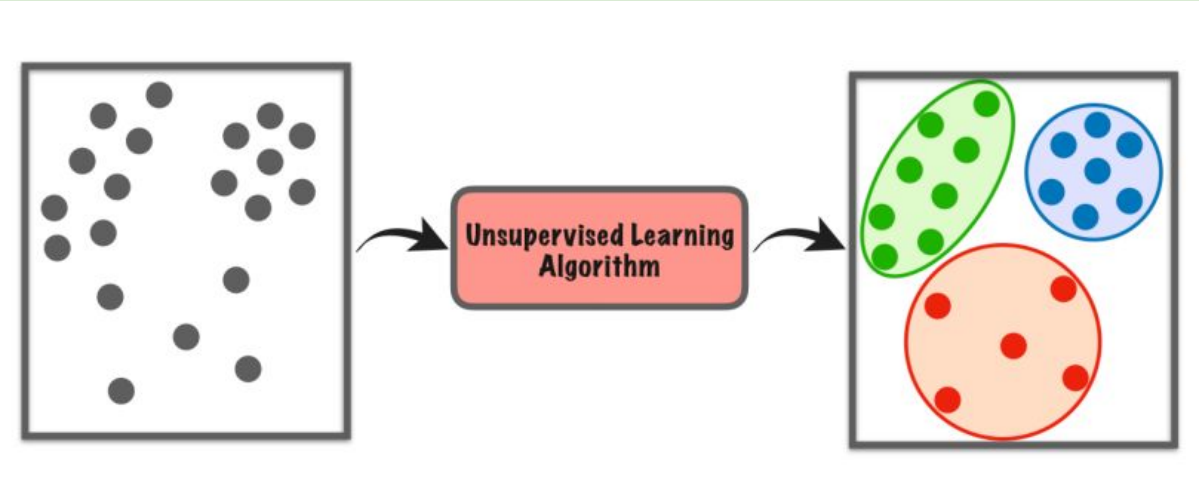
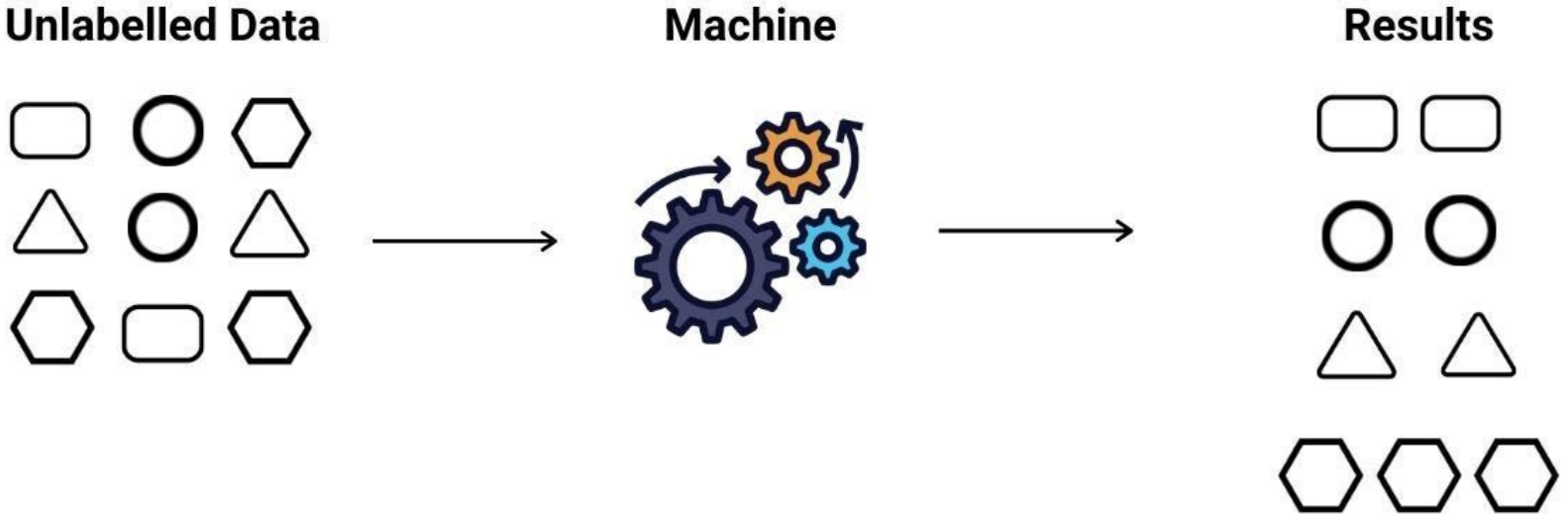
Aprendizaje No Supervisado

Mariana Pereyra

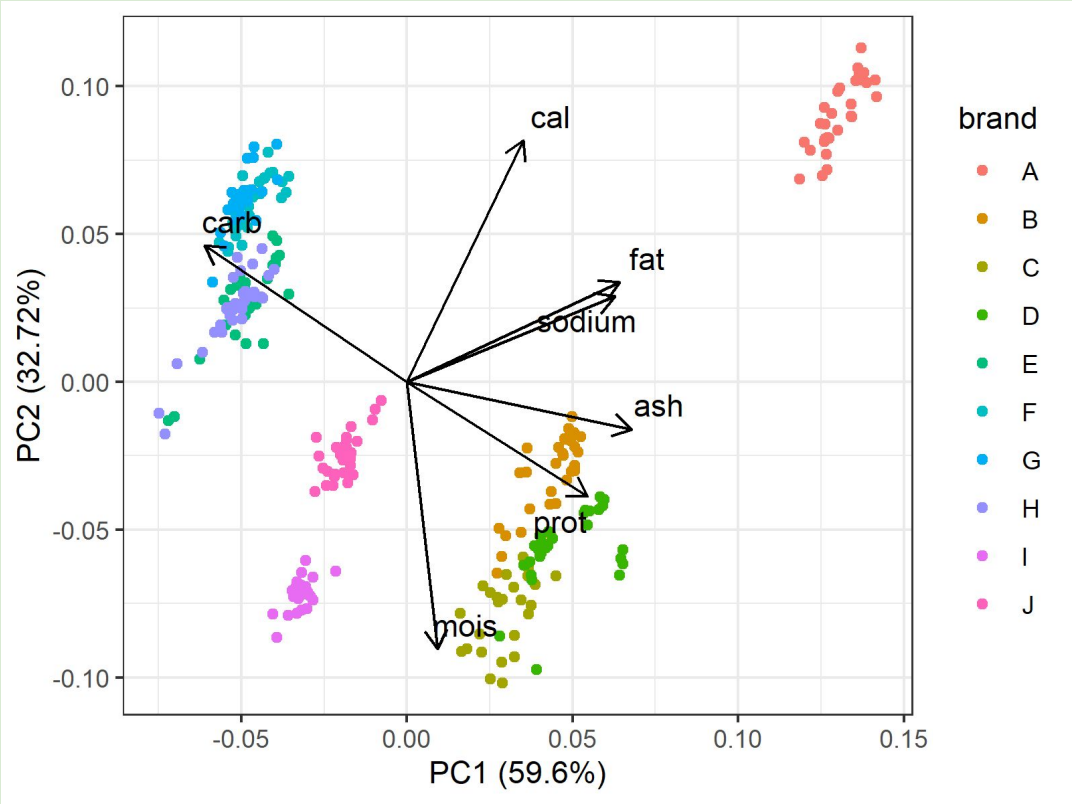
El **aprendizaje no supervisado** es un tipo de problema de aprendizaje automático en el que los datos de entrenamiento consisten en un conjunto de vectores de entrada pero sin los valores target correspondientes. La idea que subyace a este tipo de aprendizaje es agrupar la información basándose en similitudes, patrones y diferencias.



Unsupervised Learning



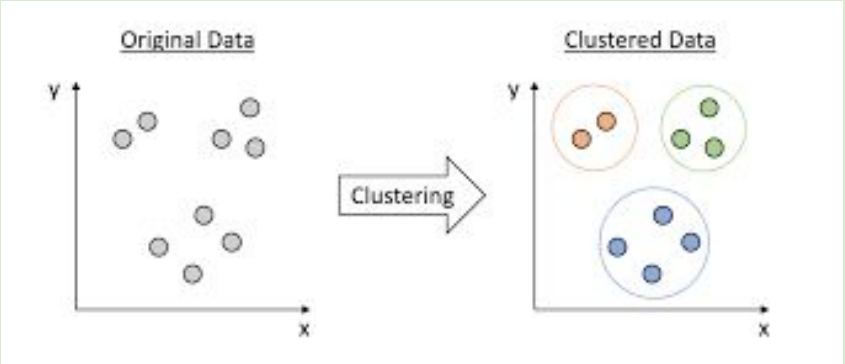
Reducción de dimensionalidad



Clustering

Los métodos de clustering consisten en agrupar datos no etiquetados en función de sus similitudes y diferencias. Cuando dos instancias aparecen en grupos diferentes, podemos inferir que tienen propiedades disímiles.

- DBSCAN
- K-Means
- Gaussian Admixture

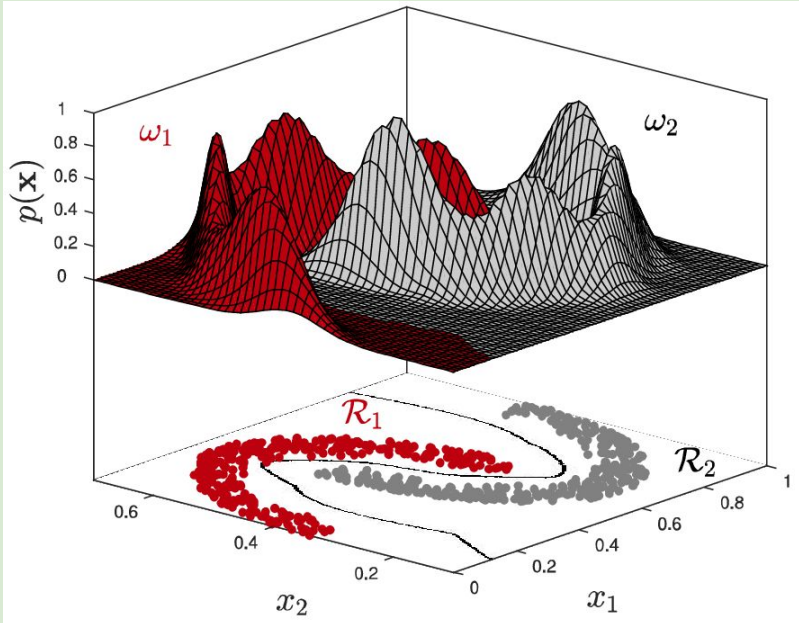


Gaussian admixture (Mezcla de Gaussianas)

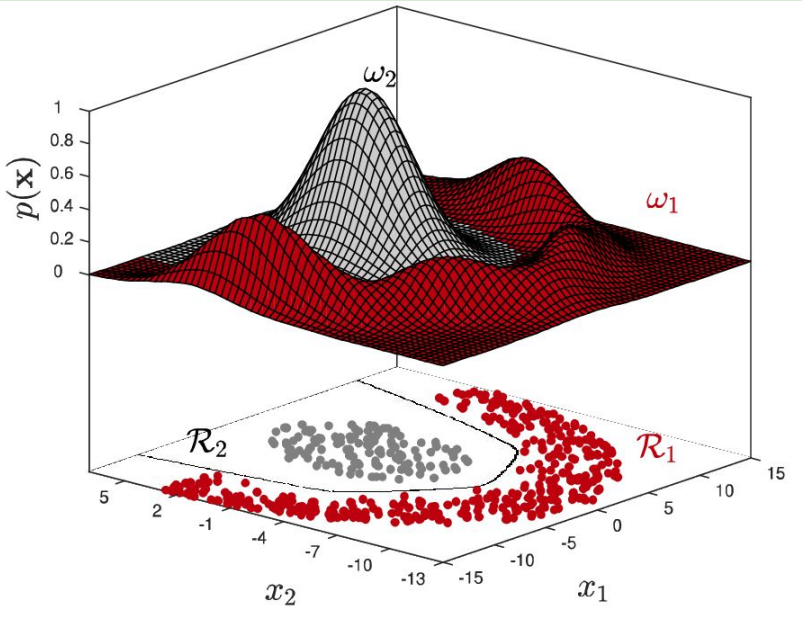
Datos numéricos (reales), producidos por una densidad mezcla de Gaussianas.

Cómo funciona:

- Se fija la cantidad de gaussianas
- Se estiman los parámetros de cada gaussiana
- Se asigna c/dato como proveniente de una de las componentes de la mezcla.
- La estimación se realiza mediante el algoritmo Expectation Maximization.


ω_1
ω_2
$p(\mathbf{x})$
1
0.8
0.6
0.4
0.2
0
$\mathcal{R}_1$
$\mathcal{R}_2$
0.6
0.4
0.2
x_2
0
0.2
0.4
0.6
0.8
1
x_1

(a)


ω_2
ω_1
$p(\mathbf{x})$
1
0.8
0.6
0.4
0.2
0
$\mathcal{R}_2$
$\mathcal{R}_1$
5
2
-1
-4
-7
-10
-13
x_2
-15
-10
-5
0
5
10
15
x_1

(b)

Parámetros

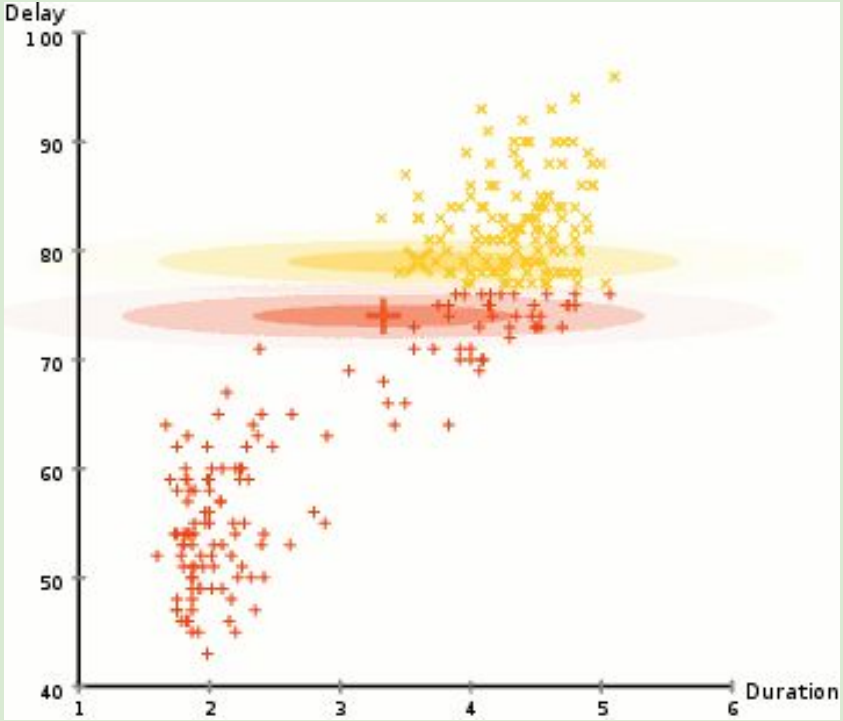
Gran problema de GMM es la determinación del número de componentes de la mezcla.

Si no se elige un buen número, el modelo parte de forma aglutinada pero los clusters pueden no tener sentido.

La otra característica que puede ser forzada de inicio es el tipo de matriz de varianza covarianza:

- 'full' (each component has its own general covariance matrix)
- 'tied' (all components share the same general covariance matrix)
- 'diag' (each component has its own diagonal covariance matrix)
- 'spherical' (each component has its own single variance).

Comenzamos con una partición aleatoria de la cual se sacan los parámetros de inicio y desde allí se itera.



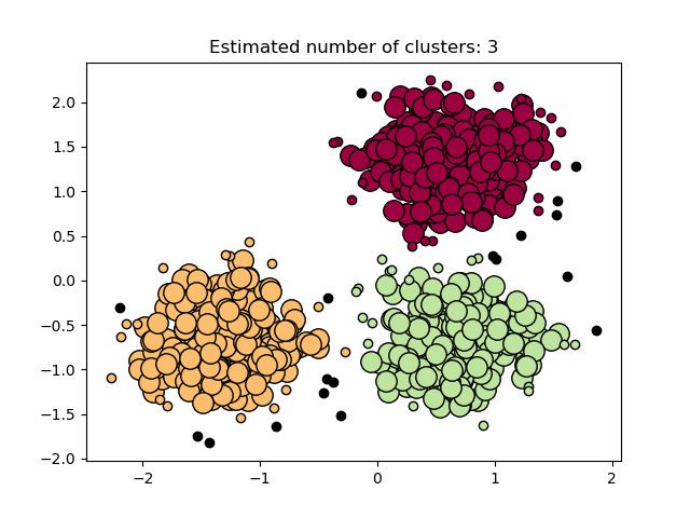
DBSCAN (Density-Based Clustering of Applications with Noise)

DBScan es un algoritmo de agrupación basado en la densidad. El hecho clave de este algoritmo es que la vecindad de cada punto de un conglomerado que se encuentre dentro de un radio determinado (R) debe tener un número mínimo de puntos (M). Este algoritmo ha demostrado ser extremadamente eficaz en la detección de valores atípicos y el tratamiento del ruido.

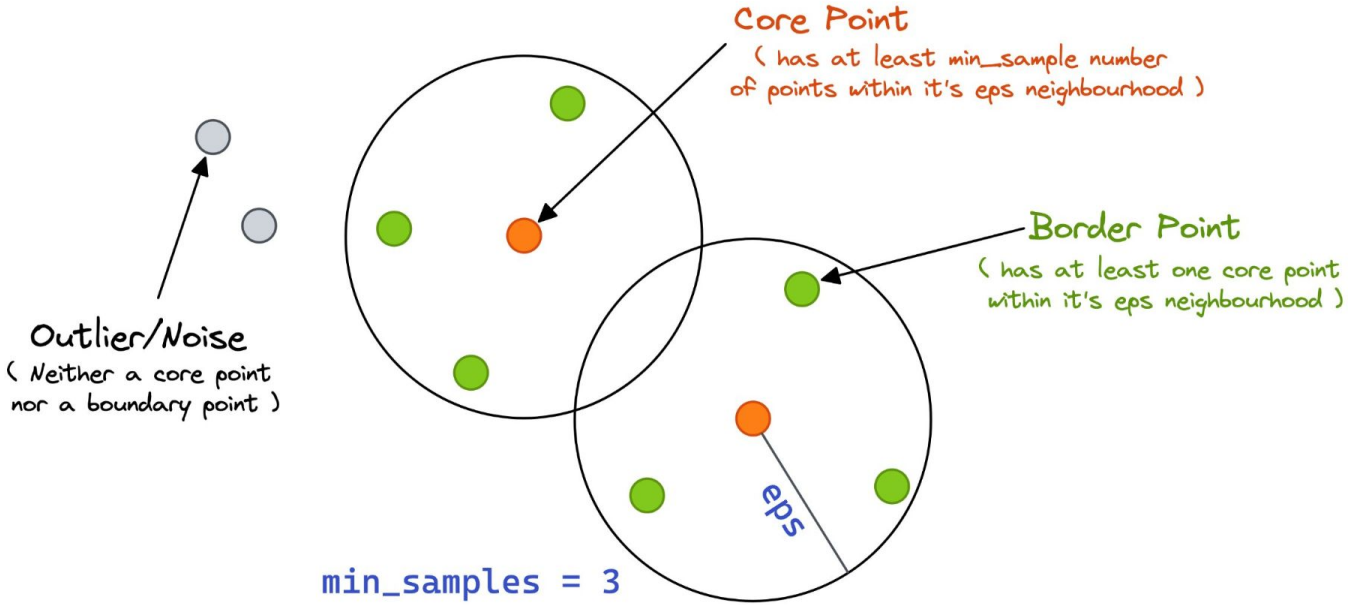
Parámetros:

- min_samples
- eps

Un min_samples más alto o un eps más bajo indican una mayor densidad necesaria para formar un conglomerado.



DBSCAN



@akshay_pachaar

K-MEANS

K-means es un algoritmo de agrupación basado en el centroide o en la partición. Este algoritmo divide todos los puntos del espacio muestral en K grupos de similitud. La similitud se suele medir utilizando la distancia euclídea.

Algoritmo:

Se colocan aleatoriamente K centroides, uno por cada cluster.

Se calcula la distancia de cada punto a cada centroide.

Cada punto de datos se asigna a su centroide más cercano, formando un cluster.

Se recalcula la posición de los K centroides.

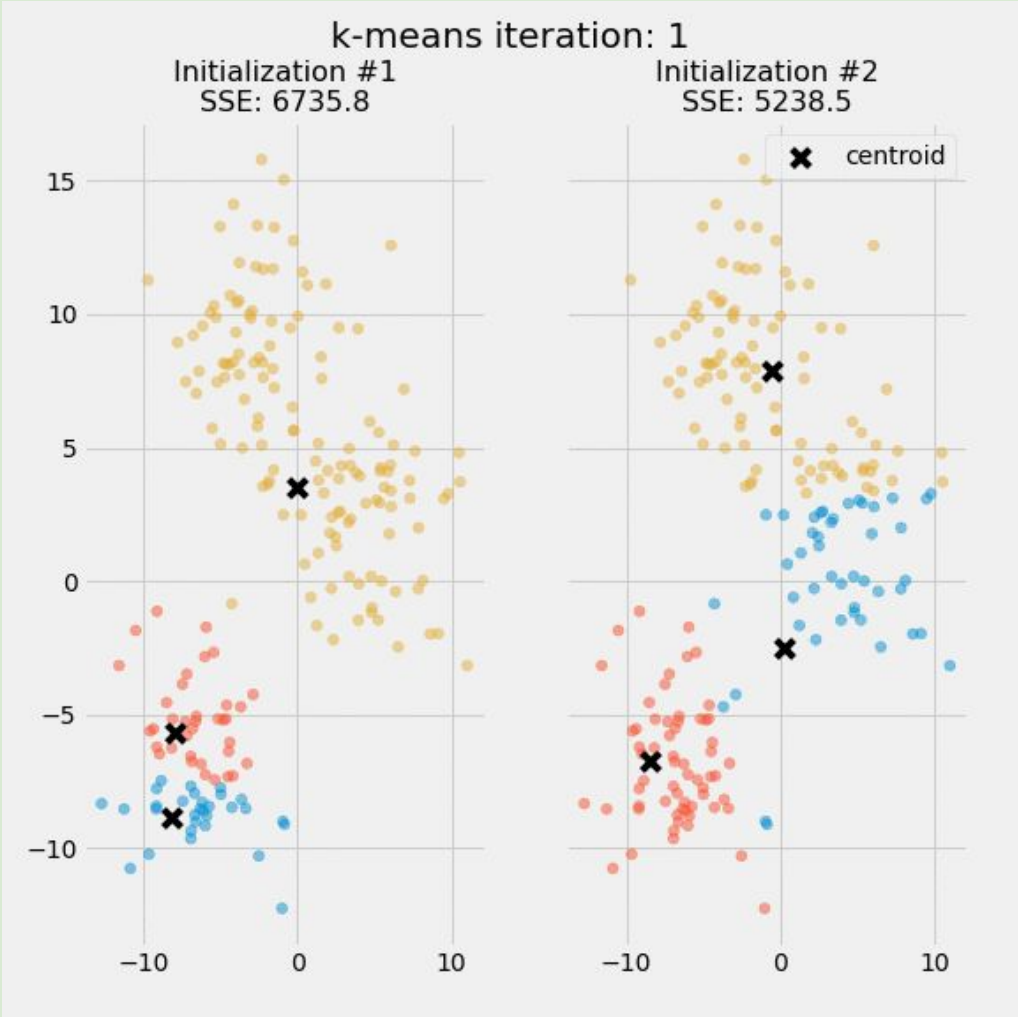
El objetivo del algoritmo K-means es elegir los centroides que minimicen la inercia, o el criterio de suma de cuadrados dentro del clúster:

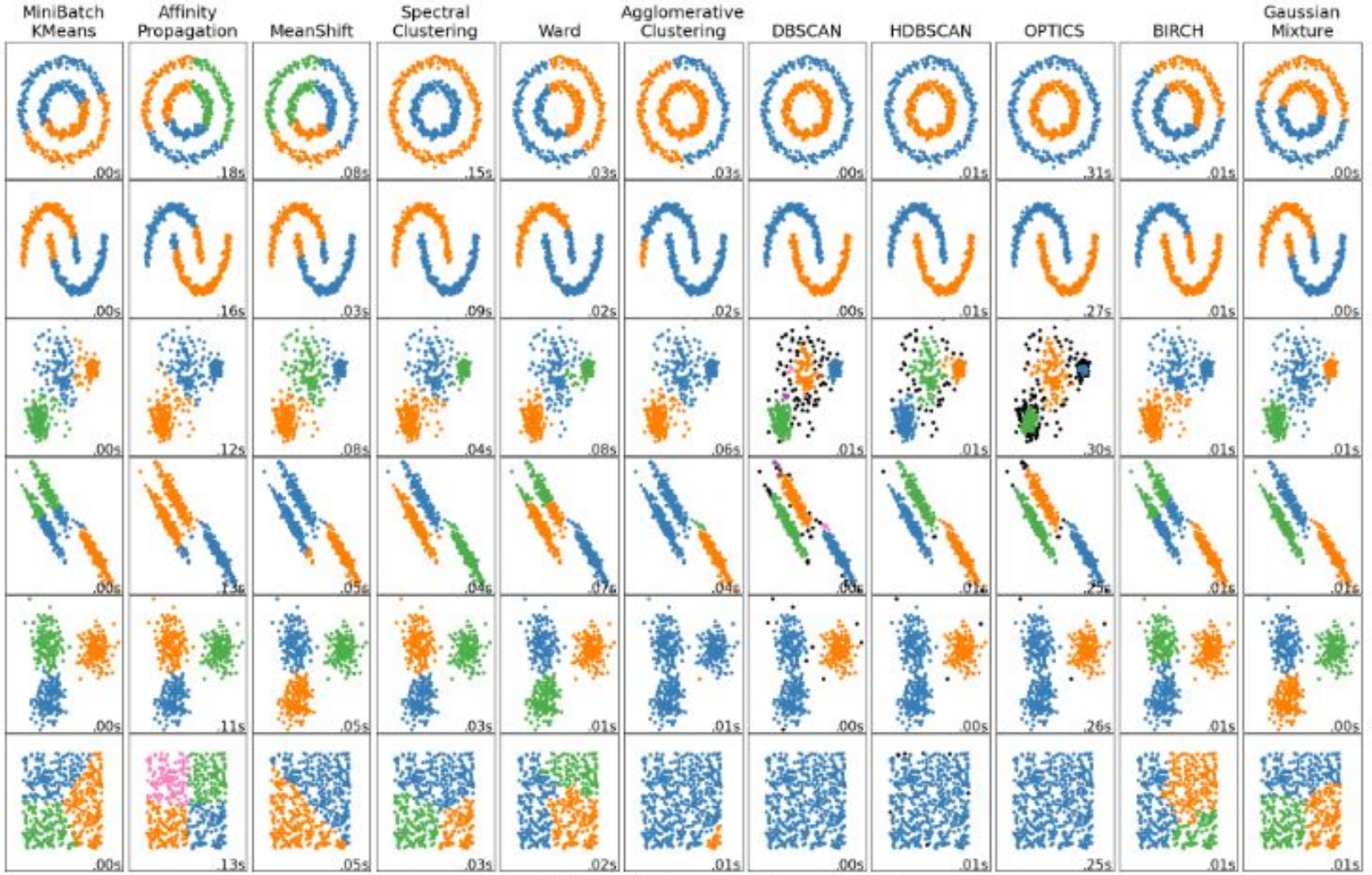
$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

La inercia puede considerarse una medida de la coherencia interna de los conglomerados.

Tiene varios inconvenientes:

- La inercia presupone que los clusters son convexos e isótropos, lo que no siempre es el caso. Responde mal a los clusters son alargados o con formas irregulares.
- La inercia no es una métrica normalizada: sólo sabemos que los valores más bajos son mejores y que cero es óptimo. Pero en espacios de dimensiones muy elevadas, las distancias euclidianas tienden a inflarse (es un caso de la llamada "maldición de la dimensionalidad"). Ejecutar un algoritmo de reducción de la dimensionalidad, antes de k-means puede aliviar este problema y acelerar los cálculos.


k-means iteration: 1
Initialization #1
SSE: 6735.8
Initialization #2
SSE: 5238.5
centroid
15
10
5
0
−5
−10
−10
0
10
−10
0
10


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
HDBSCAN
OPTICS
BIRCH
Gaussian Mixture

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters, inductive	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry, inductive	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry, inductive	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry, transductive	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters or distance threshold	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, transductive	Distances between points
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances, transductive	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes, outlier removal, transductive	Distances between nearest points
HDBSCAN	minimum cluster membership, minimum point neighbors	large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes, outlier removal, transductive, hierarchical, variable cluster density	Distances between nearest points
OPTICS	minimum cluster membership	Very large `n_samples`, large `n_clusters`	Non-flat geometry, uneven cluster sizes, variable cluster density, outlier removal, transductive	Distances between points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation, inductive	Mahalanobis distances to centers
BIRCH	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction, inductive	Euclidean distance between points
Bisecting K-Means	number of clusters	Very large `n_samples`, medium `n_clusters`	General-purpose, even cluster size, flat geometry, no empty clusters, inductive, hierarchical	Distances between points

FIN