



Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**



Clustering

Introduction to Data Science
University of Colorado Boulder

SLIDES ADAPTED FROM LAUREN HANNAH

Clustering Lab

- Review of k -means
- Work through k -means example
- Connection to GMM

k-means

```
1: procedure KMeans( $X, M$ )
2:    $s \leftarrow \infty$ 
3:    $Z \leftarrow \text{AssignToClosestCluster}(X, M)$ 
4:   while  $s > \text{Score}(X, Z, M)$  do    ▷ Iterate until score stops changing
5:      $s \leftarrow \text{Score}(X, Z, M)$     ▷ Compute score for old configuration
6:      $Z \leftarrow \text{AssignToClosestCluster}(X, M)$ 
7:     for  $k \in \{1, \dots, K\}$  do    ▷ For each cluster mean
8:        $v \leftarrow 0, \mu_k \leftarrow \vec{0}$ 
9:       for  $i \in \{1, \dots, N\}$  do    ▷ For each observation
10:        if  $z_i = k$  then ▷ If the observation is assigned to cluster  $k$ 
11:           $\mu_k \leftarrow \mu_k + x_i$     ▷ Add observation to sum
12:           $v \leftarrow v + 1$     ▷ Count points in cluster  $k$ 
13:           $\mu_k \leftarrow \frac{\mu_k}{v}$     ▷ Divide by number of observations
14:   return  $Z$ 
```

Score

1: **procedure** Score(X, Z, M)

▷ Current objective function

2: $s \leftarrow 0$

3: **for** $i \in \{1, \dots, N\}$ **do**

▷ For each observation

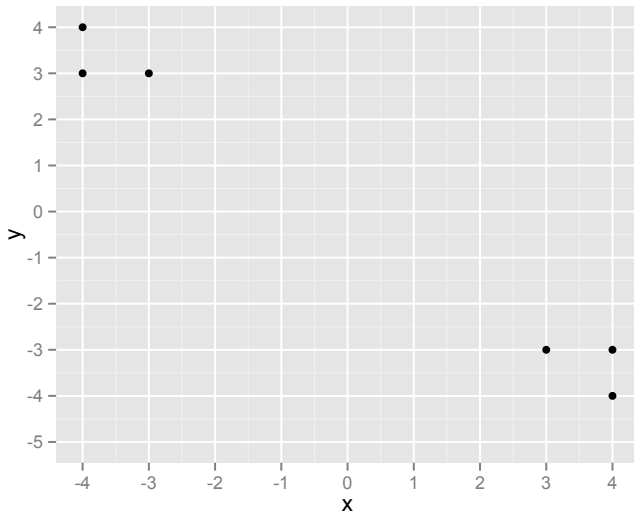
4: $s \leftarrow s + \|x_i - \mu_{z_i}\|$

▷ Accumulate how far it is from its mean

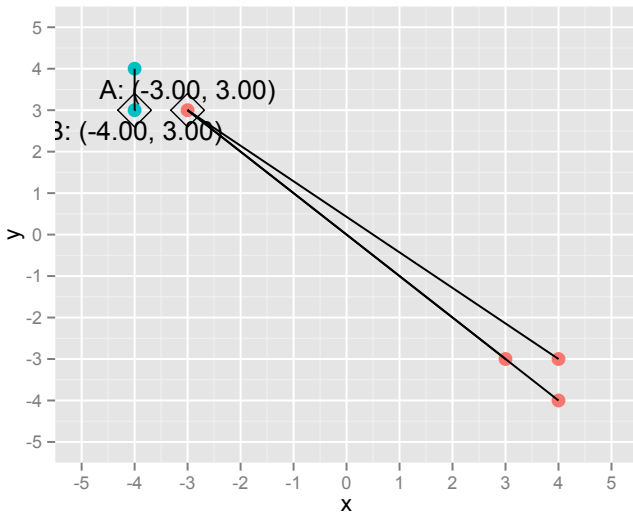
Find closest

```
1: procedure AssignToClosestCluster( $X, M$ )
2:    $Z \leftarrow \text{Vector}(N)$            ▷ Initialize assignments  $Z$  as a  $N$ -vector
3:   for  $i \in \{1, \dots, N\}$  do           ▷ For each observation
4:      $d \leftarrow -\infty$ 
5:     for  $k \in \{1, \dots, K\}$  do
6:       if  $\|x_i - \mu_k\| < d$  then
7:          $z_i \leftarrow k$ 
8:          $d \leftarrow \|x_i - \mu_k\|$ 
9:   return  $Z$ 
```

Two Points



Two Points



Two Points

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

=

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

=

Two Points

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

$$= (2, -1.75)$$

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

$$=$$

Two Points

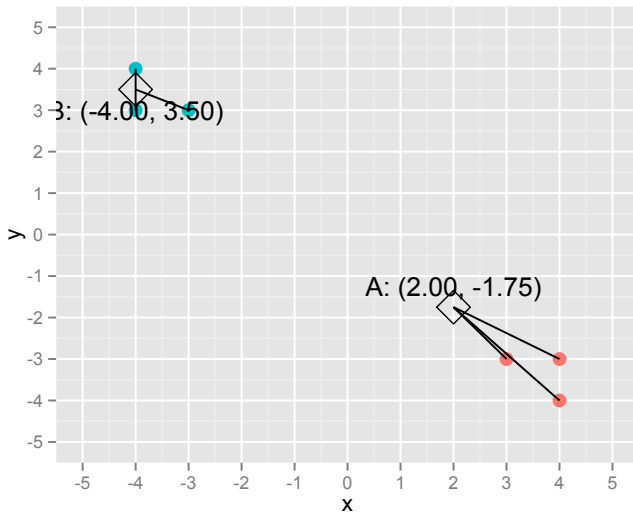
$$\mu_A = \frac{1}{4} ((-3, 3) + (3, -3) + (4, -3) + (4, -4))$$

$$= (2, -1.75)$$

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

$$= (-4, 3.5)$$

Two Points



Two Points

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

=

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

=

Two Points

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

$$= (3.67, -3.33)$$

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

$$=$$

Two Points

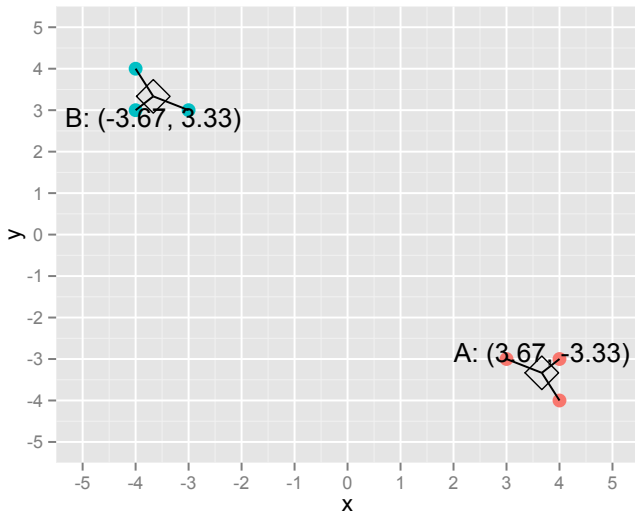
$$\mu_A = \frac{(3, -3) + (4, -3) + (4, -4)}{3}$$

$$= (3.67, -3.33)$$

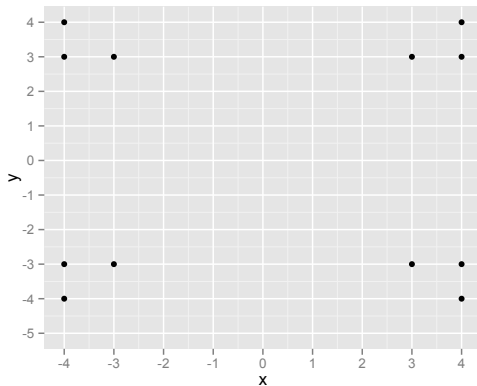
$$\mu_B = \frac{(-4, 3) + (-4, 4) + (-3, 3)}{3}$$

$$= (-3.67, 3.33)$$

Two Points



Four Points



μ_A	μ_B	μ_C	μ_D
$(-3, 3)$	$(-4, 3)$	$(3, -3)$	$(4, -3)$

Four Points

The observation at $(3,3)$ is the same distance from μ_A and μ_C . If you look at Line 10 in the algorithm, the **first** mean with the smallest distance gets the assignment. So $(3,3)$ gets assigned to cluster A .

$$\mu_A =$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

The observation at $(3,3)$ is the same distance from μ_A and μ_C . If you look at Line 10 in the algorithm, the **first** mean with the smallest distance gets the assignment. So $(3,3)$ gets assigned to cluster A.

$$\mu_A = (-1, 1)$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

The observation at $(3,3)$ is the same distance from μ_A and μ_C . If you look at Line 10 in the algorithm, the **first** mean with the smallest distance gets the assignment. So $(3,3)$ gets assigned to cluster A .

$$\mu_A = (-1, 1)$$

$$\mu_B = (-4, 0)$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

The observation at $(3,3)$ is the same distance from μ_A and μ_C . If you look at Line 10 in the algorithm, the **first** mean with the smallest distance gets the assignment. So $(3,3)$ gets assigned to cluster A .

$$\mu_A = (-1, 1)$$

$$\mu_B = (-4, 0)$$

$$\mu_C = (3, -3)$$

$$\mu_D =$$

Four Points

The observation at $(3,3)$ is the same distance from μ_A and μ_C . If you look at Line 10 in the algorithm, the **first** mean with the smallest distance gets the assignment. So $(3,3)$ gets assigned to cluster A .

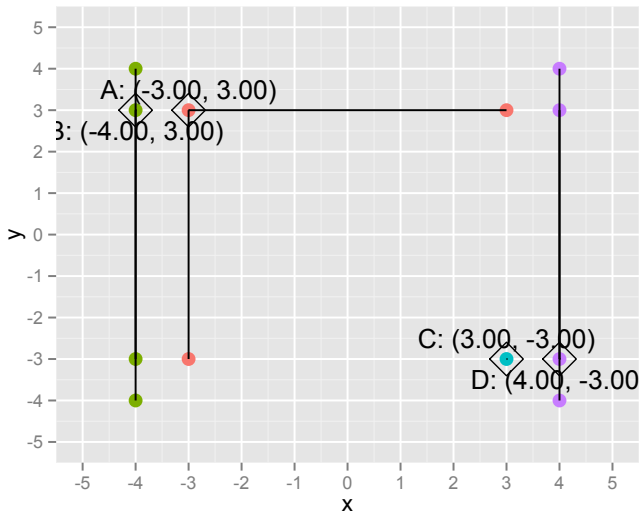
$$\mu_A = (-1, 1)$$

$$\mu_B = (-4, 0)$$

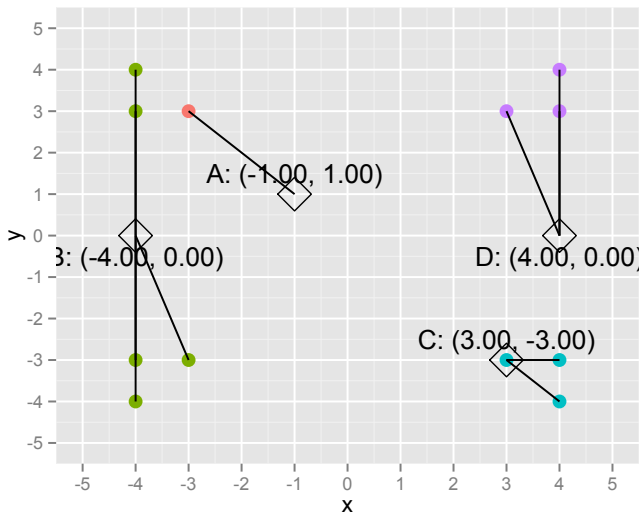
$$\mu_C = (3, -3)$$

$$\mu_D = (4, 0)$$

Four Points



Four Points



Four Points

$$\mu_A =$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

$$\mu_A = (-3, 3)$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

$$\mu_A = (-3, 3)$$

$$\mu_B = (-3.8, -0.6)$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

$$\mu_A = (-3, 3)$$

$$\mu_B = (-3.8, -0.6)$$

$$\mu_C = (3.67, -3.33)$$

$$\mu_D =$$

Four Points

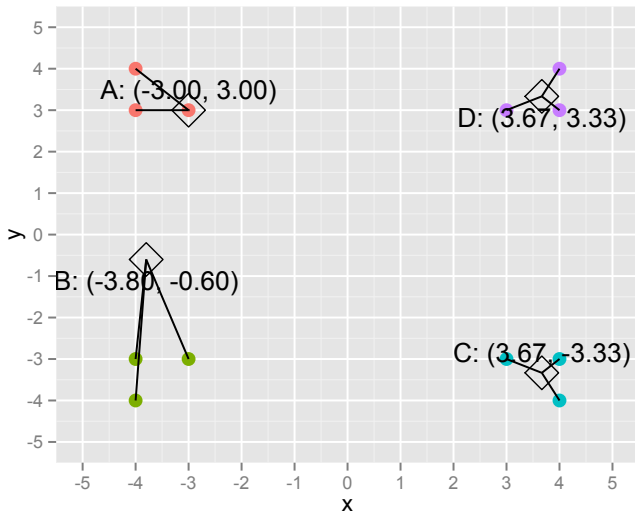
$$\mu_A = (-3, 3)$$

$$\mu_B = (-3.8, -0.6)$$

$$\mu_C = (3.67, -3.33)$$

$$\mu_D = (3.67, 3.33)$$

Four Points



Four Points

$$\mu_A =$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

$$\mu_A = (-3.67, 3.33)$$

$$\mu_B =$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

$$\mu_A = (-3.67, 3.33)$$

$$\mu_B = (-3.67, -3.33)$$

$$\mu_C =$$

$$\mu_D =$$

Four Points

$$\mu_A = (-3.67, 3.33)$$

$$\mu_B = (-3.67, -3.33)$$

$$\mu_C = (3.67, -3.33)$$

$$\mu_D =$$

Four Points

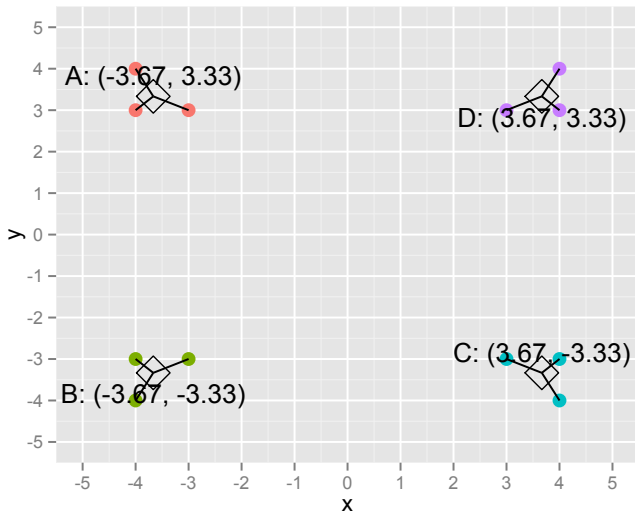
$$\mu_A = (-3.67, 3.33)$$

$$\mu_B = (-3.67, -3.33)$$

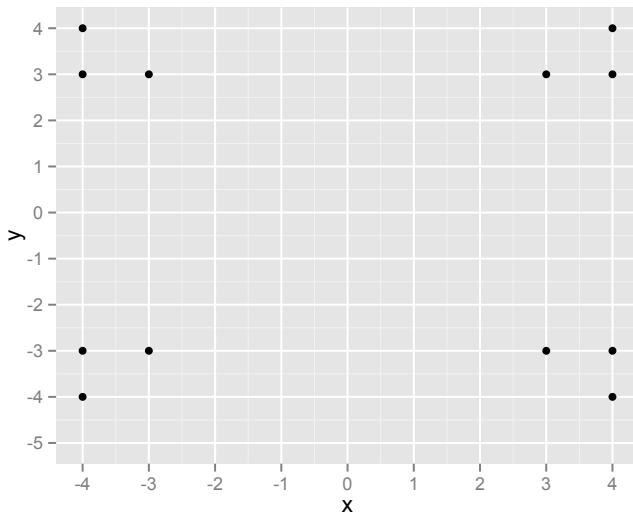
$$\mu_C = (3.67, -3.33)$$

$$\mu_D = (3.67, 3.33)$$

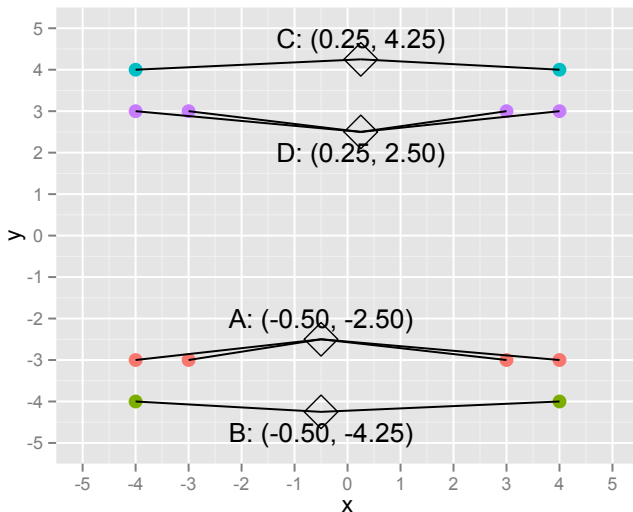
Four Points



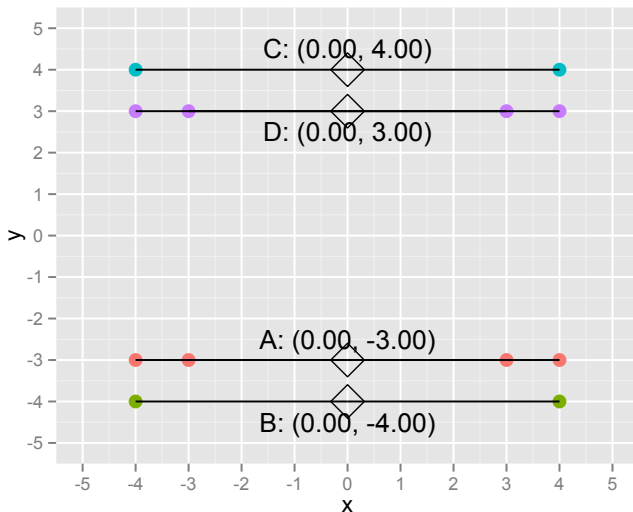
Bad Initialization



Bad Initialization



Bad Initialization



How does it change for GMM?

How does it change for GMM?

Instead of just computing mean, you also compute variance.