



Machine Learning

Prof. Adil Khan

Objectives

1. What is clustering? How is it different from classification? Why is it called unsupervised learning?
2. What is k-means? How does it work? What is its objective function? How is it motivated?
3. What is k-means++? Why is it motivated?
4. Limitations of k-means
5. K-means++
6. Hierarchical Clustering

Clustering

Recall: Supervised Learning

- Learning with a teacher
- Teacher provides the supervision
- In the context of machine learning, the labels y_i provides the supervision for x_i

Recall: Supervised Learning (2)

- Easy and well-defined
- Given $Y = f(X) + \epsilon$ and labeled dataset

$$\mathbb{D} = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^m$$

- Estimate the function f as \hat{f} , such that when a new input data x_0 is given, we can predict its outcome y_0 .

Unsupervised Learning

- We have only predictors (a.k.a inputs, or features) but no labels

$$\mathbb{D} = \{(x_i) | x_i \in \mathbb{R}^p\}_{i=1}^m$$

- Then what is the goal of learning?

Scenario (1)

- Imagine you run an online store and would like to personalize your customers' shopping experience
- You think you can do this by providing each customer with personalized recommendations
- You do not know each user's personal preferences and tastes but you have lots of data on their purchases
- How can you use this data to create recommendations?

Scenario (2)

- Imagine you are a biologist interested in studying the behavior of bees
- You have collected a lot of videos (or other data on them)
- How can you use this data to study bee behavior?

Unsupervised Learning

- In the discussed scenarios, we are not interested in prediction, because we do not have an associated response variable
- Instead, the unsupervised **learning goal** is to model the **hidden patterns** or **underlying structure** in the given input data in order to learn about the data
- This underlying structure is what we usually refer to as groups of data
- And these groups are what we refer to as **Clusters**

Example

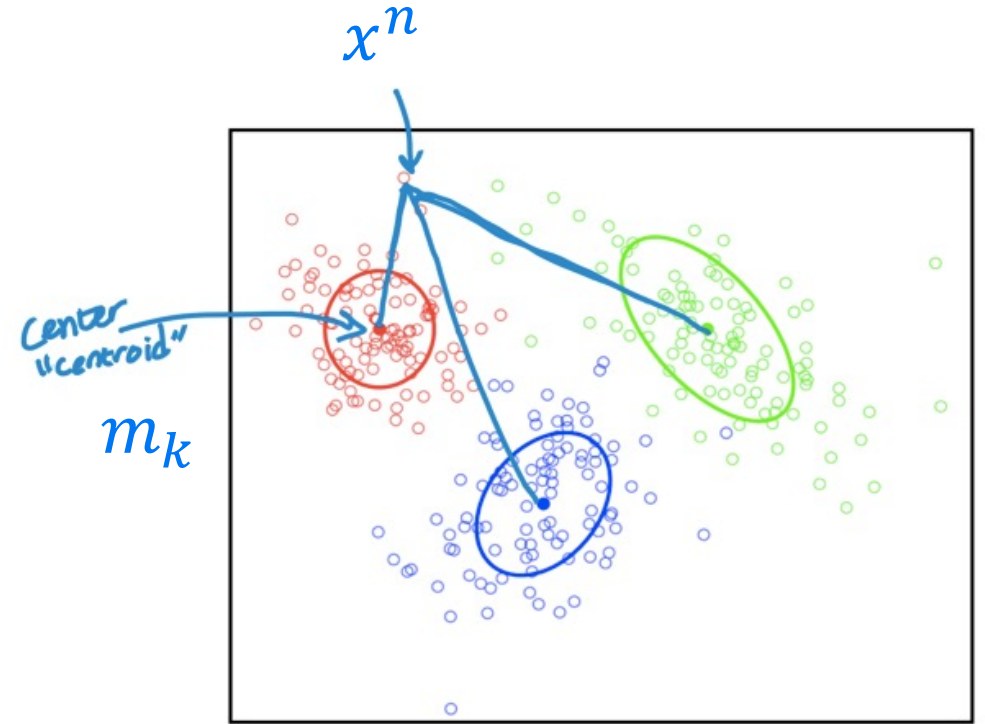
- Given a set of documents, we used their data to group them into two clusters



Let's first learn some basic
things about clustering

What defines a cluster?

- Clusters are defined by a **center** and a **spread** (which you can also call shape)
- And we assign a given point x^n to a cluster by computing its similarity to the **center** of the cluster m_k



How can we define *similarity*?

Similarity

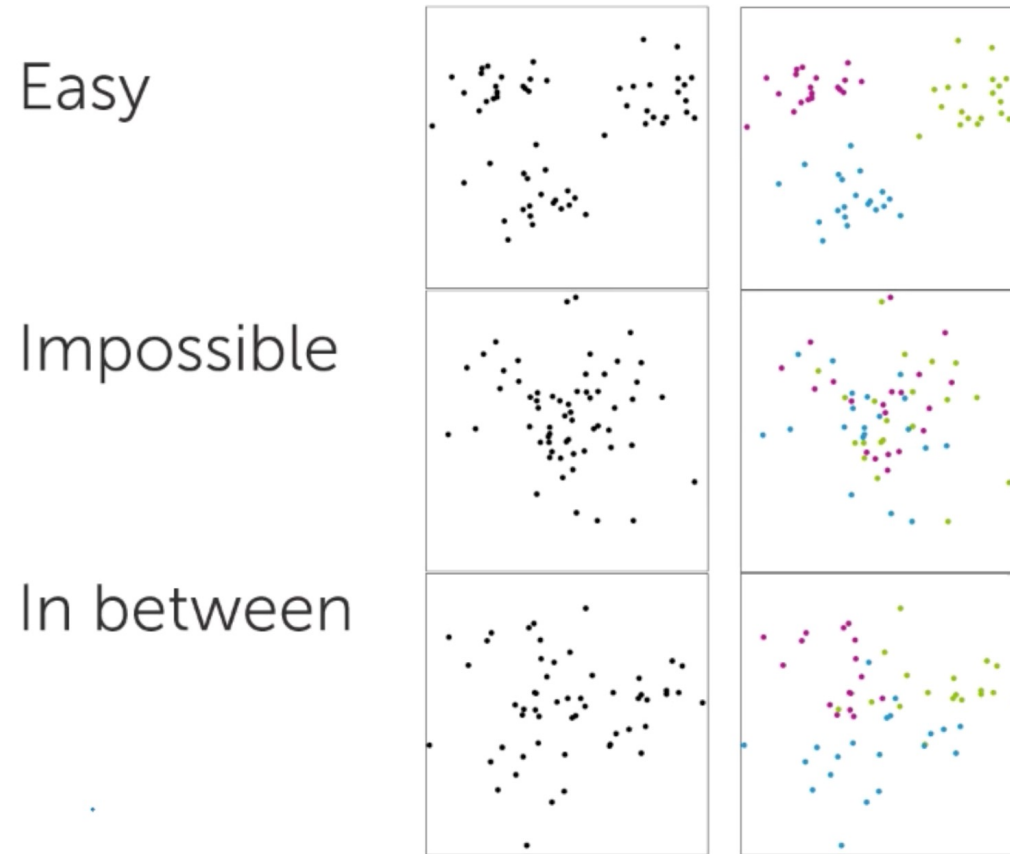
- Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered.
- A clustering method attempts to group the objects based on the definition of similarity (which is usually measured in the form of a distance from the center of a cluster) supplied to it.

Multiple Choices For Measuring Dissimilarity

- *Euclidean distance* (the most common)
- Manhattan distance
- Correlation based distances
 - Pearson correlation distance
 - Eisen cosine correlation distance
 - Spearman correlation distance
 - Kendall correlation distance
- Etc.

Read about them yourself. Materials are easily available online.

Finding Clusters is not Always “Easy” or “Difficult”



Bottom line is: clustering can be easy, hard, or in between depending on the structure of the data

Let's Formulate Clustering as an
Optimization Problem

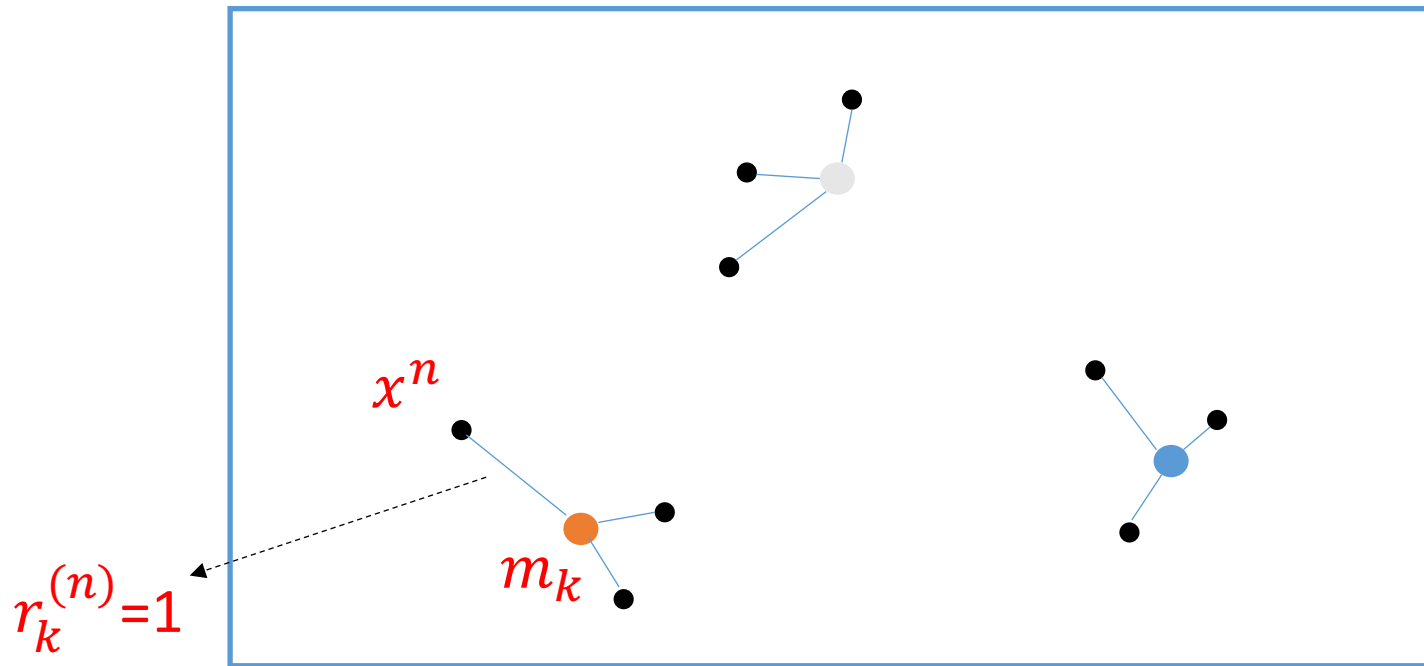
Objective

- Find cluster centers $\{\mathbf{m}_k\}_{k=1}^K$ and assignments $\{\mathbf{r}^{(n)}\}_{n=1}^N$ to minimize the sum of squared distances of data points $\{\mathbf{x}^{(n)}\}$ to their assigned centers

$$\mathbf{x}^{(n)} \in \mathbb{R}^D$$

$$\mathbf{m}_k \in \mathbb{R}^D$$

$$\mathbf{r}^{(n)} \in \mathbb{R}^K$$



$$\mathbf{r}^{(n)} = [0, \dots, 1, \dots, 0]^T$$

Mathematically

$$\min_{\{\mathbf{m}_k\}, \{\mathbf{r}^{(n)}\}} J(\{\mathbf{m}_k\}, \{\mathbf{r}^{(n)}\}) = \min_{\{\mathbf{m}_k\}, \{\mathbf{r}^{(n)}\}} \sum_{n=1}^N \sum_{k=1}^K r_k^{(n)} \|\mathbf{m}_k - \mathbf{x}^{(n)}\|^2$$

where $r_k^{(n)} = \mathbb{I}[\mathbf{x}^{(n)} \text{ is assigned to cluster } k]$, i.e., $\mathbf{r}^{(n)} = [0, \dots, 1, \dots, 0]^\top$

How to Solve the Optimization Problem?

- Optimization problem:

$$\min_{\{\mathbf{m}_k\}, \{\mathbf{r}^{(n)}\}} \sum_{n=1}^N \sum_{k=1}^K r_k^{(n)} \|\mathbf{m}_k - \mathbf{x}^{(n)}\|^2$$

- Problem is hard when minimizing jointly
- But becomes easy when we fix one and minimize over the other

How to Solve the Optimization Problem? (2)

- That is:

$$\min_{\{\mathbf{m}_k\}, \{\mathbf{r}^{(n)}\}} \sum_{n=1}^N \sum_{k=1}^K r_k^{(n)} \|\mathbf{m}_k - \mathbf{x}^{(n)}\|^2$$

- If we fix the centers, then we can easily find the optimal cluster assignments by assigning each point to the cluster with the nearest neighbor

$$r_k^{(n)} = \begin{cases} 1 & \text{if } k = \arg \min_j \|\mathbf{x}^{(n)} - \mathbf{m}_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

How to Solve the Optimization Problem? (3)

- Likewise:
 - If we fix the assignments then we can easily find the optimal cluster centers by setting each cluster's centers to the average of its assigned data points

How to Solve the Optimization Problem? (4)

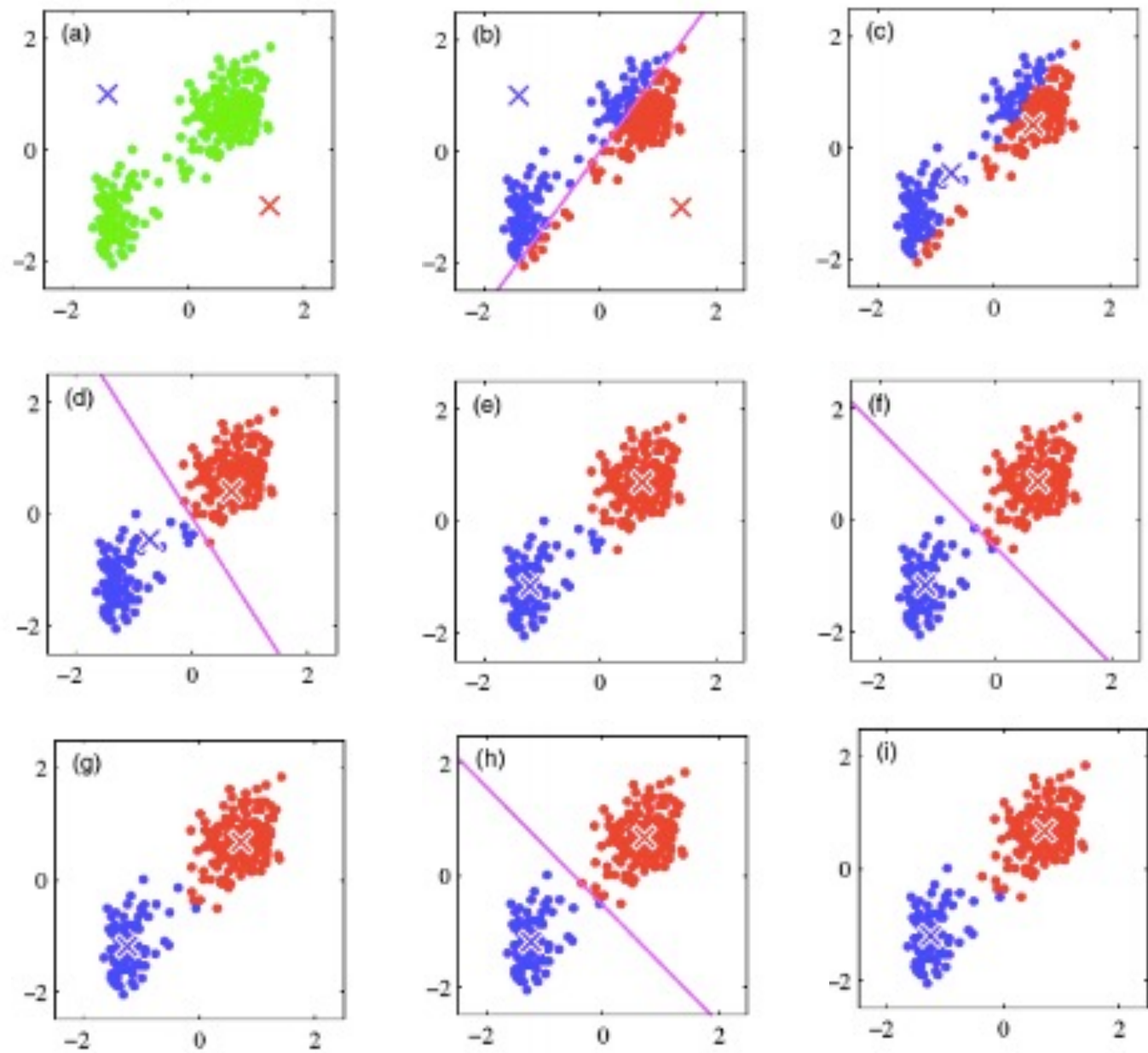
- Thus

$$\min_{\{\mathbf{m}_k\}, \{\mathbf{r}^{(n)}\}} \sum_{n=1}^N \sum_{k=1}^K r_k^{(n)} \|\mathbf{m}_k - \mathbf{x}^{(n)}\|^2$$

- We solve this optimization problem by alternating between minimizing $J(\{\mathbf{m}_k\}, \{\mathbf{r}^{(n)}\})$ with respect to $\{\mathbf{m}_k\}$ and $\{\mathbf{r}^{(n)}\}$
- This is called **alternating minimization**

This is called **k-means** Clustering

Example



K-means Clustering

- Simply works as follows
 - **Initialization**: randomly initialize cluster centers
 - The algorithm then iteratively alternates between the following two steps
 - **Assignment Step**: assign each data point to the closest center
 - **Refitting Step**: move each cluster center to the mean of the data points assigned to it

Convergence of K-means

Converges to:

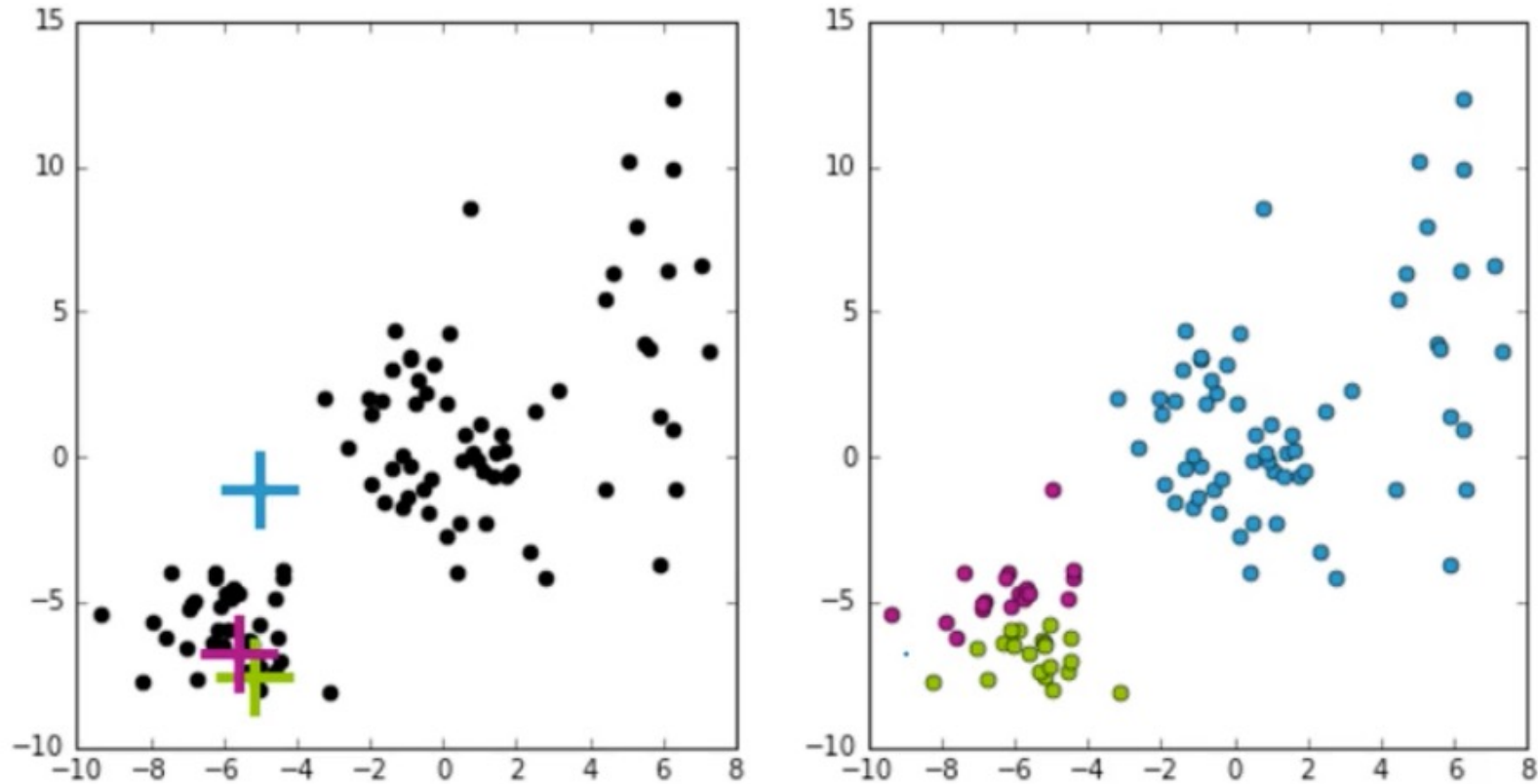
~~- Global optimum~~

- Local optimum

~~- neither~~

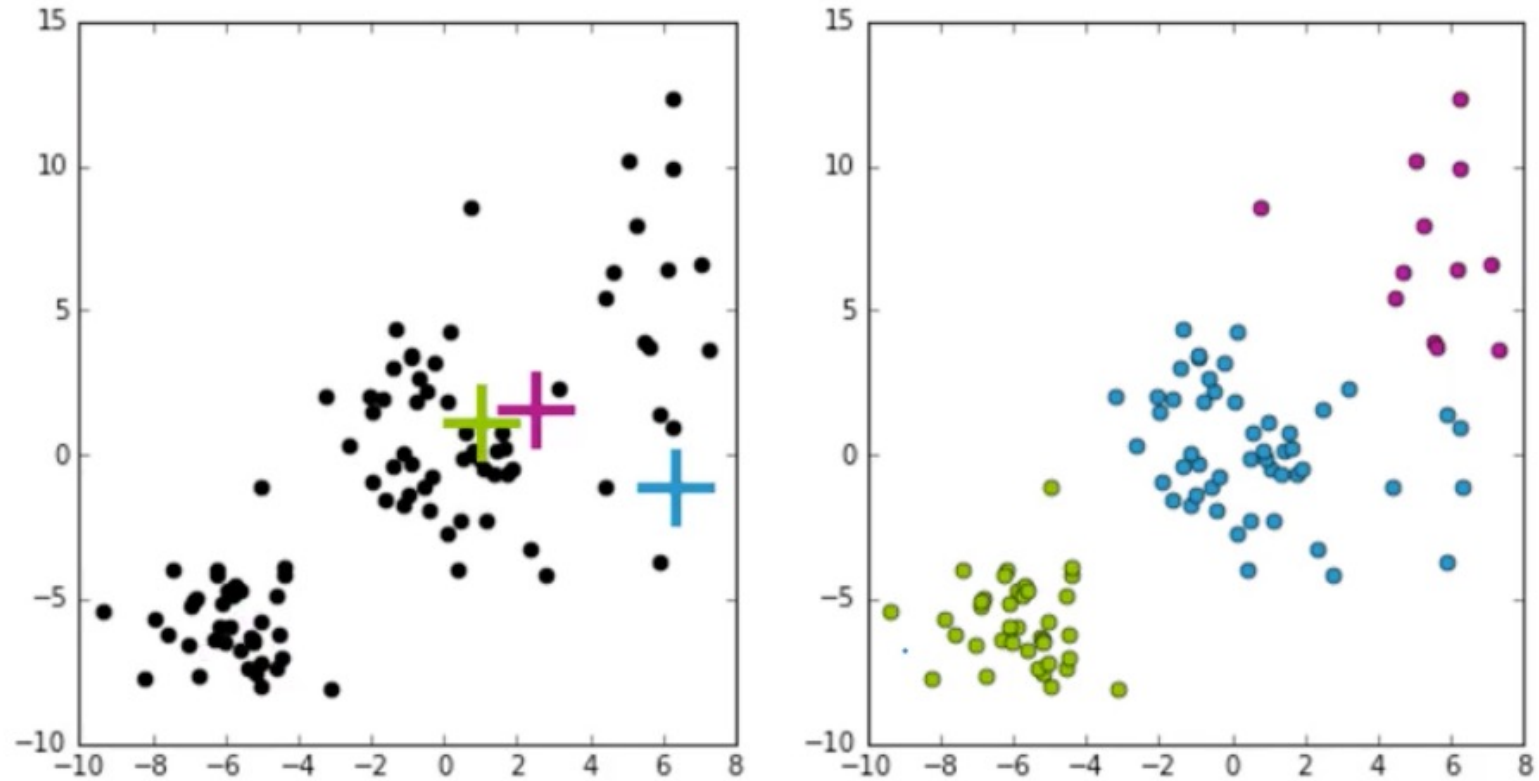
The first one does not mean that it can never converge to global optimum, it is just that there is no guarantee that it will.

Initialization Matters



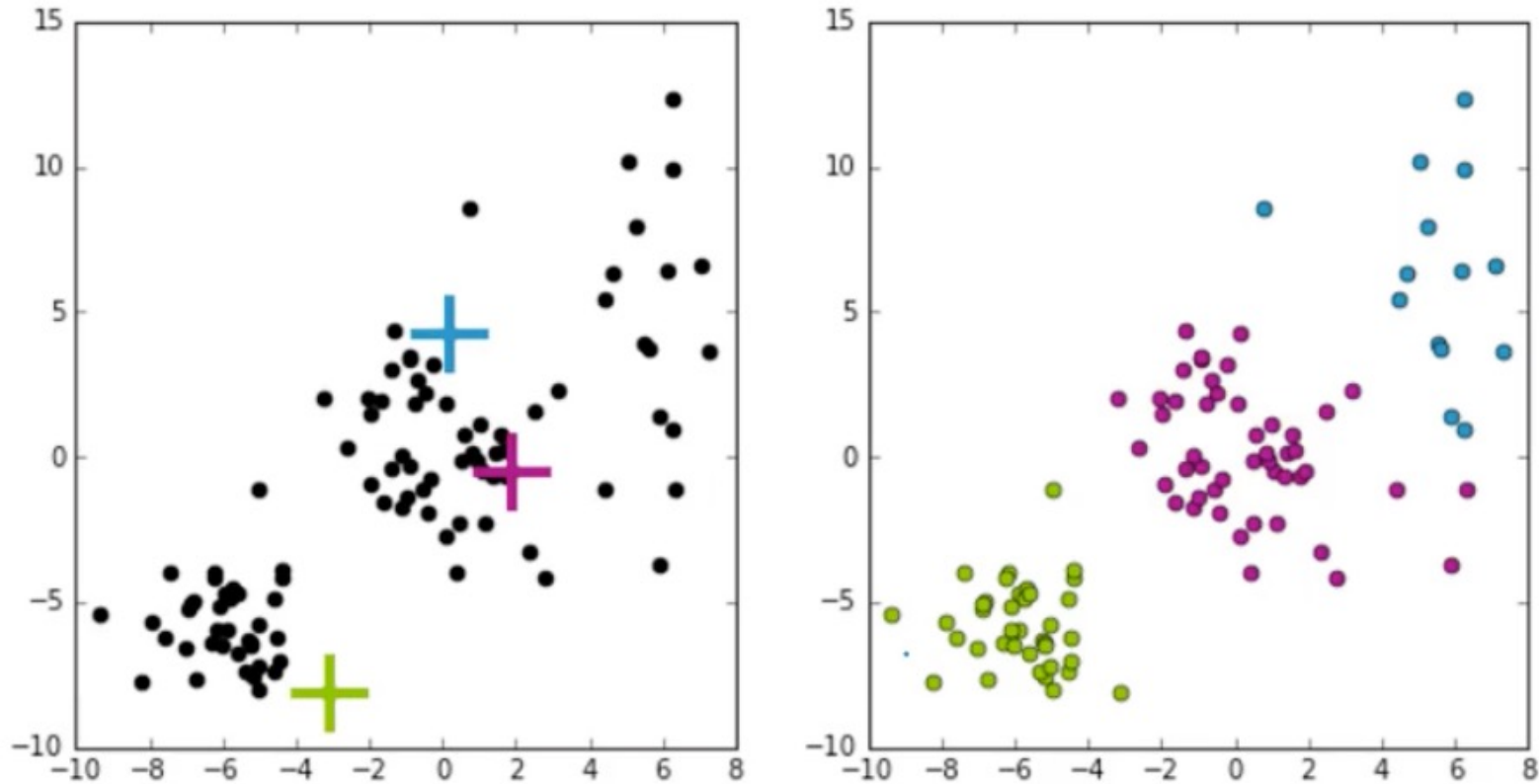
Centroids are chosen as shown in the first figure

Initialization Matters (2)



Different Initialization

Initialization Matters (3)



An even different initialization

Bottom-line

k-means is sensitive to center initialization and we can get completely different solutions, by converging to a local mode

It **can be improved** by using a specific way of choosing centers

K-means++

K-means++

- The intuition behind this approach is that **spreading out the k initial cluster centers** is a good thing
- The **first cluster center is chosen uniformly at random** from the data points that are being clustered
- After which **each subsequent cluster center is chosen from the remaining data points with probability proportional to its squared distance** from the existing cluster center.

How does K-means++ work?

- Let $D(x)$ denote the shortest distance from a data point to the closest center we have already chosen
- Then k-means ++ works as follows
 1. Take the first center m_1 , chosen uniformly at random from $\{x^{(n)}\}$
 2. Take the next center m_k , choosing $x^{(n)}$ with probability $\frac{D(x)^2}{\sum_{x^{(n)} \in X} D(x)^2}$
 3. Repeat step 2 until we have chosen K cluster centers
 4. Proceed with the standard k-means algorithm

K-means++ Summary

- Smart initialization is computationally costly relative to random initialization
- However, due to smart initialization, the subsequent k-means often converges more rapidly
- Tends to improve the quality of the local optimum

How to choose K ?

K-means Objective

$$\min_{\{\mathbf{m}_k\}, \{\mathbf{r}^{(n)}\}} \sum_{n=1}^N \sum_{k=1}^K r_k^{(n)} \|\mathbf{m}_k - \mathbf{x}^{(n)}\|^2$$

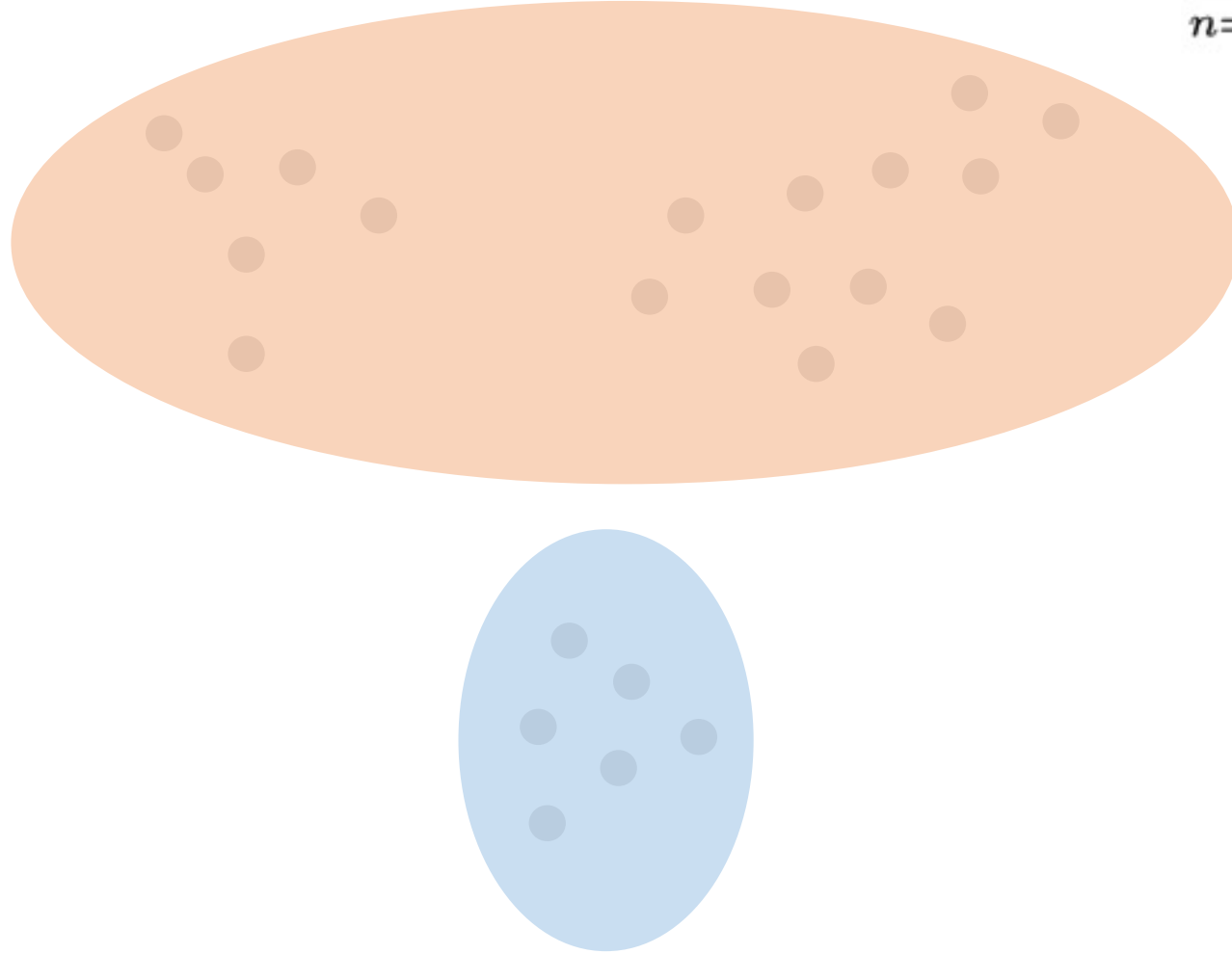
- Think about what will happen if we choose K to be some large value, like $K = N$
- The above value will converge to ZERO – But would it be an optimum solution?

Overfitting in Clustering

- Remember, we are in unsupervised learning.
 - In this case, overfitting would mean that we did not learn the true structures in the data, instead we ended up memorizing the position of the sample in the space.

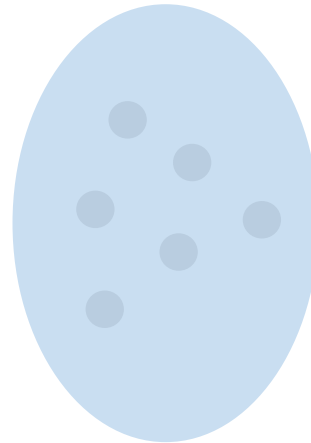
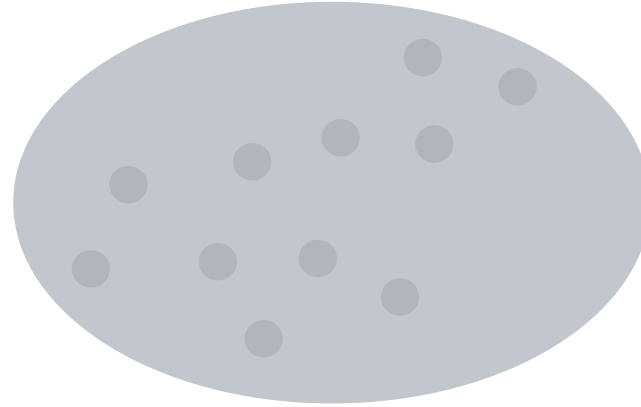
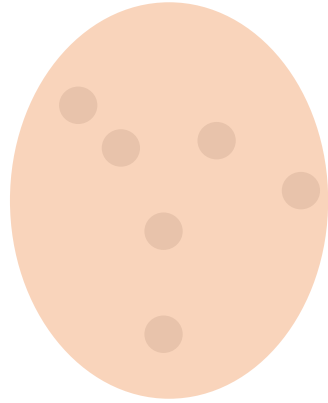
Example (K=2)

$$\sum_{n=1}^N \sum_{k=1}^K r_k^{(n)} ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$$



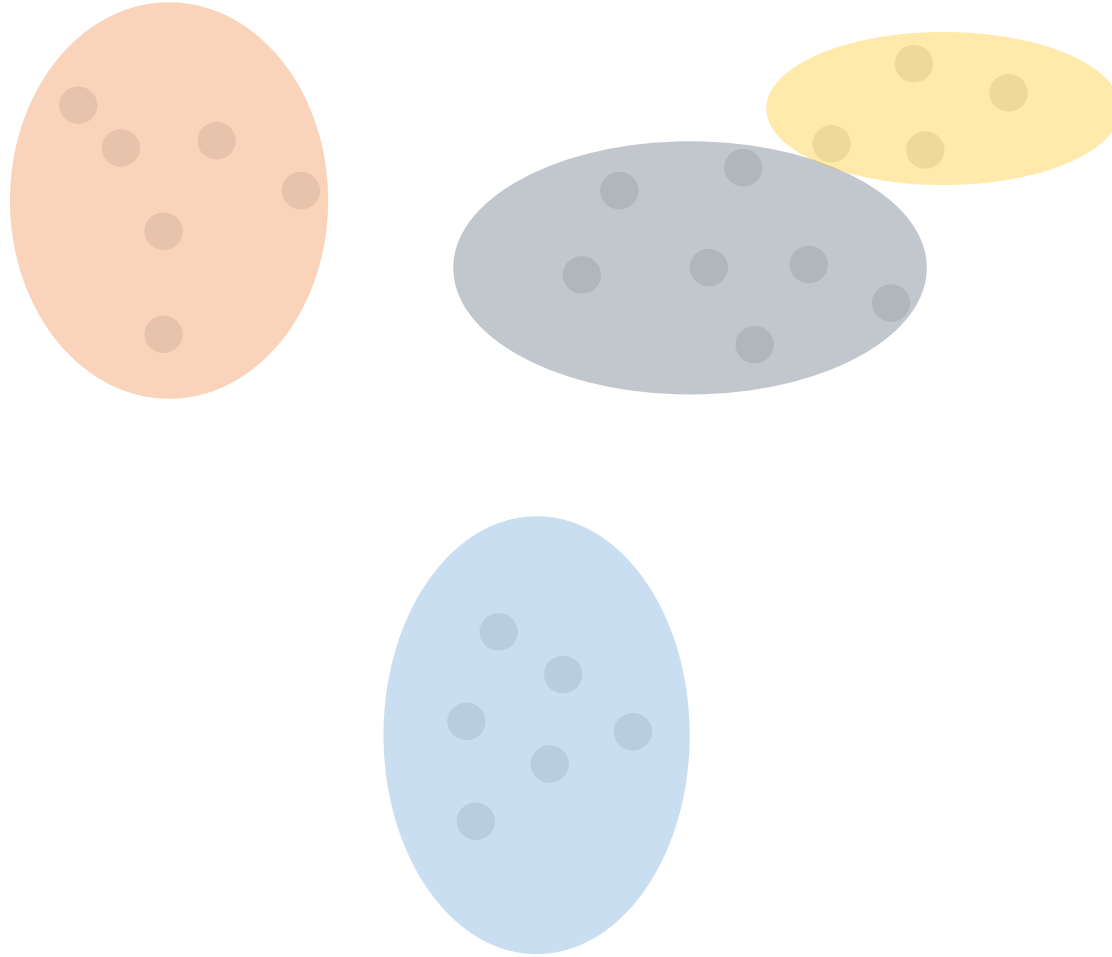
Example (K=3)

$$\sum_{n=1}^N \sum_{k=1}^K r_k^{(n)} ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$$

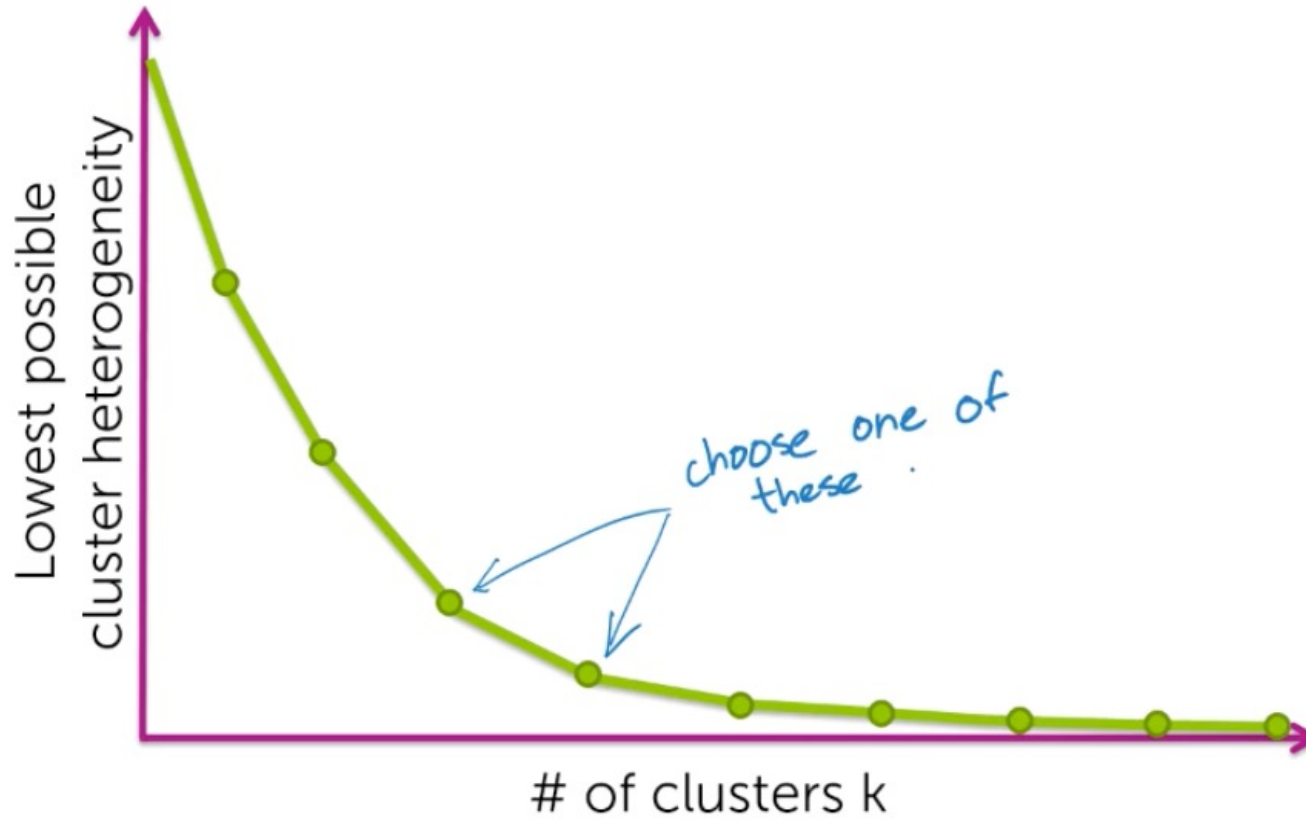


Example (K=4)

$$\sum_{n=1}^N \sum_{k=1}^K r_k^{(n)} ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$$

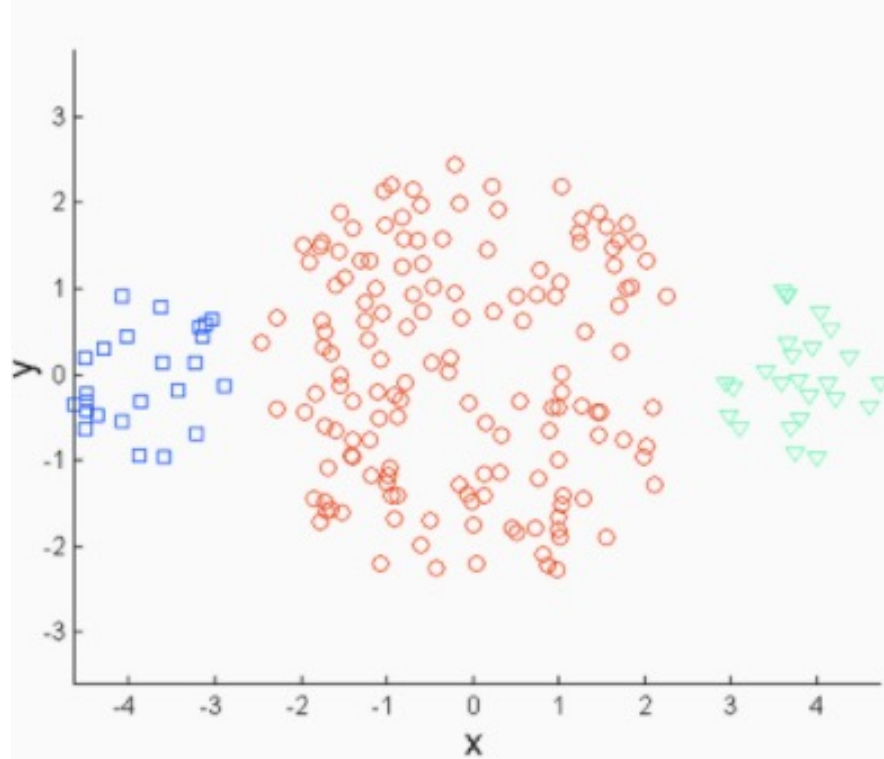


Choosing K – The Elbow Method

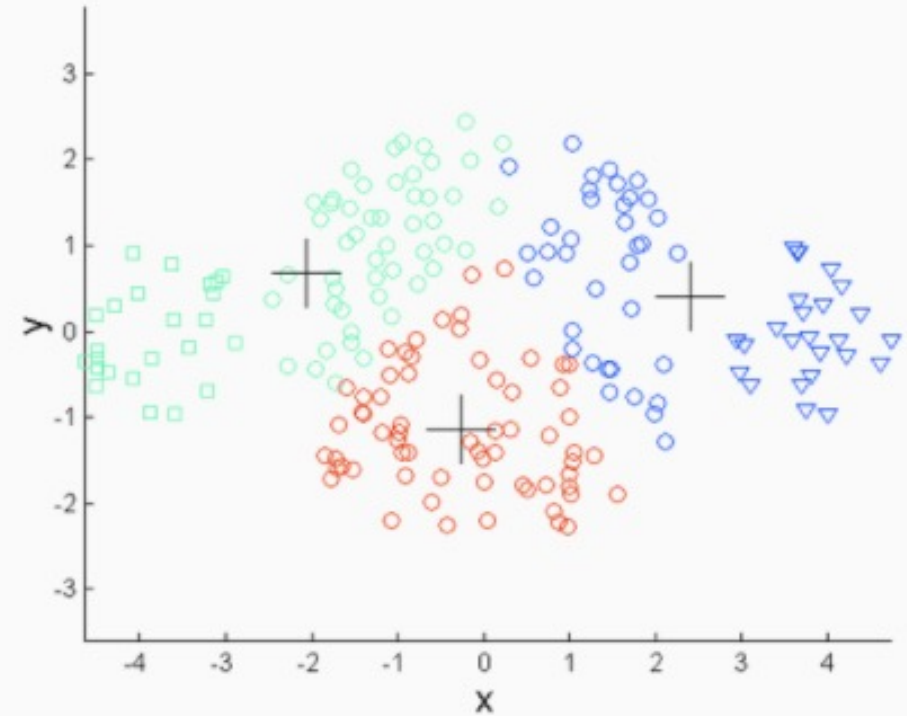


Elbow of the curve

Issues with K-means (1)

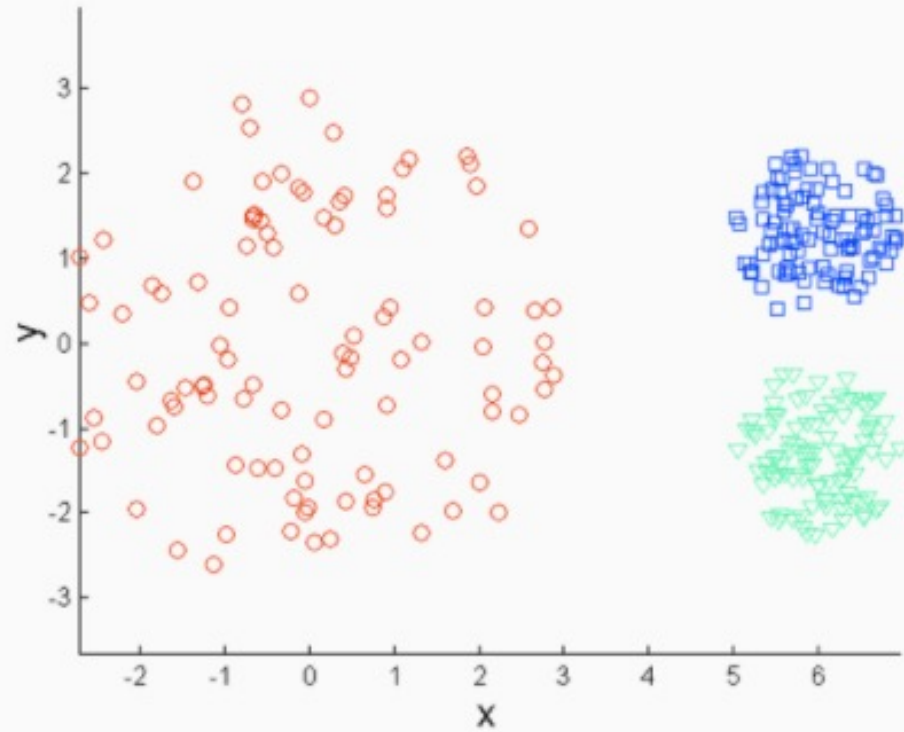


Original Points

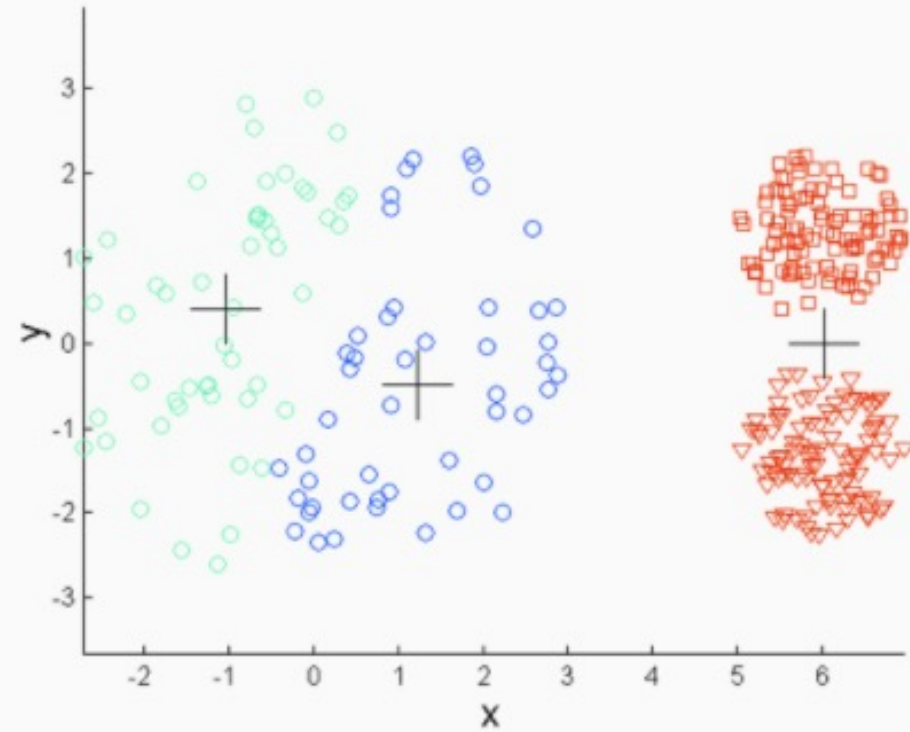


K-means (3 Clusters)

Issues with K-means (2)

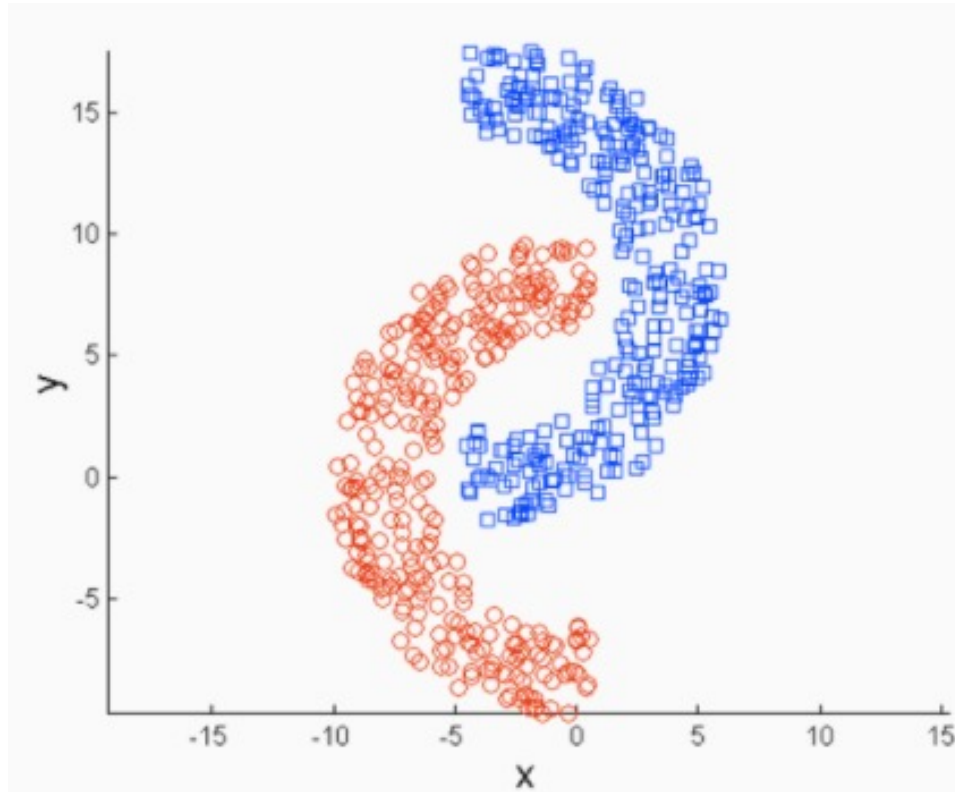


Original Points

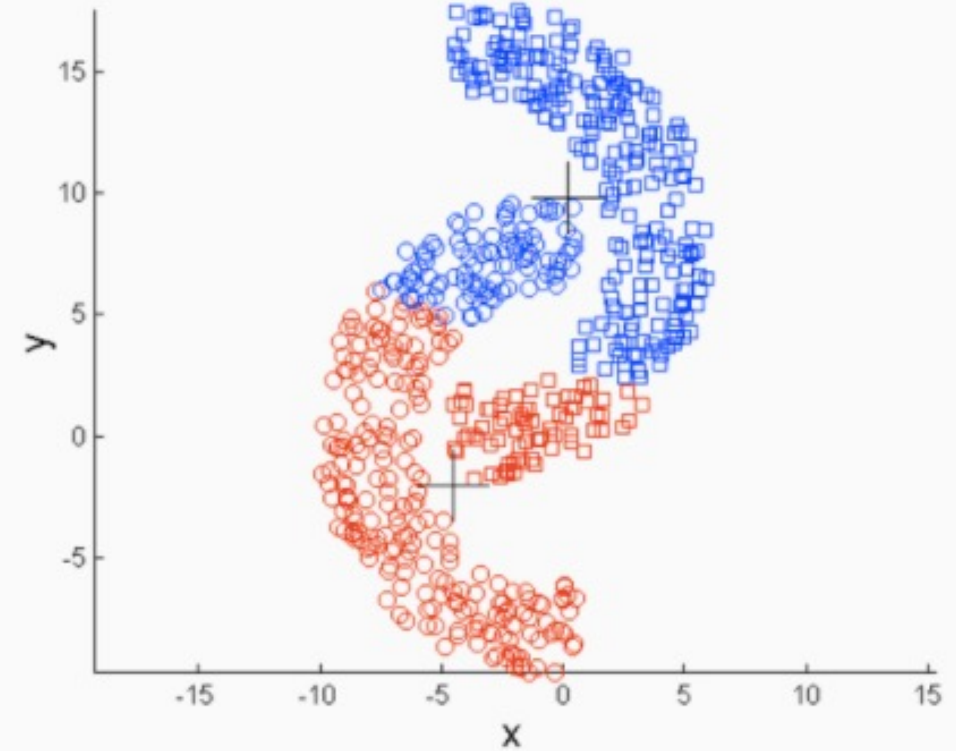


K-means (3 Clusters)

Issues with K-means (3)

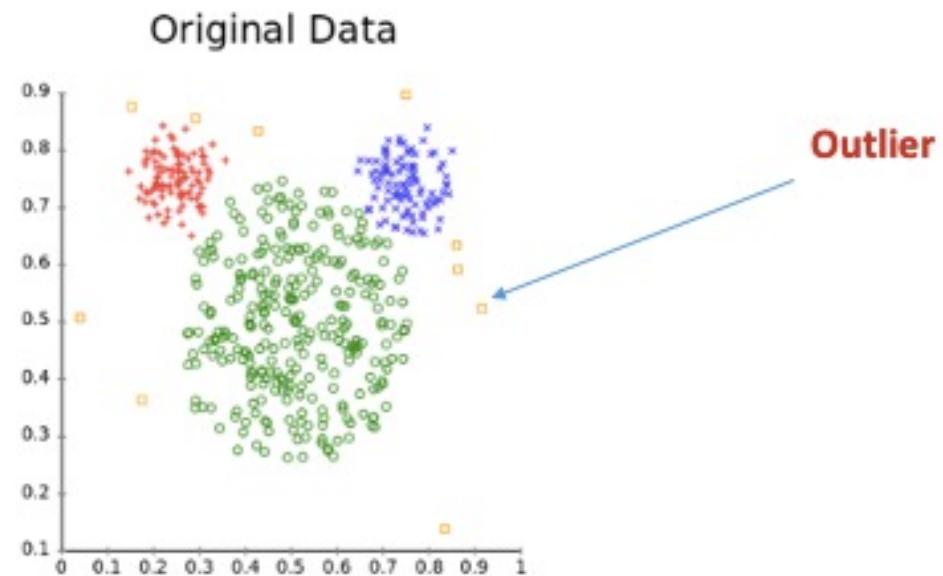


Original Points

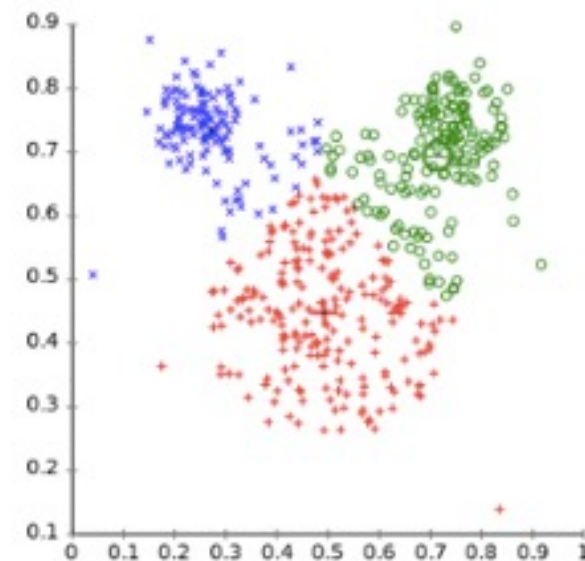


K-means (2 Clusters)

Issues with K-means (4)



k-Means Clustering



Issues with K-means

1. Not knowing the optimum value of K
2. Also, k-means clustering does not work well when
 - We want to discover clusters of varying sizes, densities and shapes
 - We do not want to include noisy points (outliers) into any clusters

Are there any Alternatives?

- Many
- But the two that I want to introduce to you are
 - Hierarchical Clustering
 - Density Based Clustering (DBSCAN)

Summary

1. Clustering: discovering structures/groups in the data
 - An unsupervised learning problem
2. Principle on which clustering works
3. K-means clustering
 - Its objective function
 - How is it motivated
 - How k-means solves its optimization problem
 - Convergence of k-means
 - K-means++
4. Limitations of k-means