UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Otros métodos de agrupamiento basados en Densidad

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 22) 25.marzo.2021

Mean-shift

Desarrollado or Fukunaga y Hostetler (1975).

Es un método no paramétrico que localiza máximos de una función de densidad. *Mean-shift* asigna los datos $\mathbf{x}_i$ a los grupos de forma iterativa al cambiar los puntos hacia la moda local. También se conoce como algoritmo de búsqueda de moda.

Dado un conjunto de puntos de datos $\mathbb{X} = (\mathbf{x}_i) \in \mathbb{R}^{n \times d}$, el algoritmo asigna iterativamente cada punto hacia el centroide del grupo más cercano y la dirección al centroide del grupo más cercano está determinada por el lugar donde se encuentran la mayoría de los puntos cercanos.

A diferencia de *k*-medias, *mean-shift* no requiere especificar el número de clústeres por adelantado, se determina en función del número de conglomerados con respecto a los datos.

Mean-shift

Comenzamos con una estimación inicial $\mathbf{x}_o$ Sea $K(\mathbf{x} - \mathbf{x}_o)$ una función de kernel (com las que se usan en estimación de densidades empíricas). Esta función determina el peso de los puntos cercanos para reestimar la media.

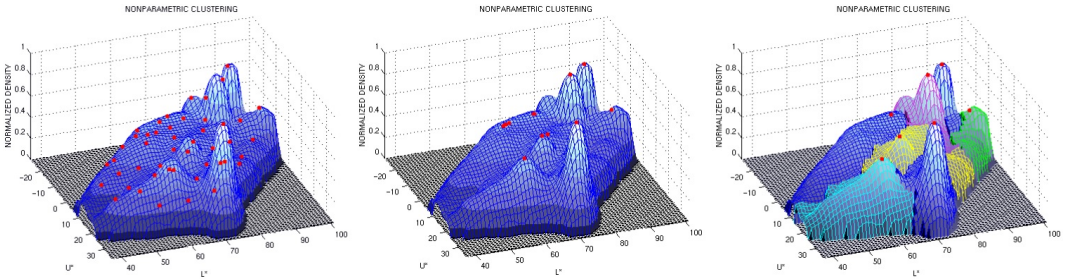
La media ponderada de la densidad en la región $N(\mathbf{x})$

$$m(\mathbf{x}) = \frac{\sum_{i \in N(\mathbf{x})} K(\mathbf{x}_i - \mathbf{x})\mathbf{x}_i}{\sum_{i \in N(\mathbf{x})} K(\mathbf{x}_i - \mathbf{x})},$$

donde $N(\mathbf{x})$ es la vecindad de $\mathbf{x}$.

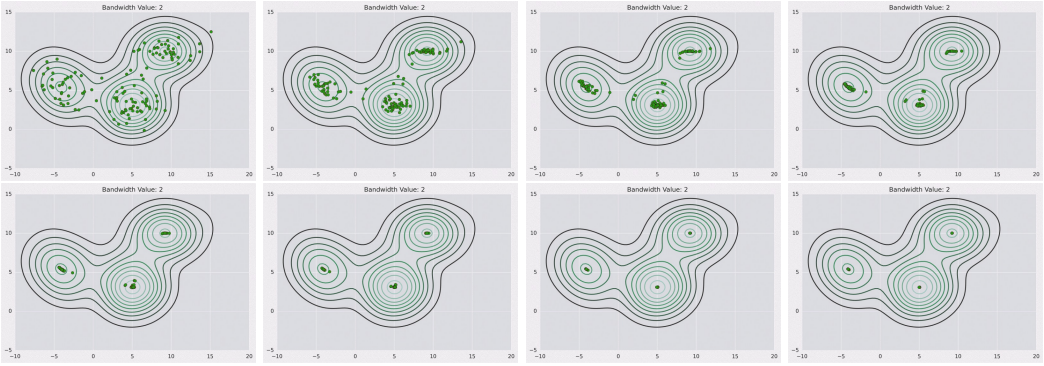
A la diferencia $m(\mathbf{x}) - \mathbf{x}$ se llama **mean-shift**. El algoritmo ahora establece $\mathbf{x} \leftarrow m(\mathbf{x})$ y repite la estimación hasta convergencia.

Mean-shift



En *mean-shift* cada punto converge hacia su moda más cercana

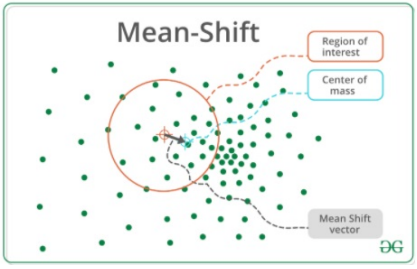
Mean-shift

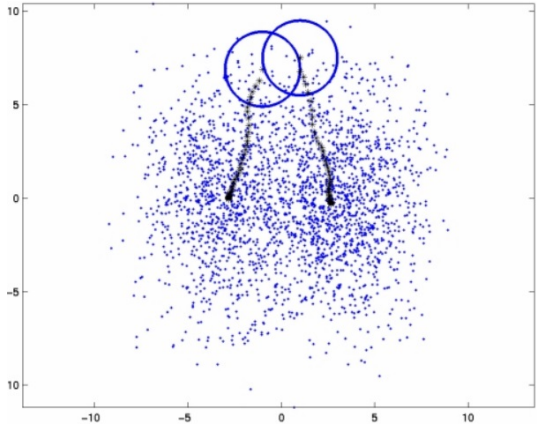

Bandwidth Value: 2
Bandwidth Value: 2
Bandwidth Value: 2
Bandwidth Value: 2
Bandwidth Value: 2
Bandwidth Value: 2
Bandwidth Value: 2
Bandwidth Value: 2

Mean-shift

Iterative Mode Search –

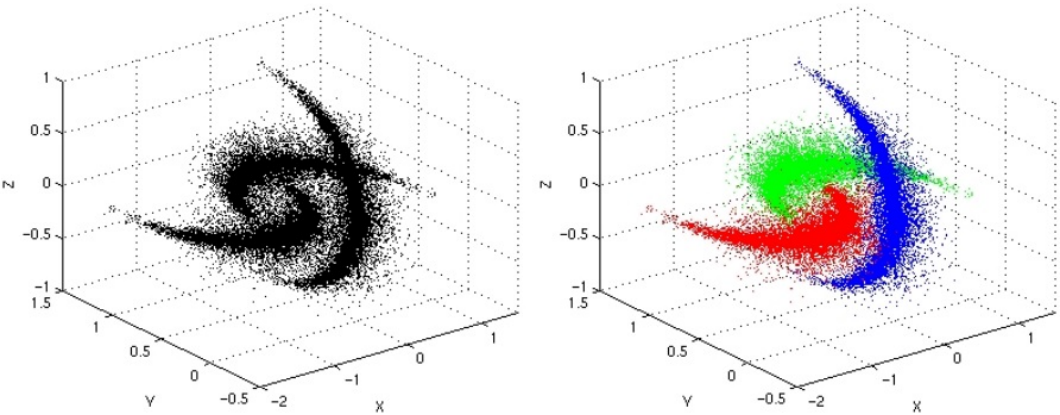
1. Initialize random seed and window W.
2. Calculate the center of gravity (mean) of W.
3. Shift the search window to the mean.
4. Repeat Step 2 until convergence.





UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Mean-shift



(a) Synthetic example of three non-linearly separable clusters (32640 points).

DBSCAN

DBSCAN = *Density-Based Spatial Clustering of Applications with Noise*
Desarrollado por Ester, Kriegel, Sander y Xu (1996).

Es un algoritmo no paramétrico de agrupamiento basado en densidad. Dado un conjunto de puntos $\mathbb{X} = (\mathbf{x}_i) \in \mathbb{R}^{n \times d}$ en algún espacio, agrupa los puntos que están muy juntos (puntos con muchos vecinos cercanos), marcando como valores atípicos los puntos que se encuentran solos en regiones de baja densidad, o cuyos vecinos más cercanos están demasiado lejos.

DBSCAN es uno de los algoritmos de agrupamiento más comunes y también el más citado en la literatura científica.

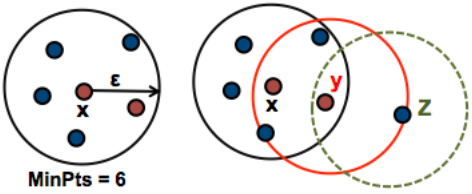
DBSCAN

Sea ε un parámetro que especifica el radio de una vecindad con $N(\mathbf{x})$ respecto a algún punto **x**. En DBSCAN, los puntos se clasifican como puntos centrales, puntos alcanzables (densidad) y valores atípicos, de la siguiente manera:

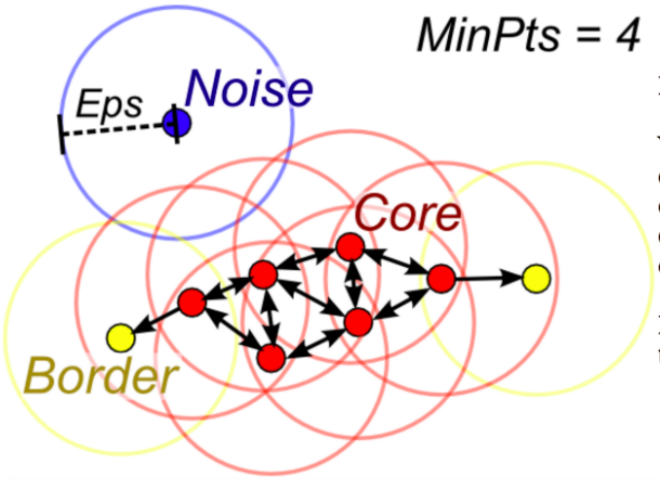
- Un punto **p** es un **punto central** si al menos *minPts* puntos están a una distancia ε de él (incluido **p**).
- Un punto **q** es **directamente accesible** desde **p** si el punto **q** está a una distancia ε del punto central **p**. Se dice que los puntos sólo son accesibles directamente desde los puntos centrales.
- Un punto **q** es **accesible** desde **p** si hay una ruta $\mathbf{p}_1, \ldots, \mathbf{p}_r$, con $\mathbf{p}_1 = \mathbf{p}$ y $\mathbf{p}_r = \mathbf{q}$, donde cada $\mathbf{p}_{i+1}$ es directamente accesible desde $\mathbf{p}_i$. El punto inicial y todos los puntos del camino deben ser puntos centrales, con la posible excepción de **q**.

DBSCAN

- Los puntos no accesibles desde cualquier otro punto son valores atípicos o puntos de ruido.
- Si **p** es punto central, forma un clúster junto con todos los puntos alcanzables desde él. Cada grupo contiene al menos un punto central. Los puntos no centrales pueden formar parte de un clúster, pero forman su **borde**, ya que no se pueden utilizar para alcanzar más puntos.



DBSCAN

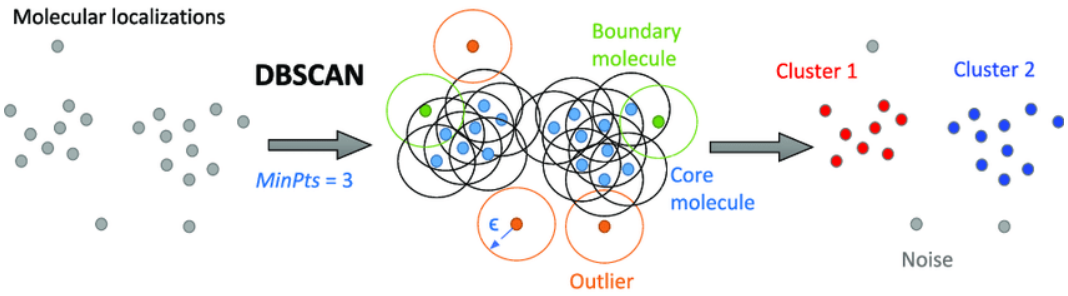


Red: Core Points

Yellow: Border points. Still part of the cluster because it's within epsilon of a core point, but not does not meet the min_points criteria

Blue: Noise point. Not assigned to a cluster

DBSCAN


Molecular localizations
DBSCAN
MinPts = 3
Boundary molecule
Core molecule
ε
Outlier
Cluster 1
Cluster 2
Noise

OPTICS

OPTICS = *Ordering Points To Identify the Clustering Structure*.
Desarrollado por Ankerst, Breunig, Kriegel y Sander (1999).

Es un algoritmo para encontrar agrupaciones basadas en densidad. La idea básica es similar a DBSCAN pero evita una de las principales debilidades de DBSCAN: el problema de detectar agrupaciones significativas en datos de densidad variable.

Para hacerlo, los puntos del conjunto de datos $\mathbb{X}$ se ordenan (linealmente) de modo que los puntos más cercanos espacialmente se conviertan en vecinos en el orden. Adicionalmente, se almacena una distancia especial para cada punto que representa la densidad que se debe aceptar para un cluster para que ambos puntos pertenezcan al mismo clúster. Esto se representa como un dendrograma.

OPTICS

Al igual que DBSCAN, OPTICS requiere dos parámetros: ε, que describe la distancia máxima (radio) a considerar, y *MinPts*, que describe el número de puntos necesarios para formar un grupo.

Un punto **p** es un punto central si al menos *MinPts* puntos se encuentran dentro de su vecindario $N_\varepsilon(\mathbf{p})$ (incluido el punto **p** mismo). A diferencia de DBSCAN, OPTICS también considera los puntos que forman parte de un grupo más denso, por lo que a cada punto se le asigna una distancia central que describe la distancia al *MinPts*-ésimo punto más cercano:

$$\text{core-dist}_{\varepsilon,\mathit{MinPts}}(\mathbf{p}) = \begin{cases} \text{UNDEFINED}, & \text{si } |N_\varepsilon(\mathbf{p})| < \mathit{MinPts}; \\ \mathit{MinPts}\text{-ésima menor distancia en } N_\varepsilon(\mathbf{p}), & \text{caso contrario.} \end{cases}$$

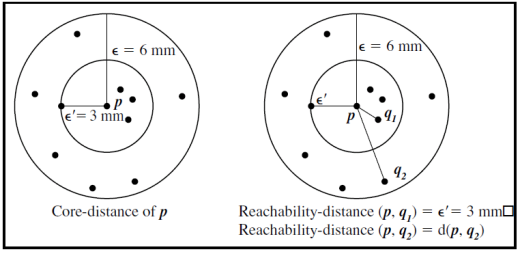
OPTICS

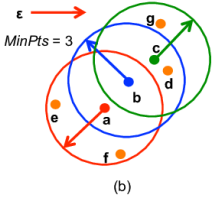
La distancia de accesibilidad desde un punto **p** a **q**, es la distancia entre **p** y **q**, o la distancia-núcleo de **p**, la que sea mayor:

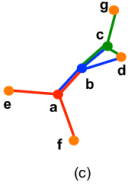
$$\text{reachability-dist}_{\varepsilon,MinPts}(\mathbf{p},\mathbf{q}) = \begin{cases} \text{UNDEFINED}, & \text{si } |N_\varepsilon(\mathbf{p})| < MinPts; \\ \max\{\text{core-dist}_{\varepsilon,MinPts}(\mathbf{p}), d(\mathbf{p},\mathbf{q})\}, & \text{caso contrario.} \end{cases}$$

Si **p** y **q** son vecinos más cercanos, este es el $\varepsilon' < \varepsilon$ que debemos asumir para que **p** y **q** pertenezcan al mismo grupo.

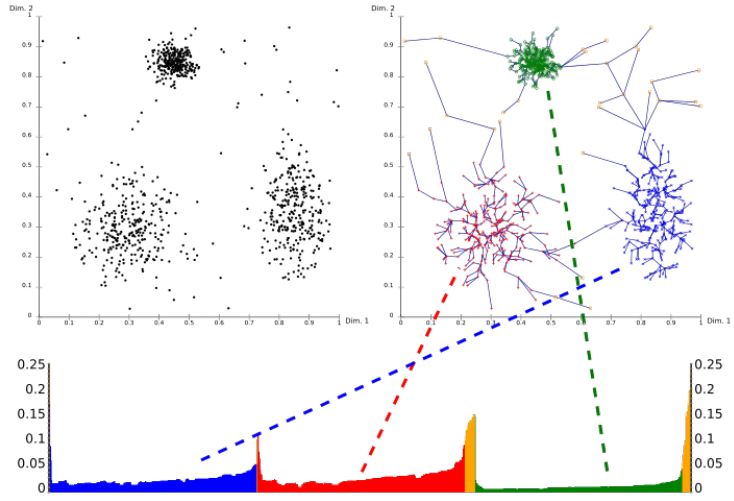
OPTICS







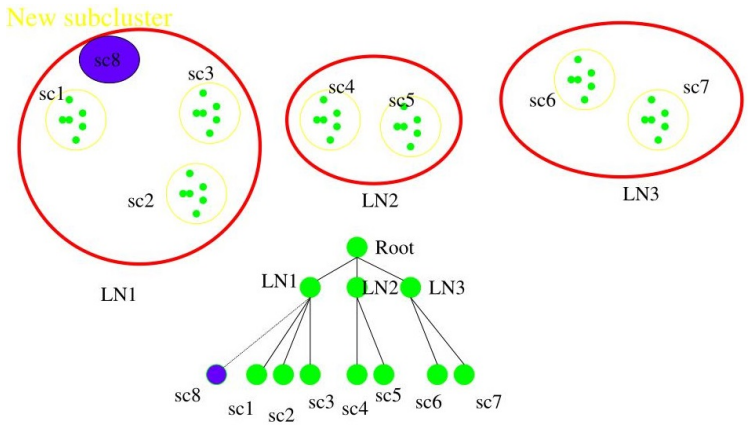
OPTICS



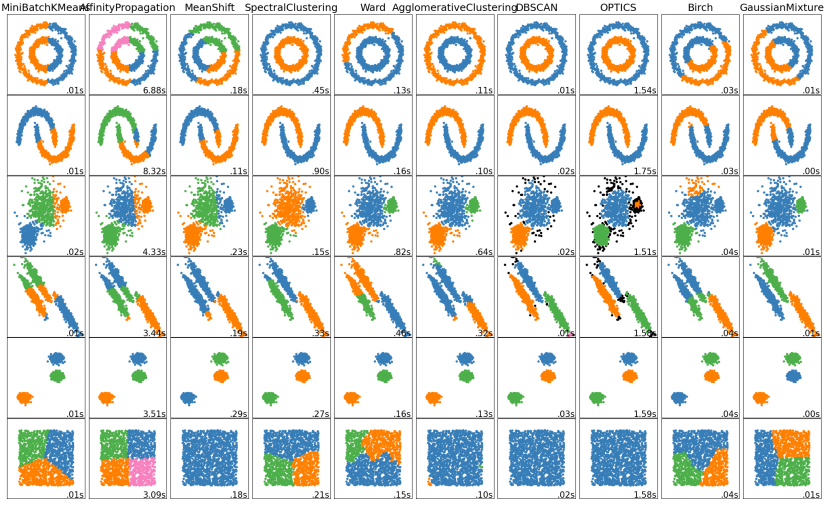
UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

BIRCH

BIRCH = *Balanced Iterative Reducing and Clustering using Hierarchies.*



Comparación de métodos



UVG UNIVERSIDAD DEL VALLE DE GUATEMALA