Clustering

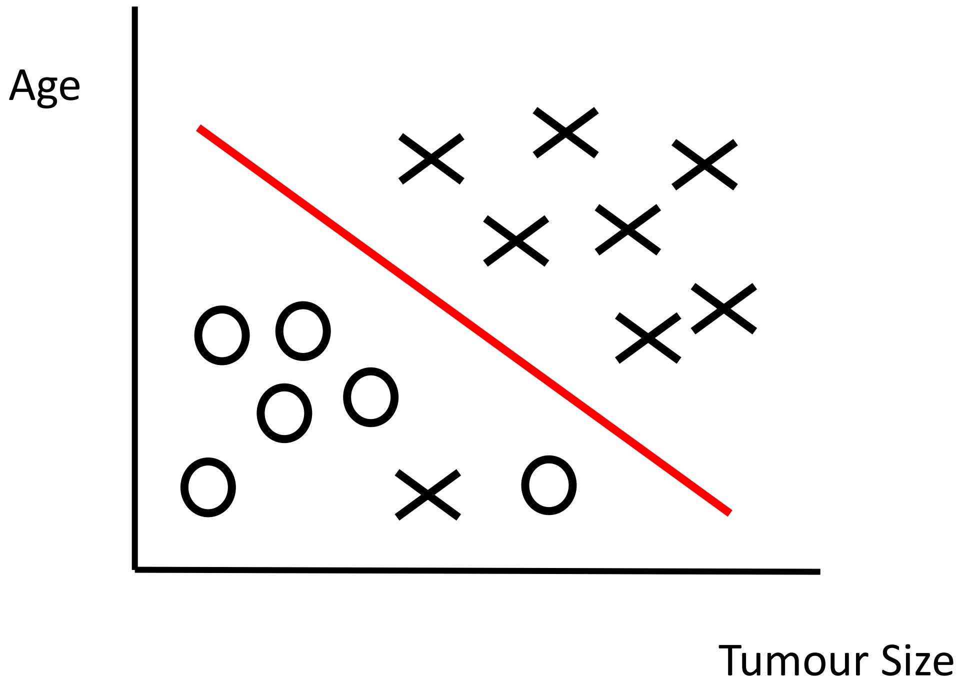
Prof. Dr. Christina Bauer

christina.bauer@th-deg.de

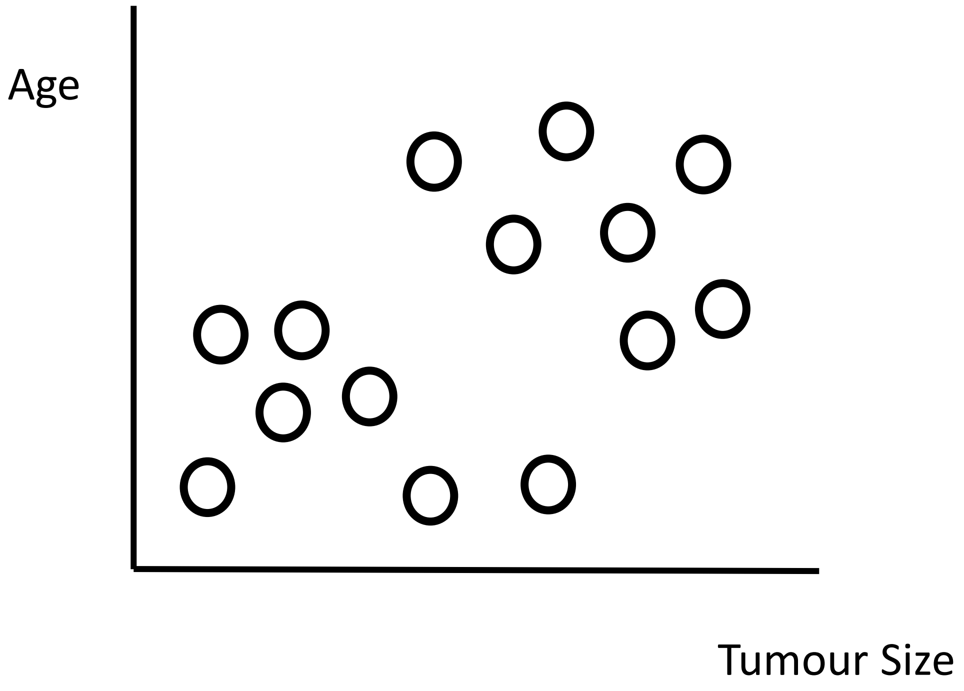
Faculty of Computer Science

Unsupervised Learning

supervised

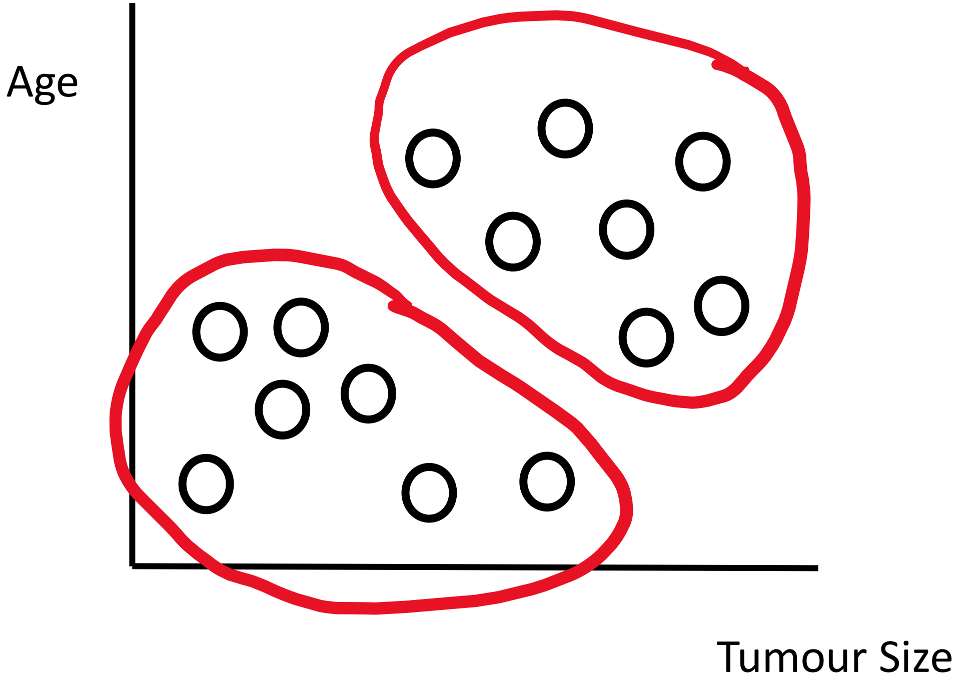


unsupervised



Training set $\{(x^1,y^1),\ldots,(x^m,y^m)\}$

Unsupervised Learning



Clustering algorithm

Unsupervised Learning

Hotel quarantine: 'It'll cost us thousands and we'll be miles from home'

BBC News · 2 hours ago

- **Inside a quarantine hotel on Heathrow's 'Isolation Row'**
  The Independent · 3 hours ago
- **Coronavirus in the UK: Quarantine loophole still exists just hours before hotel policy begins, says Michael Matheson**
  Edinburgh News · 22 hours ago
- **Hotel quarantine is another example of too little too late – it's all up to immigration officials now**
  The Independent · Yesterday · Opinion
- **Covid vaccine rollout 'an unbelievable effort' - Johnson**
  BBC News · 15 hours ago

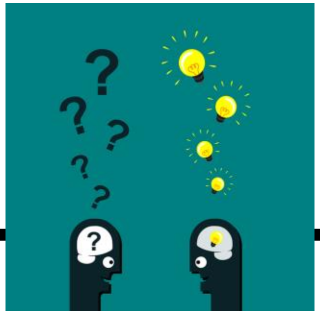
View Full coverage

news.google.com/

Unsupervised Learning

- Data centre management: Organize computer cluster
- Social science: Social network analysis
- Market segmentation – based on costumer data
- Astronomical data analysis – e. g. “how are galaxies formed”

QUESTION



Which of the following statements are true? Check all that apply.

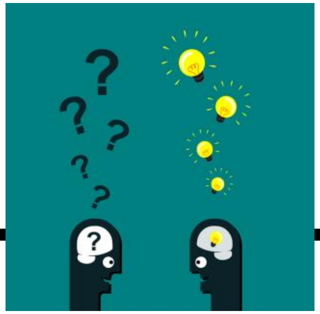
A: In unsupervised learning, the training set is of the form $\{x^{(1)},x^{(2)},\ldots,x^{(m)}\}$ without labels $y^{(i)}$.

B: Clustering is an example of unsupervised learning.

C: In unsupervised learning, you are given an unlabeled dataset and are asked to find "structure" in the data.

D: Clustering is the only unsupervised learning algorithm.

Question



Which of the following statements are true? Check all that apply.

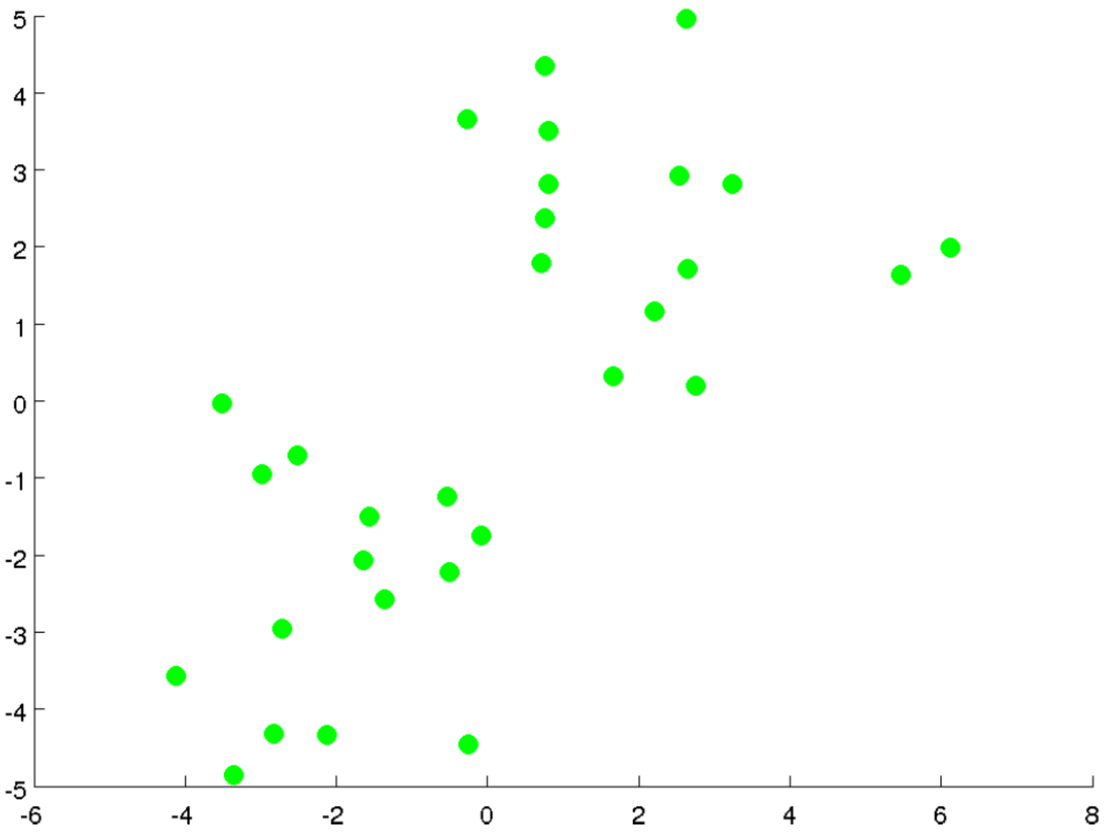
A: In unsupervised learning, the training set is of the form $\{x^{(1)},x^{(2)},\ldots,x^{(m)}\}$ without labels $y^{(i)}$.

B: Clustering is an example of unsupervised learning.

C: In unsupervised learning, you are given an unlabeled dataset and are asked to find "structure" in the data.

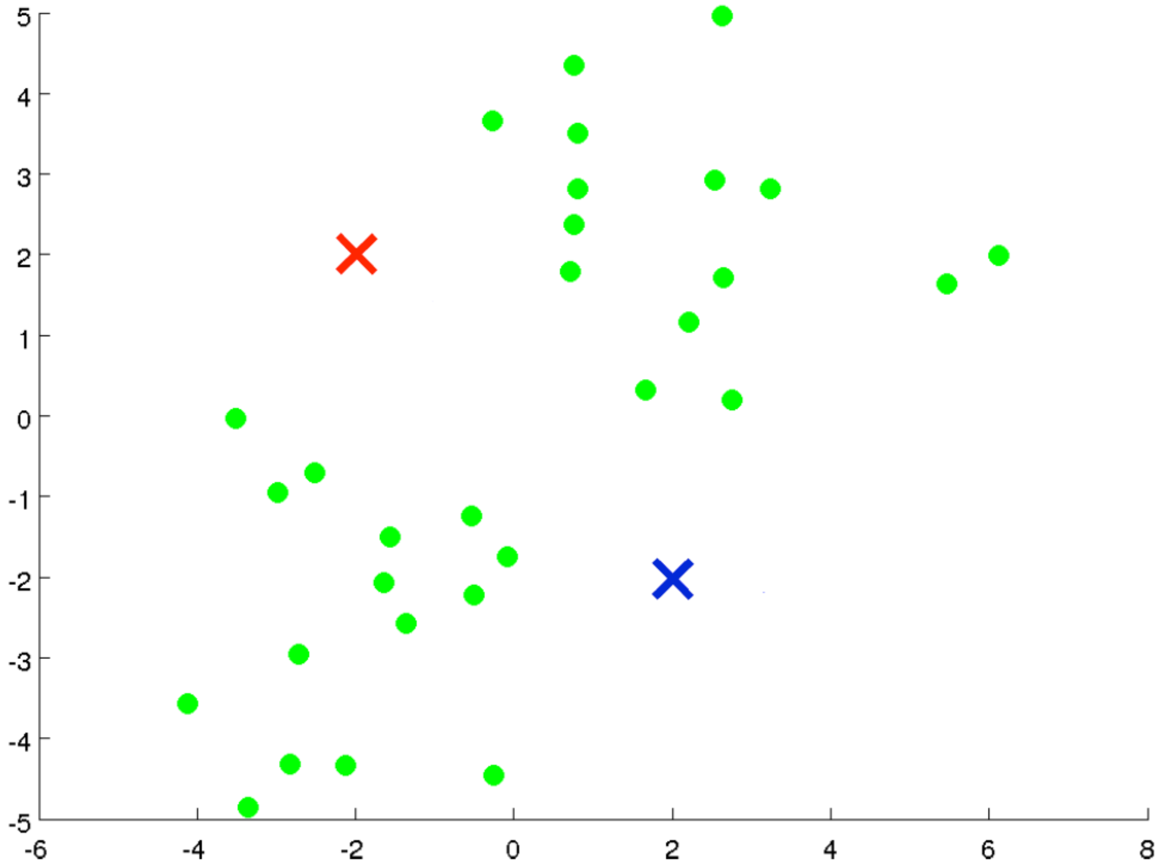
D: Clustering is the only unsupervised learning algorithm.

K-MEANS

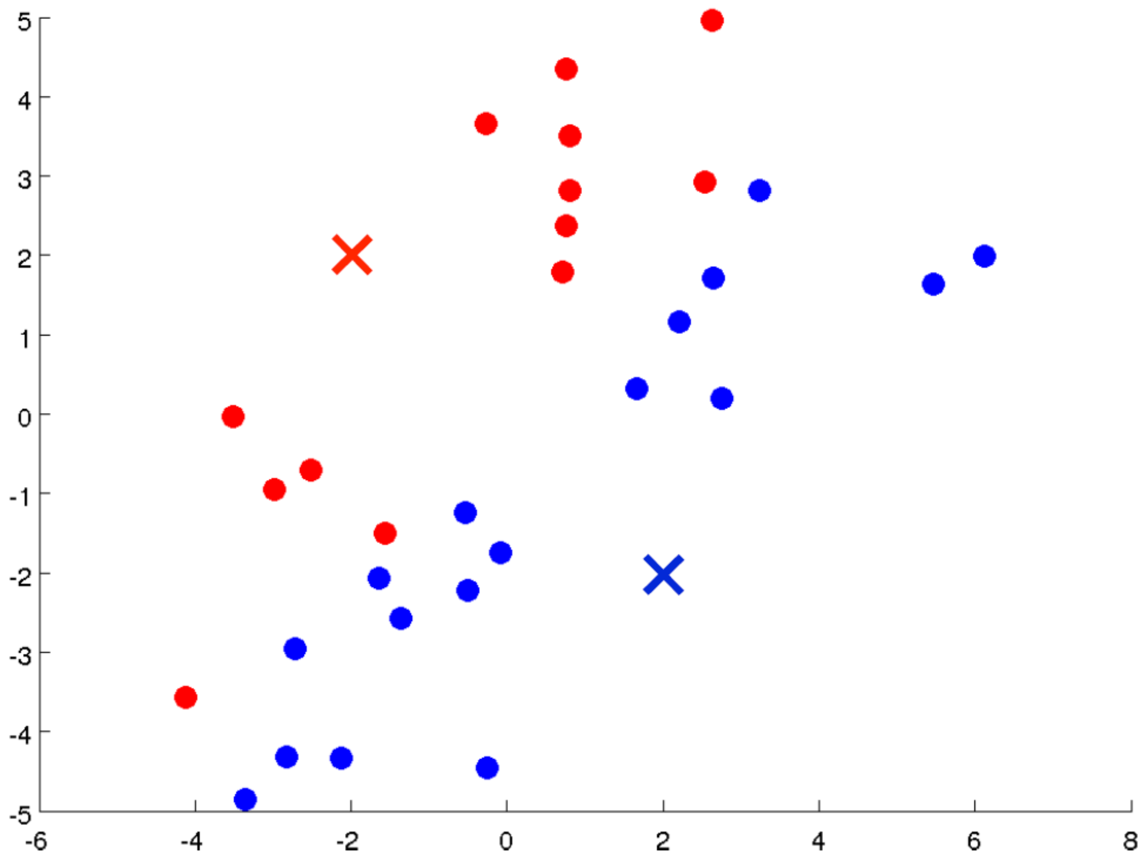


Plots by Andrew Ng

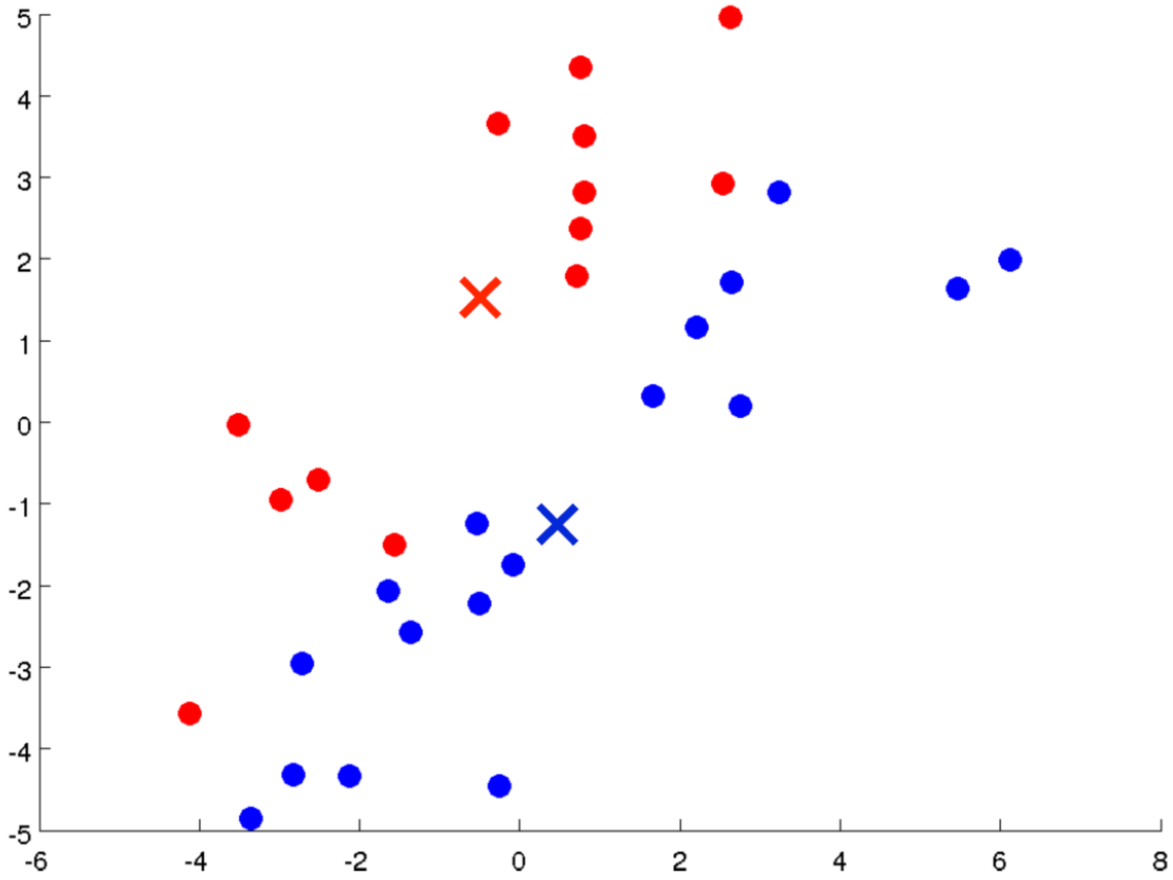
K-MEANS – CLUSTER CENTROIDS



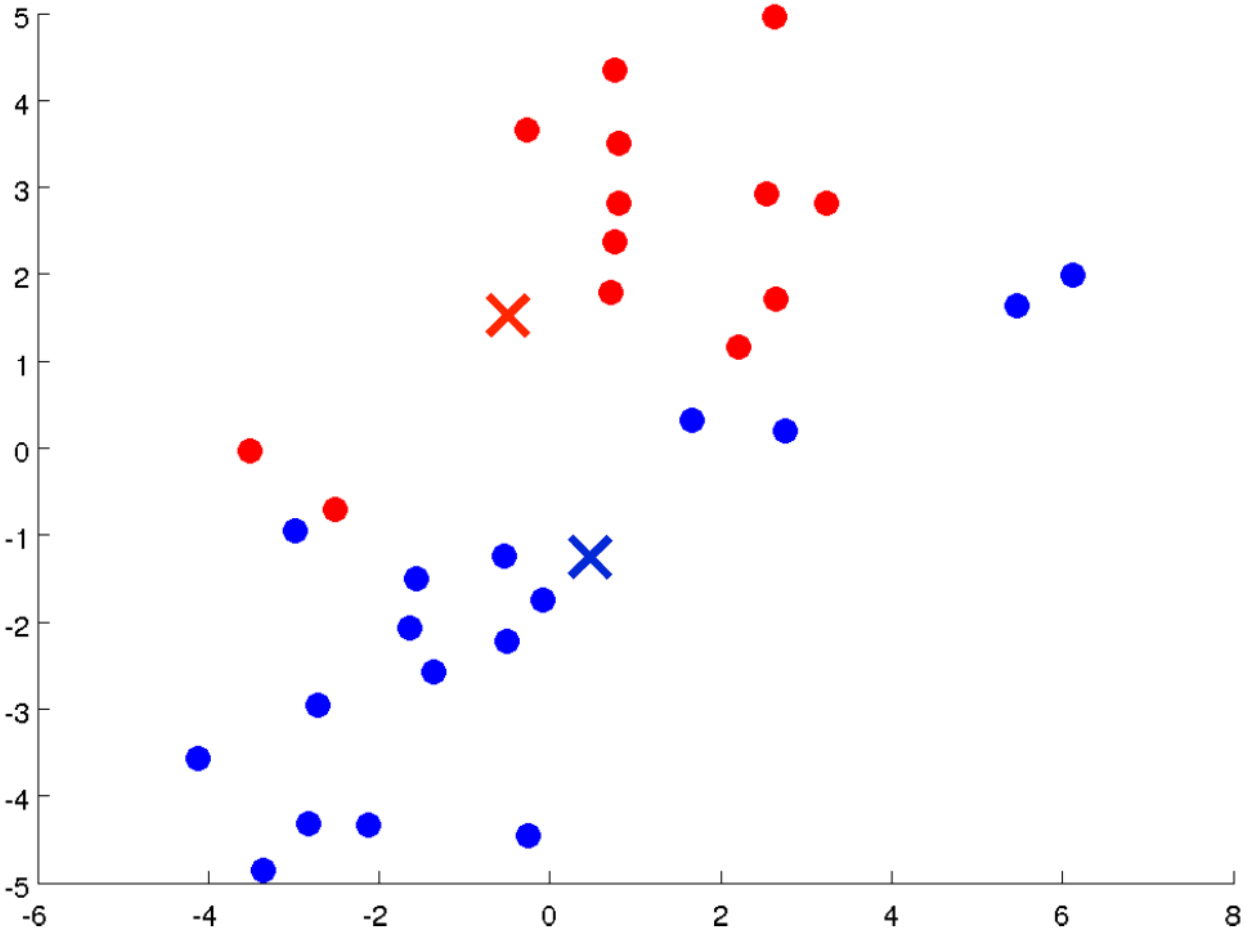
K-means – Assign data points to cluster centroids



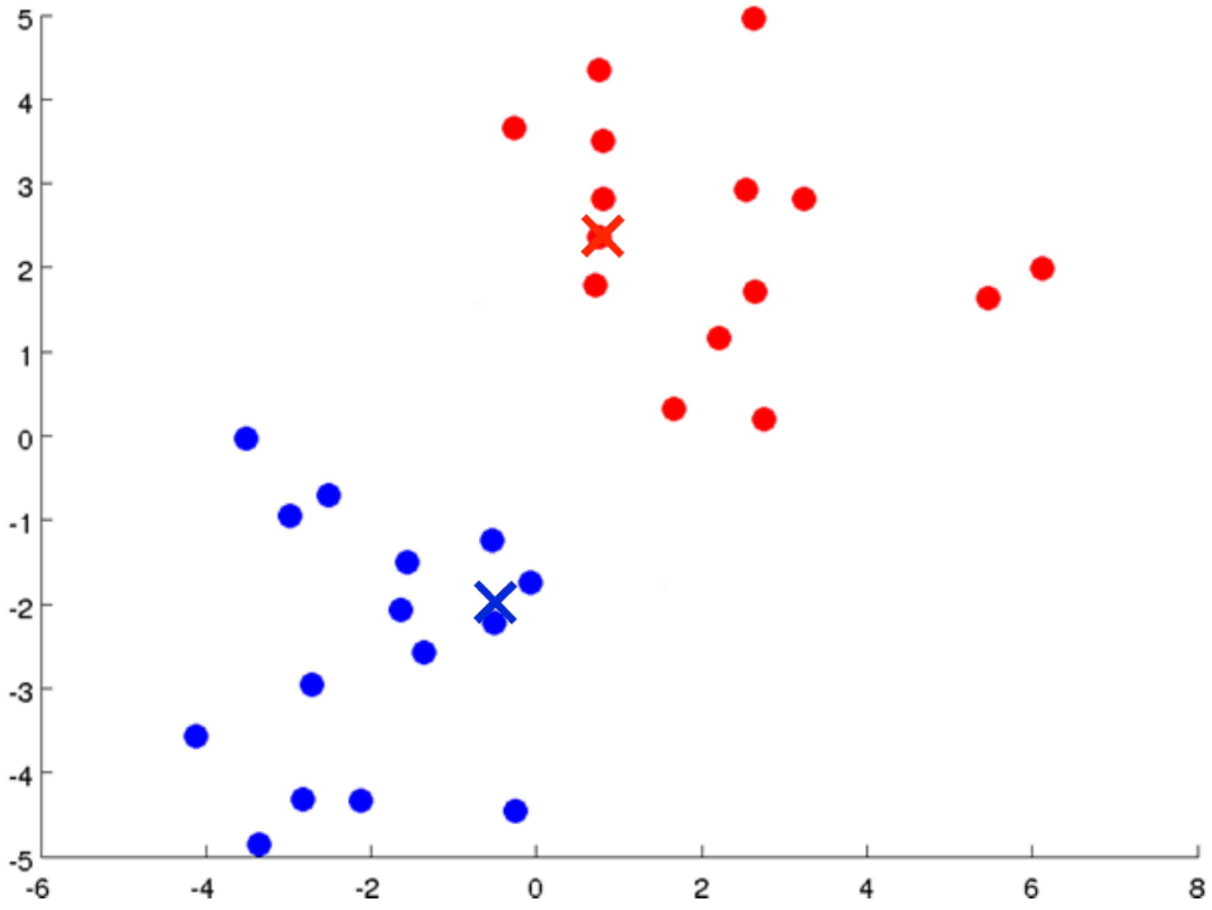
K-means – Move to Average of points of same colour



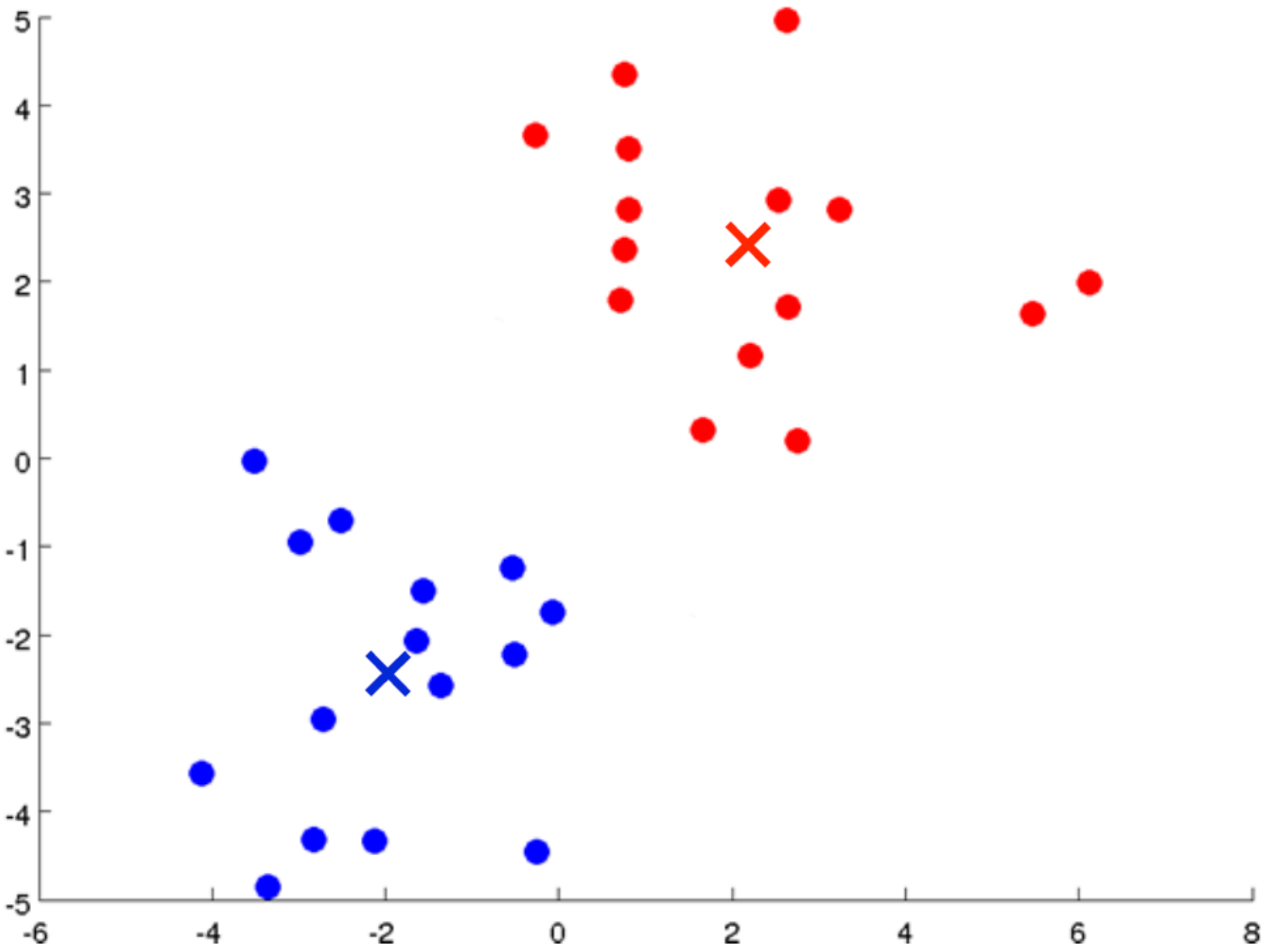
K-means – Assign data points to "new" cluster centroids



K-means – move and Assign



K-MEANS – MOVE AND ASSIGN – NO CHANGE - DONE



K-MEANS

- Input:
  - K (number of clusters)
  - Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

- $X^{(1)} \in \mathbb{R}^n$ (no $x_0 = 1$ convention, so not $\mathbb{R}^{n+1}$)

K-MEANS

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat

{

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroids closest to $x^{(i)}$

for k = 1 to K

μ_K := average (mean)of points assigned to cluster k

}

Example: $c^{(1)}=2, c^{(5)}=2, c^{(6)}=2 \rightarrow \mu_K = 1/3\ (x^{(1)} + x^{(5)} + x^{(6)}) \in \mathbb{R}^n \rightarrow$ n-dimensional vector

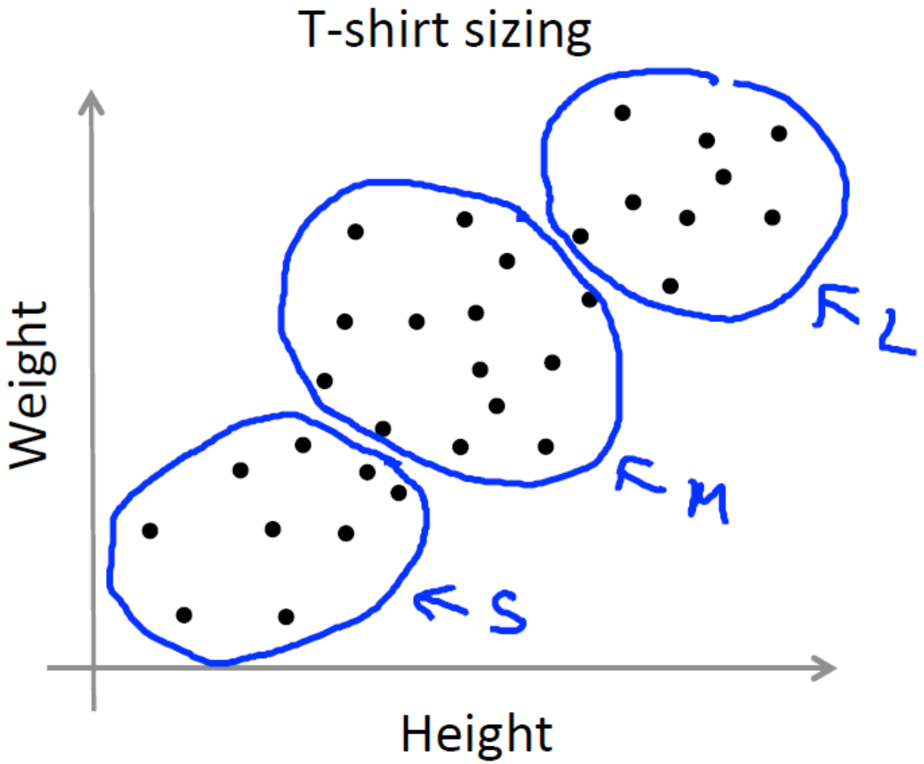
Cluster assignment step

$\mathrm{Min}_k \| x^{(i)} - \mu_K \|^2$

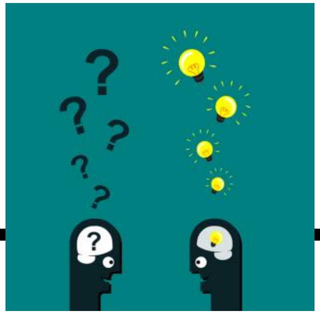
Set this value as $c^{(i)}$

Move centroid step

K-means for non-separated clusters



QUESTION



Suppose you run k-means and after the algorithm converges, you have: $c^{(1)}=3, c^{(2)}=3, c^{(3)}=5, \ldots$ Which of the following statements are true? Check all that apply.

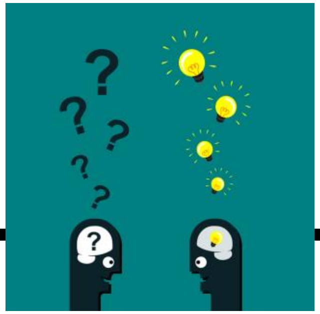
A: The third example $x^{(3)}$ has been assigned to cluster 5.

B: The first and second training examples $x^{(1)}$ and $x^{(2)}$ have been assigned to the same cluster.

C: The second and third training examples have been assigned to the same cluster.

D: Out of all the possible values of $k \in \{1,2,\ldots,K\}$ the value $k=3$ minimizes $\|x^{(2)} - \mu_k\|^2$

QUESTION



Suppose you run k-means and after the algorithm converges, you have:
$c^{(1)}=3, c^{(2)}=3, c^{(3)}=5, \ldots$ Which of the following statements are true? Check all that apply.

A: The third example $x^{(3)}$ has been assigned to cluster 5.

B: The first and second training examples $x^{(1)}$ and $x^{(2)}$ have been assigned to the same cluster.

C: The second and third training examples have been assigned to the same cluster.

D: Out of all the possible values of $k \in \{1,2,\ldots,K\}$ the value $k=3$ minimizes $\|x^{(2)}-\mu_k\|^2$

K-means Optimization objective

- $c^{(i)}$ = index of cluster (1,2,..,K) to which example $x^{(i)}$ is currently assigned
- μ_K = cluster centroid k ($\mu_K \in \mathbb{R}^n$; $k \in \{1,2,\ldots,K\}$)
- $\mu_c^{(i)}$ = cluster centorid of cluster to which example $x^{(i)}$ has been assigned (e.g. $x^{(i)} \rightarrow 5$ i.e. $c^{(i)} = 5$ i.e. $\mu_c^{(i)} = \mu_5$)

Optimization objective:

$J(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_K) = \frac{1}{m}\sum_{i=1}^{m} \| x^{(i)} - \mu_c^{(i)}\|^2$ (Distortion Cost Function)

$$\min_{\substack{c^{(1)},\ldots,c^{(m)} \\ \mu_1,\ldots,\mu_K}} J(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_K)$$

K-MEANS

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat

{

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroids closest to $x^{(i)}$

for k = 1 to K

μ_K := average (mean)of points assigned to cluster k

}

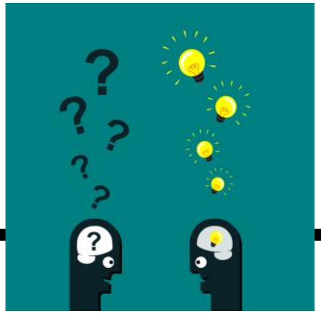
Cluster assignment step

Min J()
with respect to
$c^{(1)}, \ldots, c^{(m)}$
(holding $\mu_1, \ldots, \mu_K$
fixed)

Move centroid step

Minimize J() with respect to $\mu_1, \ldots, \mu_K$

QUESTION



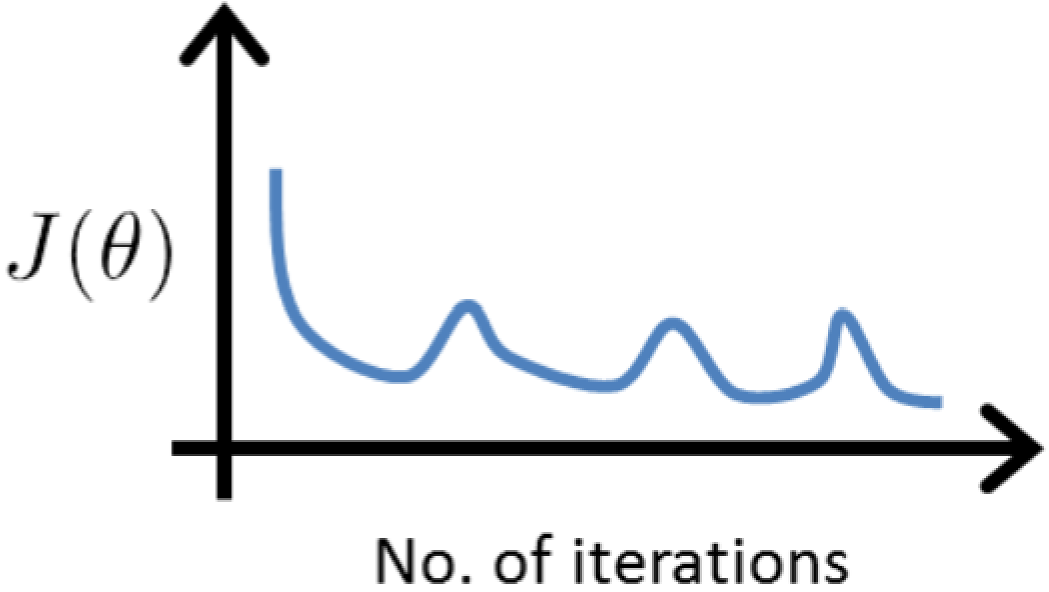
Suppose you have implemented k-means and to check that it is running correctly, you plot the cost function $J(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_K)$ as a function of the number of iterations. You get the given plot. What does this mean?

A: The learning rate is too large.

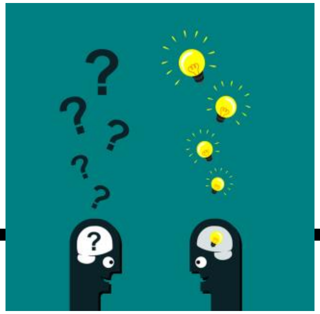
B: The algorithm is working correctly.

C: The algorithm is working, but k is too large.

D: It is not possible for the cost function to sometimes increase. There must be a bug in the code.



QUESTION



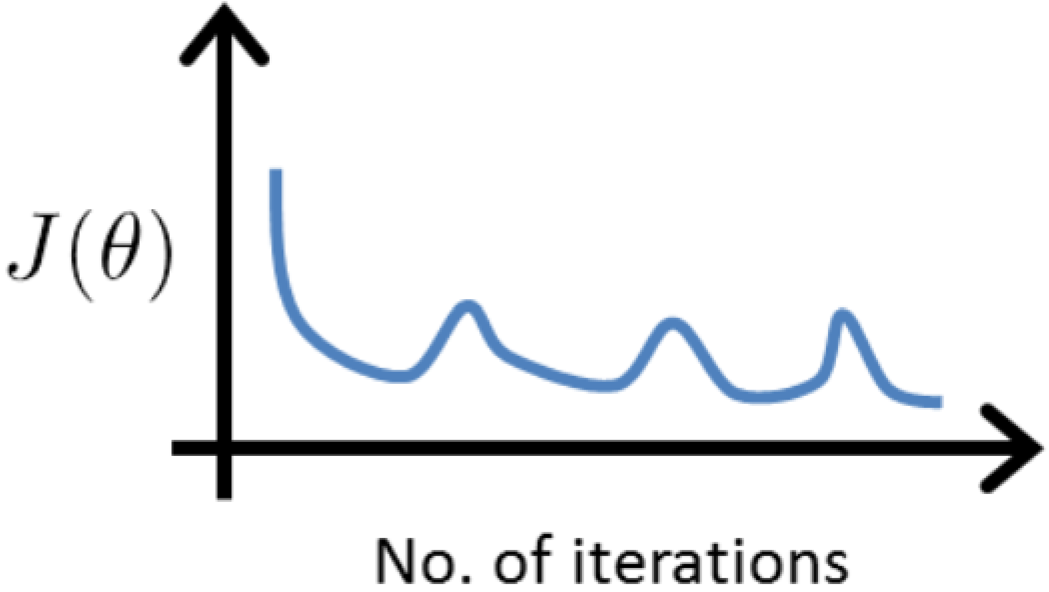
Suppose you have implemented k-means and to check that it is running correctly, you plot the cost function $J(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_K)$ as a function of the number of iterations. You get the given plot. What does this mean?

A: The learning rate is too large.

B: The algorithm is working correctly.

C: The algorithm is working, but k is too large.

D: It is not possible for the cost function to sometimes increase. There must be a bug in the code.



RANDOM INITIALIZATION

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat

{

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroids closest to $x^{(i)}$

for k = 1 to K

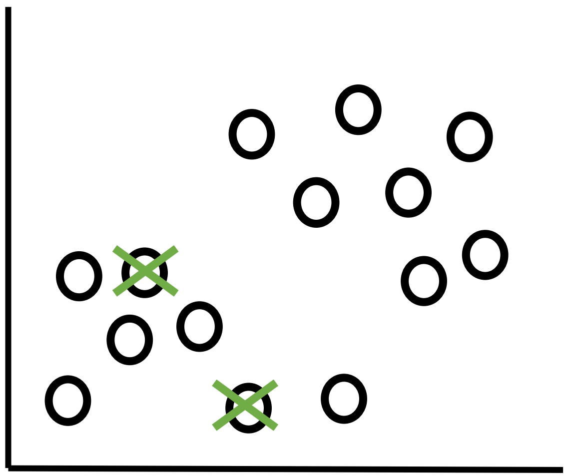
μ_K := average (mean)of points assigned to cluster k

}

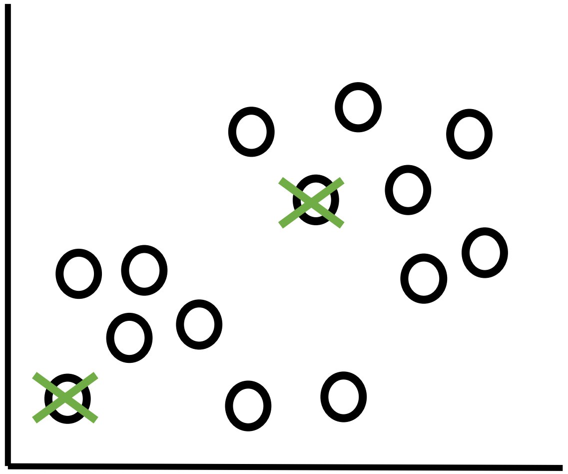
Random Initialization

- Should have K<m
- Randomly pick K training examples
- Set $\mu_1, \ldots, \mu_K$ equal to these K examples

K = 2



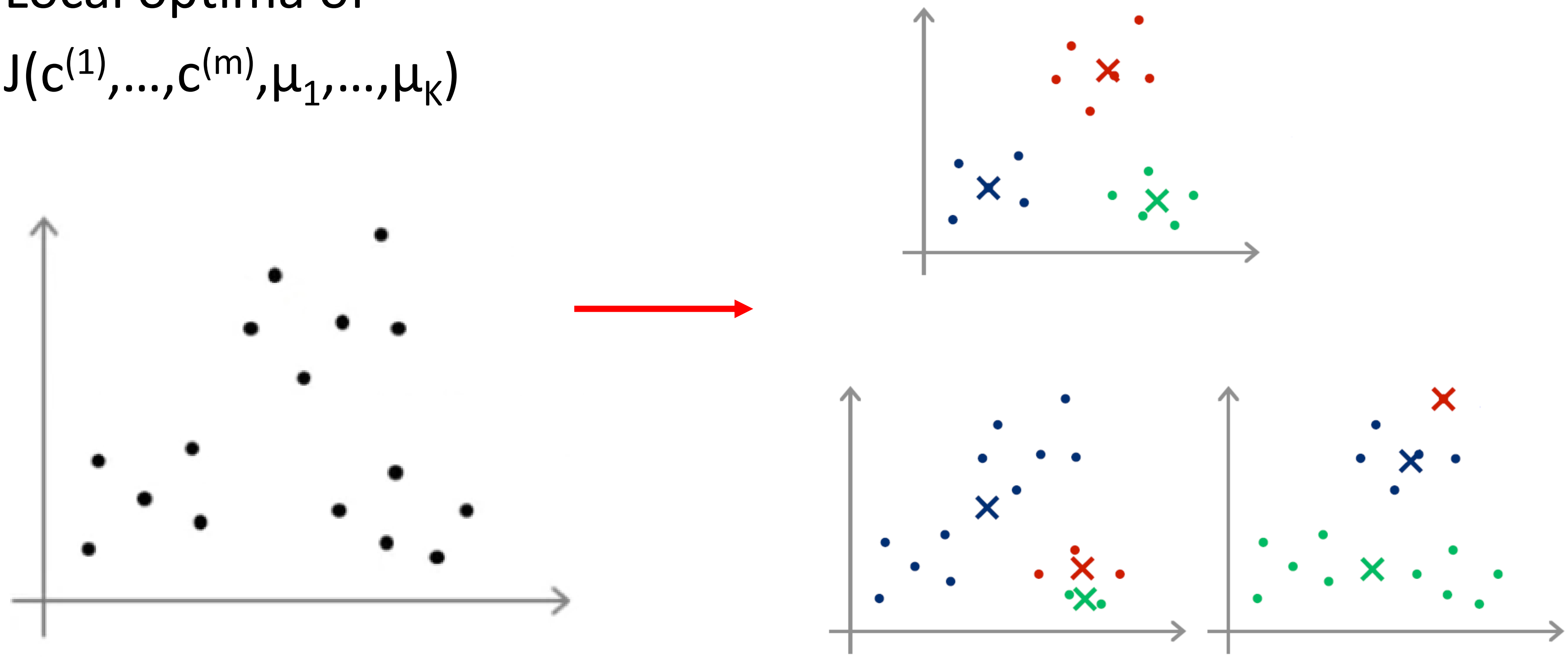
or



Random Initialization

Local optima of

$J(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_K)$



Random Initialization

For i = 1 to 100

{

Randomly initialize k-means.

Run k-means.

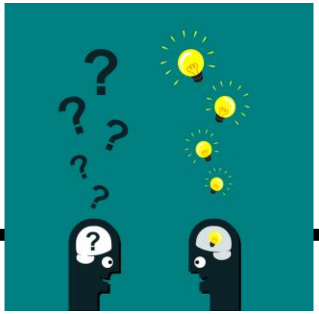
Get $c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_{K.}$

Compute the cost function $J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_K)$.

}

→ Pick the clustering with the lowest cost $J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_K)$.

QUESTION



Which of the following is the recommended way to initialize k-means?

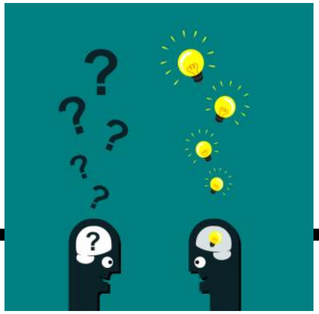
A: Pick a random integer i from {1,…,k}. Set $\mu_1=\mu_2=\ldots=\mu_k=x^{(i)}$.

B: Pick k distinct random integers $i_1,\ldots,i_k$ from {1,…,k}. Set $\mu_1=x^{(1)},\mu_2=x^{(2)},\ldots,\mu_k=x^{(k)}$.

C: Pick k distinct random integers $i_1,\ldots,i_k$ from {1,…,m}. Set $\mu_1=x^{(1)},\mu_2=x^{(2)},\ldots,\mu_k=x^{(k)}$.

D: Set every element of $\mu_i \in \mathbb{R}^n$ to a random value between $-\epsilon$ and ϵ for some small ϵ.

QUESTION



Which of the following is the recommended way to initialize k-means?

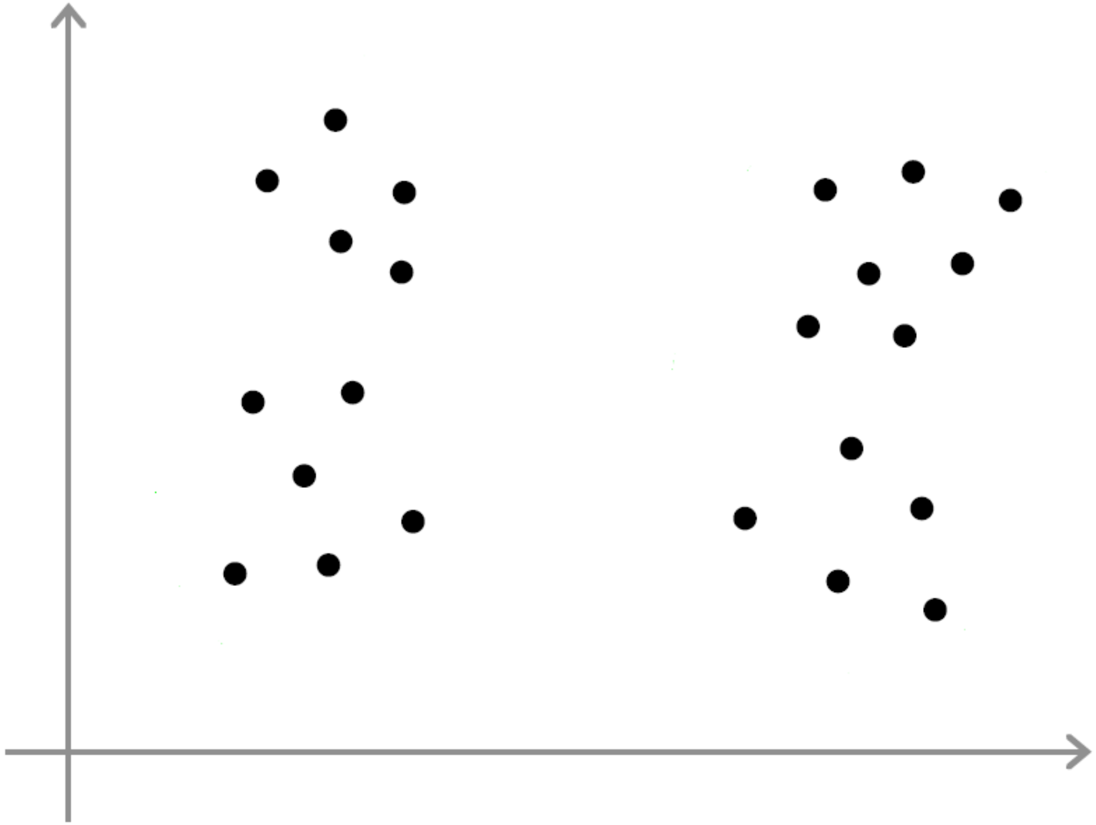
A: Pick a random integer i from {1,…,k}. Set $\mu_1=\mu_2=\ldots=\mu_k=x^{(i)}$.

B: Pick k distinct random integers $i_1,\ldots,i_k$ from {1,…,k}. Set $\mu_1=x^{(1)},\mu_2=x^{(2)},\ldots,\mu_k=x^{(k)}$.

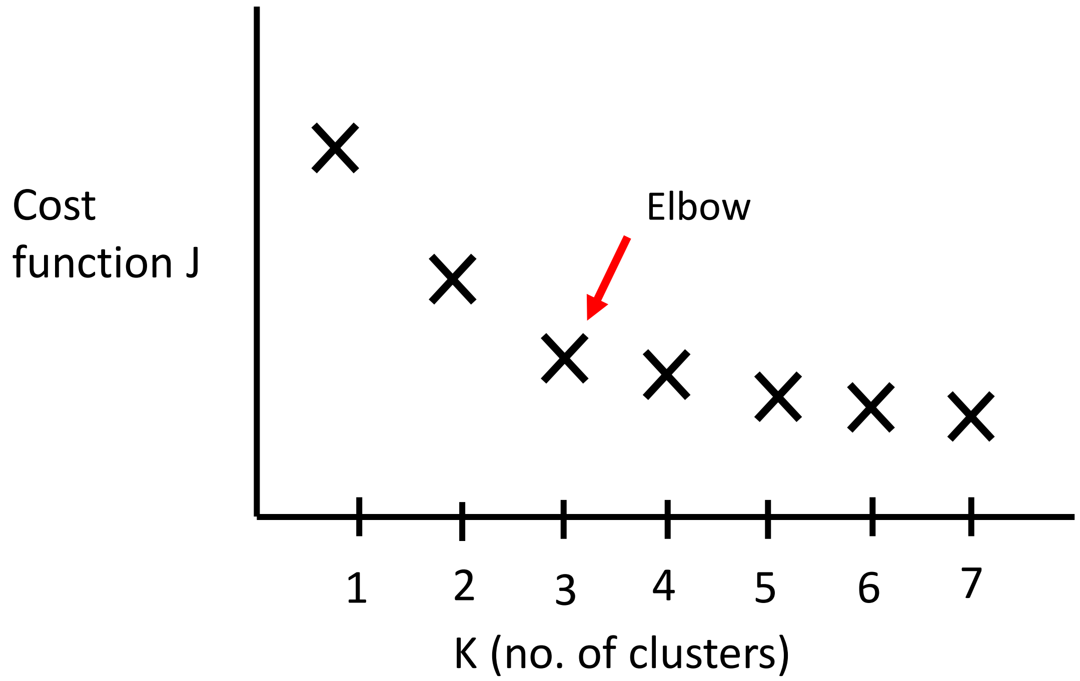
C: Pick k distinct random integers $i_1,\ldots,i_k$ from {1,…,m}. Set $\mu_1=x^{(1)},\mu_2=x^{(2)},\ldots,\mu_k=x^{(k)}$.

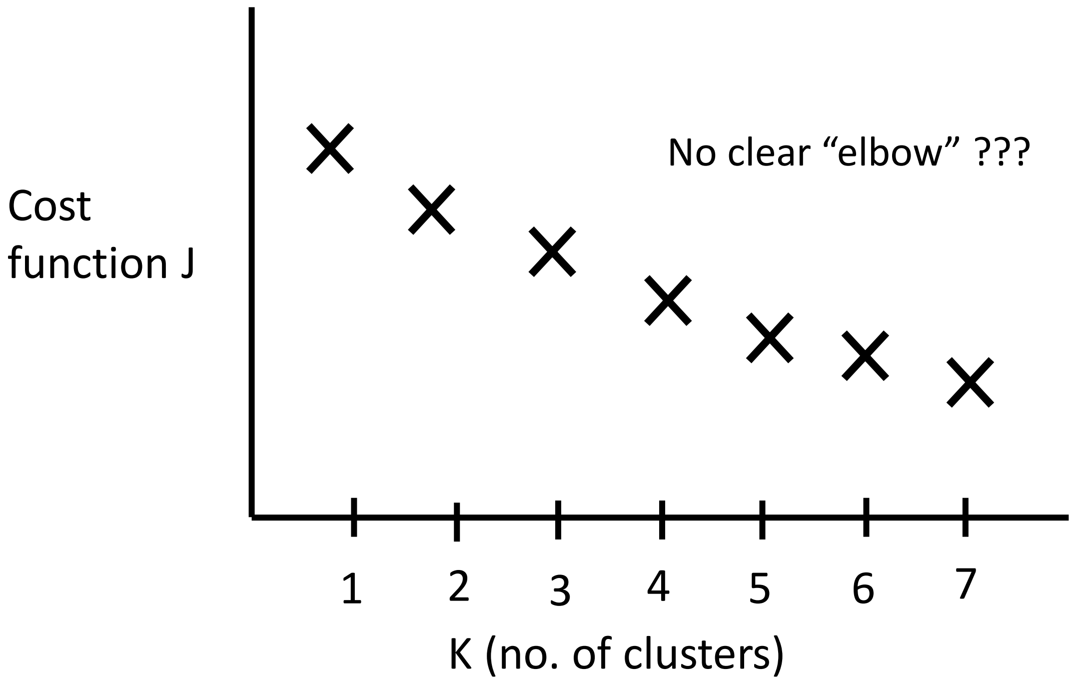
D: Set every element of $\mu_i \in \mathbb{R}^n$ to a random value between $-\epsilon$ and ϵ for some small ϵ.

HOW MANY CLUSTERS DO YOU SEE?

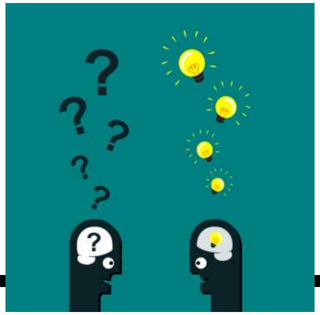


WHAT IS THE RIGHT VALUE OF K? ELBOW METHOD





QUESTION



Suppose you run k-means using k = 3 and k = 5. You find that the cost function J is much higher for k = 5 than for k = 3. What can you conclude?

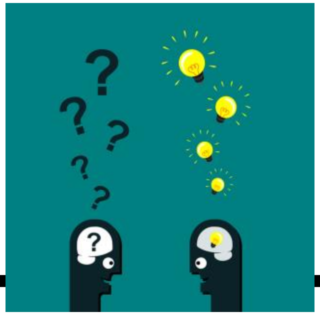
A: This is mathematically impossible. There must be a bug in the code.

B: The correct number of clusters is k = 3.

C: In the run with k = 5, k-means got stuck in a bad local minimum. You should try re-running k-means with multiple random initializations.

D: In the run with k = 3, k-means got lucky. You should try re-running k-means with k = 3 and different random initializations until it performs no better than with k = 5.

QUESTION



Suppose you run k-means using k = 3 and k = 5. You find that the cost function J is much higher for k = 5 than for k = 3. What can you conclude?

A: This is mathematically impossible. There must be a bug in the code.

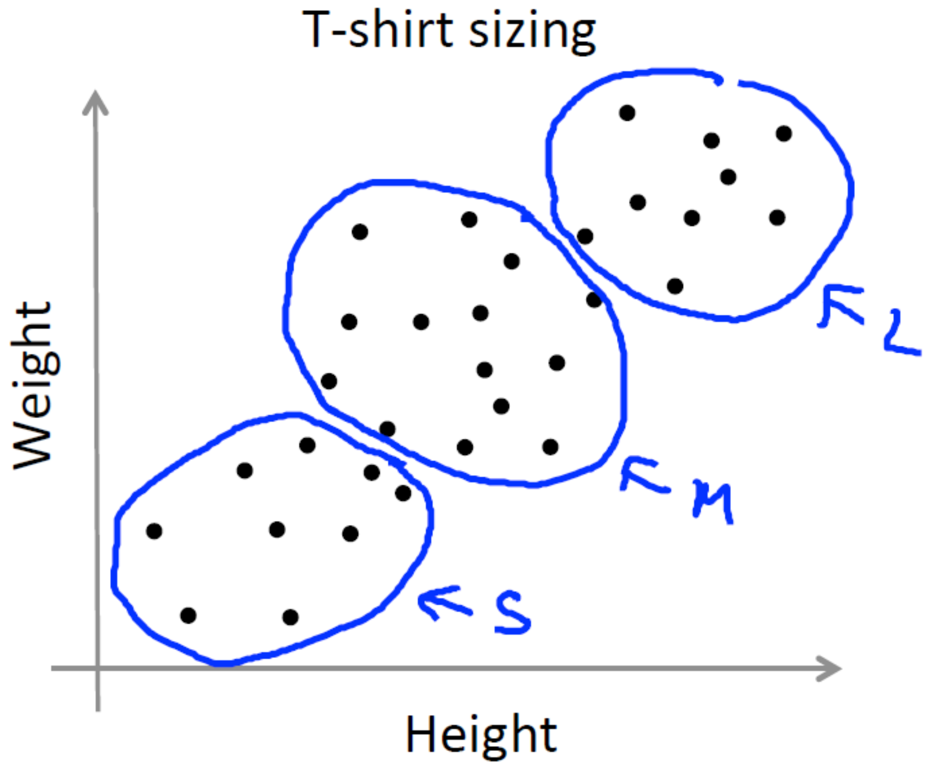
B: The correct number of clusters is k = 3.

C: In the run with k = 5, k-means got stuck in a bad local minimum. You should try re-running k-means with multiple random initializations.

D: In the run with k = 3, k-means got lucky. You should try re-running k-means with k = 3 and different random initializations until it performs no better than with k = 5.

CHOOSING THE VALUE OF K

- Often you use k-means to get clusters for a later purpose
- Evaluate your clusters based on this purpose
- Example:
  - How many t-shirt clusters do I want to have?
  - K = 3: easier to produce
  - K = 5: better fit for the customers



QUIZ - QUESTION 1

For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

A: Given a set of news articles from many different news websites, find out what are the main topics covered.

B: Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

C: Given many emails, you want to determine if they are Spam or Non-Spam emails.

D: From the user usage patterns on a website, figure out what different groups of users exist.

QUIZ - QUESTION 1

For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

A: Given a set of news articles from many different news websites, find out what are the main topics covered.

B: Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

C: Given many emails, you want to determine if they are Spam or Non-Spam emails.

D: From the user usage patterns on a website, figure out what different groups of users exist.

QUIZ - QUESTION 2

Suppose we have three cluster centroids μ_1=[1;2], μ_2=[-3;0], and μ_3=[4;2]. Furthermore, we have a training example $x^{(i)}$=[−2;1]. After a cluster assignment step, what will $c^{(i)}$?

A: $c^{(i)} = 3$.

B: $c^{(i)}$ is not assigned.

C: $c^{(i)} = 1$

D: $c^{(i)} = 2$

QUIZ - QUESTION 2

Suppose we have three cluster centroids $\mu_1=[1;2]$, $\mu_2=[-3;0]$, and $\mu_3=[4;2]$. Furthermore, we have a training example $x^{(i)}=[-2;1]$. After a cluster assignment step, what will $c^{(i)}$?

A: $c^{(i)} = 3$.

B: $c^{(i)}$ is not assigned.

C: $c^{(i)} = 1$

D: $c^{(i)} = 2$

$\mu_1=[1;2] \rightarrow [-2;1] - [1;2] = |[-3,-1]|$

$\mu_2=[-3;0] \rightarrow [-2;1] - [-3;0] = \underline{|[\mathbf{1,1}]|}$

$\mu_3=[4;2] \rightarrow [-2;1] -[4;2] = |[-6,-1]|$

Quiz - Question 3

K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

A: The cluster centroid assignment step, where each cluster centroid μ_i is assigned (by setting $c^{(i)}$) to the closest training example $x^{(i)}$.

B: Move each cluster centroid μ_k by setting it to be equal to the closest training example $x^{(i)}$.

C: The cluster assignment step, where the parameters $c^{(i)}$ are updated.

D: Move the cluster centroids, where the centroids μ_k are updated.

Quiz - Question 3

K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

A: The cluster centroid assignment step, where each cluster centroid μ_i is assigned (by setting $c^{(i)}$) to the closest training example $x^{(i)}$.

B: Move each cluster centroid μ_k by setting it to be equal to the closest training example $x^{(i)}$.

C: The cluster assignment step, where the parameters $c^{(i)}$ are updated.

D: Move the cluster centroids, where the centroids μ_k are updated.

REMINDER: K-MEANS

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat

{

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroids closest to $x^{(i)}$

for k = 1 to K

μ_K := average (mean)of points assigned to cluster k

}

Example: $c^{(1)}=2, c^{(5)}=2, c^{(6)}=2 \rightarrow \mu_K = 1/3\,(x^{(1)} + x^{(5)} + x^{(6)}) \in \mathbb{R}^n \rightarrow$ n-dimensional vector

Cluster assignment step

$\mathrm{Min}_k \| x^{(i)} - \mu_K \|^2$
Set this value as $c^{(i)}$

Move centroid step

QUIZ - QUESTION 4

Suppose you have an unlabeled dataset $\{x^{(1)},\dots,x^{(m)}\}$. You run K-means with 50 different random initializations, and obtain 50 different clusterings of the data. What is the recommended way for choosing which one of these 50 clusterings to use?

A: The answer is ambiguous, and there is no good way of choosing.

B: For each of the clusterings, compute $\frac{1}{m}\sum_{i=1}^{m} \| x^{(i)} - \mu_c^{(i)} \|^2$ and pick the one that minimizes this.

C: Always pick the final (50th) clustering found, since by that time it is more likely to have converged to a good solution.

D: The only way to do so is if we also have labels $y^{(i)}$ for our data.

QUIZ - QUESTION 4

Suppose you have an unlabeled dataset $\{x^{(1)},\ldots,x^{(m)}\}$. You run K-means with 50 different random initializations, and obtain 50 different clusterings of the data. What is the recommended way for choosing which one of these 50 clusterings to use?

A: The answer is ambiguous, and there is no good way of choosing.

B: For each of the clusterings, compute $\frac{1}{m}\sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2$ and pick the one that minimizes this.

C: Always pick the final (50th) clustering found, since by that time it is more likely to have converged to a good solution.

D: The only way to do so is if we also have labels $y^{(i)}$ for our data.

QUIZ - QUESTION 5

Which of the following statements are true? Select all that apply.

A: If we are worried about K-means getting stuck in bad local optima, one way to reduce this problem is if we try using multiple random initializations.

B: For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.

C: The standard way of initializing K-means is setting $\mu_1=\ldots=\mu_k$ to be equal to a vector of zeros.

D: Since K-Means is an unsupervised learning algorithm, it cannot overfit the data, and thus it is always better to have as large a number of clusters as is computationally feasible.

QUIZ - QUESTION 5

Which of the following statements are true? Select all that apply.

A: If we are worried about K-means getting stuck in bad local optima, one way to reduce this problem is if we try using multiple random initializations.

B: For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.

C: The standard way of initializing K-means is setting $\mu_1=\ldots=\mu_k$ to be equal to a vector of zeros.

D: Since K-Means is an unsupervised learning algorithm, it cannot overfit the data, and thus it is always better to have as large a number of clusters as is computationally feasible.