Clustering

PHYS591000 2022.03.08

Outline

- Unsupervised learning and K-means clustering
- Dimensionality reduction: Principal component analysis (PCA)
- Remarks on other clustering algorithms

Warming up

- Access control of the building has been granted.

- As usual, take 3 mins to introduce yourself to your teammate for this week!
  – “Were you OK during the blackout last Thursday?”
  – “Let’s work together to make this week nice and easy!”

Unsupervised learning

- The training data are not **labeled** (no information of the **ground truth** given to the model).

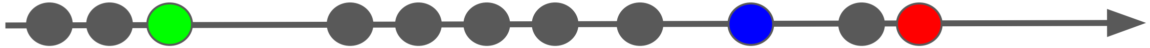
- Physics example: Divide the stars/astronomical objects into different groups according to their similarities given the observation data (Lab for today) – **Clustering**

K-means Clustering

- K-means is one of the ‘classic’ methods for clustering:

  Step 0. Choose K = number of clusters you wish to assign

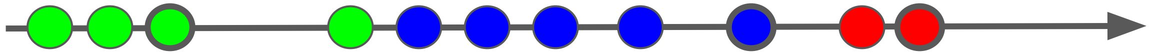
  Step 1. *Randomly* select K data points as centers of initial clusters (“centroids”)



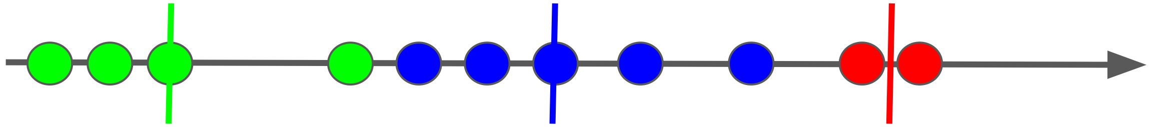
K-means Clustering

- K-means is one of the ‘classic’ methods for clustering:

  Step 2. Assign each point to its closest centroid



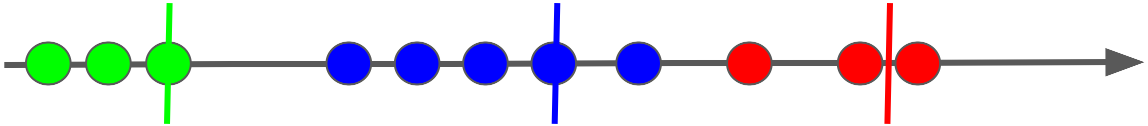
  Step 3. Re-compute the new means (centers) of the clusters



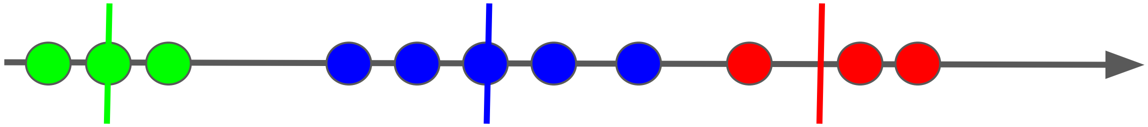
K-means Clustering

- K-means is one of the 'classic' methods for clustering:

  Step 4. Assign each point using the new centers

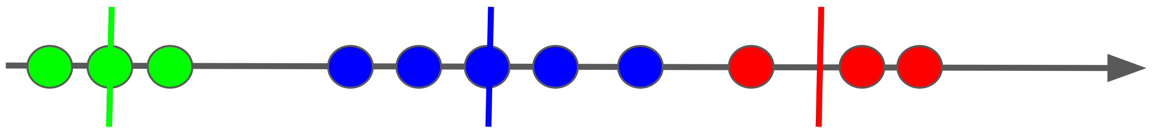


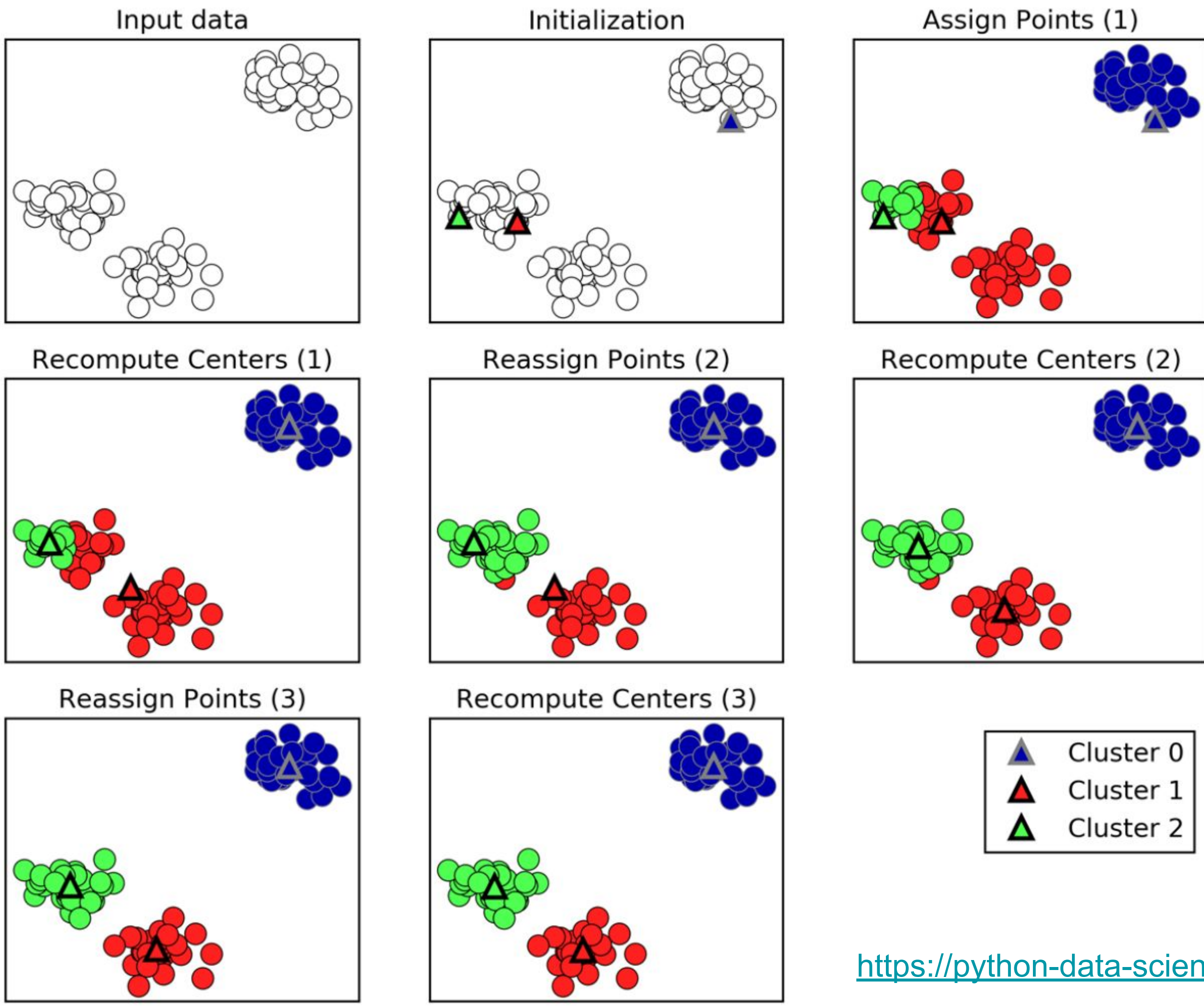
  Step 5. Re-compute the new means (centers) of the clusters



K-means Clustering

- Repeat ‘finding new means’ → ‘reassign points according to new means’ a few times until the centers converge (the positions no longer change).





https://python-data-science.readthedocs.io/

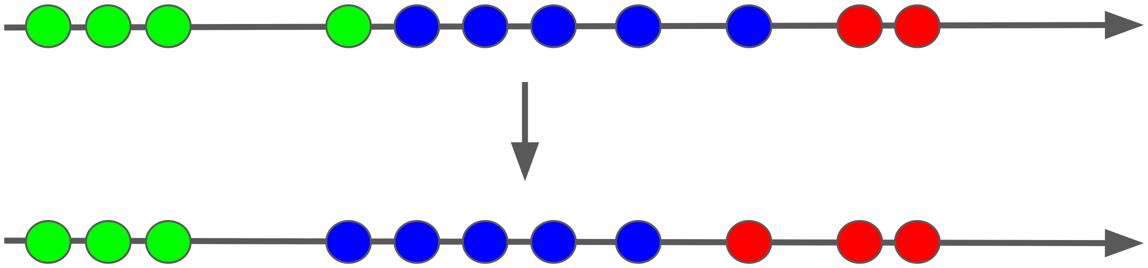
K-means Clustering

- Q: How to find the ‘right’ K (number of clusters)?
  - Domain knowledge
  - Evaluate the performance of different choices of K

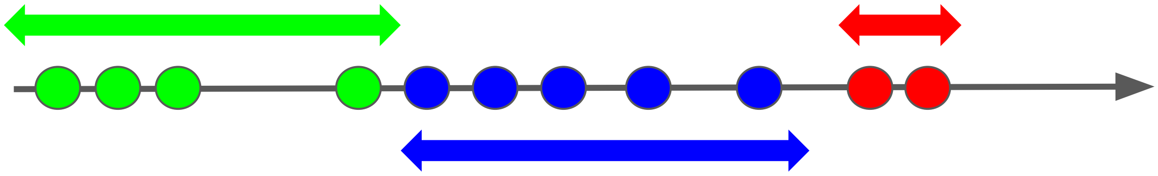
- One metric is the sum of (squared) distances of each point to its closest centroid.

K-means Clustering

- Sum of distances (intuitively, ‘sizes of clusters’) can be a way to evaluate the performance of clustering. Compare the initial and final clustering of the previous case:



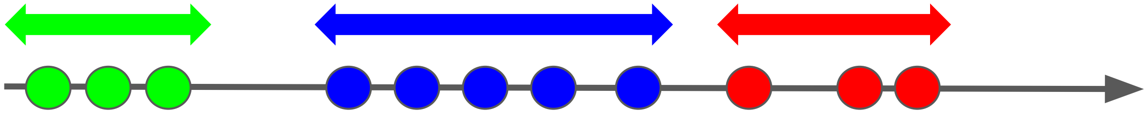
K-means Clustering



Sum of ‘sizes’ of clusters = 

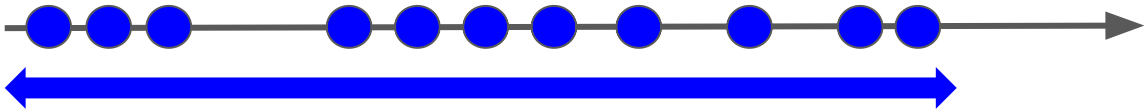
Sum of ‘sizes’ of clusters = 

Considered as
‘better’
for the same K

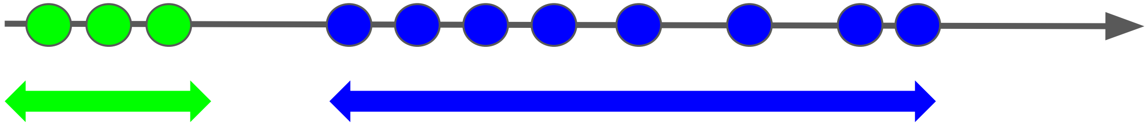


K-means Clustering

- Compare the sum for different K: K=1 is the largest

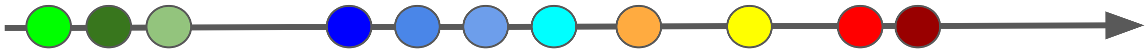


K=2 will make it smaller:



K-means Clustering

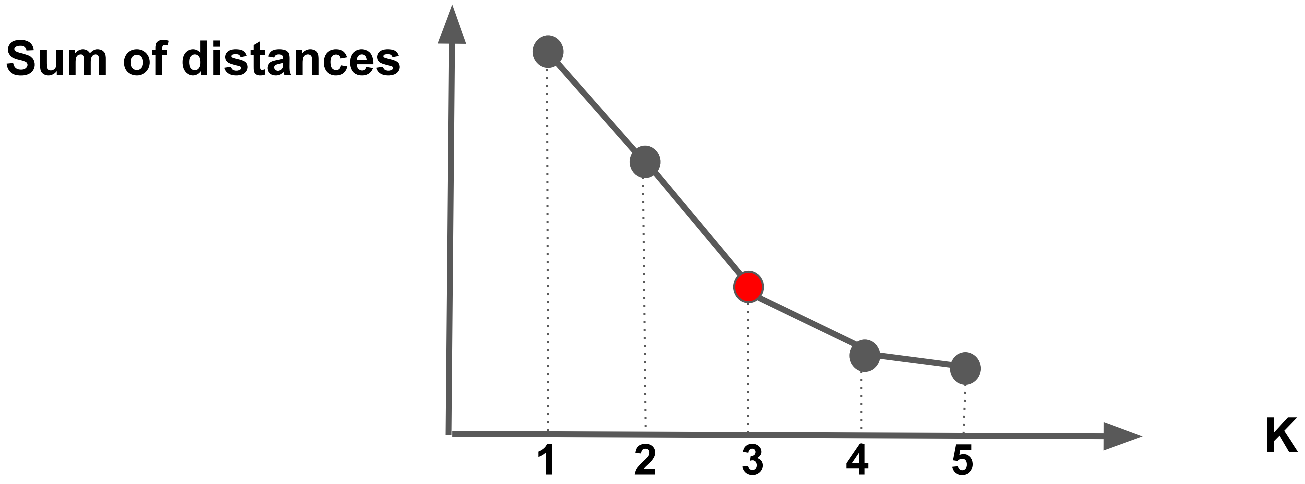
- And the sum will be 0 when K=number of points



Apparently that won't be useful for any purpose…

K-means Clustering: Elbow Plot

- If we plot the sum v.s. K there is usually an ‘elbow’ point beyond which the reduction is less significant.



Data Representation

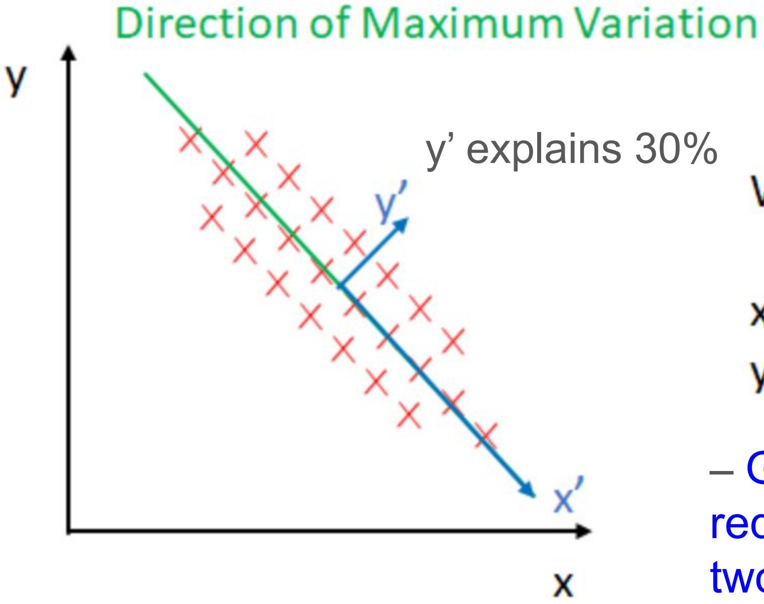
- We often collect a lot of information in our data, e.g. for each star we record 10 quantities we can observe.

- It will be difficult to visualize and/or do clustering in this 10-dimensional feature space!

- We can simplify the data by using a lower dimensional representation → **Dimensionality reduction**

Dimensionality Reduction

- One way is to do a Principal Component Analysis (PCA)



x' explains 70% of the variation

y' explains 30%

With:

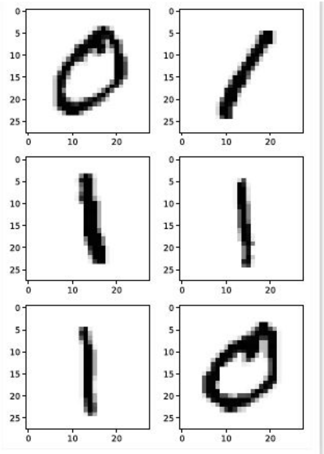
x' → Principal Component 1
y' → Principal Component 2

– Generalize to higher dimensions, e.g. reduce O(10) features to the most important two axes x' and y'

In-class exercise for this week

- Let's turn to the in-class exercise this week: We'll use 0's and 1's from the MNIST dataset again!
    - x_train: 28*28 images; each pixel is associated with a number from 0-255 (~ the amount of 'ink')

- We'll use PCA to plot the data, and do K-means clustering to see if it can separate the data into the two 'correct' clusters.

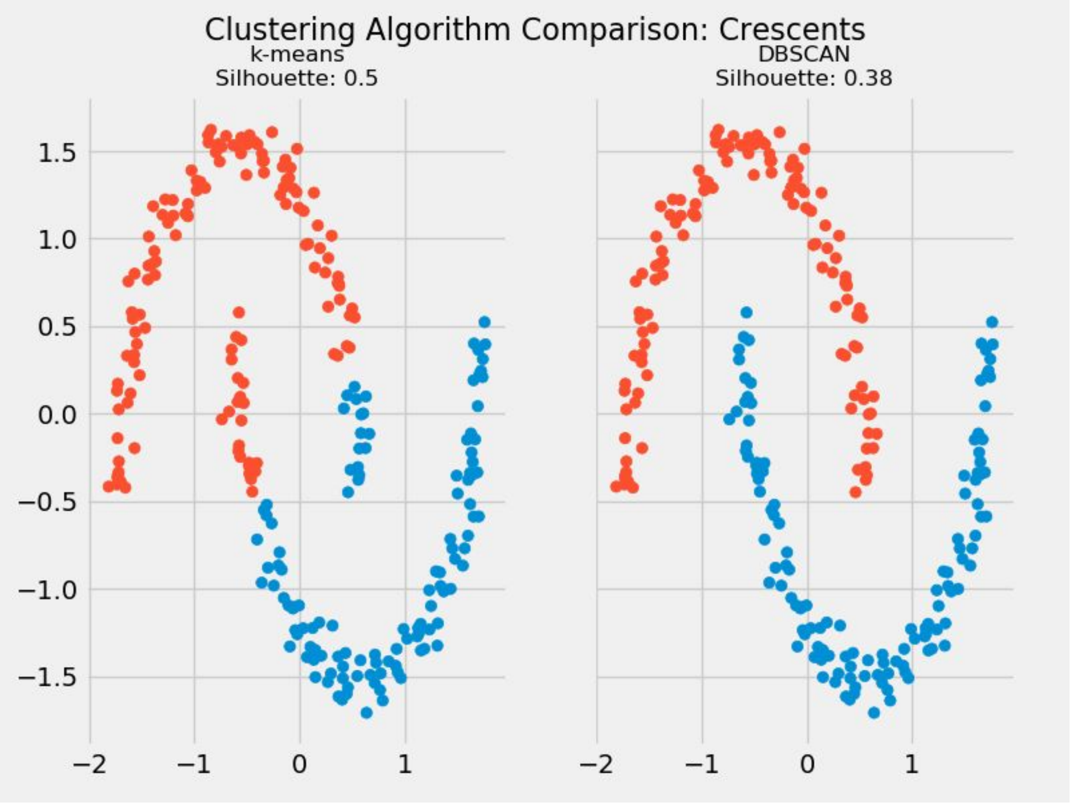


Remarks on other clustering algorithms

- K-Means do not work well for non-spherical clusters.
- Instead one can use density-based algorithms e.g. DBSCAN

Remarks on other clustering algorithms

K-Means



DBSCAN

Lab for this week

- For the Lab this week, we’ll use astrophysics data from "Spitzer From Molecular Cores to Planet-Forming Disks (C2D)" project.

- It contains spectrum energy data of three kinds of astronomical objects. We’ll do
    - K-means clustering (unsupervised learning)
    - KNN classification (supervised learning)