые.
- Если K верное или достаточно маленькое, то кластеризации, полученные с помощью этого алгоритма, как правило, устойчивые.

The actual K-means algorithm - 1



The actual K-means algorithm - 2

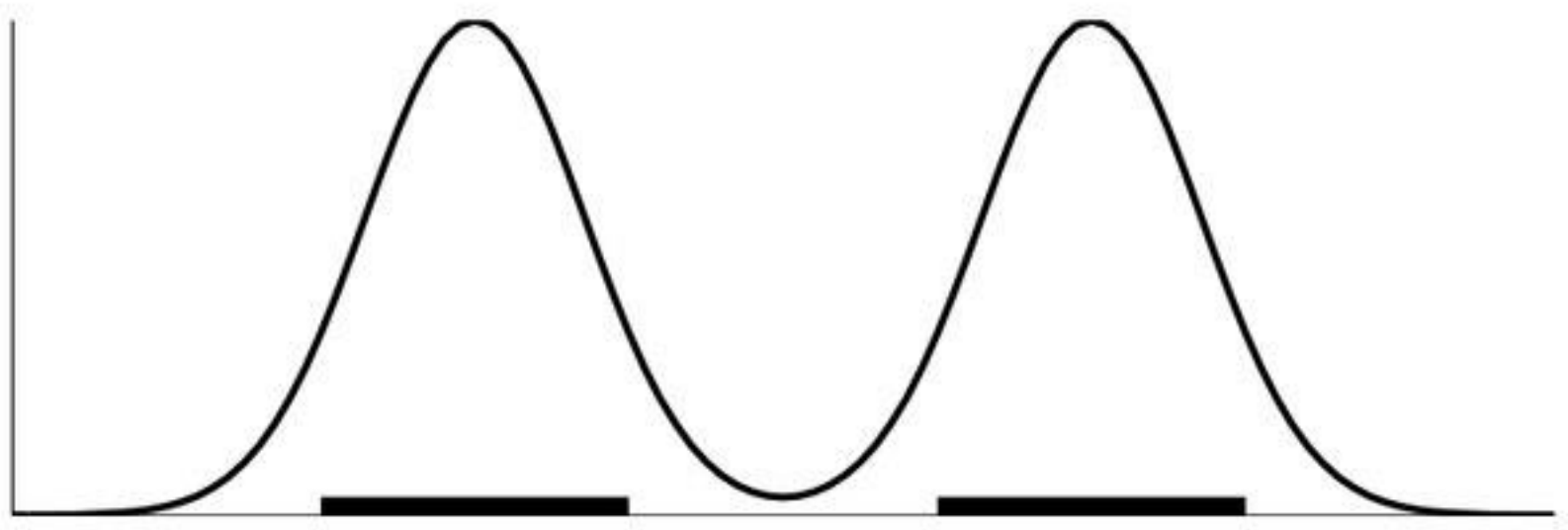
Лемма 3 (стабильность the actual K-means algorithm)

Пусть распределение P является смесью двух хорошо разделенных гауссиан на $\mathbb{R}$ со средними μ_1 и μ_2.

- Пусть мы запустили алгоритм с $K = 2$, и пусть мы используем инициализацию, которая с высокой вероятностью размещает каждый первоначальный центр в истинный кластер. Тогда алгоритм будет стабильным в том смысле, что с высокой вероятностью в итоге один центр будет близким к μ_1, а второй к μ_2.

- Пусть мы запустили алгоритм с $K = 3$, и пусть мы используем инициализацию, которая с высокой вероятностью размещает каждый первоначальный центр в истинный кластер. Тогда алгоритм будет нестабильным в том смысле, что с вероятностью, близкой к 0.5, одним кластером будет первая гауссиана, а вторая разделится на две части, и с такой же вероятностью может быть обратная ситуация.

The actual K-means algorithm - 3



The actual K-means algorithm - 4

Инициализация (I) центров для леммы 3:

- Выберем произвольно L центров, $L \approx Klog(K)$;
- Выполним один шаг алгоритма;
- Удалим все центры, в кластерах которых меньше $p_0 \approx \frac{1}{L}$ элементов;
- Среди оставшихся выберем K центров:
  - Первый центр выбираем произвольно;
  - Выбираем следующий центр, который максимизирует минимальное расстояние до уже выбранных центров.

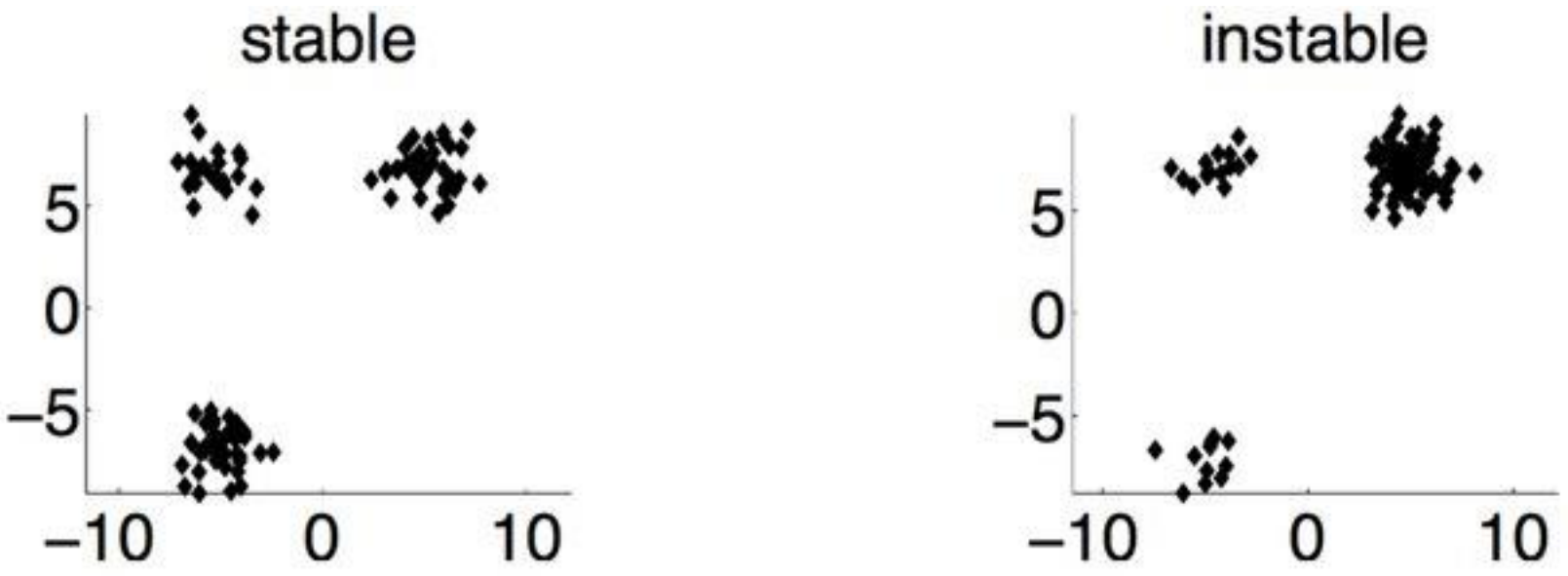
The actual K-means algorithm - 5

Лемма 4 (Инициализация)

Пусть дана смесь K_{true} хорошо разделенных гауссиан на $\mathbb{R}$ со средними μ_i. Если выбирать K_{init} центров с помощью (I), то $\exists$ K_{true} непересекающихся областей A_k с $\mu_k \in A_k$: все K_{init} центры находятся в одной из A_k, причем:

- $K_{init} = K_{true} \Longrightarrow A_k$ содержит только один центр;
- $K_{init} < K_{true} \Longrightarrow A_k$ содержит не более одного центра;
- $K_{init} > K_{true} \Longrightarrow A_k$ содержит как минимум один центр.

The actual K-means algorithm - 6



The actual K-means algorithm - 7

Утверждение 4 (Стабильность the actual K-means algorithm)

Пусть в исходном распределении есть K_{true} хорошо разделенных кластеров, и пусть мы используем (I) для инициализации K_{init}. Тогда:

- $K_{init} = K_{true} \Rightarrow$ w.h.p. в каждом кластере один центр;
- $K_{init} > K_{true} \Rightarrow$ наблюдаем нестабильность;
- $K_{init} > K_{true} \Rightarrow$ либо стабильность, либо нестабильность.

General clustering algorithms

Пусть Q – целевая функция кластеризации, и пусть есть идеализированный алгоритм, который ее глобально минимизирует.

Тогда справедливы:

- Лемма 1
- Утверждение 1

Спасибо за внимание!