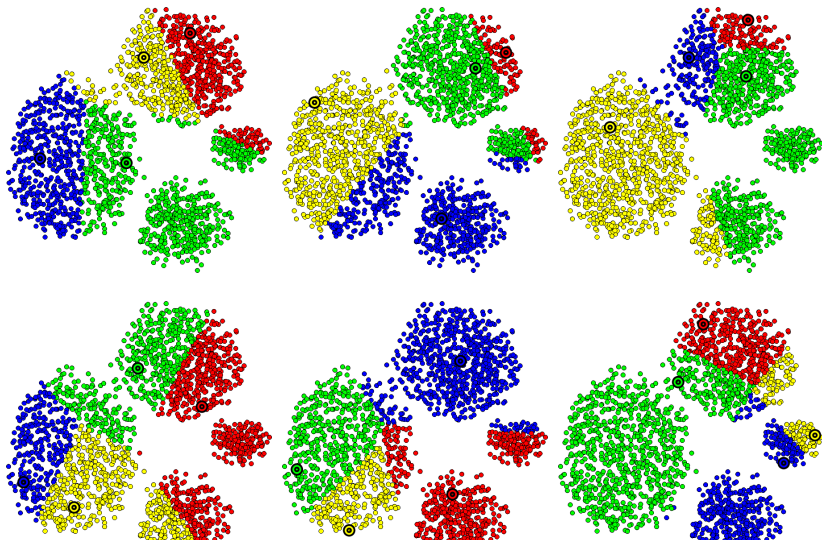


Agrupamiento basado en particiones

Inicializar K-means



Agrupamiento basado en particiones

Inicializar K-means

Método basado en repeticiones

- ▶ Inicializar aleatoriamente y ejecutar K -means R veces
- ▶ Medir la bondad de los R diferentes agrupamientos
- ▶ Devolver como resultado el mejor agrupamiento

Agrupamiento basado en particiones

Inicializar K-means

K-means++

Inicialización avanzada del método K -means favoreciendo la separación de los centros iniciales

Intuición

Elección individual (y dependiente) de los centros

- ▶ **Centros:** Muestreo aleatorio no uniforme
- ▶ **Probabilidad:** Un ejemplo tiene mayor probabilidad de ser escogido como centro (inicial) cuanto mayor sea su distancia con los centros

Agrupamiento basado en particiones

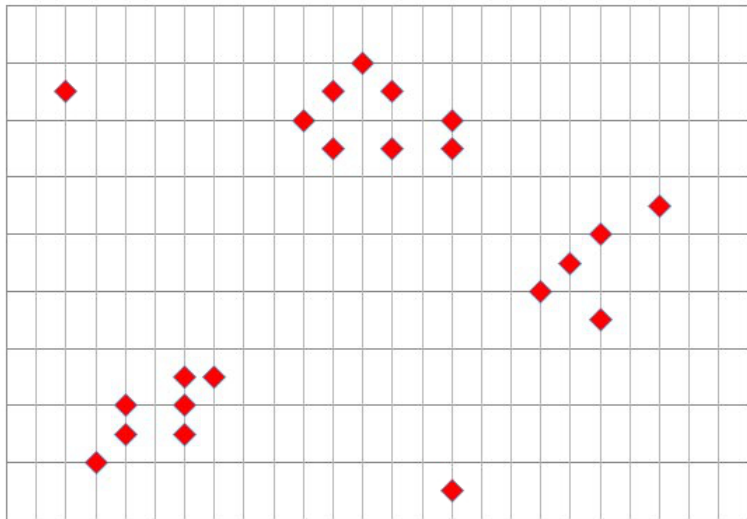
K-means++

La probabilidad de un ejemplo es proporcional al cuadrado de su distancia mínima a un centro ya incluido, $D(\mathbf{x}, S)$

$$D(\mathbf{x}, S) = \min_{k \in \{1, \dots, |S|\}} \|\mathbf{x} - \bar{\mathbf{x}}_k\|^2$$

Agrupamiento basado en particiones

K-means++



Agupamiento basado en particiones

K-means++

K-means++ (inicialización)

Recibe: Conjunto de entrenamiento, $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$; número de clústeres, K

1. Elección (aleatoria) de 1 punto del conjunto de entrenamiento como primer centro, $S = \{\bar{\mathbf{x}}_1\}$.

2. Mientras $|S| < K$, repetir

2.1. Para todos los ejemplos de entrenamiento, calcular $D(\mathbf{x}_i, S)$, la distancia al centro más cercano: $D(\mathbf{x}_i, S) = \min_{k \in \{1, \dots, |S|\}} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2$

2.2. Muestrear un nuevo caso \mathbf{x}' del conjunto de entrenamiento, donde el caso \mathbf{x} tiene probabilidad $D(\mathbf{x}, S)^2 / \sum_{i=1}^n D(\mathbf{x}_i, S)^2$ y añadir a S : $S = S \cup \{\mathbf{x}'\}$

Devuelve: Conjunto de centros, $\{\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_K\}$
