

# Exercises for K-Means

Jordan Boyd-Graber  
Digging into Data

April 7, 2013

## 1 Clustering Example



## 2 K-means Algorithm

**Input:** A set of data points  $X \equiv \{x_1, x_2, \dots, x_N\}$ , a number of clusters  $K$ , and initial means  $M = \{\mu_1, \dots, \mu_K\}$ .

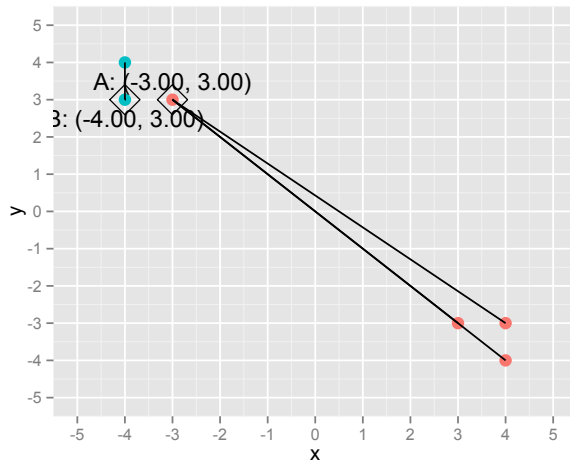
**Output:** Updated means  $\{\mu_1, \dots, \mu_K\}$  and cluster assignments  $\{z_1, \dots, z_N\}$ . Such that  $f(X, Z, M) = \sum_i \|x_i - \mu_{z_i}\|$  is minimized.

While  $f(X, Z, M)$  has not converged:

1. Assign each  $x_i$  to the mean  $c$  such that  $d(\mu_c, x_i)$  is smallest ( $z_i = c$ ).
2. Recompute each of the means:  $\mu_j = \frac{1}{V_j} \sum_i \mathbb{1}[z_i = j] x_i$

## 3 K-means Warmup

### 3.1 Two Cluster 1



Updated Means:

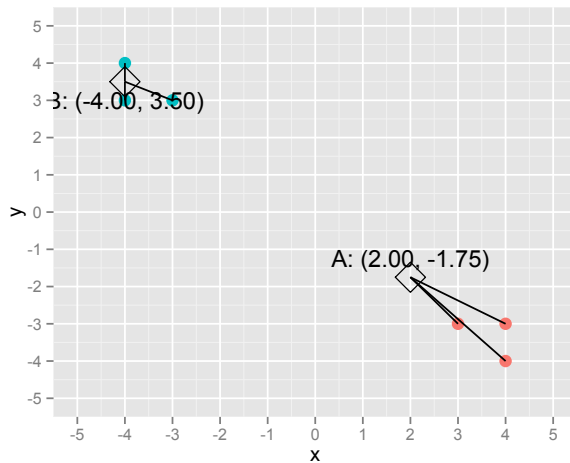
$$\mu_A = \frac{1}{4} ((-3, 3) + (3, -3) + (4, -3) + (4, -4))$$

=

$$\mu_B = \frac{(-4, 3) + (-4, 4)}{2}$$

=

### 3.2 Two Cluster 2



Updated Means:

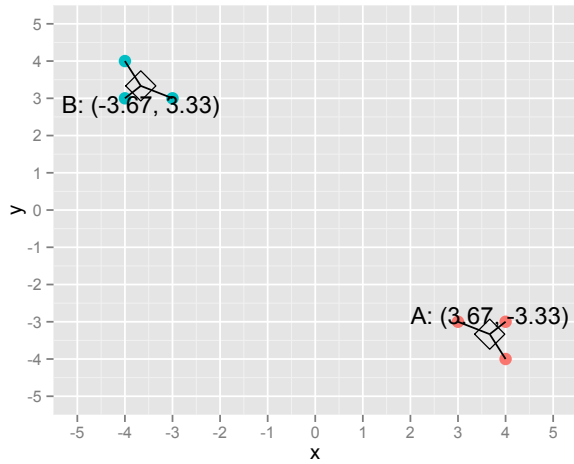
$$\mu_A = \frac{(3, -3) + (4, -3) + (4, -4)}{3}$$

=

$$\mu_B = \frac{(-4, 3) + (-4, 4) + (-3, 3)}{3}$$

=

### 3.3 Two Cluster 3



Updated Means:

$$\mu_A = \frac{(3, -3) + (4, -3) + (4, -4)}{3}$$

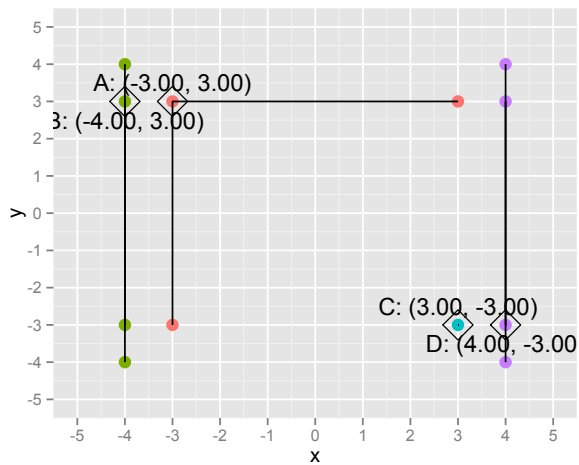
=

$$\mu_B = \frac{(-4, 3) + (-4, 4) + (-3, 3)}{3}$$

=

## 4 Four Clusters

### 4.1 Four Clusters 1



Updated Means:

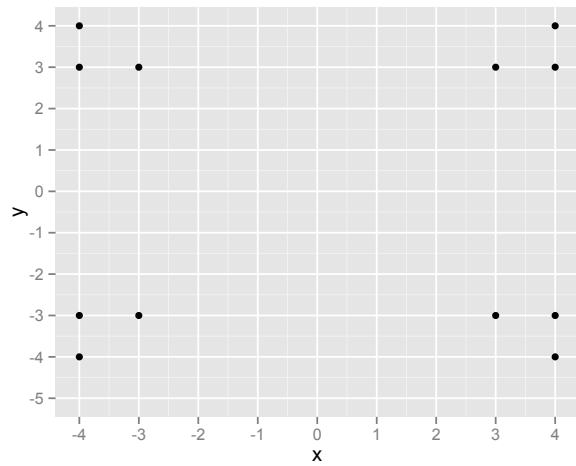
$$\mu_A =$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

## 4.2 Four Clusters 2



Updated Means:

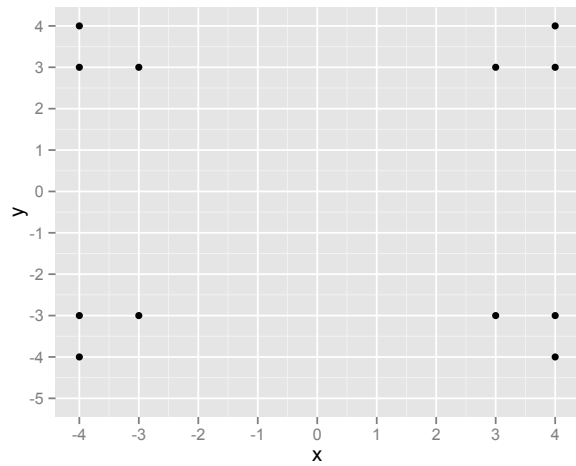
$$\mu_A =$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

## 4.3 Four Clusters 3



Updated Means:

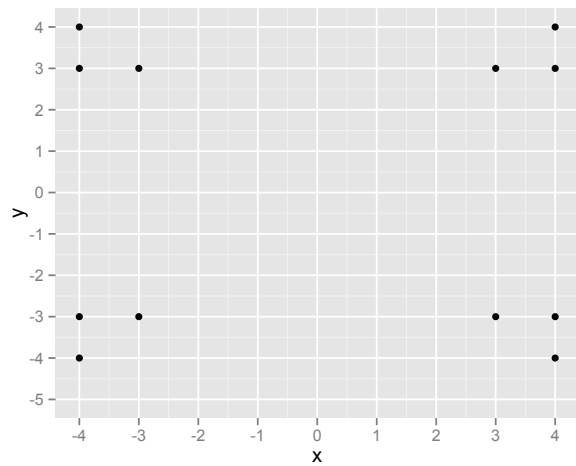
$$\mu_A =$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

## 4.4 Four Clusters 4



Updated Means:

$$\mu_A =$$

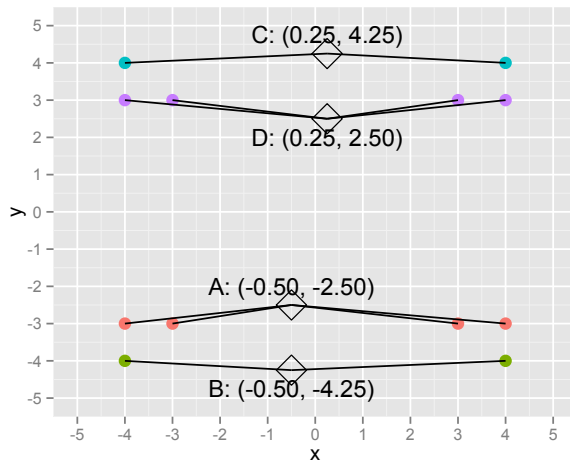
$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

## 5 Strange Initialization

### 5.1 Strange Initialization 1



Updated Means:

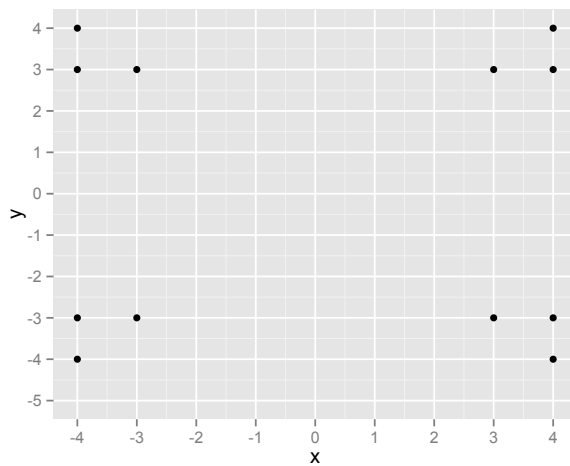
$$\mu_A =$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

### 5.2 Strange Initialization 2



Updated Means:

$$\mu_A =$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$