Clustering Lab

Data Science: Jordan Boyd-Graber
University of Maryland
APRIL 2, 2018

Clustering Lab

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

k-means

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

Score

procedure Score(X, Z, M) ▹ Current objective function

$s \leftarrow 0$

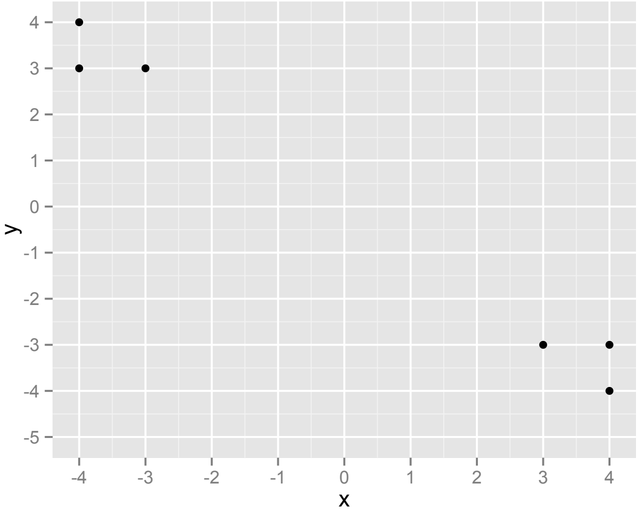
for $i \in \{1, \ldots, N\}$ **do** ▹ For each observation

$s \leftarrow s + ||x_i - \mu_{z_i}||$ ▹ Accumulate how far it is from its mean

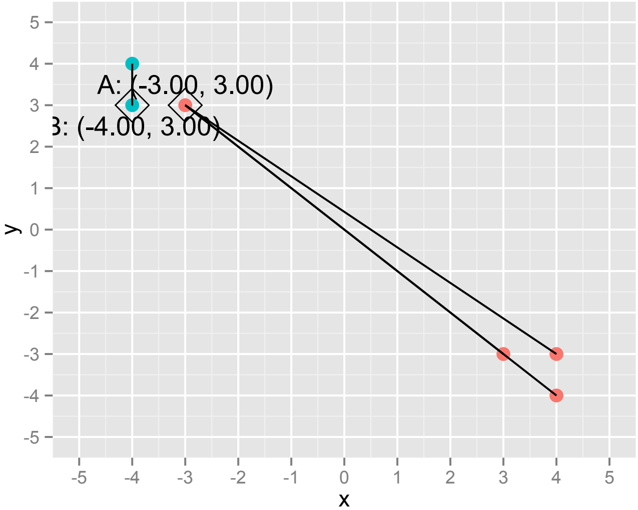
Find closest

procedure AssignToClosestCluster(X, M)
  $Z \leftarrow$ Vector(N) ▷ Initialize assignments Z as a N-vector
  for $i \in \{1, \ldots, N\}$ **do** ▷ For each observation
    $d \leftarrow -\infty$
    for $k \in \{1, \ldots, K\}$ **do**
      if $||x_i - \mu_k|| < d$ **then**
        $z_i \leftarrow k$
        $d < -||x_i - \mu_k||$
  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

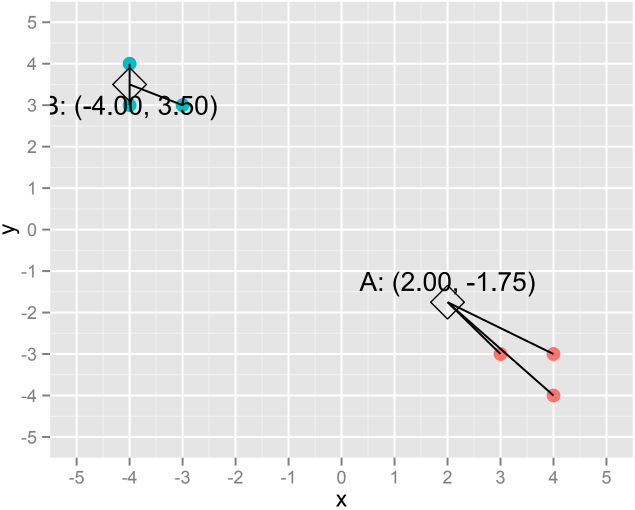
$$\mu_A = \frac{1}{4}((-3,3)+(3,-3)+(4,-3)+(4,-4))$$

$$= (2,-1.75)$$

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

$$= (-4,3.5)$$

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

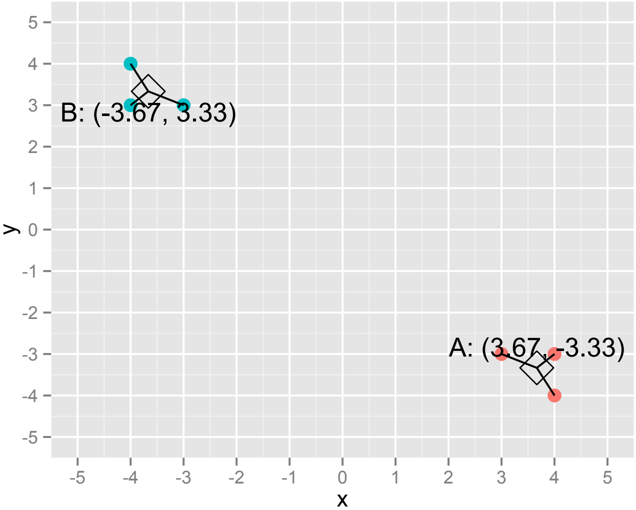
$$\mu_A = \frac{(3,-3)+(4,-3)+(4,-4)}{3}$$

$$= (3.67,-3.33)$$

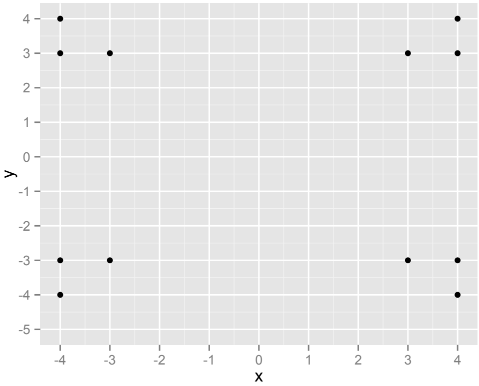
$$\mu_B = \frac{(-4,3)+(-4,4)+(-3,3)}{3}$$

$$= (-3.67,3.33)$$

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.

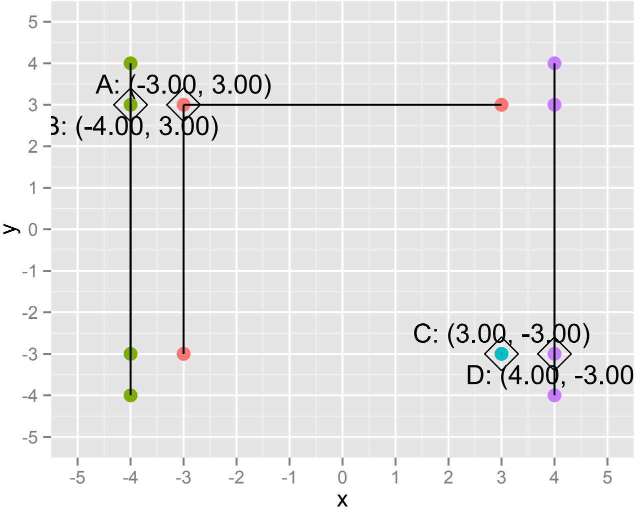
$$\mu_A = (-1,1)$$
$$\mu_B = (-4,0)$$
$$\mu_C = (3,-3)$$
$$\mu_D =$$

Four Points

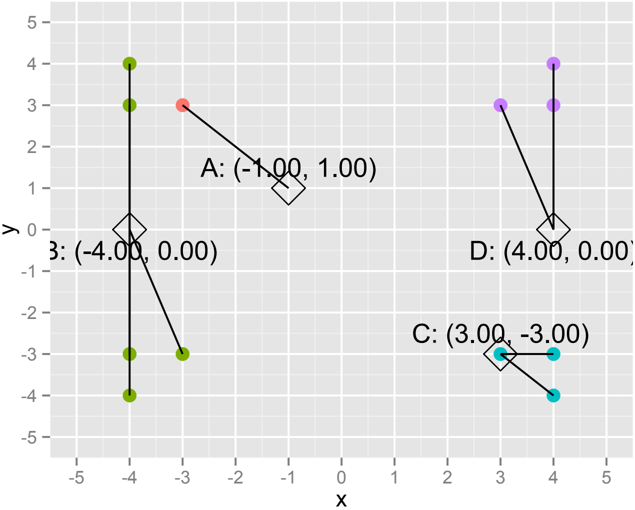
The observation at $(3,3)$ is the same distance from μ_A and μ_C. If you look at Line 10 in the algorithm, the **first** mean with the smallest distance gets the assignment. So $(3,3)$ gets assigned to cluster A.

$$\mu_A = (-1,1)$$
$$\mu_B = (-4,0)$$
$$\mu_C = (3,-3)$$
$$\mu_D = (4,0)$$

Four Points


A: (-3.00, 3.00)
3: (-4.00, 3.00)
C: (3.00, -3.00)
D: (4.00, -3.00
y
x
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

Four Points



A: (-1.00, 1.00)
3: (-4.00, 0.00)
D: (4.00, 0.00
C: (3.00, -3.00)
y
x
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

Four Points

$$\mu_A =$$
$$\mu_B =$$
$$\mu_C =$$
$$\mu_D =$$

Four Points

$$\mu_A = (-3, 3)$$
$$\mu_B =$$
$$\mu_C =$$
$$\mu_D =$$

Four Points

$$\mu_A = (-3, 3)$$
$$\mu_B = (-3.8, -0.6)$$
$$\mu_C =$$
$$\mu_D =$$

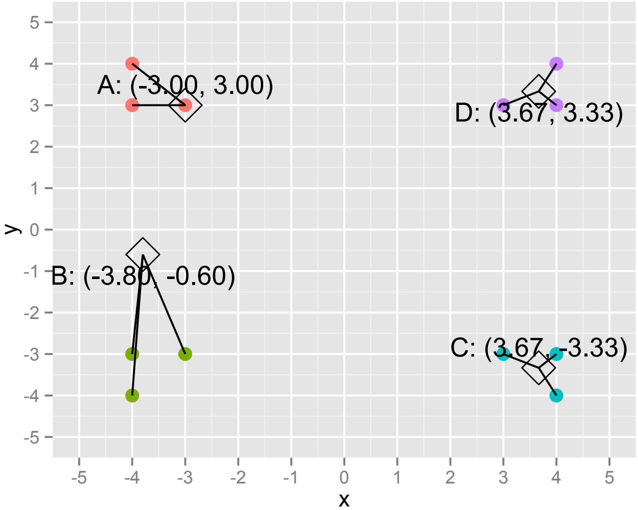
Four Points

$$\mu_A = (-3, 3)$$
$$\mu_B = (-3.8, -0.6)$$
$$\mu_C = (3.67, -3.33)$$
$$\mu_D =$$

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



A: (-3.00, 3.00)
D: (3.67, 3.33)
B: (-3.80, -0.60)
C: (3.67, -3.33)
y
x
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

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

$$\mu_A = (-3.67, 3.33)$$
$$\mu_B = (-3.67, -3.33)$$
$$\mu_C =$$
$$\mu_D =$$

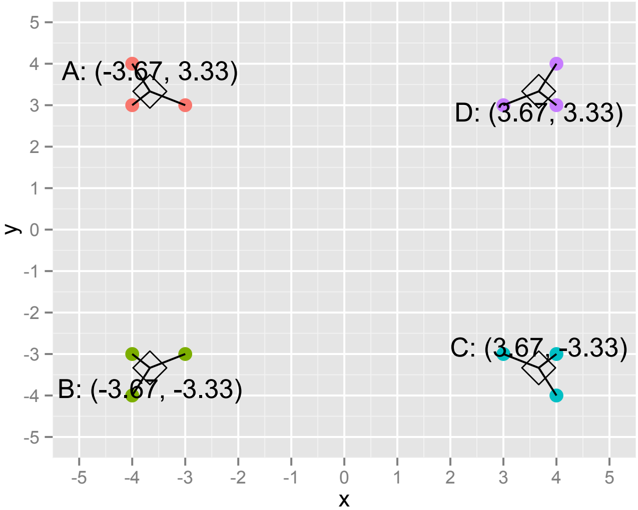
Four Points

$$\mu_A = (-3.67, 3.33)$$
$$\mu_B = (-3.67, -3.33)$$
$$\mu_C = (3.67, -3.33)$$
$$\mu_D =$$

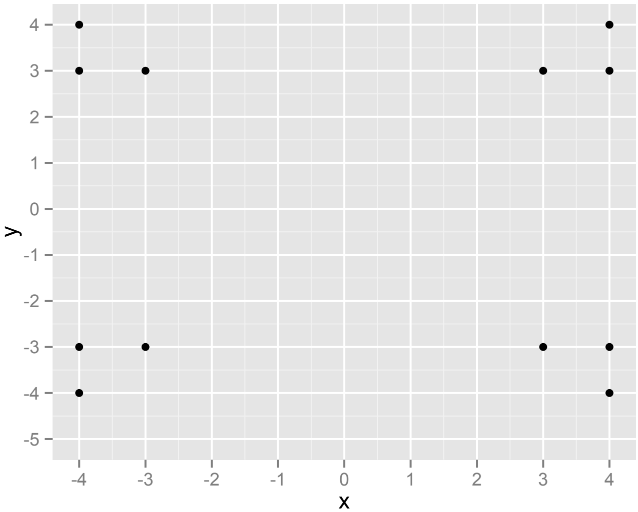
Four Points

$$\mu_A = (-3.67, 3.33)$$
$$\mu_B = (-3.67, -3.33)$$
$$\mu_C = (3.67, -3.33)$$
$$\mu_D = (3.67, 3.33)$$

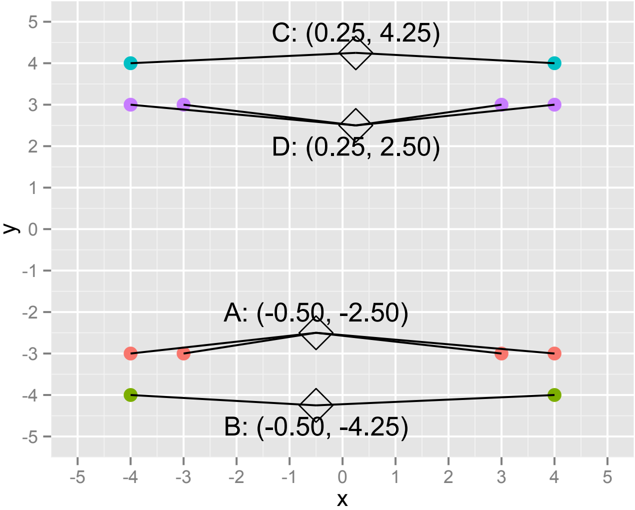
Four Points


A: (-3.67, 3.33)
D: (3.67, 3.33)
C: (3.67, -3.33)
B: (-3.67, -3.33)
y
x
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

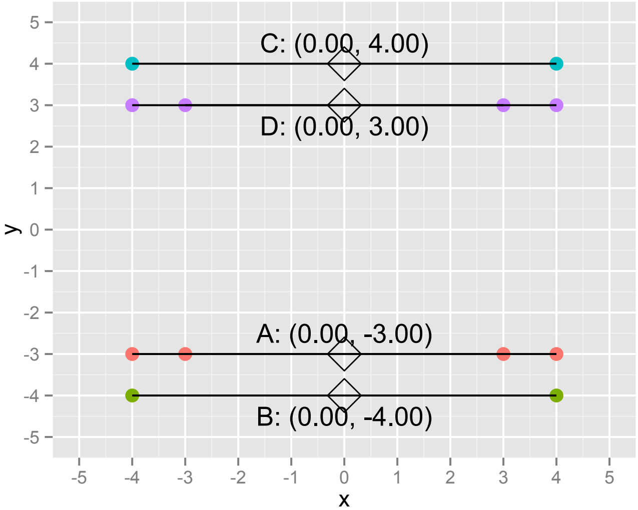
Bad Initialization



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.