Clustering

Shusen Wang

Clustering Task

Tasks

clustering

Methods

K-means

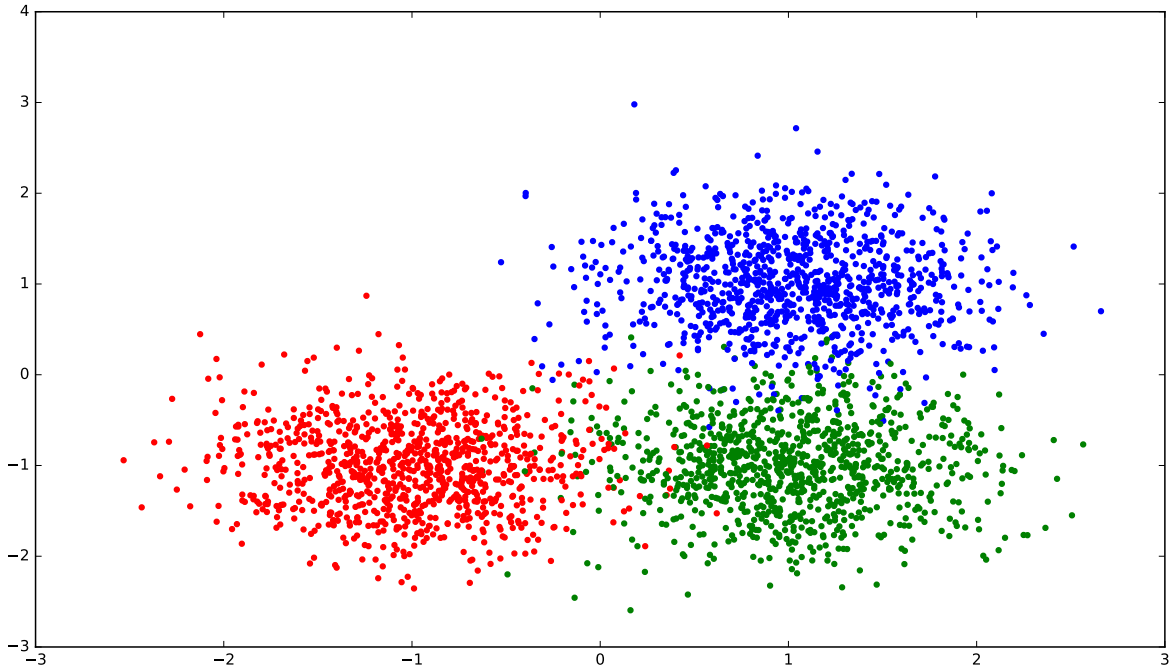
Algorithms

Lloyd's algorithm

Clustering Task

Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

Output: predicted class labels $y_1, \cdots, y_n \in \{1, 2, \cdots, k\}$.

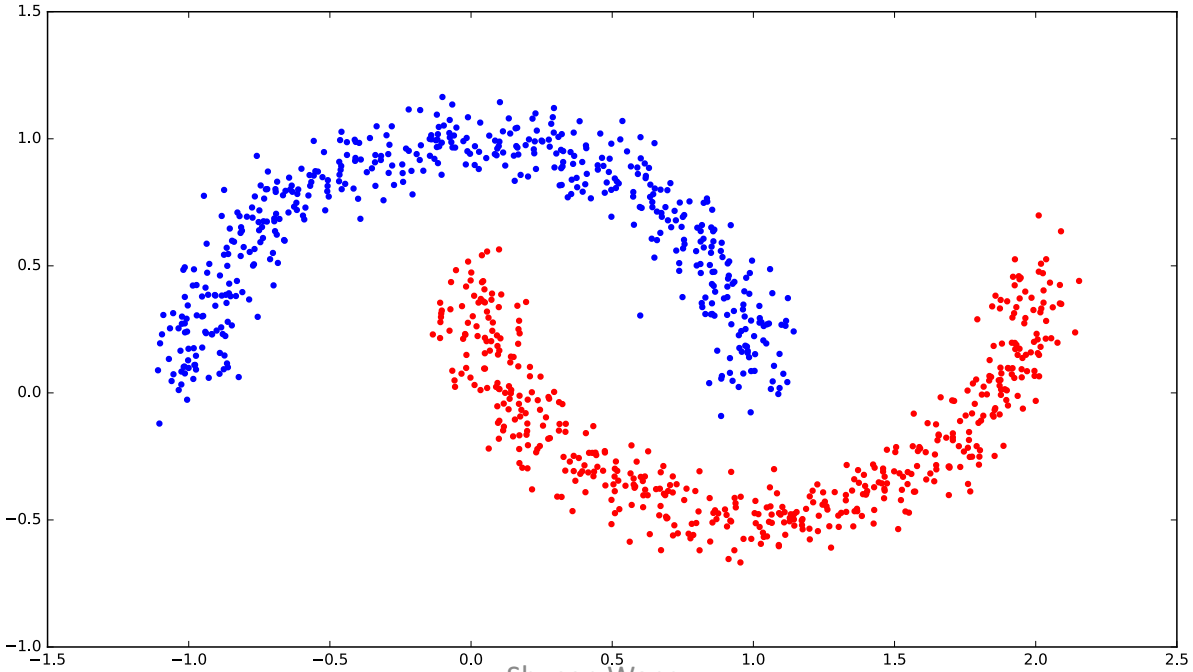


$k = 3$

Clustering Task

Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

Output: predicted class labels $y_1, \cdots, y_n \in \{1, 2, \cdots, k\}$.

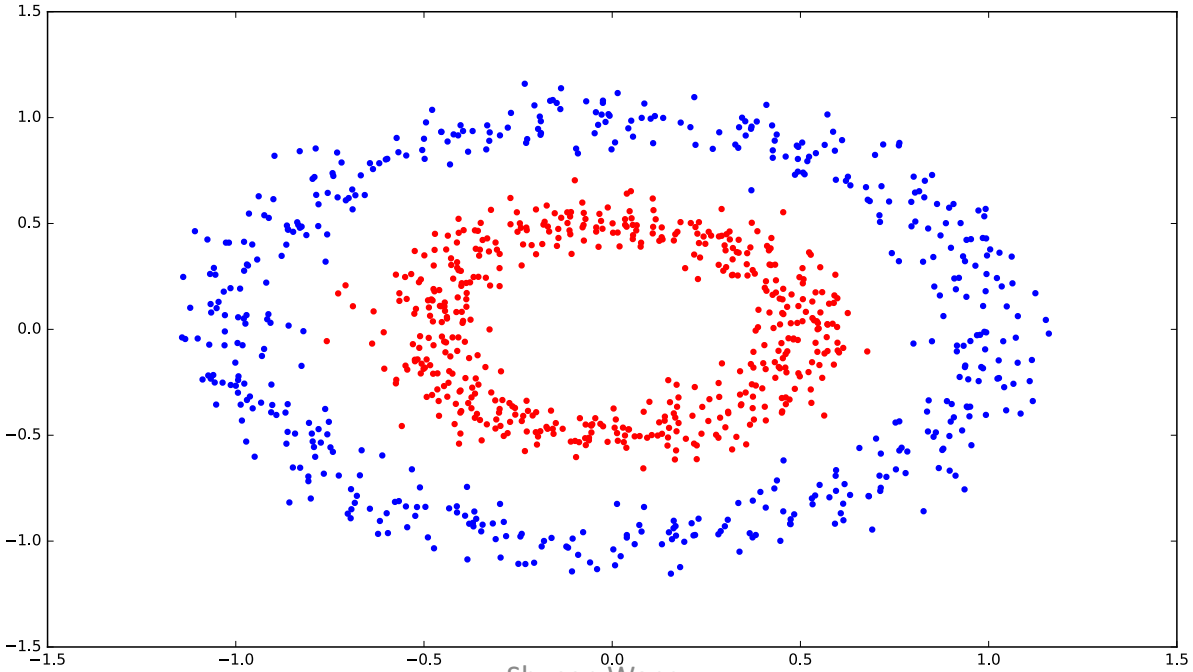


$k = 2$

Clustering Task

Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

Output: predicted class labels $y_1, \cdots, y_n \in \{1, 2, \cdots, k\}$.



$k = 2$

Clustering Task

Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

Output: predicted class labels $y_1, \cdots, y_n \in [k]$.

denote $[k] = \{1, 2, \cdots, k\}$

Clustering Task

Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

Output: predicted class labels $y_1, \cdots, y_n \in [k]$.

Equivalent definition:

Clustering: partition $[n]$ to k disjoint subsets $S_1, \cdots, S_k$,

- $S_1 \cup \cdots \cup S_k = [n]$,
- $S_i \cap S_j = \emptyset$, for all $i \neq j$.

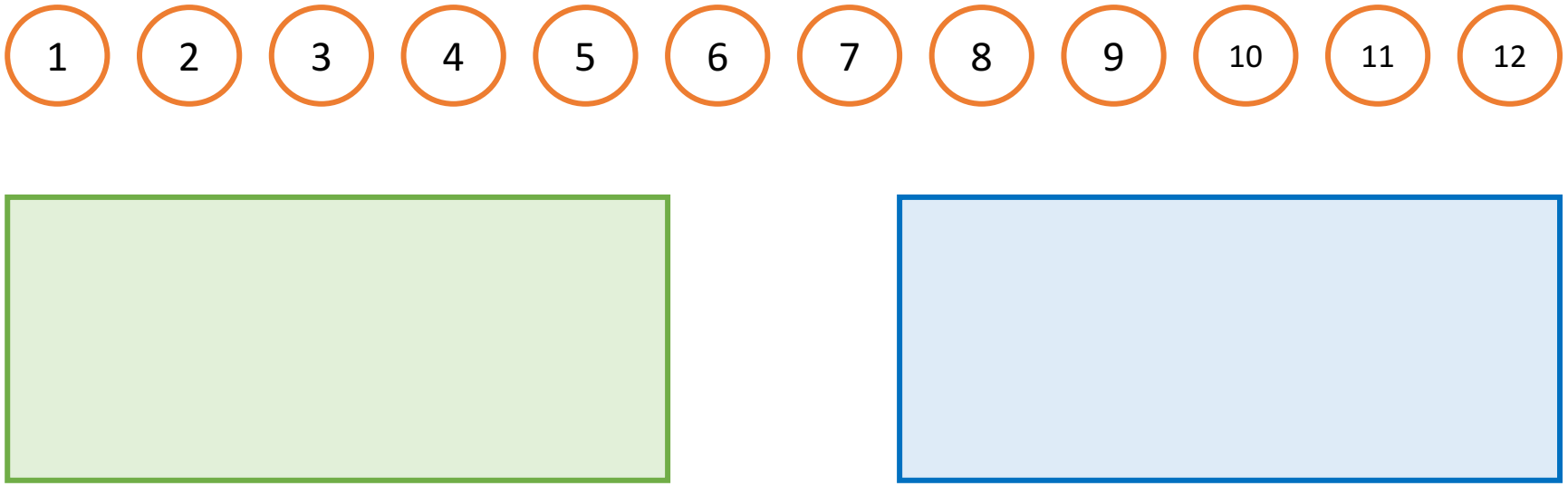
Clustering Task

Clustering: partition $[n]$ to k disjoint subsets $S_1, \cdots, S_k$,

- $S_1 \cup \cdots \cup S_k = [n]$,
- $S_i \cap S_j = \emptyset$, for all $i \neq j$.

Example

- $n = 12$
- $k = 2$



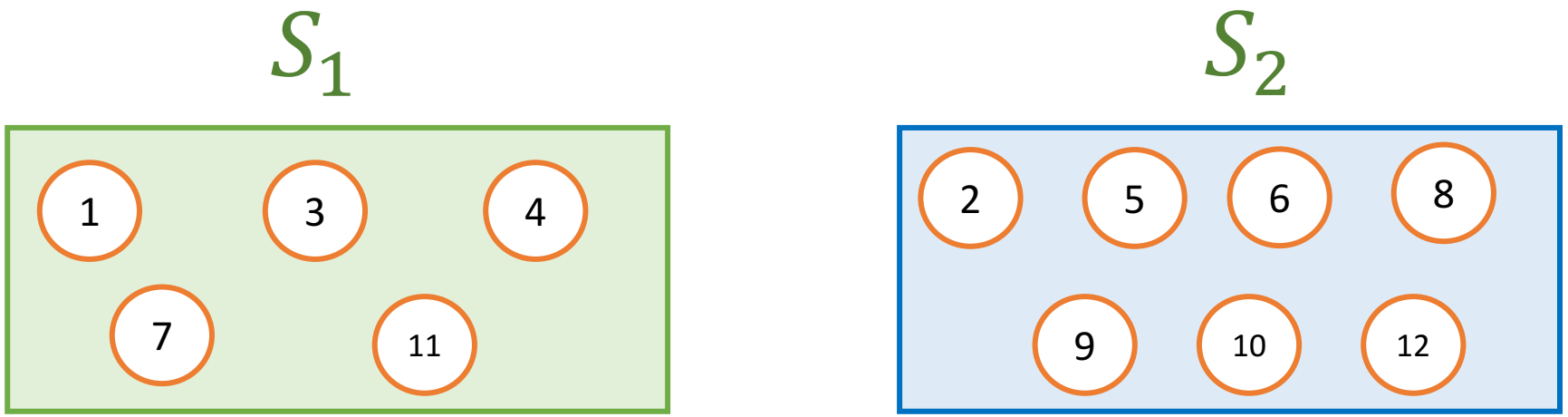
Clustering Task

Clustering: partition $[n]$ to k disjoint subsets $S_1, \cdots, S_k$,

- $S_1 \cup \cdots \cup S_k = [n]$,
- $S_i \cap S_j = \emptyset$, for all $i \neq j$.

Example

- $n = 12$
- $k = 2$



K-Means Clustering Method

Tasks	Methods	Algorithms
clustering	K-means	Lloyd's algorithm

K-Means Clustering Method

Clustering: partition $[n]$ to k disjoint subsets $S_1, \cdots, S_k$,

- $S_1 \cup \cdots \cup S_k = [n]$,
- $S_i \cap S_j = \emptyset$, for all $i \neq j$.

K-Means Model:

$$\min_{S_1,\cdots,S_k} \sum_{i=1}^{k} \sum_{j \in S_i} \left\| \mathbf{x}_j - \frac{1}{|S_i|} \sum_{l \in S_i} \mathbf{x}_l \right\|_2^2 .$$

- $\mathbf{x}_j$ indexed by S_i

K-Means Clustering Method

Clustering: partition $[n]$ to k disjoint subsets $S_1, \cdots, S_k$,

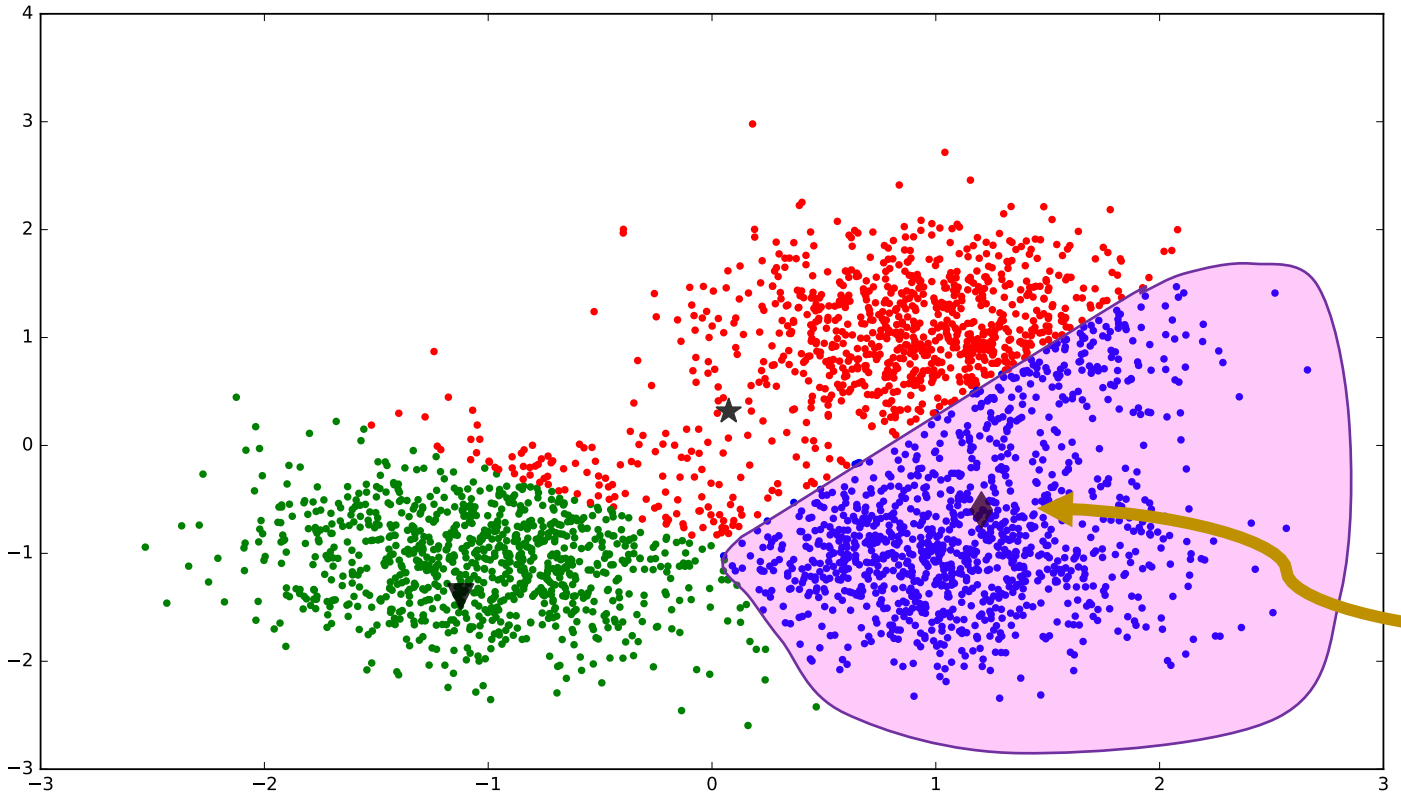
- $S_1 \cup \cdots \cup S_k = [n]$,
- $S_i \cap S_j = \emptyset$, for all $i \neq j$.

K-Means Model:

$$\min_{S_1,\cdots,S_k} \sum_{i=1}^{k} \sum_{j \in S_i} \left\| \mathbf{x}_j - \frac{1}{|S_i|} \sum_{l \in S_i} \mathbf{x}_l \right\|_2^2 .$$

- $\mathbf{x}_j$ indexed by S_i
- The centroid of the i-th cluster.

K-Means Clustering Method



- $\mathbf{x}_j$ indexed by S_i

- The centroid of the i-th cluster.

K-Means Clustering Method

Clustering: partition $[n]$ to k disjoint subsets $S_1, \cdots, S_k$,

- $S_1 \cup \cdots \cup S_k = [n]$,
- $S_i \cap S_j = \emptyset$, for all $i \neq j$.

K-Means Model:

$$\min_{S_1,\cdots,S_k} \sum_{i=1}^{k} \sum_{j \in S_i} \left\| \mathbf{x}_j - \frac{1}{|S_i|} \sum_{l \in S_i} \mathbf{x}_l \right\|_2^2$$

- Squared Euclean distance between $\mathbf{x}_j$ and its cluster centroid.

K-Means Clustering Method

Clustering: partition $[n]$ to k disjoint subsets $S_1, \cdots, S_k$,

- $S_1 \cup \cdots \cup S_k = [n]$,
- $S_i \cap S_j = \emptyset$, for all $i \neq j$.

K-Means Model:

$$\min_{S_1,\cdots,S_k} \sum_{i=1}^{k} \sum_{j \in S_i} \left\| \mathbf{x}_j - \frac{1}{|S_i|} \sum_{l \in S_i} \mathbf{x}_l \right\|_2^2 .$$

- It is a combinatorial optimization problem.
- NP-hard!

Lloyd's Algorithm

Tasks	Methods	Algorithms
clustering	K-means	Lloyd's algorithm

Lloyd's Algorithm

Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

1. Initialize cluster centroids $\mathbf{c}_1, \cdots, \mathbf{c}_k \in \mathbb{R}^d$.

Example 1: Randomly select k points from $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ as the cluster centroids .

Lloyd's Algorithm

Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

1. Initialize cluster centroids $\mathbf{c}_1, \cdots, \mathbf{c}_k \in \mathbb{R}^d$.

Example 2:

- Randomly select one point from $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ as the 1st centroid, $\mathbf{c}_1$.
- Select the point from $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ as $\mathbf{c}_2$ by maximizing $\left|\left|\mathbf{x}_j - \mathbf{c}_1\right|\right|_2^2$.
- Select the point from $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ as $\mathbf{c}_3$ by maximizing $\left|\left|\mathbf{x}_j - \mathbf{c}_1\right|\right|_2^2 + \left|\left|\mathbf{x}_j - \mathbf{c}_2\right|\right|_2^2$.
- And so on…

Lloyd's Algorithm

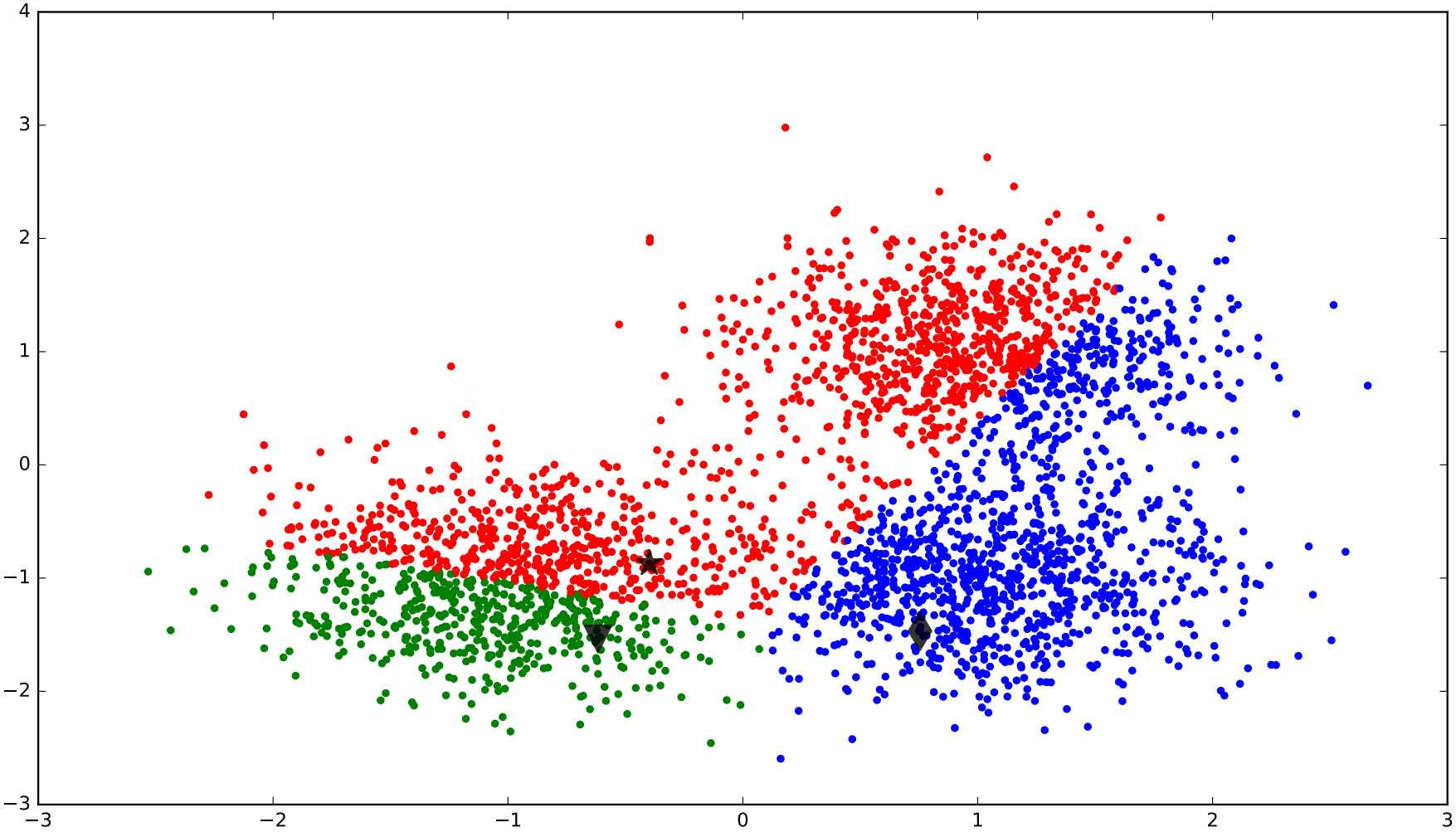
Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

1. Initialize cluster centroids $\mathbf{c}_1, \cdots, \mathbf{c}_k \in \mathbb{R}^d$.
2. Repeat
   i. For $j = 1$ to n, move j to set S_i, where i optimizes

$$\min_{i \in [k]} || \mathbf{x}_j - c_i | \Big|_2^2 .$$

Assign $\mathbf{x}_j$ to its nearest centroid.

Lloyd's Algorithm



Lloyd's Algorithm

Input: vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

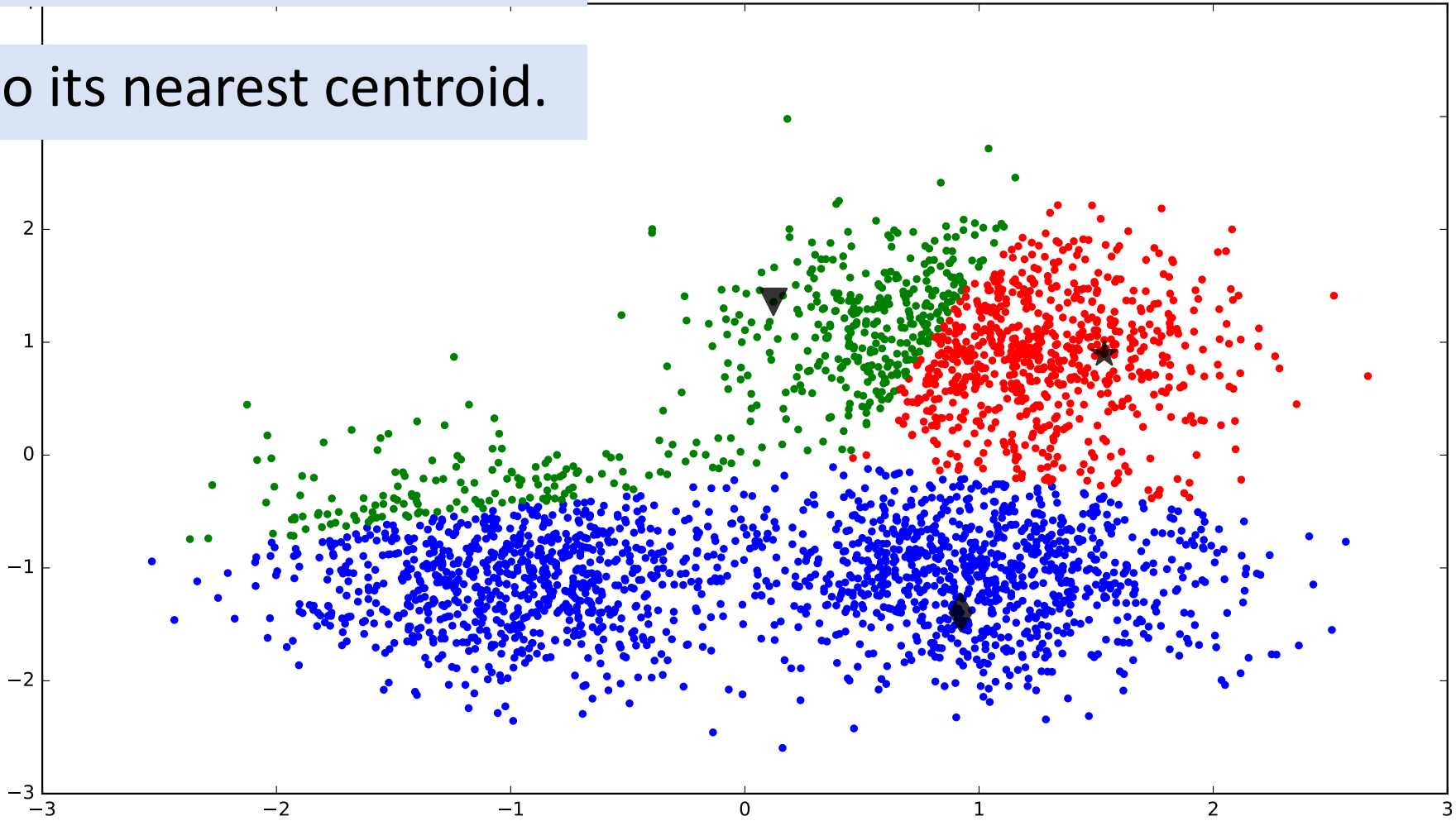
1. Initialize cluster centroids $\mathbf{c}_1, \cdots, \mathbf{c}_k \in \mathbb{R}^d$.
2. Repeat
   i. For $j = 1$ to n, move j to set S_i, where i optimizes
   $$\min_{i \in [k]} || \mathbf{x}_j - c_i | \Big|_2^2 .$$
   ii. For $i = 1$ to k, re-compute centroids by $\tilde{\mathbf{c}}_i = \frac{1}{|S_i|} \sum_{j \in S_i} \mathbf{x}_j$.
   - New centroid $\mathbf{c}_i$: the feature vector (among $\mathbf{x}_1, \cdots, \mathbf{x}_n$) closest to $\tilde{\mathbf{c}}_i$.

move centroids

Lloyd's Algorithm

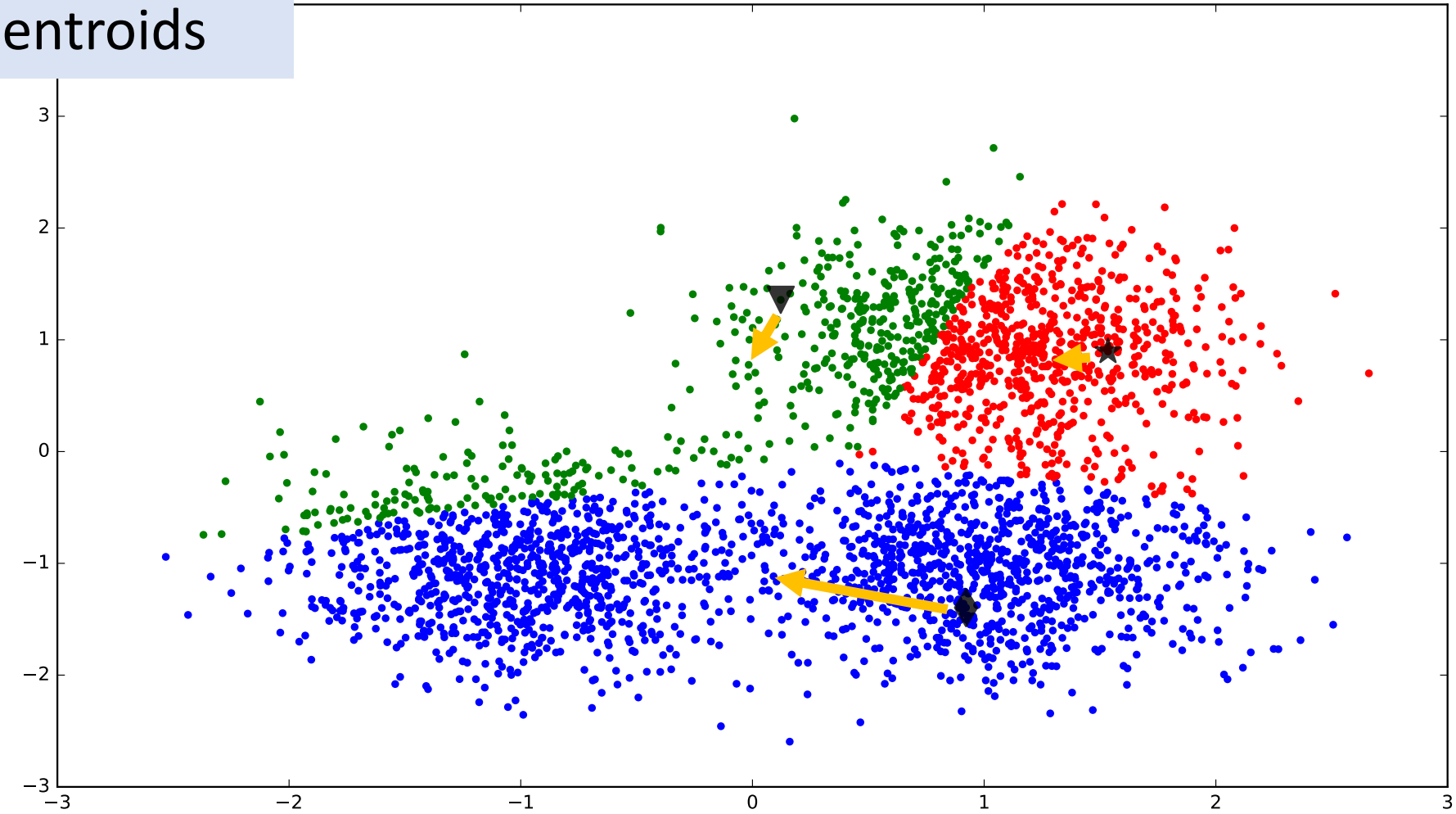
Random initialization.

Assign $\mathbf{x}_j$ to its nearest centroid.



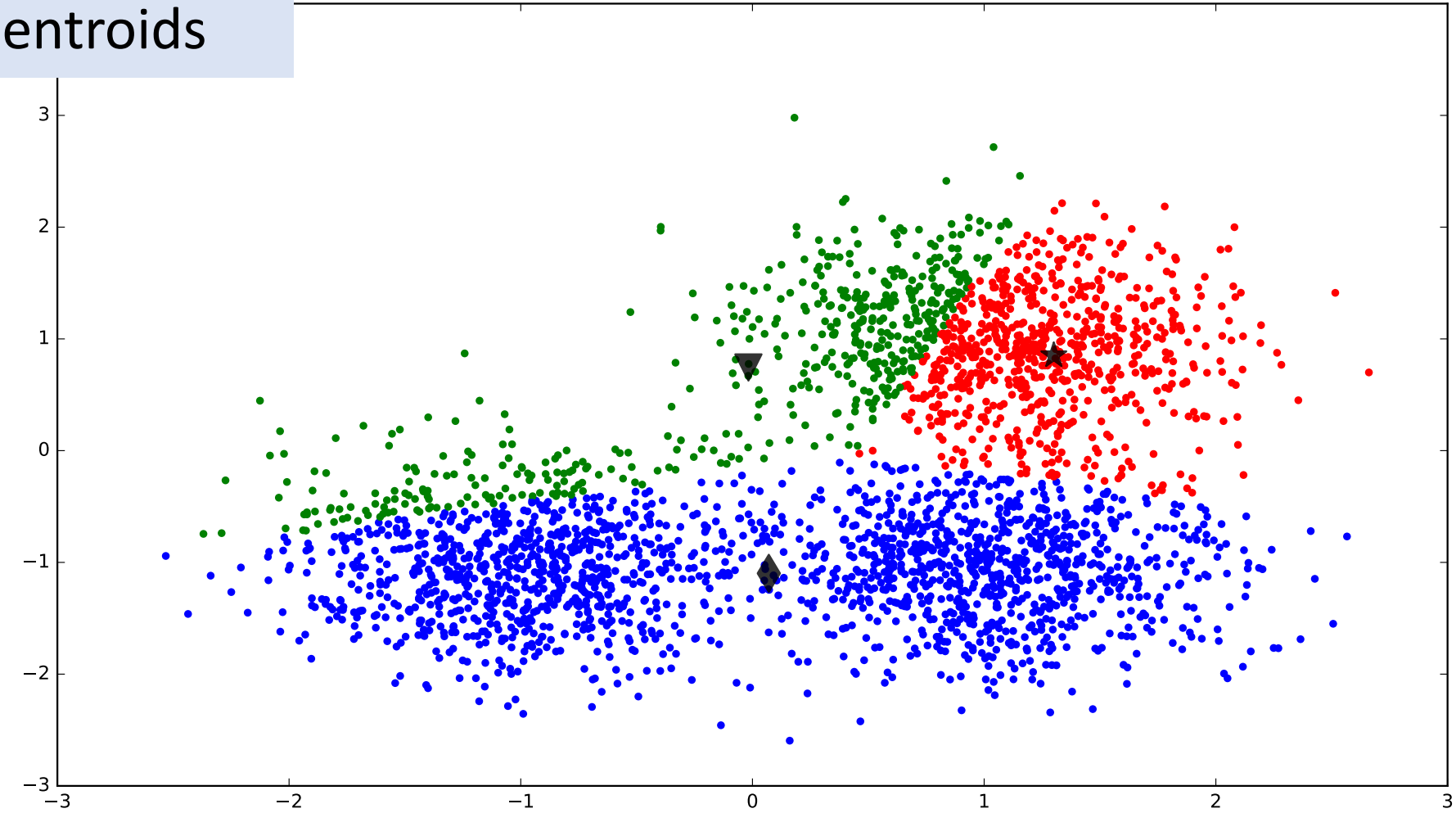
Lloyd's Algorithm

move centroids



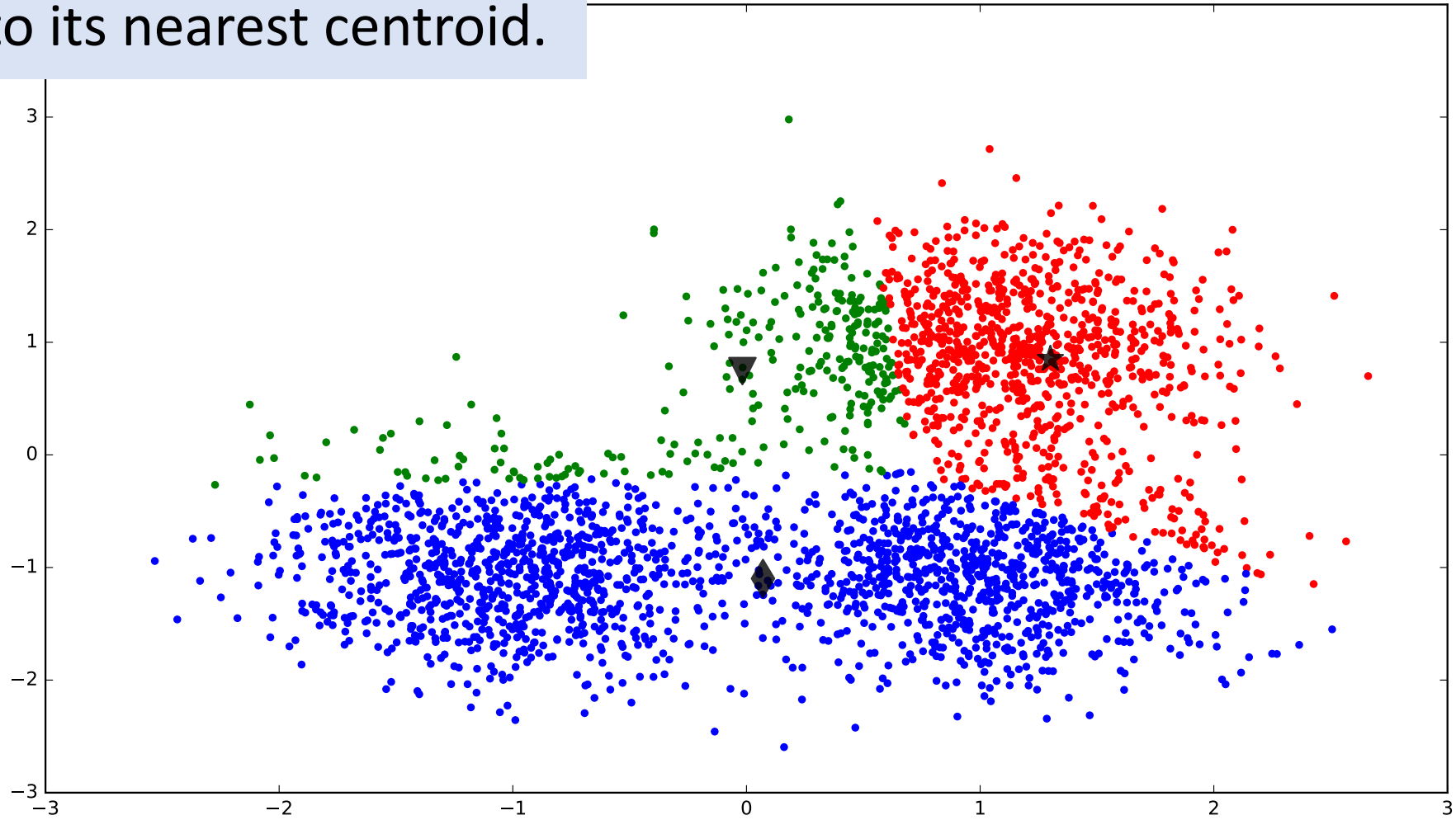
Lloyd's Algorithm

move centroids



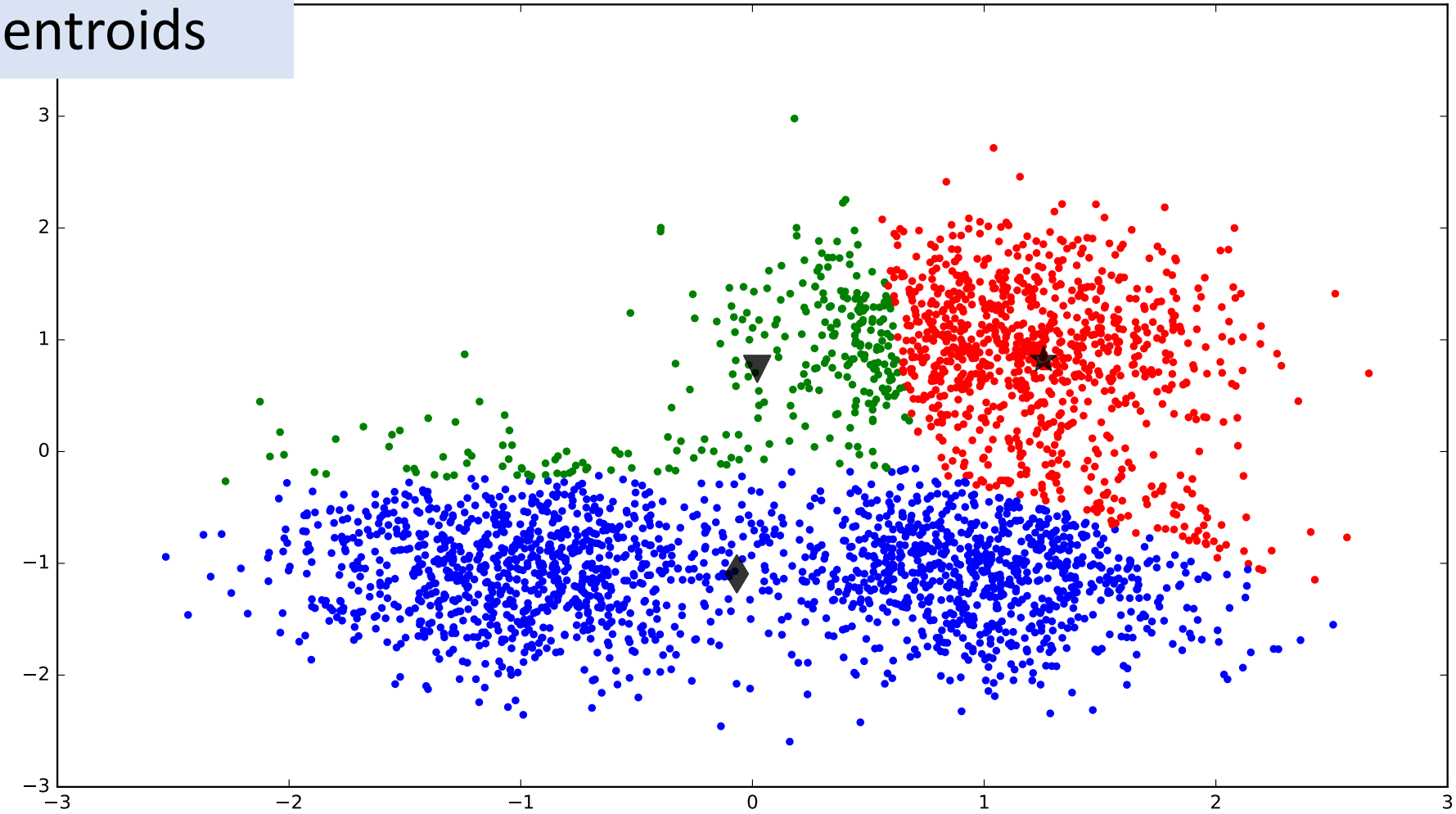
Lloyd's Algorithm

Assign $\mathbf{x}_j$ to its nearest centroid.



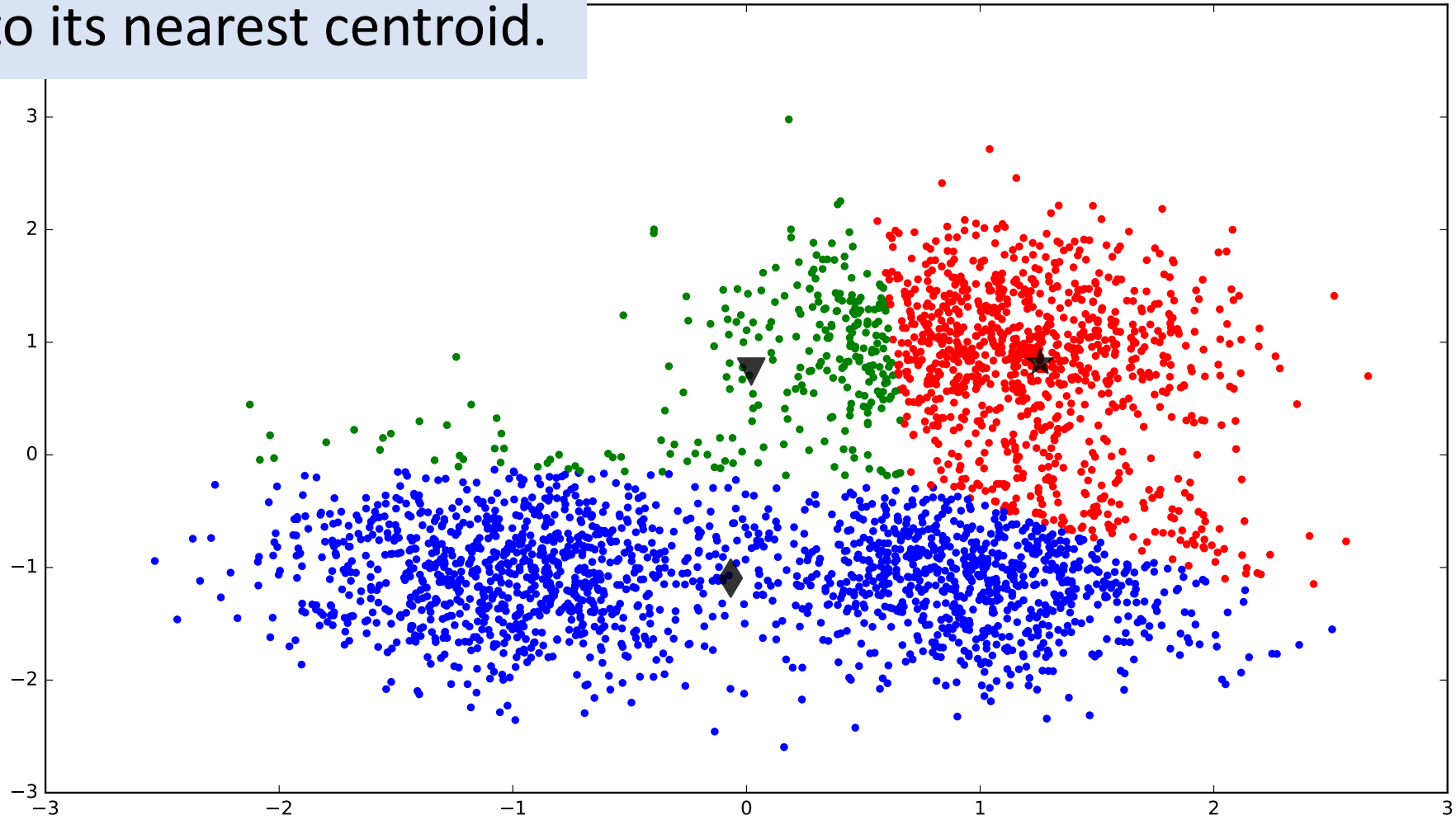
Lloyd's Algorithm

move centroids



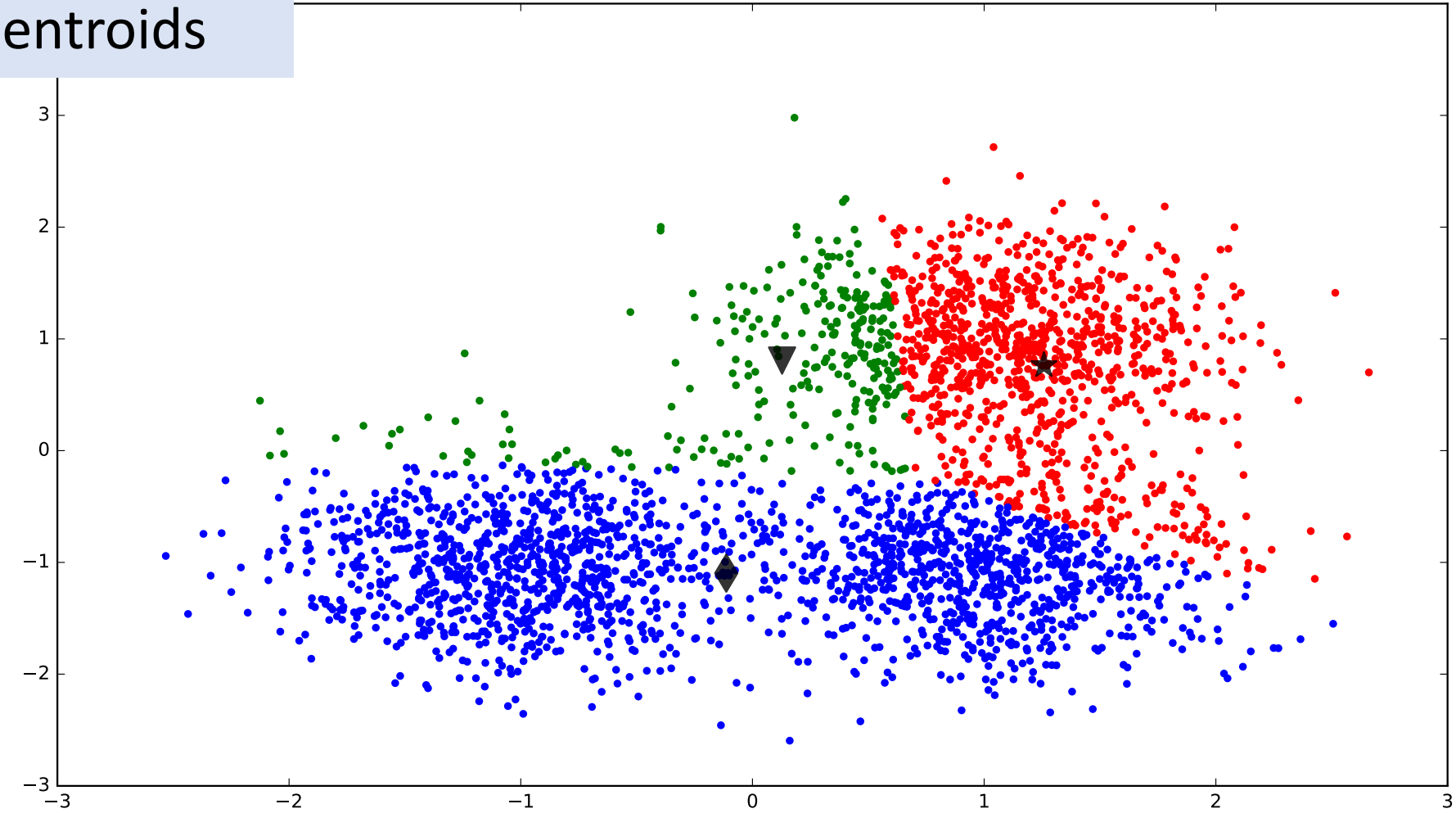
Lloyd's Algorithm

Assign $\mathbf{x}_j$ to its nearest centroid.



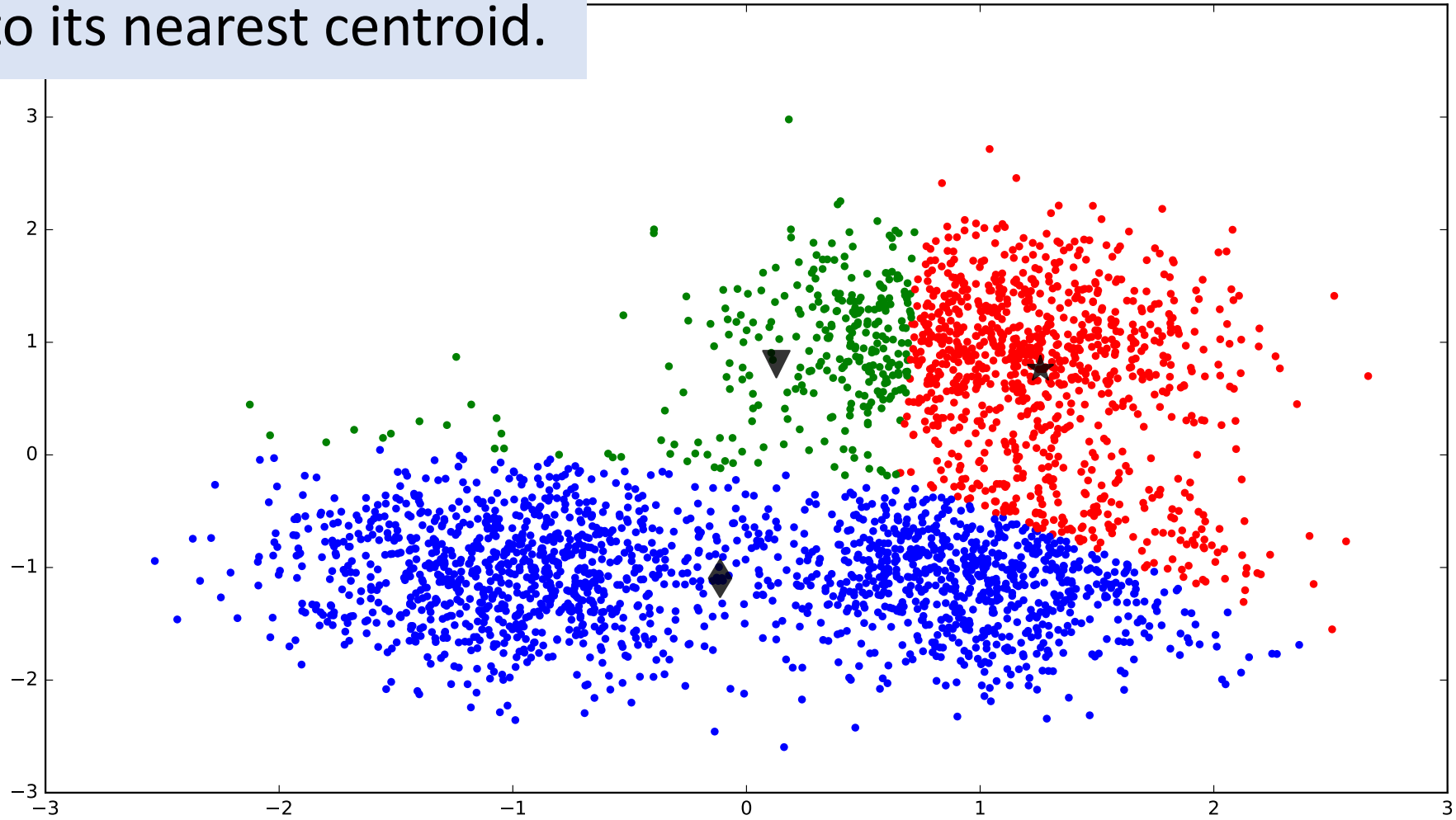
Lloyd's Algorithm

move centroids



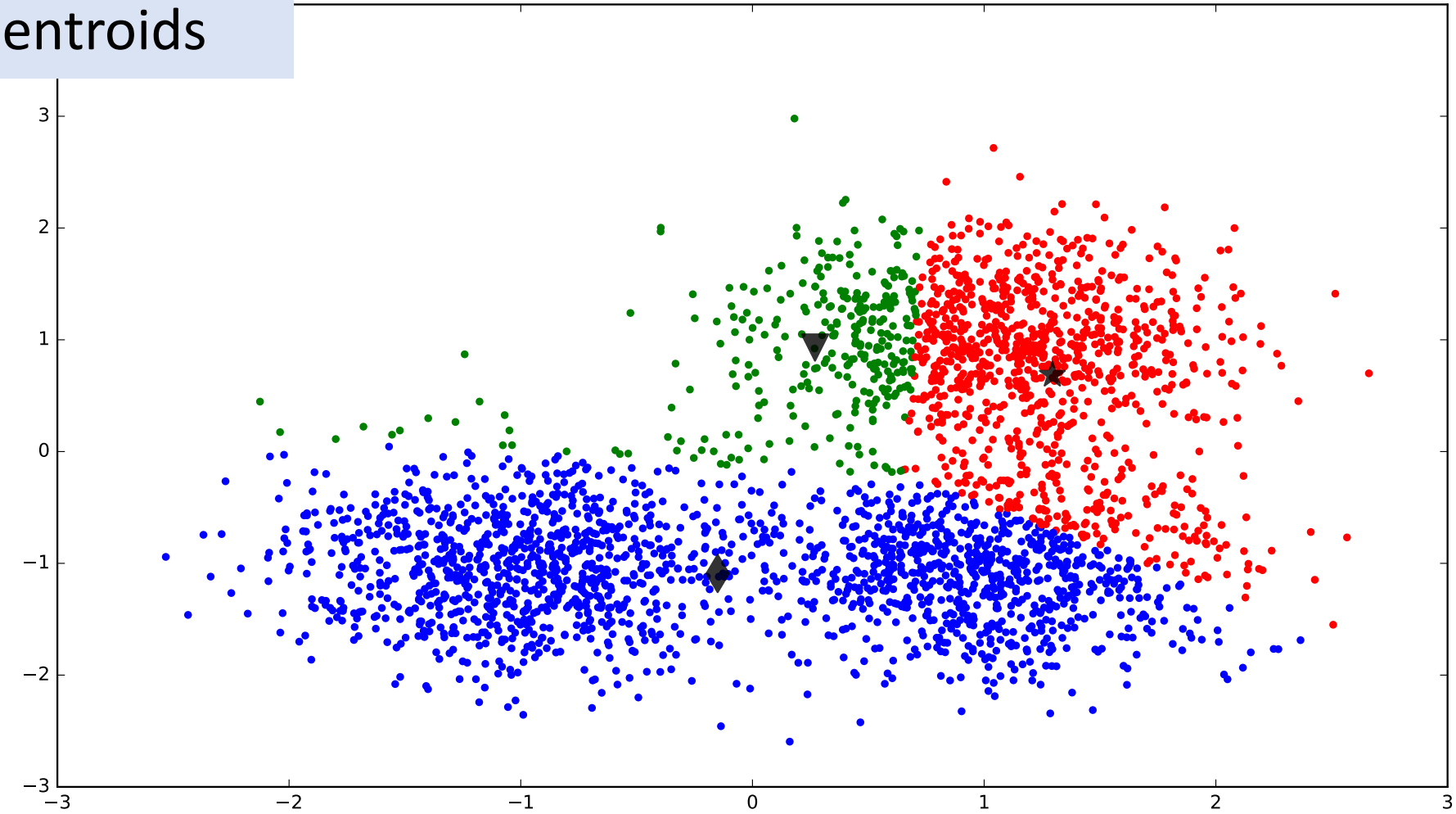
Lloyd's Algorithm

Assign $\mathbf{x}_j$ to its nearest centroid.



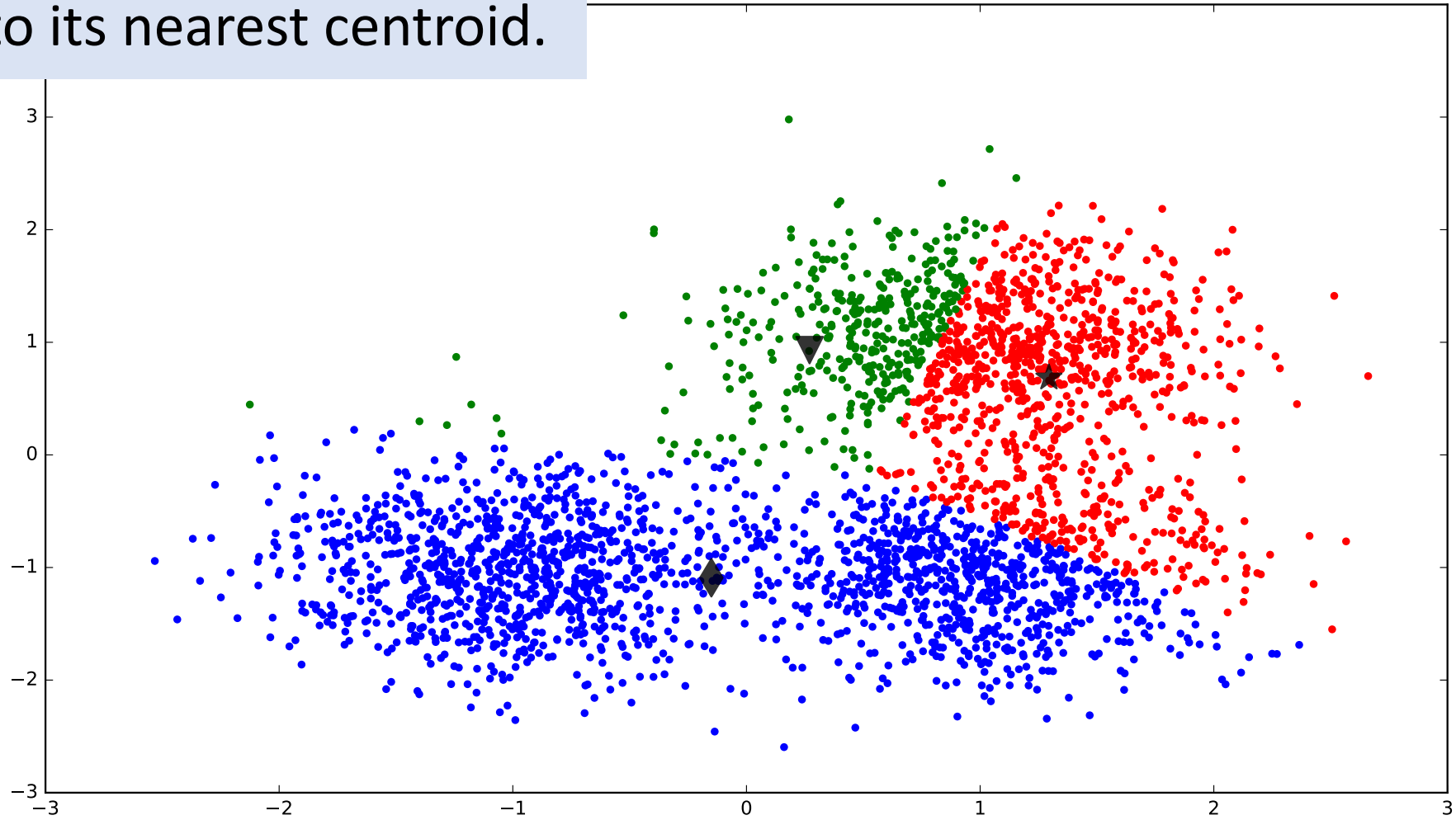
Lloyd's Algorithm

move centroids



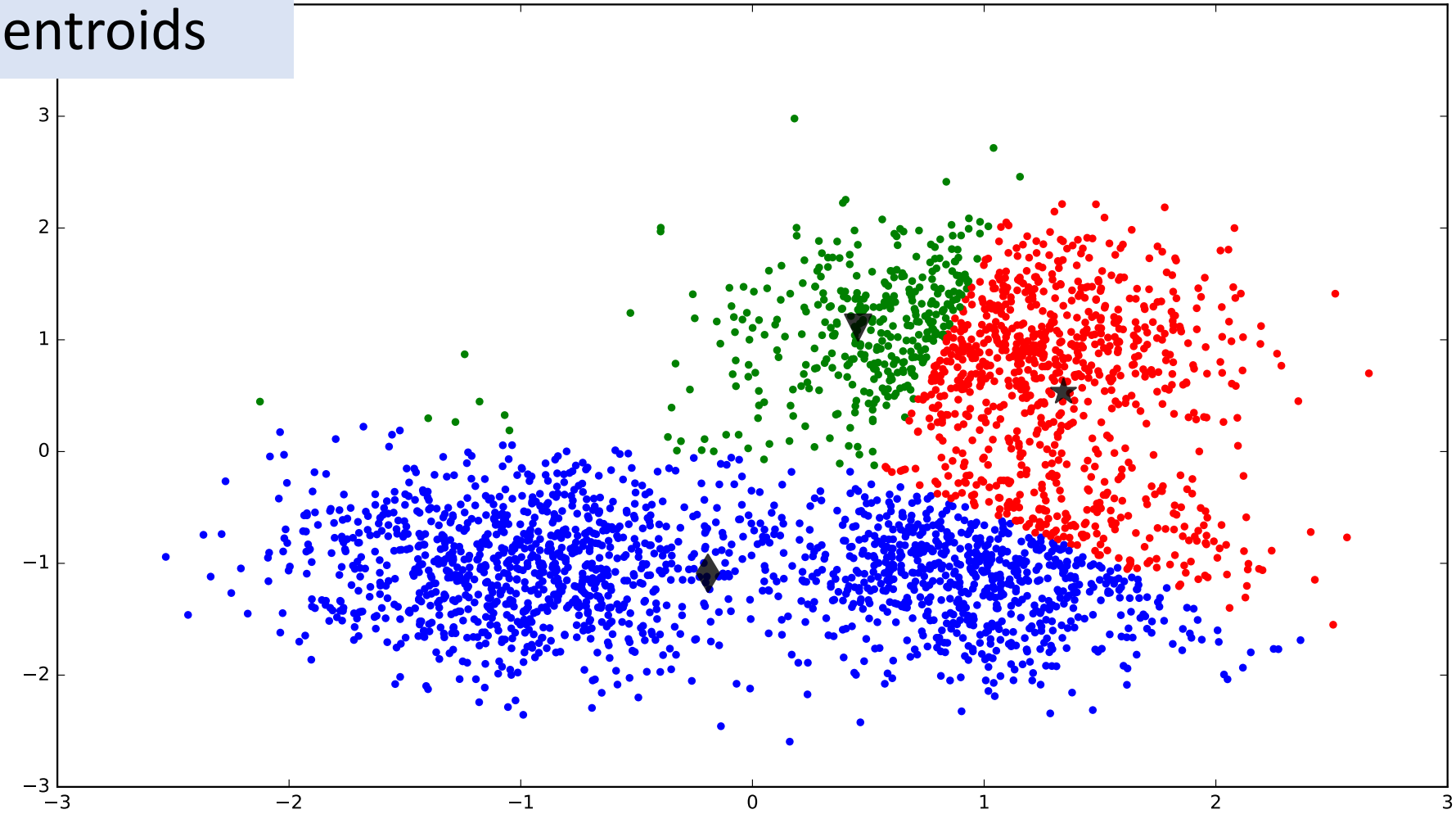
Lloyd's Algorithm

Assign $\mathbf{x}_j$ to its nearest centroid.



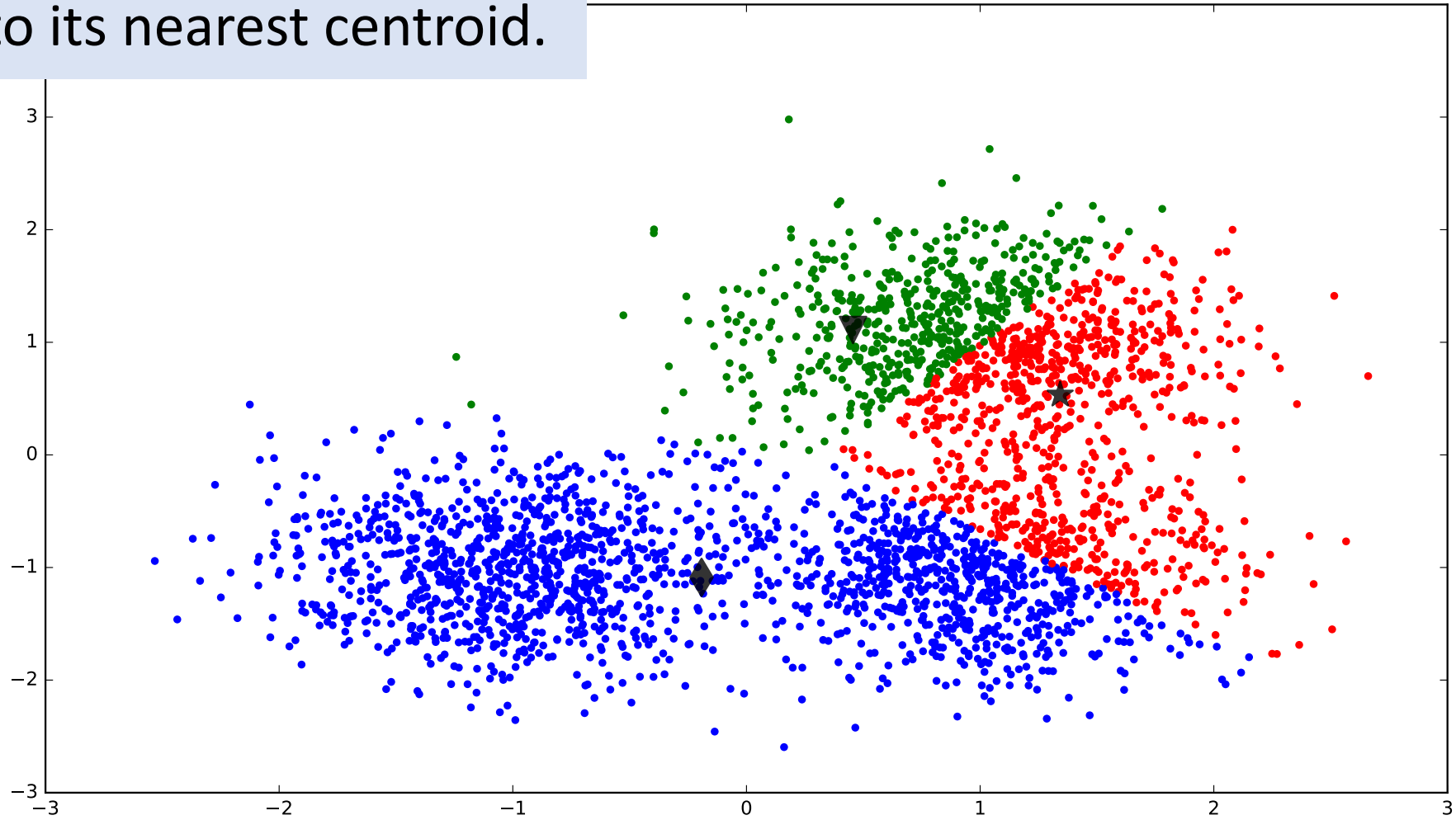
Lloyd's Algorithm

move centroids



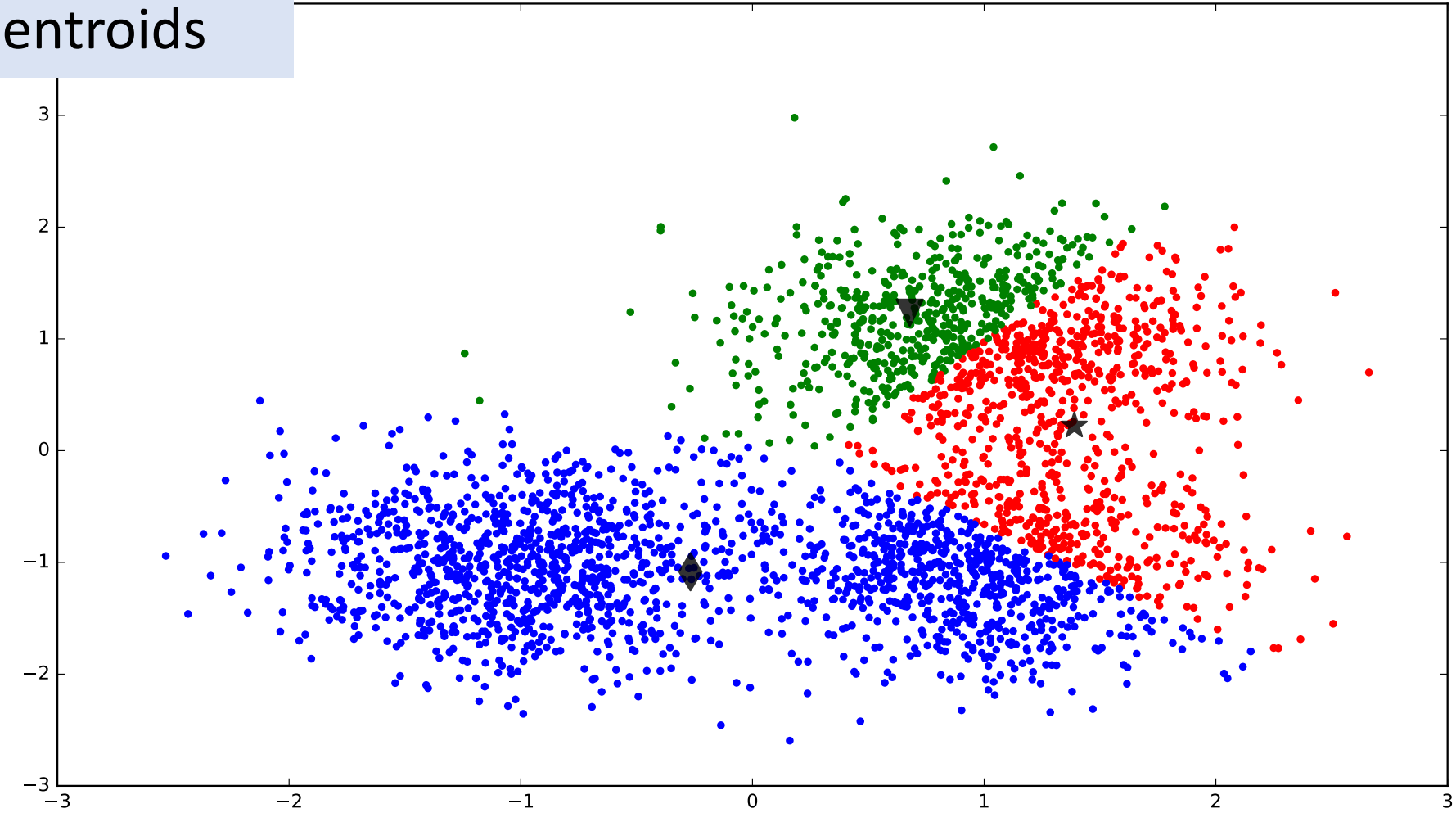
Lloyd's Algorithm

Assign $\mathbf{x}_j$ to its nearest centroid.



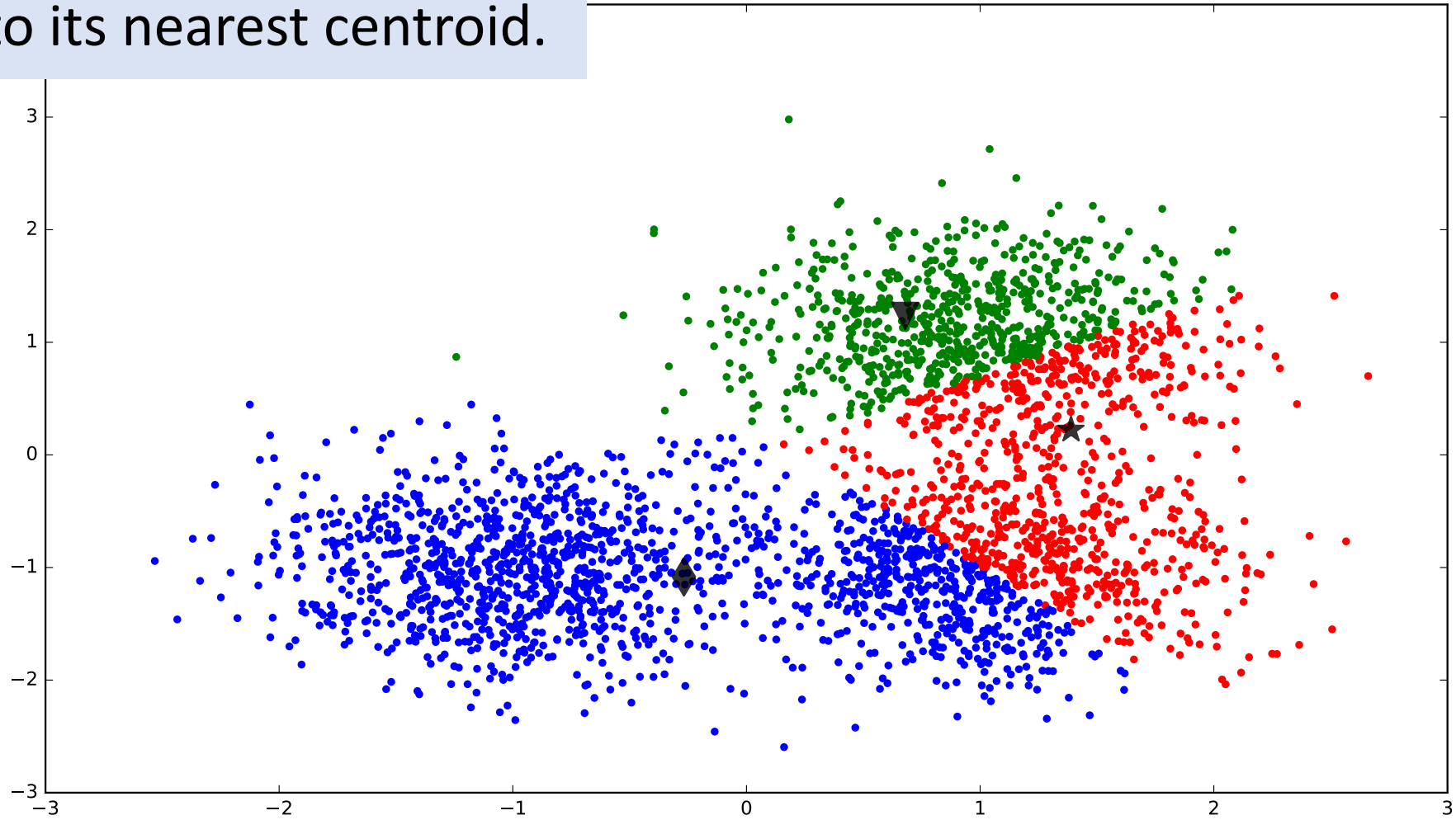
Lloyd's Algorithm

move centroids



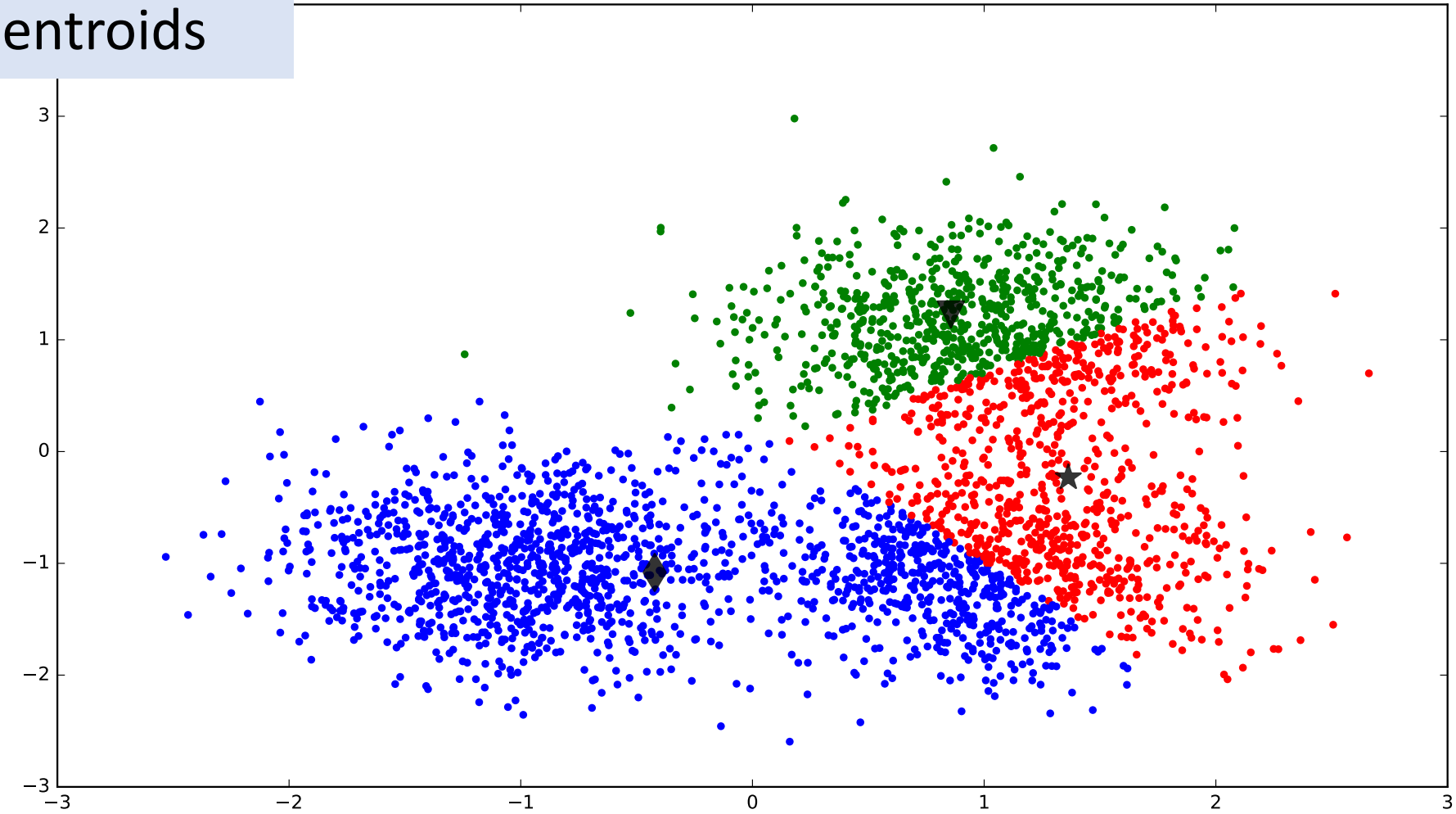
Lloyd's Algorithm

Assign $\mathbf{x}_j$ to its nearest centroid.



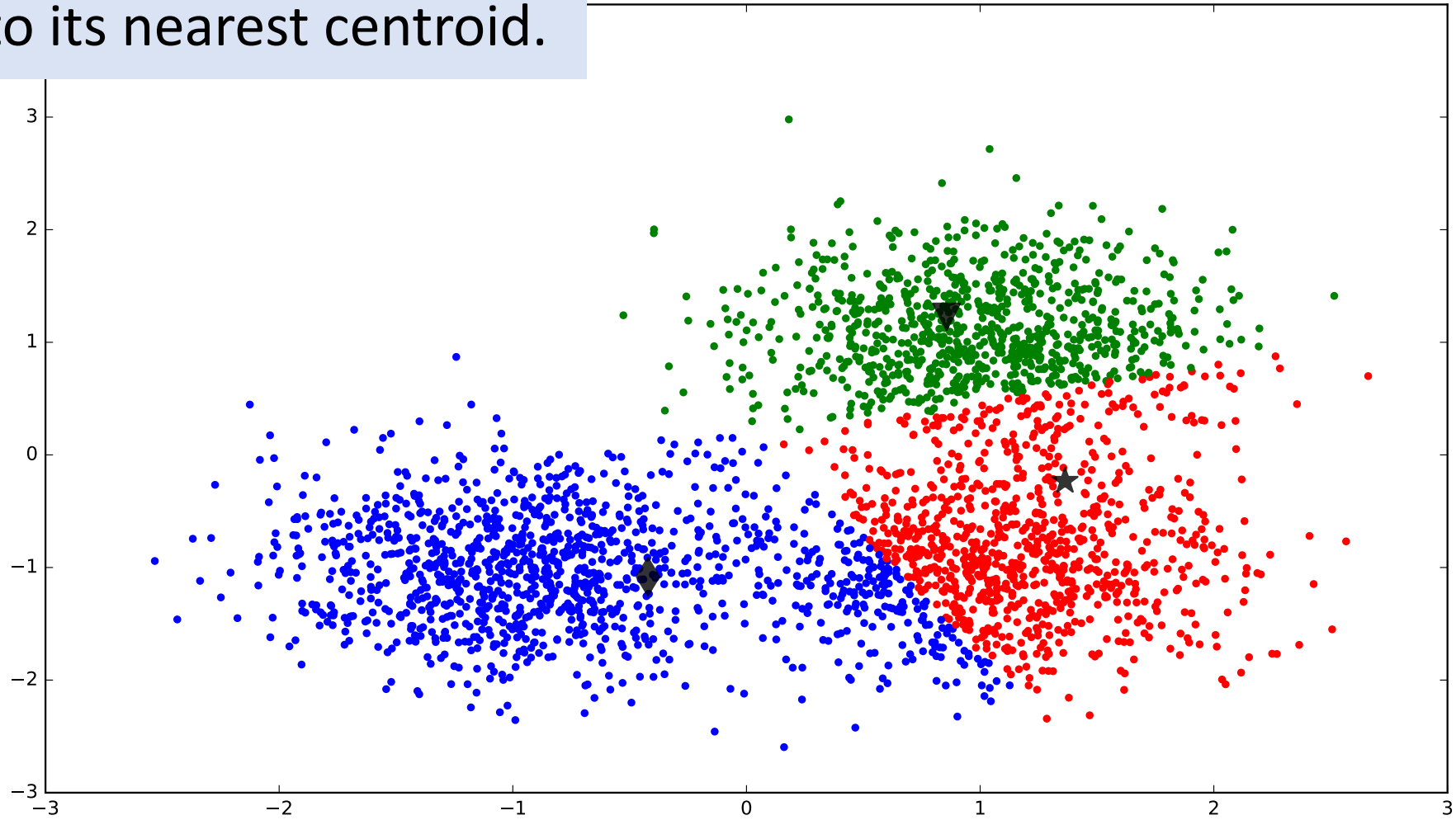
Lloyd's Algorithm

move centroids



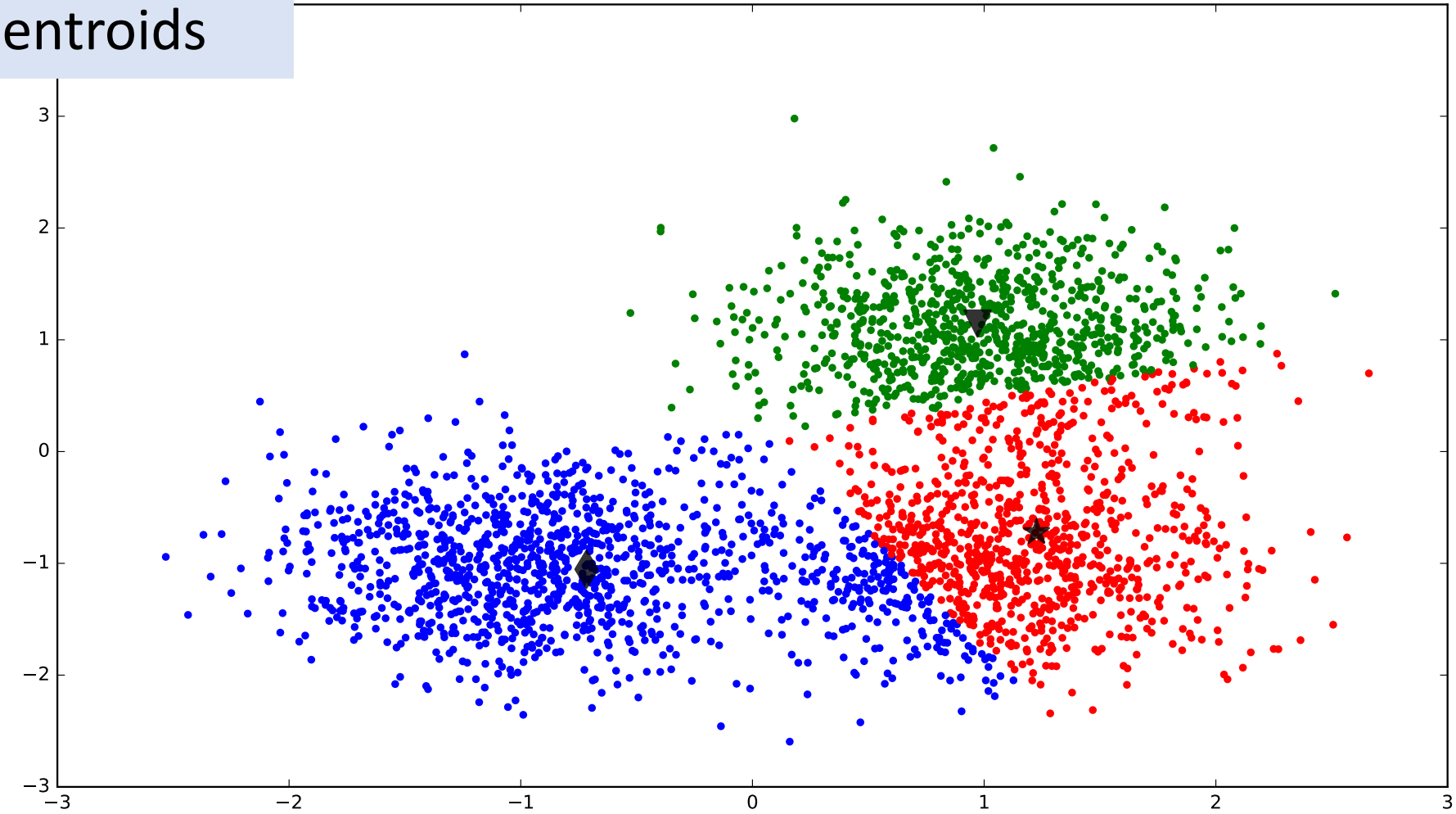
Lloyd's Algorithm

Assign $\mathbf{x}_j$ to its nearest centroid.



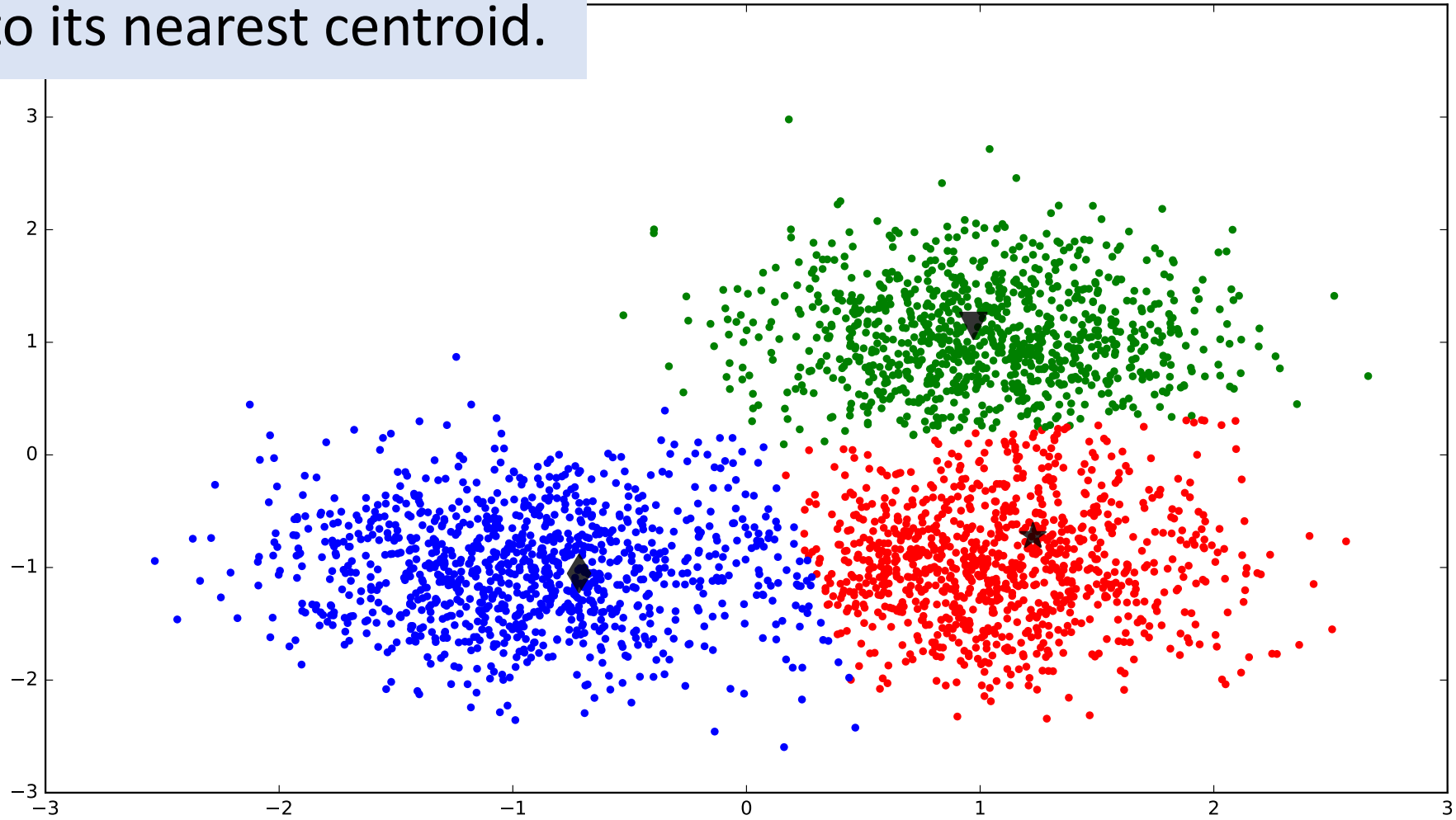
Lloyd's Algorithm

move centroids



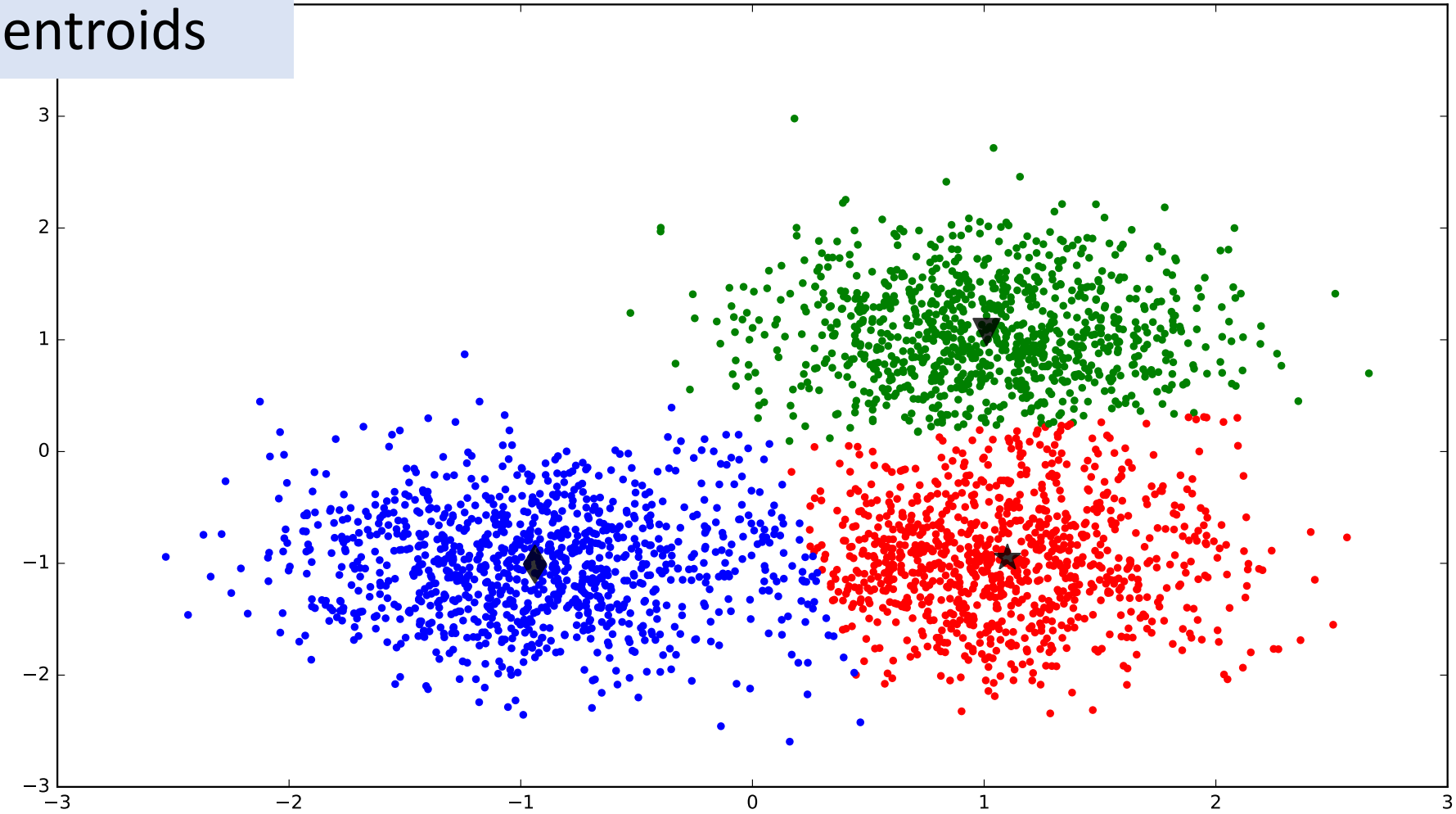
Lloyd's Algorithm

Assign $\mathbf{x}_j$ to its nearest centroid.



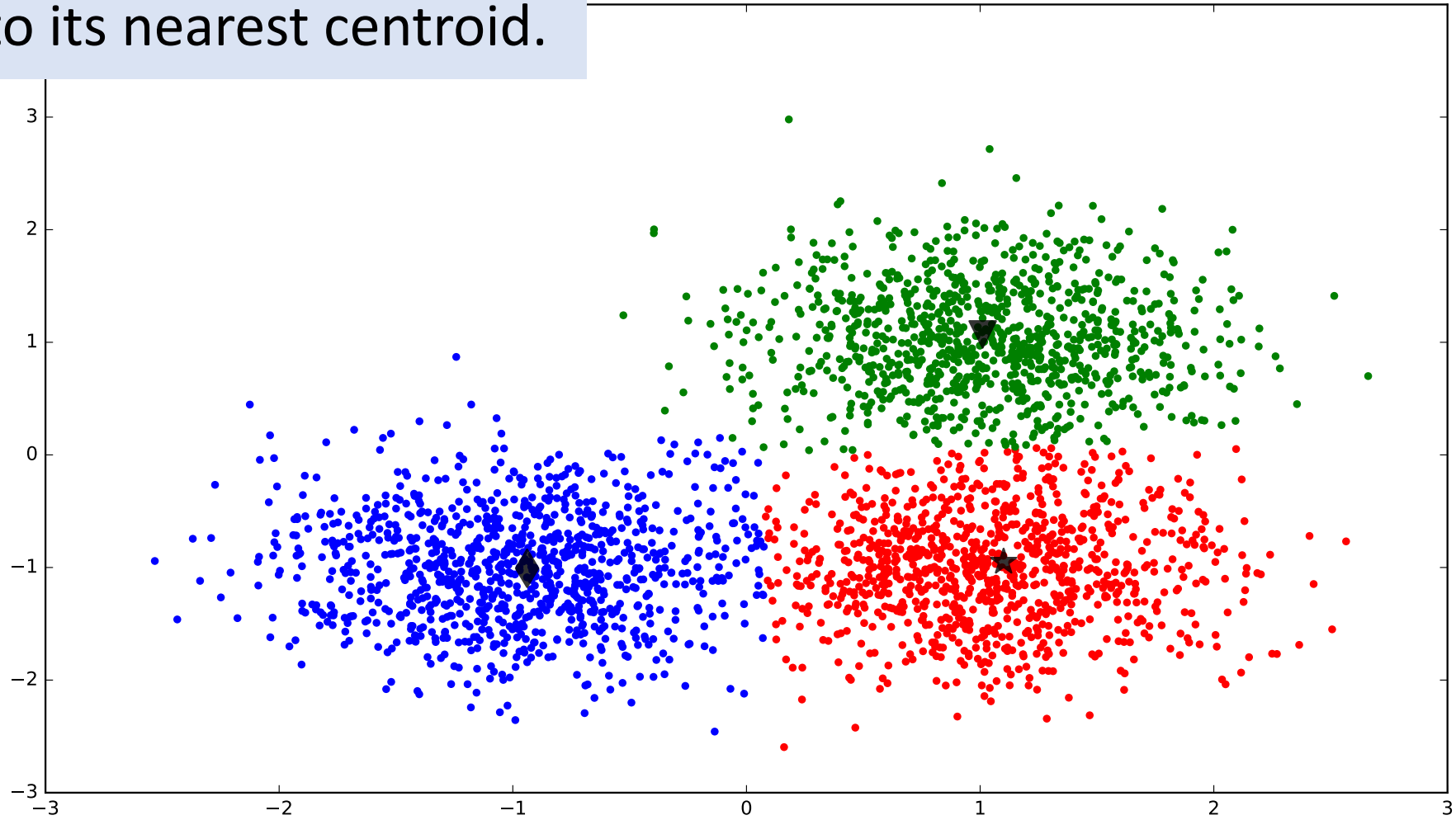
Lloyd's Algorithm

move centroids



Lloyd's Algorithm

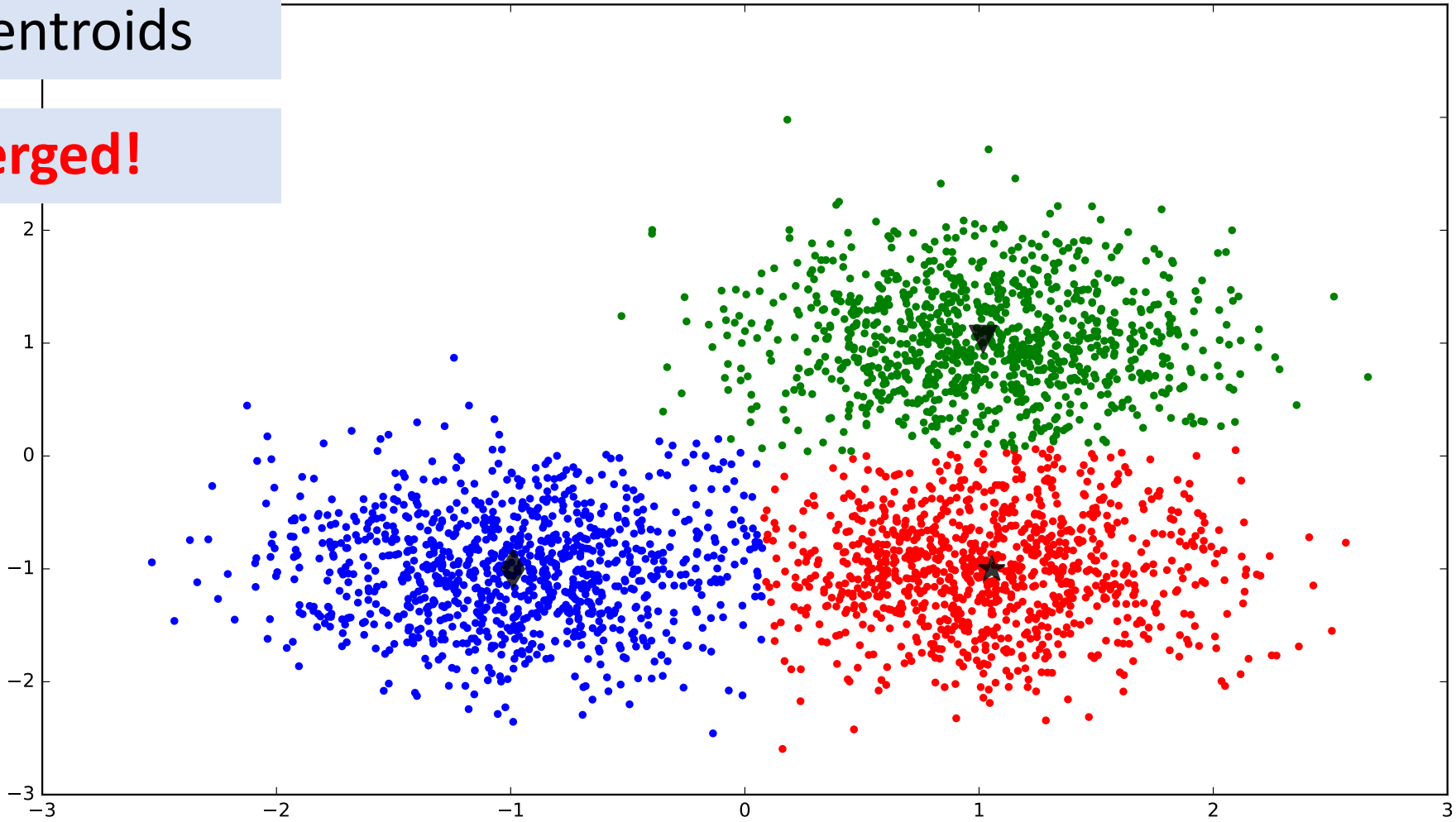
Assign $\mathbf{x}_j$ to its nearest centroid.



Lloyd's Algorithm

move centroids

converged!



Summary

Summary

- Clustering task: partition $[n]$ into k subsets according to the feature vectors.
- K-means model: $\min_{S_1,\cdots,S_k} \sum_{i=1}^{k} \sum_{j\in S_i} \left\| \mathbf{x}_j - \frac{1}{|S_i|} \sum_{l\in S_i} \mathbf{x}_l \right\|_2^2$.
- Lloyd's algorithm for solving the model (approximately).
- Other algorithms: Forgy, MacQueen, and Hartigan.

Tasks

clustering

Methods

K-means

Algorithms

Lloyd's algorithm

Summary

- Clustering task: partition $[n]$ into k subsets according to the feature vectors.
- K-means model: $\min_{S_1,\cdots,S_k} \sum_{i=1}^{k} \sum_{j\in S_i} \left|\left| \mathbf{x}_j - \frac{1}{|S_i|} \sum_{l\in S_i} \mathbf{x}_l \right|\right|_2^2$.
- Lloyd's algorithm for solving the model (approximately).
- Other algorithms: Forgy, MacQueen, and Hartigan.

- There is not such a thing called "k-means algorithm"!

Application: Space Partitioning for KNN

Evaluation Metrics

- Objective function of k-means.
- Accuracy.
- Normalized mutual information (NMI).

See scikit-learn: http://scikit-learn.org/stable/modules/classes.html#clustering-metrics