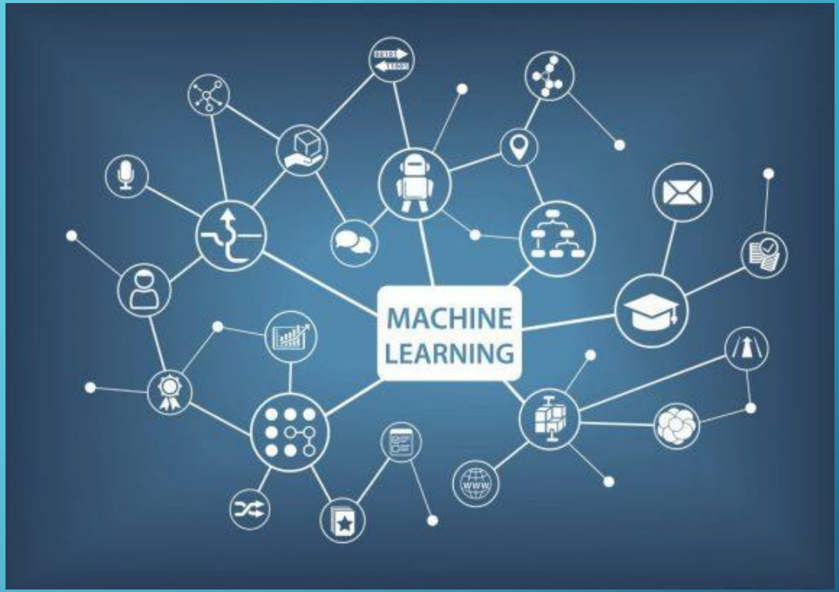
CS5056
TIME SERIES CLUSTERING



HÉCTOR G. CEBALLOS, FRANCISCO J. CANTU

CEBALLOS@TEC.MX, FCANTU@TEC.MX

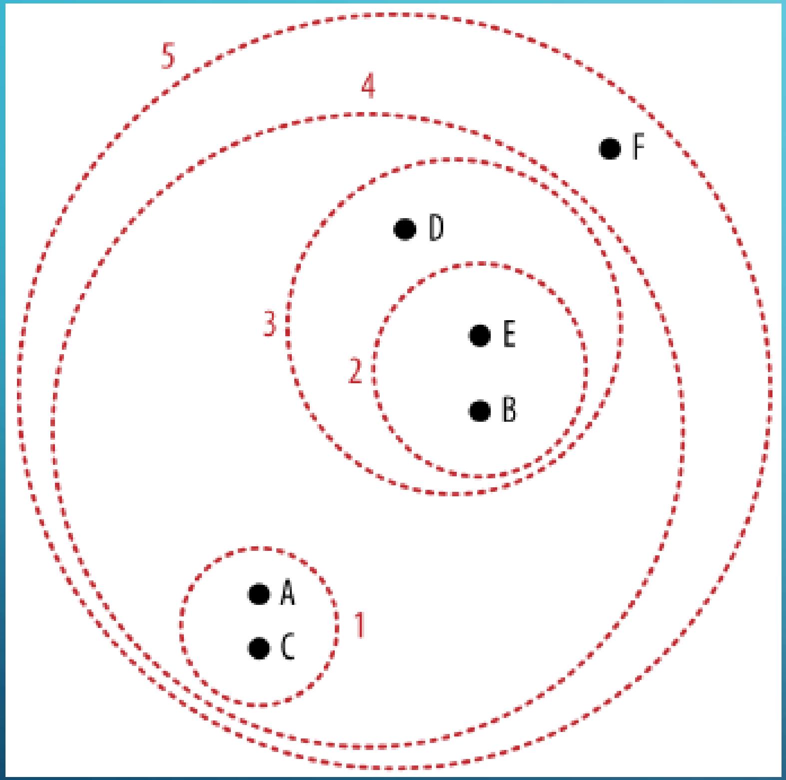
Image: https://www.mba-madrid.com/empresas/impacto-del-machine-learning-ambito-empresarial/

THEORY

TIME SERIES CLUSTERING

Image: https://www.muycomputerpro.com/2018/02/07/glosario-terminos-basicos-machine-learning

HIERARCHICAL CLUSTERING

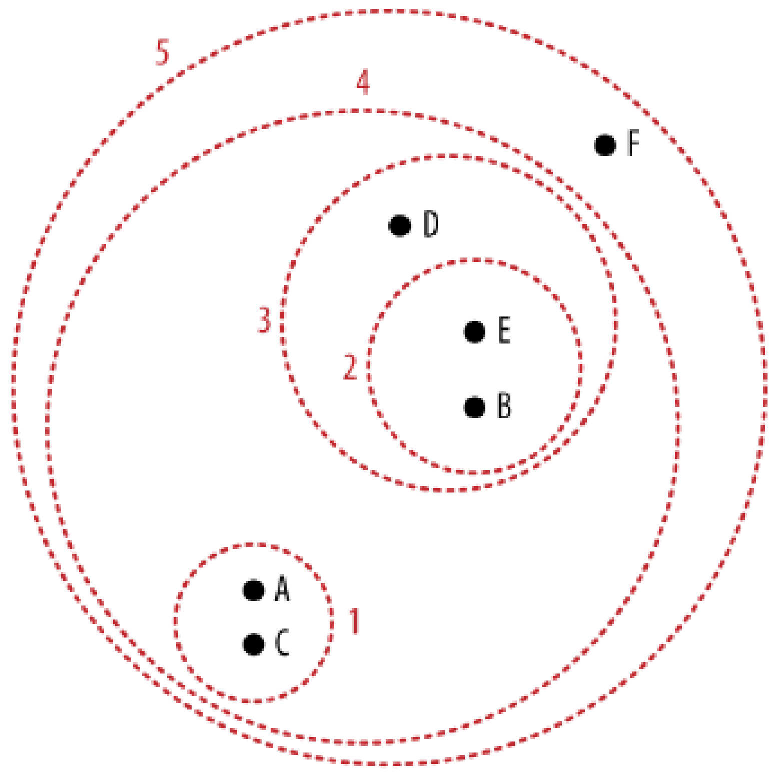


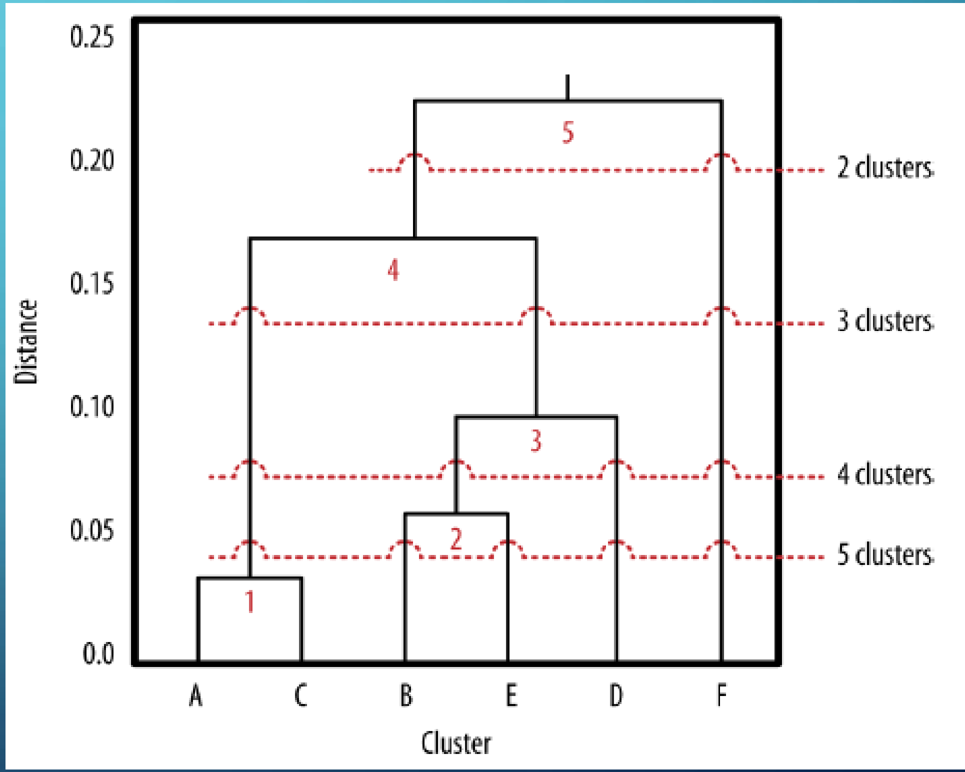
6 objects to classify

- Looks for finding a set of **clusters**, where elements of each **cluster** is distinct from elements of each other **cluster**, and the elements within each **cluster** are broadly similar to each other.

DENDOGRAMS

- Height of horizontal lines indicates how distant is a set of objects from other





SINGLE-LINKAGE CLUSTERING

- Agglomerative (bottom-up) clustering method.
- At each step combines two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other.
- Drawback: nearby elements of the same cluster have small distances, but elements at opposite ends of a cluster may be much farther from each other than two elements of other clusters.

https://en.wikipedia.org/wiki/Single-linkage_clustering

SINGLE-LINKAGE CLUSTERING
METHOD

- In the beginning, each element is in a cluster of its own.
- The clusters are then sequentially combined into larger clusters, until all elements end up being in the same cluster.
- At each step, the two clusters separated by the shortest distance are combined.
- The function used to determine the distance between two clusters, known as the *linkage function*, is what differentiates the agglomerative clustering methods.

SINGLE-LINKAGE CLUSTERING
DISTANCE FUNCTION

- In single-linkage clustering, the distance between two clusters is determined by a single pair of elements: those two elements (one in each cluster) that are closest to each other.

Mathematically, the linkage function – the distance $D(X,Y)$ between clusters X and Y – is described by the expression

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y),$$

where X and Y are any two sets of elements considered as clusters, and $d(x,y)$ denotes the distance between the two elements x and y.

LINKAGE CLUSTERING

SCIPY.CLUSTER.HIERARCHY.LINKAGE

- method='single' assigns

$$d(u,v) = \min(dist(u[i], v[j]))$$

  for all points i in cluster u and j in cluster v. This is also known as the Nearest Point Algorithm.
- method='complete' assigns

$$d(u,v) = \max(dist(u[i], v[j]))$$

  for all points i in cluster u and j in cluster v. This is also known by the Farthest Point Algorithm or Voor Hees Algorithm.
- method='average' assigns

$$d(u,v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)}$$

  for all points i and j where $|u|$ and $|v|$ are the cardinalities of clusters u and v, respectively. This is also called the UPGMA algorithm.
- method='weighted' assigns

$$d(u,v) = (dist(s,v) + dist(t,v))/2$$

  where cluster u was formed with cluster s and t and v is a remaining cluster in the forest (also called WPGMA).
- method='centroid' assigns

$$dist(s,t) = ||c_s - c_t||_2$$

  where c_s and c_t are the centroids of clusters s and t, respectively. When two clusters s and t are combined into a new cluster u, the new centroid is computed over all the original objects in clusters s and t. The distance then becomes the Euclidean distance between the centroid of u and the centroid of a remaining cluster v in the forest. This is also known as the UPGMC algorithm.
- method='median' assigns $d(s,t)$ like the `centroid` method. When two clusters s and t are combined into a new cluster u, the average of centroids s and t give the new centroid u. This is also known as the WPGMC algorithm.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

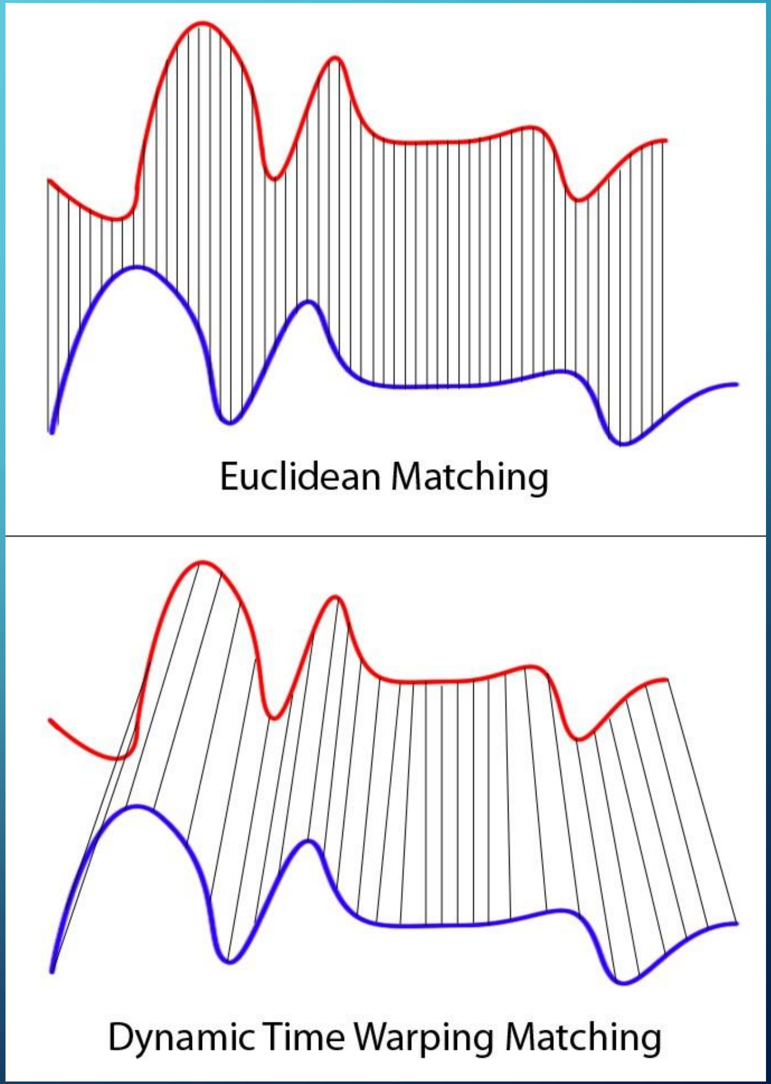
DYNAMIC TIME WARPING

- Dynamic Time Warping (DTW) is one of the algorithms for measuring similarity between two temporal sequences, which may vary in speed.
- DTW has been applied to temporal sequences of video, audio, and graphics data.

https://towardsdatascience.com/dynamic-time-warping-3933f25fcdd

DYNAMIC TIME WARPING

- The idea to compare arrays with different length is to build one-to-many and many-to-one matches so that the total distance can be minimized between the two.
- These two series follow the same pattern, but the blue curve is longer than the red.



DYNAMIC TIME WARPING

- DTW is calculated as the squared root of the sum of squared distances between each element in X and its nearest point in Y. Note that DTW(X, Y) ≠ DTW(Y, X).

$$DTW(x, y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2}$$

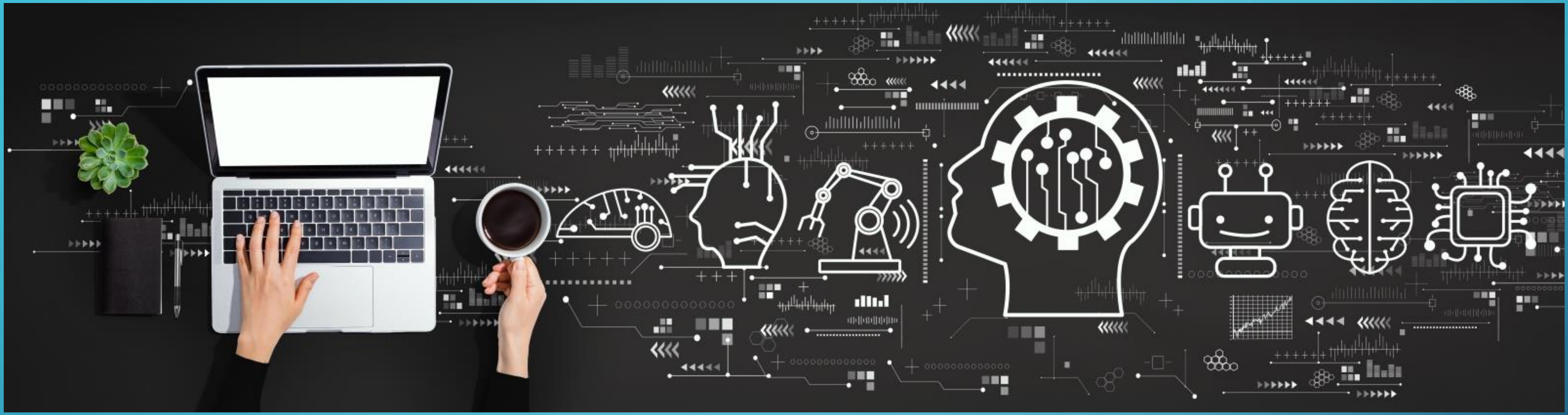
where $\pi = [\pi_0, \dots, \pi_K]$ is a path that satisfies the following properties:

- it is a list of index pairs $\pi_k = (i_k, j_k)$ with $0 \le i_k < n$ and $0 \le j_k < m$
- $\pi_0 = (0, 0)$ and $\pi_K = (n-1, m-1)$
- for all $k > 0$, $\pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows:
  - $i_{k-1} \le i_k \le i_{k-1} + 1$
  - $j_{k-1} \le j_k \le j_{k-1} + 1$

https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3

TIME SERIES CLUSTERING

- In clustering, time series become the data points.
- A distance metric is used for determining if two time series must be grouped together.
- Any clustering method based on metrics can be used, e.g. Linkage and K-means.



PRACTICE

DATA SERIES CLUSTERING

Credits: https://expertsystem.com/machine-learning-definition/

DATA SERIES CLUSTERING
HANDS-ON EXERCISE

- Generate 6 time series
- Apply linkage clustering to detect the 6 clusters
- Use Pearson and Spearman correlation as distance metric
- Plot the dendogram
- Using the dendogram, predict the composition of clusters on k = [2,3,4,5,7,8]
- Use DTW for linkage clustering.
- Jupyter Notebook: TimeseriesClustering.ipynb