キーメアン

k-means

-CLUSTERING-

optimize your SUSHI DELIVERY!

ABOUT K-MEANS

This unsupervised algorithm allows you to tackle problems without having any idea of what the results will look like. It enables an in-depth understanding of data, **revealing hidden patterns.**

ABOUT CLUSTERS

Clustering is a type of data mining techniques that analyzes data to **create groups based on patterns** generated by various criteria, such as minimum distances, or density of data points.

WHY WE NEED K-MEANS CLUSTERING

Everyone loves sushi. You're about to open three sushi stores in Milan and you need to figure out the **best location** in the city to **minimize the delivery distances**.

Here, the algorithm analyzes the geographical coordinates of the sushi demands and provides the best solution according to their location.

This algorithm is widely used in customer segmentation, anomaly detection, and delivery path optimization.

I CASE STUDY – *Legend*

*Find the **stores' best location** to **minimize** the **delivery distance***

3 x Stores to open (CENTROIDS)

- **Store 1** *(Centroid C_1 of the cluster1)*
- **Store 2** *(Centroid C_2 of the cluster2)*
- **Store 3** *(Centroid C_3 of the cluster3)*

13 x Delivery request (DATASET)

II INITIALIZATION *of the algorithm*

#1 COMPUTE DISTANCE

In this algorithm, clusters are **about minimizing distances** between **datapoints** (**customers**) and **centroids** (**stores**). So, you need to represent the unlabelled data relying on **specific numeric features**

*In this case, we are working with **geographical coordinates**, using **km** as a unit of measure of **distance**. Here is the datacloud in which every point represents a sushi delivery demand in Milan*

#2 INITIALIZE CENTROIDS

- First, you need to **define** the **number of clusters** into which divide the dataset.
- Then, **randomly initialize centroids** (**stores**): they must be far enough apart (or may have computational problems).

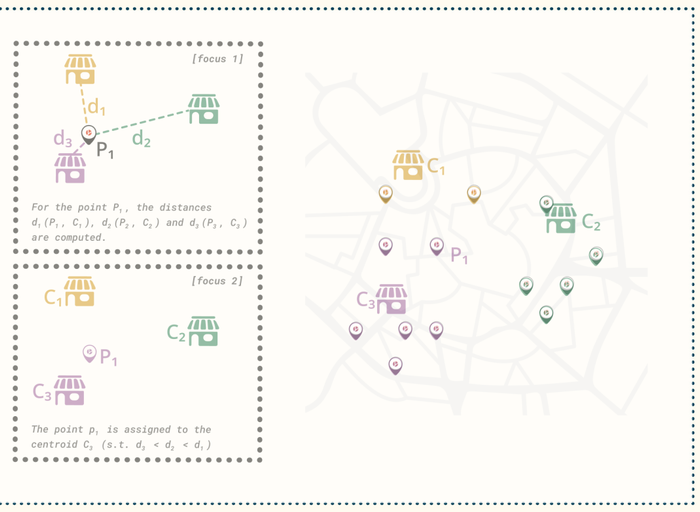
*The sushi company wants to open **three new shops** and you needs to **find suitable locations** to optimise the delivery service by minimising the distance between the shops and customer orders.*

III ITERATION PROCESS: *core of the algorithm*

The algorithm **iteratively calculates distances** to **assign** the **closest datapoints** to the **same cluster**: ITERATION NUMBER → it.1

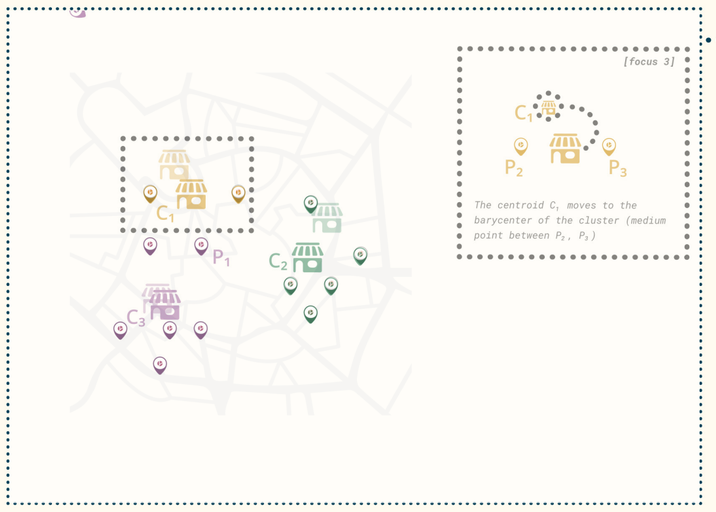
#3 CLUSTER ASSIGNMENT

- **Calculate the distance** between datapoints and centroids ***[focus 1]***
- **Assign** each **datapoint** to the minimum distant **centroid** ***[focus 2]***



#4 MOVE CENTROIDS

Reassign the centroids by computing the **barycentre** of the new clusters ***[focus 3]***



#5 OPTIMIZATION

SPIN

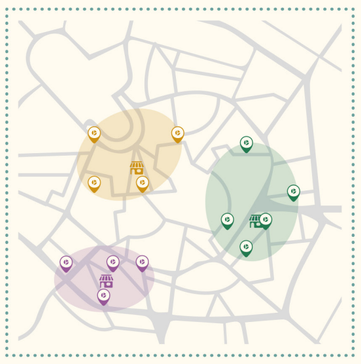
Spin if datapoints have been reasigned (iteration #3-#4)

Stop if the datapoints no longer reassigned.

Warning! To reset the wheel, spin to it. 1

IV OUTPUT

#6 FINAL OUTPUT



Congratulations, you **successfully divided** the **dataset** into **3 clusters**!

As you can see, there are 3 areas that have the highest demand for sushi, and since the **centroids** are now **static**, you just found out which are the **best locations** in Milan to **open new sushi stores!**

* HOW MANY STORES? *Threshold method*

*What if you don't know how many store to open? Try the ***Threshold method***, a process to **compute the optimal number of clusters**

WHAT IT IS

It plots the dispersion of the data points and translates it into a mathematical function.

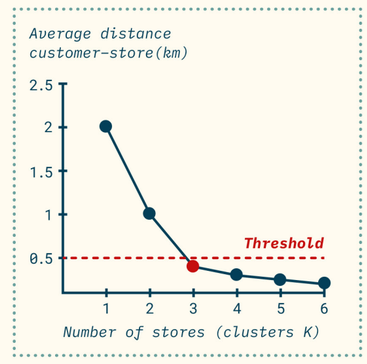
- ***X axis*** *= number of stores*
- ***Y axis*** *= average distance customer – store (km)*

The result is a curve where: **more stores = less distance** (curve distortion decreases)

HOW IT WORKS

The sushi company has decided that the **best customer-store distance** is **0,500 km**

*So, you should **trace** a **threshold** around that measure on the graph*



RESULT

The point where the line intersects with the curve (limit of the **threshold**) represents the **optimal number of stores** to divide the deliveries into

*Here you can also see that after a certain number of clusters, the kilometres remain almost unchanged: **non-optimal solutions***

VISUAL EXPLANATIONS OF STATISTICAL METHODS

K-means clustering

AUTHORS

Silvia Altamura
Ana Doric
Wanlin Li
Jessica Moreschi
Martina Paggi
Matteo Repetto
Lucrezia Valentini

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Arianna Bellantuono

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2022/2023

DEN-SITY GN+

POLITECNICO MILANO 1863
SCHOOL OF DESIGN