K-Means++ Clustering

Algorithm Introduction

Yuwen Chen
Mingliang Qi
Mingyuan Wu

29.04.2014

Outline

Outline

K-Means Algorithm

Data: a set of observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, number of clusters k, Convergence Delta ξ
Result: a set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

Initialization:

select uniformly k data points $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$ as the centroids of clusters

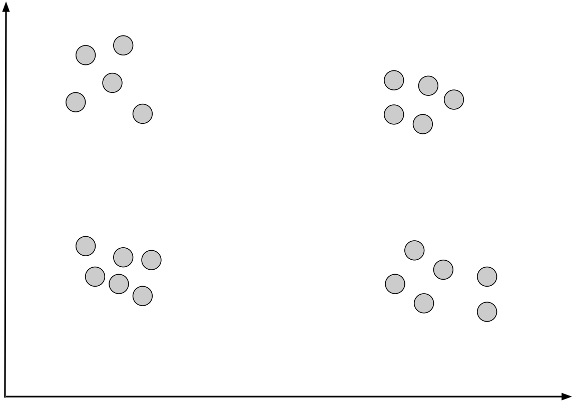
Compute:

repeat
Form k clusters by assigning each point to its closest centroid;

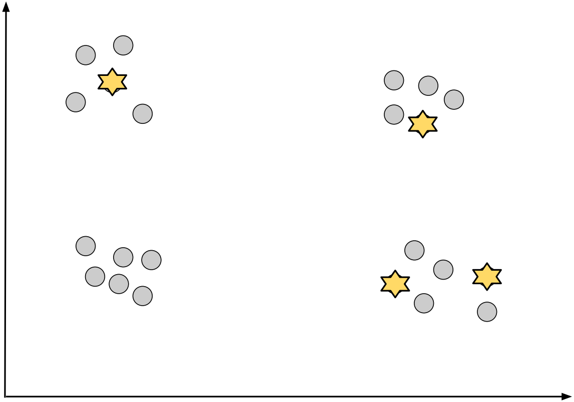
Recompute the center of each cluster;
until $\Delta\mathcal{C} < \xi$;

Outline

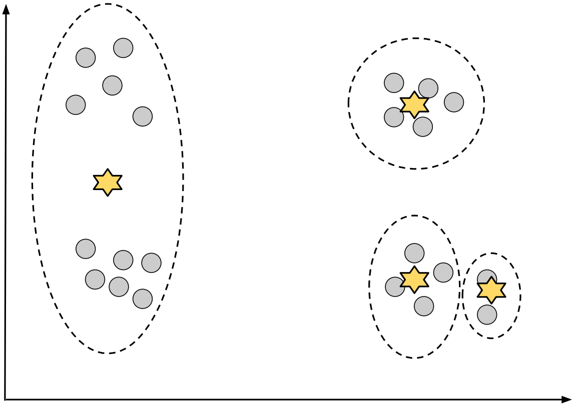
Problem: Poor Initial Centroids



Problem: Poor Initial Centroids



Problem: Poor Initial Centroids



Outline

Motivation: Better Seeding

- **Increased accuracy**
  Potential to obtain smaller SSE (i.e. BETTER Result)
- **Faster Convergence**
  Terminates faster than poor initialized K-means

Outline

K-Means++ Algorithm

Data: a set of observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$,
number of clusters k, Convergence Delta ξ

Result: a set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

Target Function: $\phi = min \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} D(\mathbf{x})^2$

Initialization:

Take a centroid $\mathbf{c}_1$, chosen uniformly at random from $\mathcal{X}$

repeat

Take a new center $\mathbf{c}_i$, choosing $\mathbf{x} \in \mathcal{X}$ with probability $\frac{D(\mathbf{x})^2}{\sum_{\mathbf{x} \in \mathcal{X}} D(\mathbf{x})^2}$

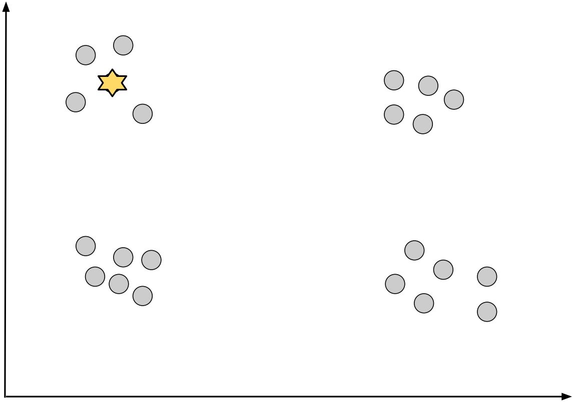
until *k centroids generated* ;

Compute:

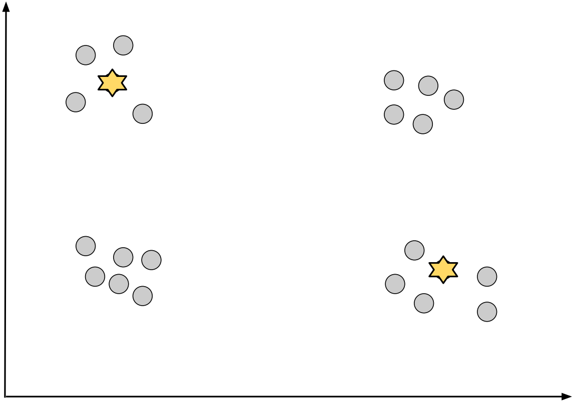
Proceed as with the standard k-means algorithm

Outline

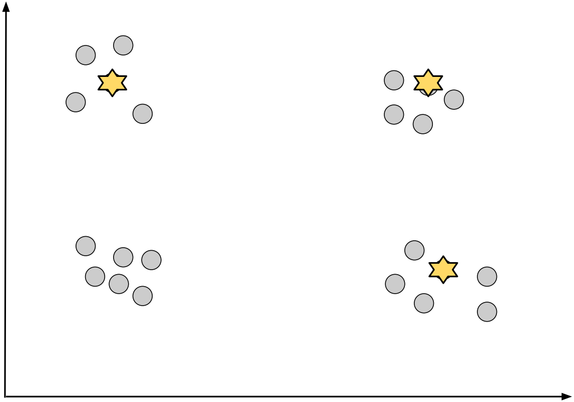
Example: Improvement of k-means



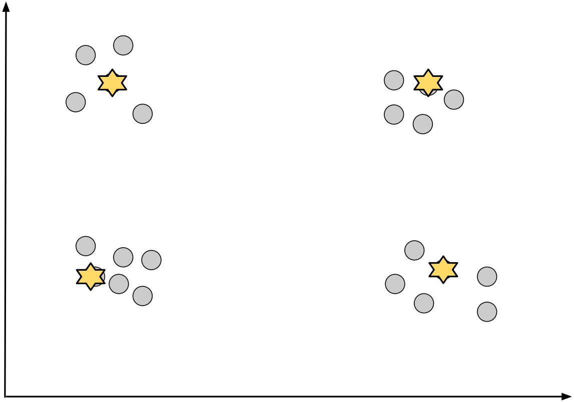
Example: Improvement of k-means



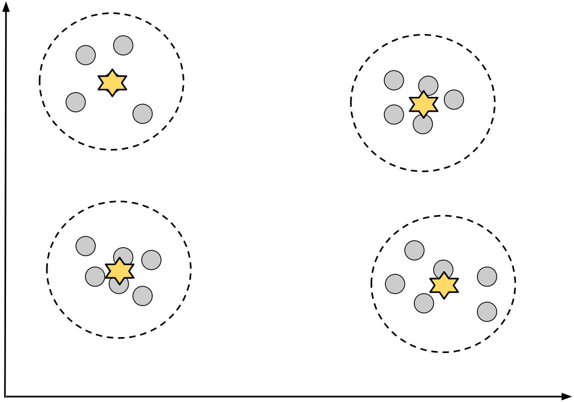
Example: Improvement of k-means



Example: Improvement of k-means



Example: Improvement of k-means



Outline

Comparison of Time & Space Complexity

- n - number of data points
- d - dimension
- k - number of target clusters
- I - number of iterations

	k-means	k-means++
Time Complexity	$O(Iknd)$	$O(knd) + O(Iknd)$
Space Complexity	$O((n+k)d)$	$O((n+k)d)$

Comparison of Experimental Results

ϕ - $min \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} D(\mathbf{x})^2$

T - Execution Time

	Average ϕ		Minimum ϕ		Average T	
k	k-means	k-means++	k-means	k-means++	k-means	k-means++
10	7553.5	6151.2	6139.45	5631.99	0.12	0.05
25	3626.1	2064.9	2568.2	1988.76	0.19	0.09
50	2004.2	1133.7	1344	1088	0.27	0.17

Figure: Experimental Results with data set $n = 1024, d = 10$ [2]

Outline

Existing Problems

- **Scalability**

  $\boxed{O(knd)} + O(Iknd)$

  $\downarrow$

  extra k iterations over all data points
  recompute distance distribution in each iteration

- **Confusion From Outliers**

  Outliers get chosen more easily

  $\downarrow$

  Converges slower

Outline

K-Means||: Improvement of K-Means++

pick more than one centroid **independently** in each round **with a larger probability**

$\downarrow$

- Reduce the number of iterations
- Less computation cost of Distance distribution
- Cover more data points

K-Means|| Algorithm

Data: a set of observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, number of clusters k, Convergence Delta ξ, Oversampling Factor f, number of iterations R

Result: a set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

Initialization:

Take a centroid $\mathbf{c}_1$, chosen uniformly at random from $\mathcal{X}$

for *R rounds* **do**

Sample each data point independently with probability $f \frac{D(\mathbf{x})^2}{\sum_{\mathbf{x} \in \mathcal{X}} D(\mathbf{x})^2}$

Add all sampled points to $\mathcal{C}$

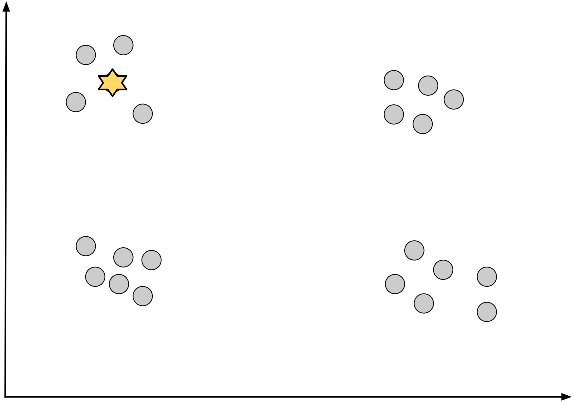
end

Recompute $\mathcal{C}$ to k clusters, use the centroid of each cluster as the initial centroid for k-means

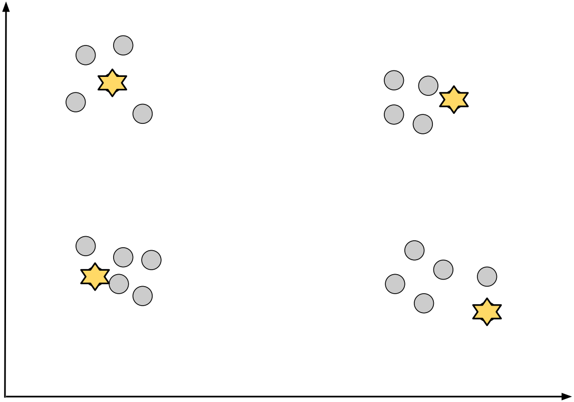
Compute:

Proceed as with the standard k-means algorithm

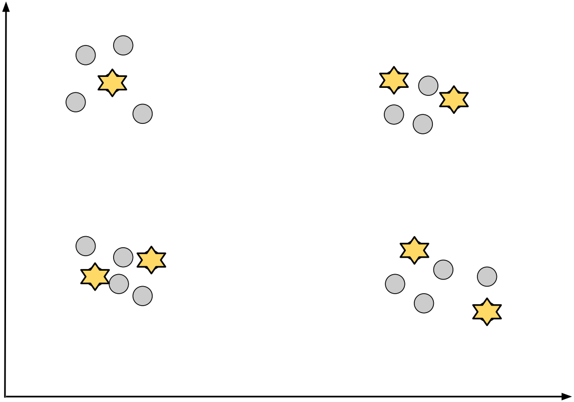
Example: K-Means||



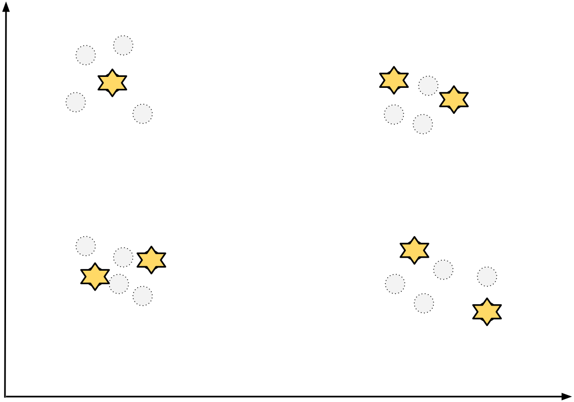
Example: K-Means||



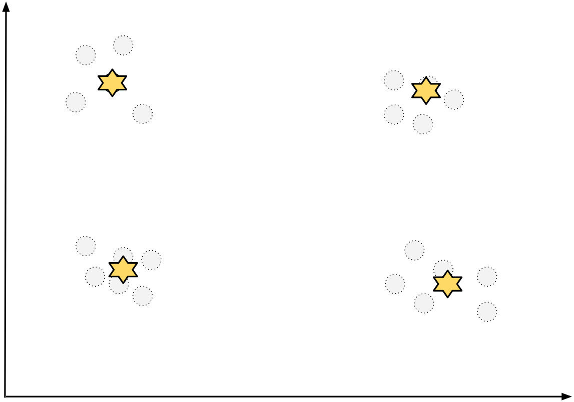
Example: K-Means||



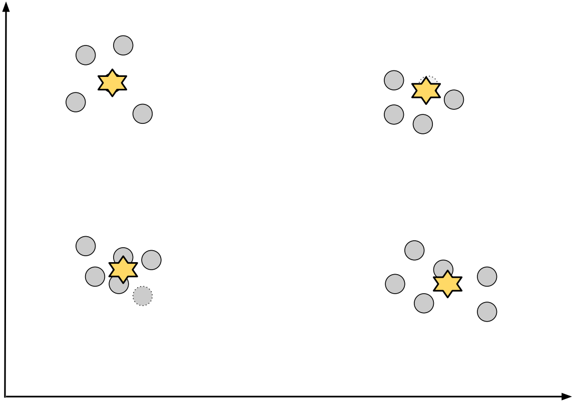
Example: K-Means||



Example: K-Means||



Example: K-Means||



Thank You

For Further Reading I

- Kumar, Vipin, Pang-Ning Tan, and Michael Steinbach. "Cluster analysis: basic concepts and algorithms." Introduction to data mining (2006): 487-586.
- Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.
- Bahmani, Bahman, et al. "Scalable k-means++." Proceedings of the VLDB Endowment 5.7 (2012): 622-633.