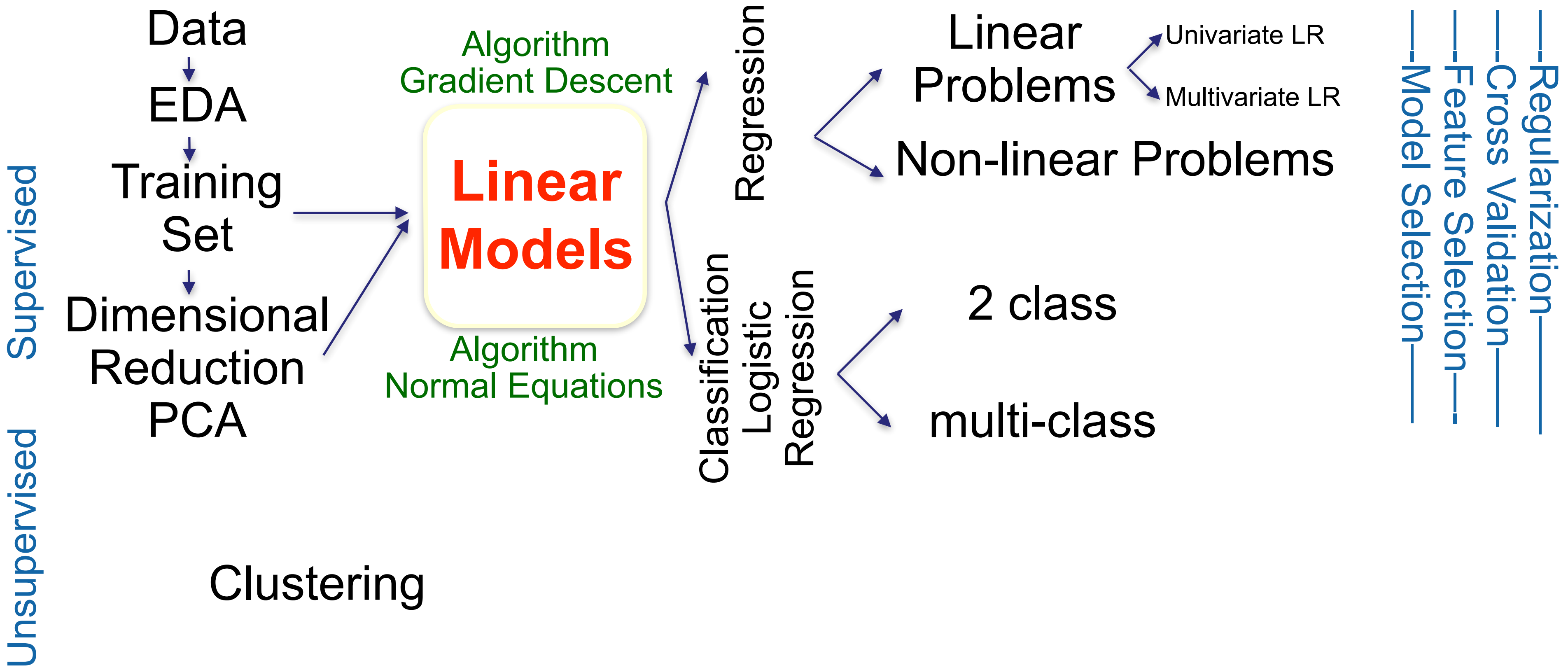
INTRO TO DATA SCIENCE
LECTURE 10:UNSUPERVISED LEARNING

Where are we on the data science road-map?



KEY CONCEPTS - UNSUPERVISED LEARNING

- Unsupervised learning
  - No requirement for labelled data or known outcomes
  - No y

- Is there some underlying structure in the data?
  - Do any sub-populations exist in the data?
  - If so, how many are there?
  - If so, how big are they?
  - What are their common properties?
  - Are there outliers?

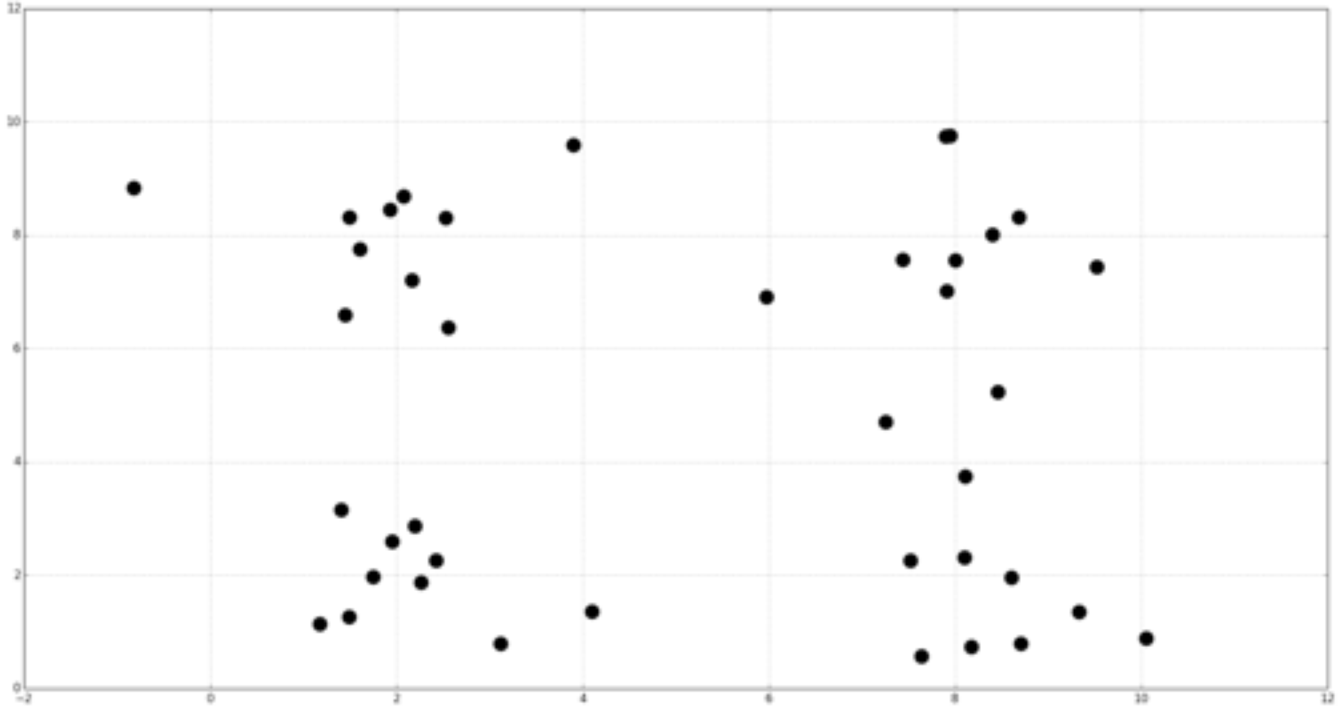
Clustering, or **cluster analysis**, is the task of grouping observations such that members of the same group, or **cluster**, are more similar to each other by some metric than they are to the members of the other clusters

- What is a cluster?

- A cluster comprises a group of data points whose inter-point distances are ‘small’ compared with the distances to points outside the cluster

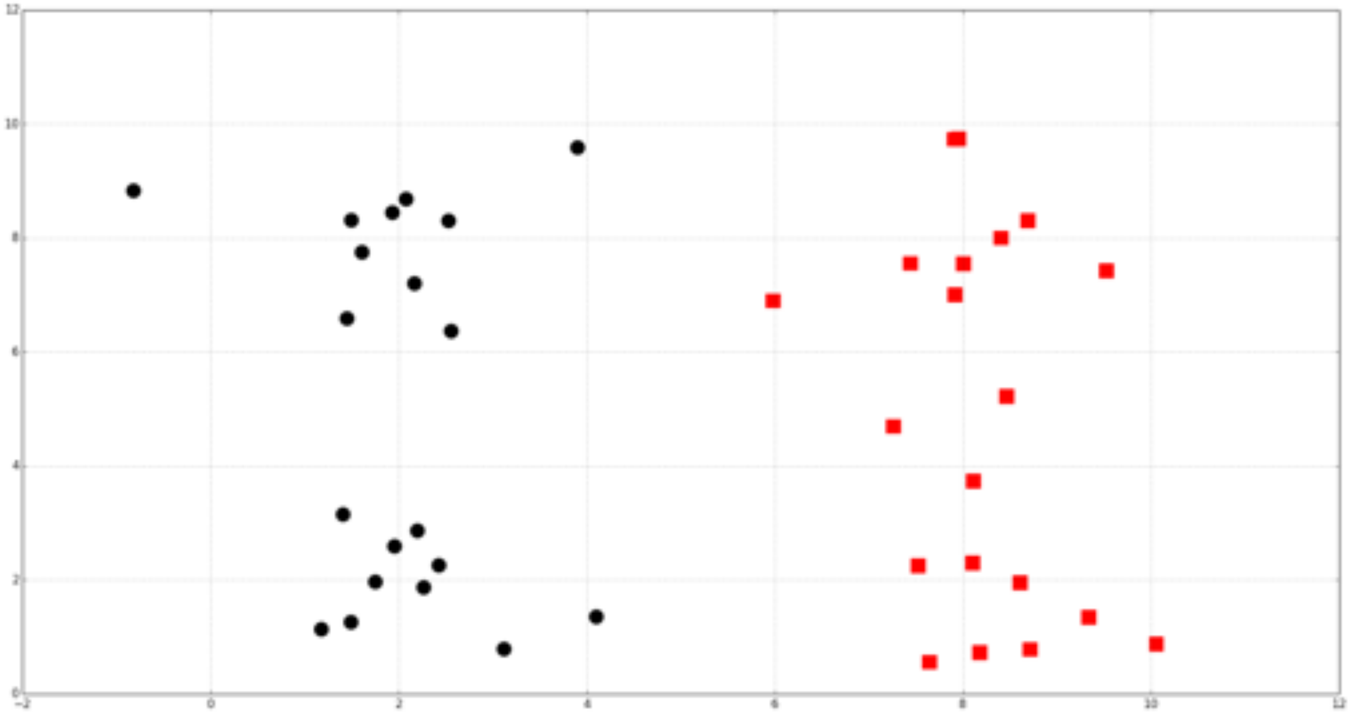
- We need an algorithmic solution to finding these clusters

How many clusters are there?

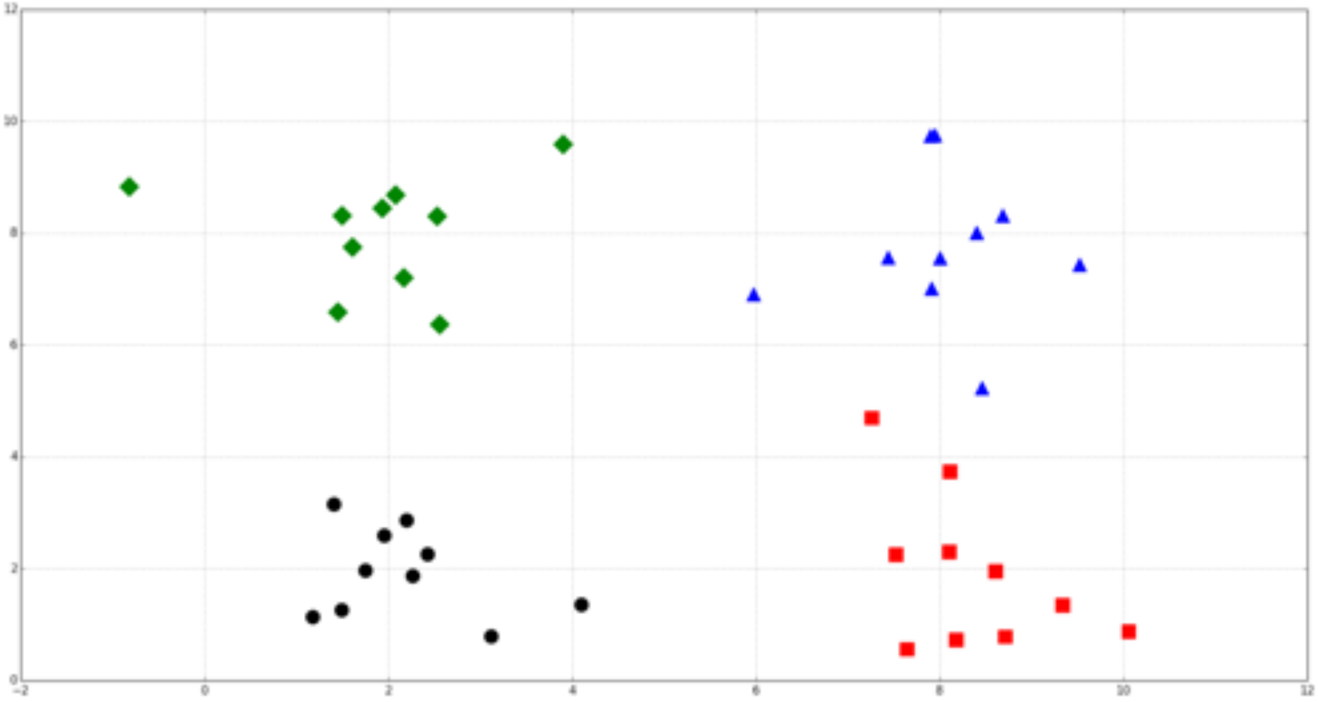


Key Concepts - Clustering

Three?



Four?



Key Concepts - Clustering

- Data exploration
- Identify communities, connections in social networks
- Customer segmentation
- Find groups of genes with similar expression patterns
- Recommendation systems
- Image compression

KEY CONCEPTS - K-MEANS CLUSTERING

The aim of K-means is to partition your dataset, consisting of m observations, each of which has N dimensions, into k clusters.

KEY CONCEPTS - K-MEANS CLUSTERING

- Possibly the most popular clustering algorithm
- It's simple, and quick – but needs to be used correctly
- It is an iterative algorithm
- If used inappropriately can give poor results
- K-means is affected by starting conditions, and therefore, as a result of 'unlucky' starting conditions it can get 'stuck' in local minima

Key Concepts - K-means Clustering Algorithm

Step 1: Choose the number of clusters
- user provided input to the algorithm

Step 2: Initialize the cluster centroids
- the centroids are found locations within the dataset

Step 3: Assign the data points to their “closest” cluster centroid

Step 4: Using the data points associated with each cluster recalculate the cluster centroids

Step 5: Repeat Step 3, 4 & 5 until convergence

Key Concepts - K-means Algorithm Choosing the Locations for the Centroids

- Initialization is best done by selecting some data points from your dataset, at random, and using the location of those points for the initial positions of the cluster centroid

KEY CONCEPTS - K-MEANS ALGORITHM ASSIGNING TO CENTROIDS

- Having set the initial positions for the centroids, each data point is 'assigned' to a centroid based upon a 'distance' measure

- Euclidean Distance is the most commonly used measure of distance

$$\text{Euclidean\_Distance}\ (c, x) = \sqrt{\sum_{i=1}^{N} (c_i - x_i)^2}$$

- Be aware of the Jaccard Similarity Score
- A popular metric in text mining problems or any problems with sparse binary data. E.g. vector may contain a 0 or 1 depending on the absence or presence of a word within a document

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Key Concepts - K-means Algorithm Re-locating the Centroids

- Each centroid now has a portion of the data set assigned to it
- Using the cluster of points each centroid now finds the mean of the cluster.
- This implies finding the mean position in every dimension
- Once the mean is found the algorithm iterates by re-assigning the dataset to the newly located centroids

- K-means has a cost function, which is the optimization objective
- The K-means cost function is known as the K-means Distortion Function

- The K-means cost function is the sum of squared distances of each point to its assigned cluster

$$\sum_{k}^{K} \sum_{i=1, x_i \in C_i}^{m} e\_dist(C_k, x_i)^2$$

KEY CONCEPTS - K-MEANS ALGORITHM CONVERGENCE

- After each iteration the cost function is evaluated
- The algorithm has converged when there is no further reduction in the cost function
- Convergence is assured!!, each iteration will lower the value of J
- But there is no guaranteed global minimum

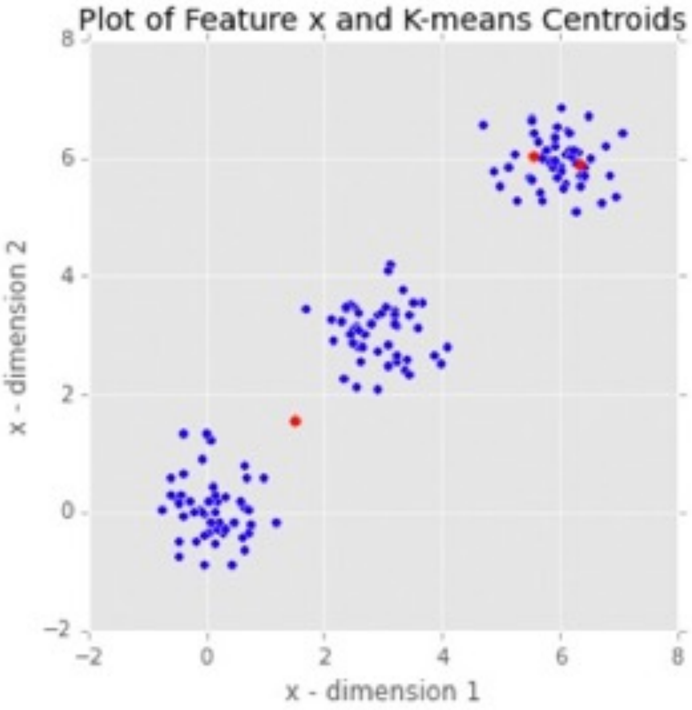
KEY CONCEPTS - K-MEANS ADVANTAGES

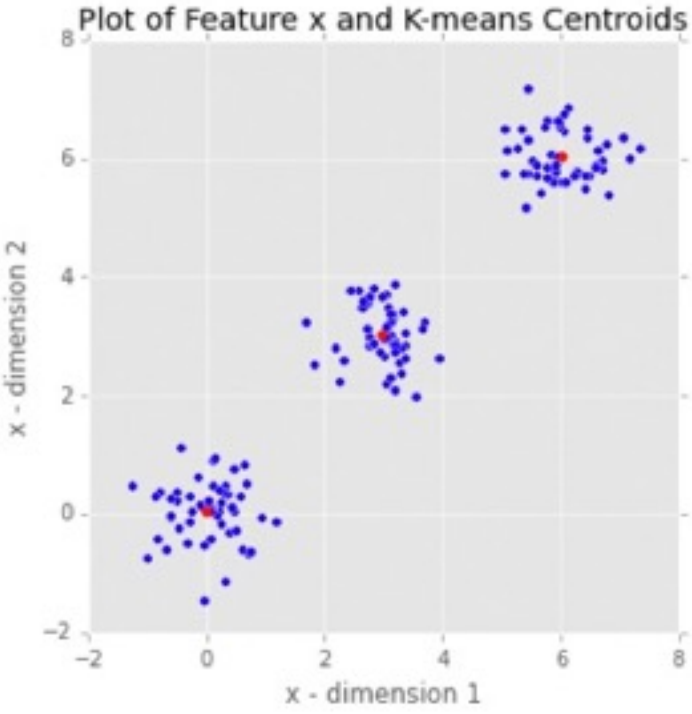
- K-Means is relatively fast
- Can be scaled to large data sets when using mini-batches
- Excellent for general-purpose clustering

KEY CONCEPTS - K-MEANS ADVANTAGES

- The use of euclidean distance makes the algorithm susceptible to outliers
- You have to find a good value for k
- K-means is subject to the local minima problem.
  - An un-lucky initial randomization of centroids may yield a poor clustering result

KEY CONCEPTS - K-MEANS - LOCAL OPTIMA





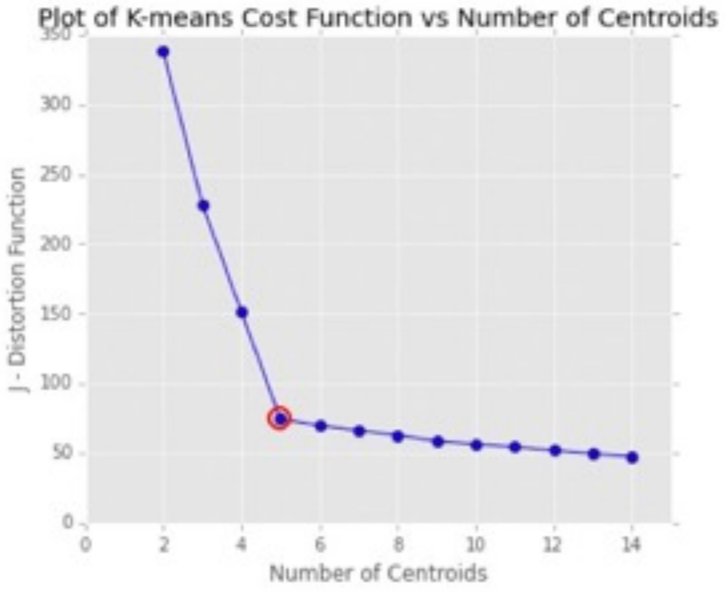
Key Concepts - K-means Algorithm Optimizing - Finding a Good Value for K

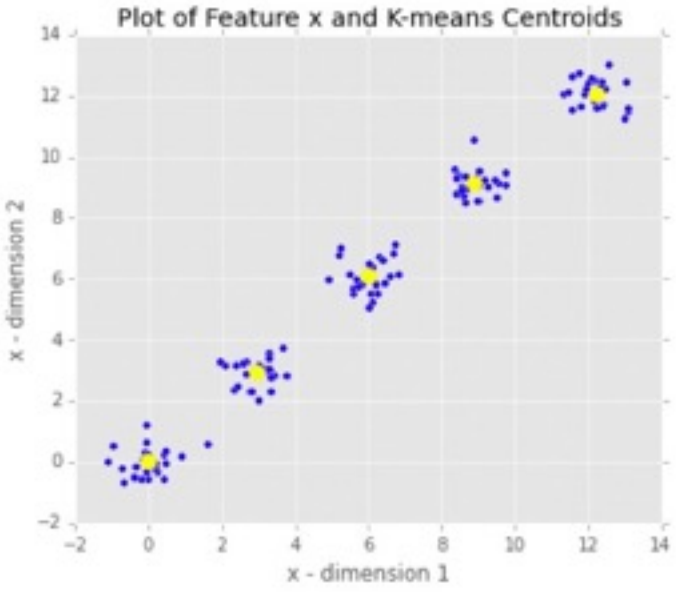
- The elbow (or knee-of-the-curve) method plots the value of the cost function produced by different values of k

- As k increases, the average distances (and hence J) will decrease; each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids

- However, the improvement to J will decline as k increases. The value of k at which the improvement to J declines the most is called the elbow

Key Concepts - K-means Algorithm Optimizing - Finding a Good Value for K





Key Concepts - K-means Algorithm Optimizing - Avoiding Local Minima

- Because K-means is a relatively fast algorithm you can usually run 50, 100, or even 200 runs, each with a different random initialization, so as to avoid poor solutions

- For each choice of K, you would run the algorithm 50, 100 or even 200 times with a different random starting configuration

- This, in general, ‘solves’ the local minima problem

OTHER CLUSTERING ALGORITHMS

- Connectivity models - e.g.Hierarchical
- Distribution models - e.g. Gaussian Mixture Model
- Density models - e.g DBSCAN
- Graph-based models - e.g HCS clustering algorithm