DSC 40A

Theoretical Foundations of Data Science I

Announcements

- Homework 7 due 12/6. - no slip day
- SET (currently = 50% < 80%)
- Review session on Friday 4-6pm

* Rebecca + Owen OH on Friday 6-7pm

* OH on Saturday, Monday, Tuesday will be announced on Ed

Question

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!

If the direct link doesn't work, click the "Lecture Questions" link in the top right corner of dsc40a.com.

Lloyds Algorithm, or k-Means Clustering

1. Randomly initialize the k centroids.
2. Keep centroids fixed. Update groups.

   Assign each point to the nearest centroid.

3. Keep groups fixed. Update centroids.

   Move each centroid to the center of its group.

4. Repeat steps 2 and 3 until done.

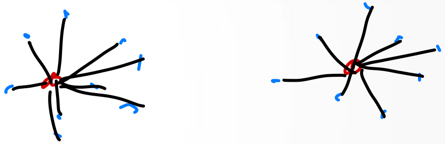
iterate

Outline

- Why does k-means clustering work?
- What are some practical considerations when using this algorithm?

Why does k-means clustering work?

Cost($\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, ... , $\boldsymbol{\mu}_k$) = total squared distance of each data point x_i to its nearest centroid $\boldsymbol{\mu}_j$



- Argue why updating the groups and centroids according to the algorithm reduces the cost with each iteration.
- With enough iterations, cost will be sufficiently small.

Why does k-means clustering work?

$\text{Cost}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k) =$ total squared distance of each data point x_i to its nearest centroid $\boldsymbol{\mu}_j$

$$\text{Cost} = \sum_{j=1}^{k} \sum_{x_i \in C_j} (\vec{x}_i - \vec{\mu}_j)^2 \qquad \vec{x}_i, \vec{\mu}_j \in \mathbb{R}^d$$

$x_i \in C_j$ → points $\vec{x}_i$ that belong to cluster j

$\text{Cost}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k) = \text{Cost}(\boldsymbol{\mu}_1) + \text{Cost}(\boldsymbol{\mu}_2) + \dots + \text{Cost}(\boldsymbol{\mu}_k)$ where

$\text{Cost}(\boldsymbol{\mu}_j) =$ total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

Why does k-means clustering work?

$\text{Cost}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k) = \text{Cost}(\boldsymbol{\mu}_1) + \text{Cost}(\boldsymbol{\mu}_2) + \ldots + \text{Cost}(\boldsymbol{\mu}_k)$ where

$\text{Cost}(\boldsymbol{\mu}_j) =$ total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

1. Randomly initialize the k centroids. ← sets initial cost (before the process begins)

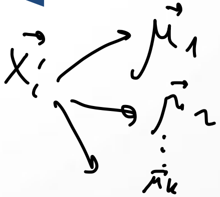
Why does k-means clustering work?

$\text{Cost}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k) = \text{Cost}(\boldsymbol{\mu}_1) + \text{Cost}(\boldsymbol{\mu}_2) + \ldots + \text{Cost}(\boldsymbol{\mu}_k)$ where

$\text{Cost}(\boldsymbol{\mu}_j) =$ total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

2. Fix the centroids. Update the groups.

consider an arbitrary iteration



Certainly $\text{Cost}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k)$ decreases in this step because assigning each point to the **closest** centroid is best.

Why does k-means clustering work?

$\text{Cost}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k) = \text{Cost}(\boldsymbol{\mu}_1) + \text{Cost}(\boldsymbol{\mu}_2) + \ldots + \text{Cost}(\boldsymbol{\mu}_k)$ where

$\text{Cost}(\boldsymbol{\mu}_j) =$ total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

3. Fix the groups. Update the centroids. ← consider an arbitrary iteration

Argue that $\text{Cost}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k)$ decreases in this step because for each group j, $\text{Cost}(\boldsymbol{\mu}_j)$ is minimized when we update the centroid.

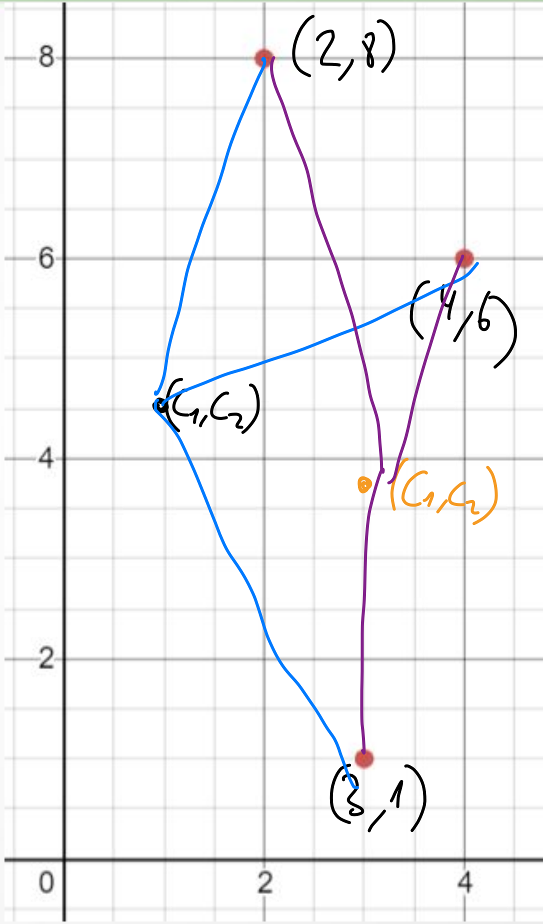
Why does k-means clustering work?

Cost($\boldsymbol{\mu}_j$) = total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

Example: group j contains (4, 6), (2, 8), (3,1) $\in \mathbb{R}^2$

How to place centroid $\vec{\boldsymbol{\mu}}_j = (c_1, c_2)$ to minimize cost?

$$\text{Cost}(\mu_j) = \left(\sqrt{(c_1-4)^2+(c_2-6)^2}\right)^2 + \left(\sqrt{(c_1-2)^2+(c_2-8)^2}\right)^2 + \left(\sqrt{(c_1-3)^2+(c_2-1)^2}\right)^2$$

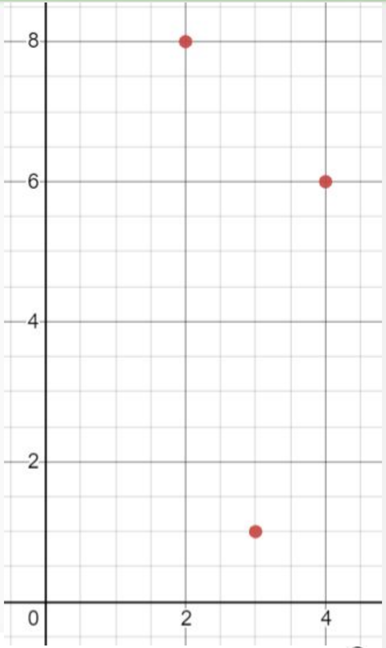


Why does k-means clustering work?

Cost($\boldsymbol{\mu}_j$) = total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

Example: group j contains (4, 6), (2, 8), (3,1)

How to place centroid $\boldsymbol{\mu}_j$ = (c_1, c_2) to minimize cost?



$$\text{Cost}\,(\vec{\mu_j}) = \left(\sqrt{(4-c_1)^2+(6-c_2)^2}\right)^2 + \left(\sqrt{(2-c_1)^2+(8-c_2)^2}\right)^2 + \left(\sqrt{(3-c_1)^2+(1-c_2)^2}\right)^2$$
$$= (4-c_1)^2+(6-c_2)^2+(2-c_1)^2+(8-c_2)^2+(3-c_1)^2+(1-c_2)^2 = f(c_1,c_2)$$

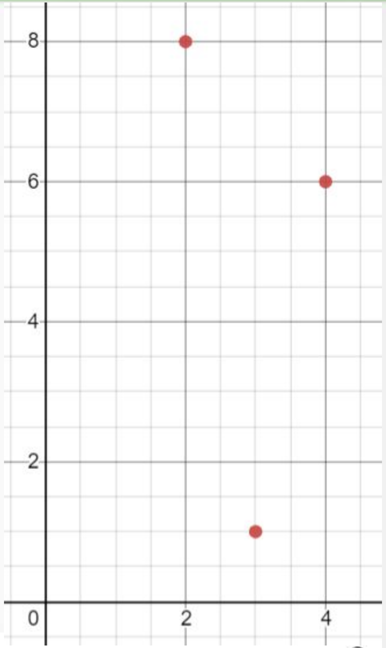
$$\frac{\partial \text{cost}}{\partial c_1} = -2(4-c_1) - 2(2-c_1) - 2(3-c_1) = -8-4-6+6c_1 = -18+6c_1$$

Why does k-means clustering work?

Cost($\boldsymbol{\mu}_j$) = total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

Example: group j contains (4, 6), (2, 8), (3,1)

How to place centroid $\boldsymbol{\mu}_j = (c_1, c_2)$ to minimize cost?



$$\text{Cost}\,(\mu_j) = \left(\sqrt{(4-c_1)^2+(6-c_2)^2}\right)^2 + \left(\sqrt{(2-c_1)^2+(8-c_2)^2}\right)^2 + \left(\sqrt{(3-c_1)^2+(1-c_2)^2}\right)^2$$
$$= (4-c_1)^2+(6-c_2)^2+(2-c_1)^2+(8-c_2)^2+(3-c_1)^2+(1-c_2)^2$$

$$\frac{\partial \text{Cost}(\mu_j)}{\partial c_1} = 2(c_1-4)+2(c_1-2)+2(c_1-3)$$

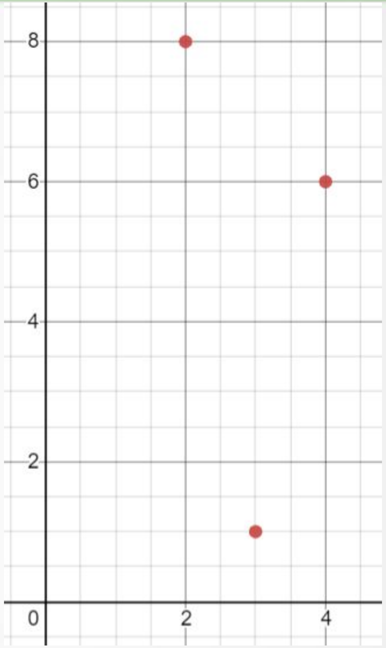
$$\frac{\partial \text{Cost}(\mu_j)}{\partial c_2} = 2(c_2-6)+2(c_2-8)+2(c_2-1)$$

Why does k-means clustering work?

Cost($\boldsymbol{\mu}_j$) = total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

Example: group j contains (4, 6), (2, 8), (3,1)

How to place centroid $\boldsymbol{\mu}_j = (c_1, c_2)$ to minimize cost?



find critical point: set equal to zero

$$\frac{\partial \text{Cost}(\mu_j)}{\partial c_1} = 2(c_1 - 4) + 2(c_1 - 2) + 2(c_1 - 3)$$

$$0 = 2(c_1 - 4) + 2(c_1 - 2) + 2(c_1 - 3)$$

$$0 = c_1 - 4 + c_1 - 2 + c_1 - 3$$

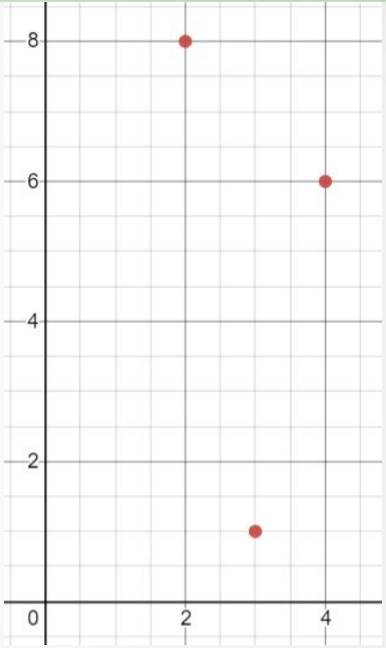
$$3c_1 = 4 + 2 + 3$$

$$c_1 = \frac{4+2+3}{3} = \frac{9}{3} = 3$$

⟹ average of 1st coord of data points

Why does k-means clustering work?

Cost($\boldsymbol{\mu}_j$) = total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$



Example: group j contains (4, 6), (2, 8), (3,1)

How to place centroid $\boldsymbol{\mu}_j = (c_1, c_2)$ to minimize cost?

$$\frac{\partial \text{Cost}(\mu_j)}{\partial c_2} = 2(c_2 - 6) + 2(c_2 - 8) + 2(c_2 - 1) = 0$$

$$0 = 2(c_2 - 6) + 2(c_2 - 8) + 2(c_2 - 1)$$

$$0 = c_2 - 6 + c_2 - 8 + c_2 - 1$$

$$3c_2 = 6 + 8 + 1$$

$$c_2 = \frac{6 + 8 + 1}{3} = \frac{15}{3} = 5$$

⟹ average of 2nd coord of datapoints

Why does k-means clustering work?

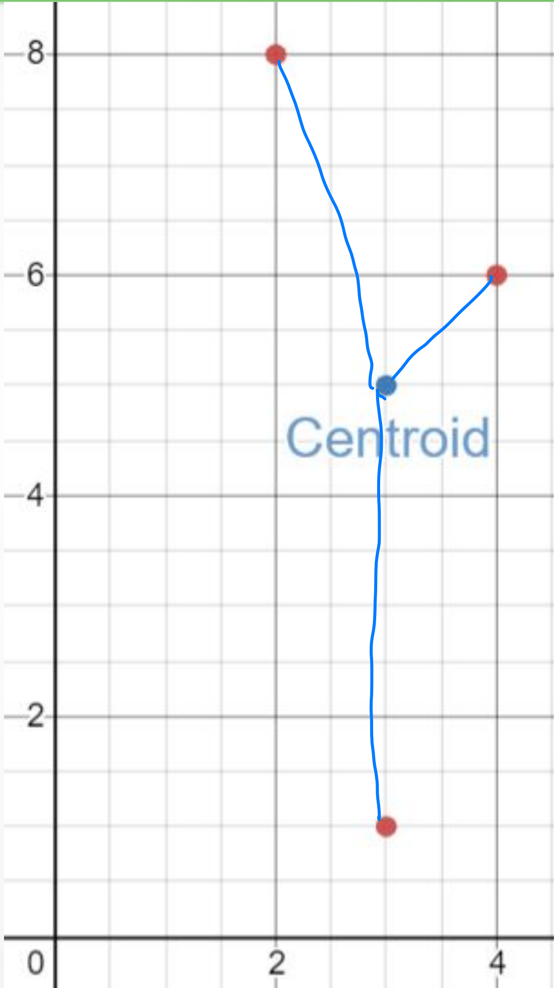
Cost($\boldsymbol{\mu}_j$) = total squared distance of each data point x_i in group j to centroid $\boldsymbol{\mu}_j$

Example: group j contains (4, 6), (2, 8), (3,1)

How to place centroid $\boldsymbol{\mu}_j = (c_1, c_2)$ to minimize cost?

$$(c_1, c_2) = \left(\frac{4+2+3}{3}, \frac{6+8+1}{3}\right) = (3, 5)$$

Minimize cost by averaging in each coordinate.



Cost, Loss, and Risk

The cost of placing the centroid at (c_1, c_2) is

$$\text{Cost}\,(\vec{\mu_j}) = \left(\sqrt{(4-c_1)^2+(6-c_2)^2}\right)^2 + \left(\sqrt{(2-c_1)^2+(8-c_2)^2}\right)^2 + \left(\sqrt{(3-c_1)^2+(1-c_2)^2}\right)^2$$
$$= (4-c_1)^2+(6-c_2)^2+(2-c_1)^2+(8-c_2)^2+(3-c_1)^2+(1-c_2)^2$$

$$\text{cost}(\vec{\mu_j}) = \underbrace{(4-c_1)^2+(2-c_1)^2+(3-c_1)^2}_{f(c_1)} + \underbrace{(6-c_2)^2+(8-c_2)^2+(1-c_2)^2}_{g(c_2)}$$

$$f(c_1) = (4-c_1)^2+(2-c_1)^2+(3-c_1)^2 \qquad \frac{df}{dc_1}=0$$

MSE for constant model h?

$$R_{sq}(h) = \frac{1}{3}\left((4-h)^2+(2-h)^2+(3-h)^2\right) \longrightarrow \frac{dR_{sq}(h)}{dh}=0$$

minimizer is mean of $\{4,2,3\}$

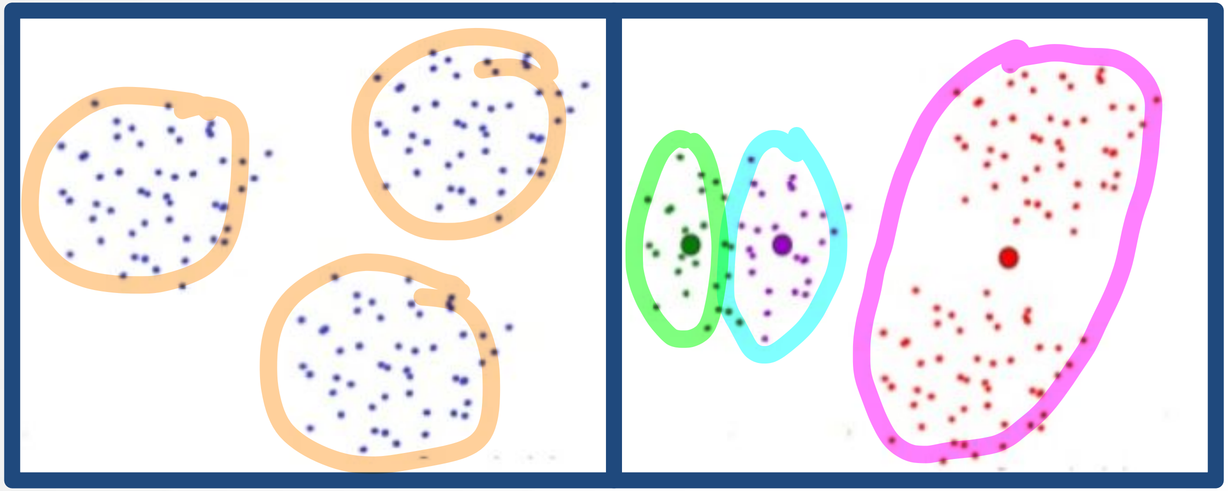
Why does k-means clustering work?

Cost($\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, ... , $\boldsymbol{\mu}_k$) = total squared distance of each data point x_i to its nearest centroid $\boldsymbol{\mu}_j$

- Argue why updating the groups and centroids according to the algorithm reduces the cost with each iteration.
- With enough iterations, cost will be sufficiently small.

k-Means Clustering in Practice: Initialization

Can get unlucky with random initialization.



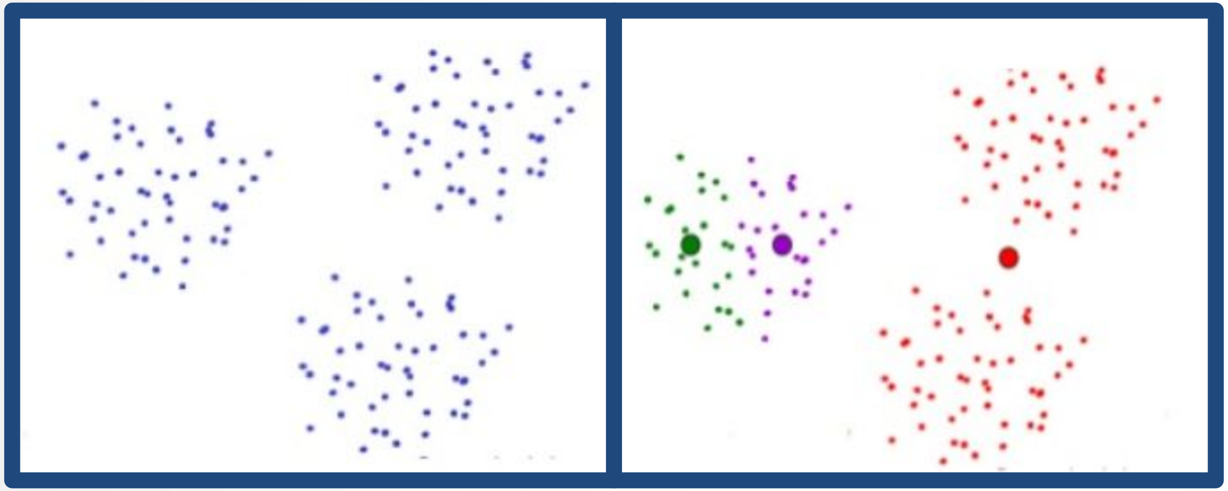
In general, how do we assess which result is the best?

A. Clusters appear how we expect them to
B. Clusters are evenly sized
C. Cost function is lowest

cost func for these centroids will be higher

k-Means Clustering in Practice: Initialization

Can get unlucky with random initialization.



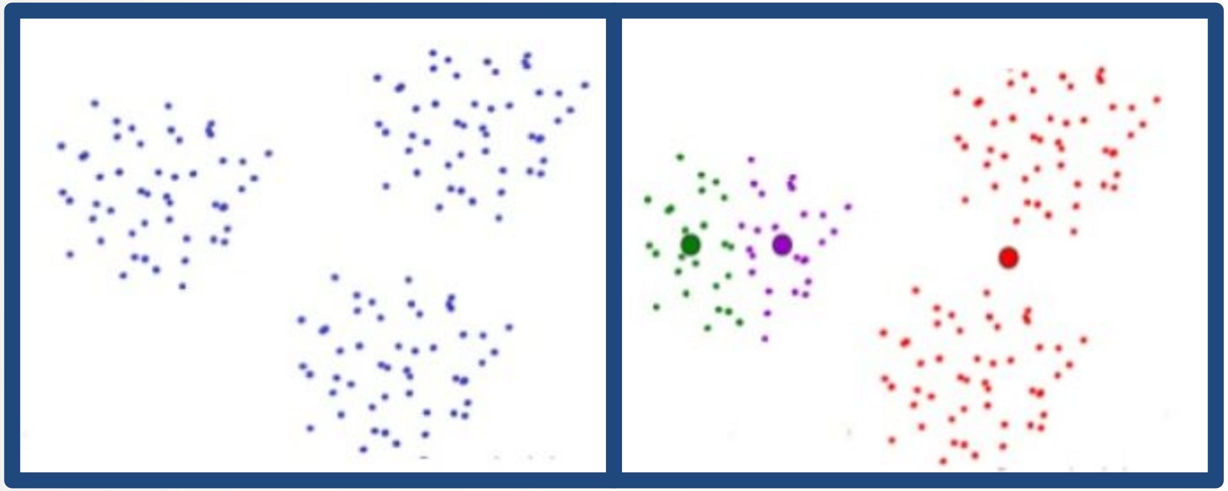
In general, how do we assess which result is the best?

A. Clusters appear how we expect them to
B. Clusters are evenly sized
C. Cost function is lowest

Solution?
- Try algorithm several times, pick the best result.
- Similar approach used in gradient descent.

k-Means Clustering in Practice: Initialization

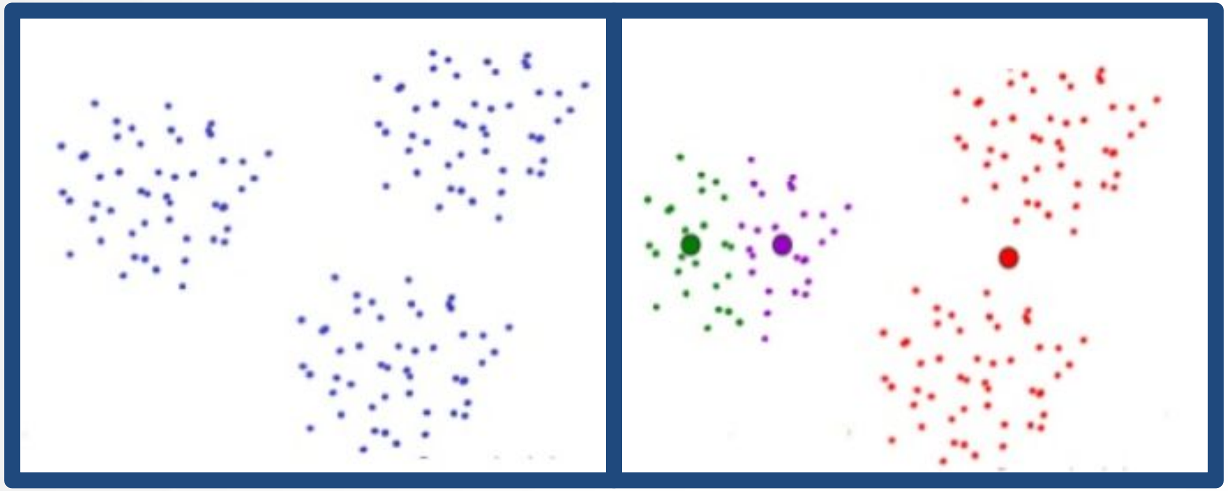
Can get unlucky with random initialization.



- No guarantees of a satisfactory solution with this algorithm.
- Brute force algorithm would try all assignments of points to clusters and choose the one with the lowest cost.

k-Means Clustering in Practice: Initialization

Can get unlucky with random initialization.



How many ways to assign n points to k clusters?

k = # colors

k k k ... k k k

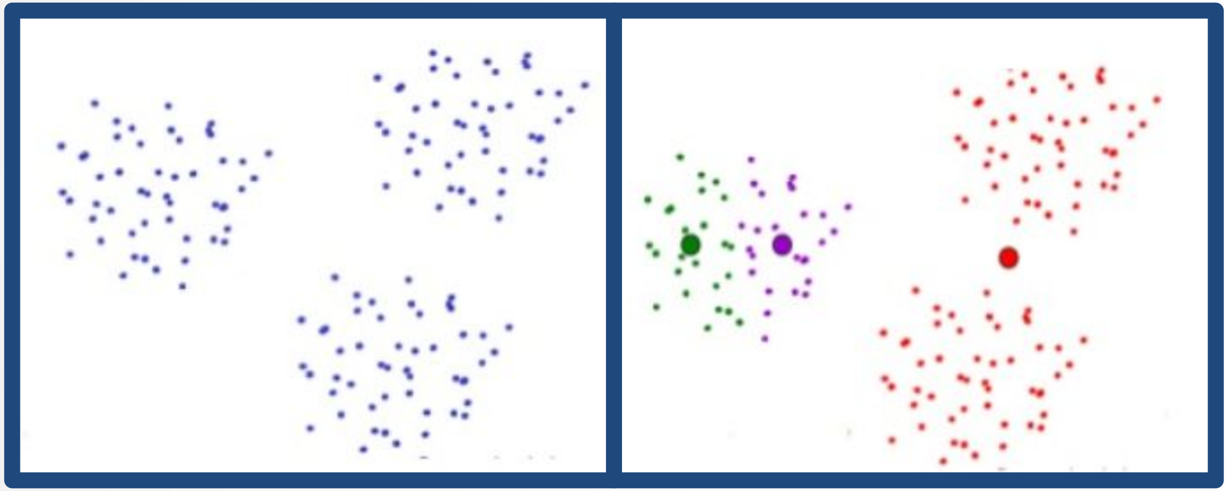
n points

$= k^n$

- No guarantees of a satisfactory solution with this algorithm.
- Brute force algorithm would try all assignments of points to clusters and choose the one with the lowest cost.

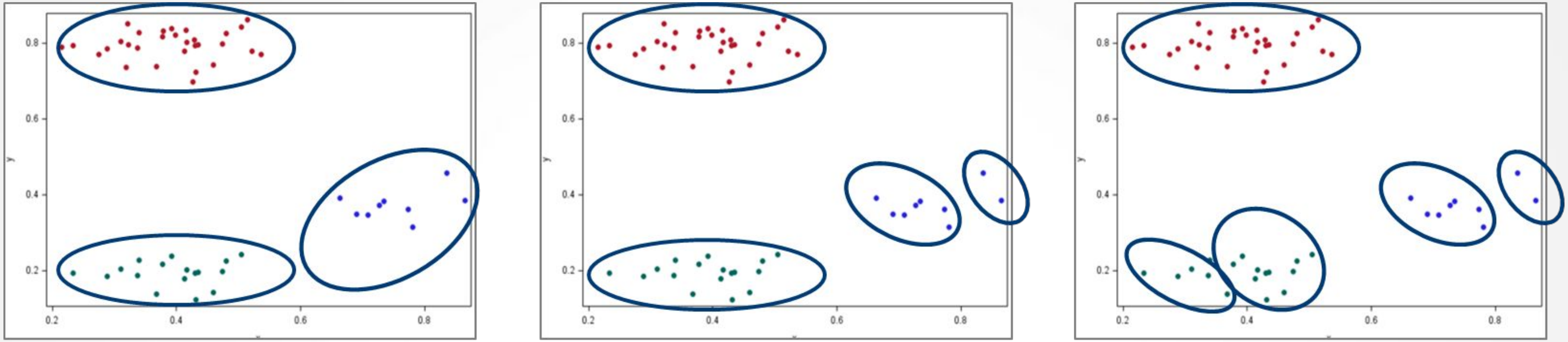
k-Means Clustering in Practice: Initialization

Can get unlucky with random initialization.



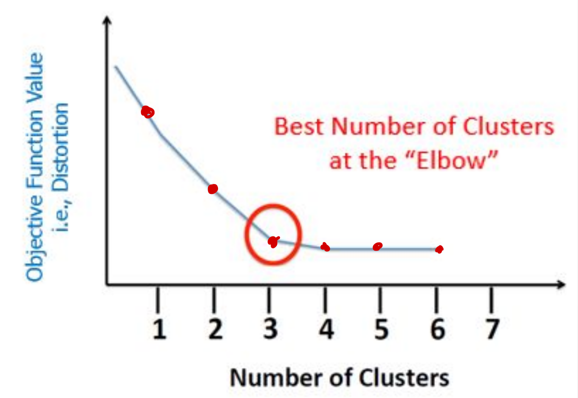
- No guarantees of a satisfactory solution with this algorithm.
- Any algorithm that is guaranteed to find the best coloring of data points takes exponential time (computationally infeasible).

k-Means Clustering in Practice: Choosing k



k=3 k=4 k=5

- Most commonly done by hand (visualizations, trial and error)
- Elbow method
- Context or domain knowledge
- Use a different clustering algorithm



What if a centroid has no points in its group?

What should we do if a centroid has no points in its group?

A. Terminate the algorithm.
B. Wait for points get added to the group in a subsequent iteration.
C. Set the centroid to be a data point, chosen at random.
D. Set the centroid to be one of the other centroids, chosen at random.

What if a centroid has no points in its group?

What should we do if a centroid has no points in its group?

A. Terminate the algorithm.
B. Wait for points get added to the group in a subsequent iteration.
C. Set the centroid to be a data point, chosen at random.
D. Set the centroid to be one of the other centroids, chosen at random.

Two options:

- Eliminate that centroid and find k-1 clusters instead
- Randomly re-initialize that centroid

Summary

- We saw that k-means clustering works because each step of the algorithm reduces the cost function, which measures the quality of a set of centroids.
- We discussed some practical considerations, including random initialization and choice of k.

Better initialization: kmeans++