

Fastbook 02

Analítica de Cliente & Predictive Analytics

Teoría general de segmentación (II)



02. Teoría general de segmentación (II)

Si en nuestro primer fastbook vimos qué es una segmentación (concretamente una segmentación de clientes), en este segundo fastbook veremos las múltiples **utilidades** que puede tener un análisis de segmentación, qué **modelos** existen y cuáles emplearemos.

De este modo, cubriremos toda la teoría que necesitamos sobre análisis de segmentación y dedicaremos los fastbook 03 y 04 a ver ejemplos reales con R. Pero para ello necesitamos saber solo un poco más de teoría, así que jáinimo!

Autor: Miguel Ángel Fernández

☰ Utilidades de la segmentación

☰ Conocimiento

☰ Modelos de segmentación

☰ Conceptos técnicos

☰ Conclusiones

Utilidades de la segmentación

X Edix Educación

Empecemos a calentar neuronas con la siguiente pregunta: **¿qué utilidades puede tener una segmentación?**

La respuesta que viene a la mente es sencillamente ‘segmentar’, porque ¿para qué va a servir un modelo de segmentación si no es para segmentar? Pues bien, segmentar es lo que hace un modelo de segmentación, pero eso no responde a ‘¿para qué me sirve a mí segmentar a mis clientes?’

Segmentar es lo que vamos a hacer, pero en todo buen análisis hay que tener claro siempre qué es lo que nos va a aportar dicho análisis y para qué nos va a servir.

Las utilidades son múltiples y las iremos observando en detalle en las siguientes secciones.

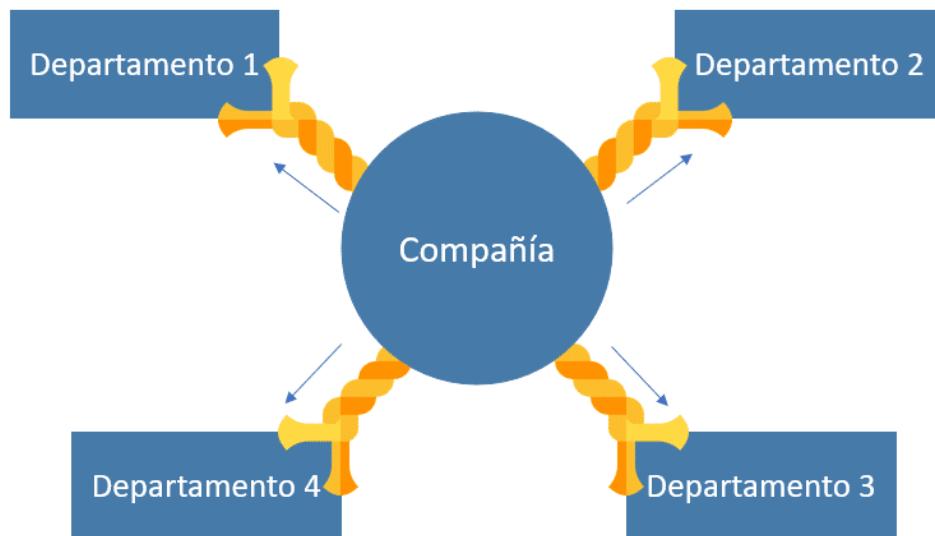
Conocimiento

X Edix Educación

En plena ‘era del dato’ hay muchas compañías que no conocen realmente a sus clientes, y en parte es natural, ya que no es una tarea sencilla. Siempre hay muchas perspectivas distintas, **muchos prismas con los que mirar a un negocio y a sus datos.**

Al departamento de Negocio pueden interesarle las visitas o los leads de los clientes para conseguir sus objetivos, mientras que el departamento de Contabilidad puede estar más interesado en los gastos/consumos que realizan los clientes en nuestro negocio.

Esto puede hacer que en ocasiones se genere **tensión entre departamentos**, pareciendo que en vez de remar en la misma dirección se produzcan ciertos pulsos, como ilustra la siguiente imagen:



Para todo analista de cualquier compañía este es un asunto clave a tener siempre en cuenta. Es tan importante que puede ser lo que mande tu análisis a la alta esfera de tu empresa o a la papelera de reciclaje.

Es tarea previa de cualquier análisis (incluida la segmentación) **señalar claramente cuál es el marco en el que se va a colocar nuestro análisis**, las bases sobre las que se sostiene para evitar críticas y dudas lo más pronto posible.

Entonces, supongamos que queremos segmentar a nuestros clientes por su actividad: ¿qué datos usamos? ¿Las visitas a nuestras tiendas/web para contentar al departamento X o el consumo en nuestras tiendas/web para contentar al departamento Y?

Habrá ocasiones en las que podamos satisfacer las demandas de ambos y, en este ejemplo, una posibilidad es usar las dos fuentes de datos (visitas y consumos) en nuestro modelo de segmentación. No obstante, habrá otras ocasiones en las que esto no será posible, ya que los enfoques que le interesan a cada uno de los departamentos son completamente opuestos y tendremos que elegir.

Es en estos casos en los que nos están pidiendo dos enfoques distintos, tendremos que hacer, inevitablemente, dos análisis distintos.

Parece obvio pero es muy común que a medida que vayamos iterando con diferentes áreas queramos convertir nuestro análisis en una ‘navaja multiusos’ y empecemos, por lo tanto, a añadir variables de todo tipo y añadir criterios difíciles de casar unos con otros para cumplir con todo el mundo.

Una vez tenemos claro sobre qué datos queremos que se base nuestra segmentación (para más ejemplos, ver sección *Ejemplos de segmentación* del fastbook 01) podremos proceder a segmentar, y ¿cuál es la primera utilidad que nos da una segmentación? Pues la de convertir los datos en información, y con ella los analistas podemos convertirla en conocimiento.

De esta forma partimos de **datos**, el modelo de segmentación nos proporciona **información** y nosotros como analistas convertimos esta en **conocimiento**.

Más adelante veremos un ejemplo donde estos conceptos quedan más claros.



Los conceptos de *datos*, *información* y *conocimiento* suelen intercambiarse unos con otros en una conversación, pero **son muy distintos** y cabe señalar qué es cada uno. Ahora vamos a ver las definiciones teóricas con las que es muy posible que estos conceptos no terminen de ser claros, pero después veremos dos ejemplos con los que no nos quedará ninguna duda.

Los datos son la unidad mínima que tenemos de información, pero están tan desagregados y hay tantos que por sí solos no nos son útiles. **Se hace necesario procesar los datos y darles un sentido para convertirlos en información.**

De esta forma, podemos definir la **información** como un conjunto de datos procesados y con un significado.

En este punto, la información sí nos es útil para tomar decisiones. Pero podemos ir un paso más allá y convertir la información en conocimiento, que se define como una **mezcla de experiencia, valores e información**.

Para convertir la información en conocimiento debemos realizar, por ejemplo, algunas de las siguientes acciones:

- Comparar con otros elementos.
- Contemplar posibles consecuencias.
- Dar contexto con otras fuentes de información.

Básicamente, se trata de darle una vuelta más a la información.

Estas definiciones son muy teóricas, pero veremos ahora unos ejemplos para entender a qué nos referimos cuando hablamos de ‘datos’, ‘información’ y ‘conocimiento’.

Ejemplo 1. Supongamos que queremos hacer un análisis descriptivo de nuestros clientes por género. Para ello, lo primero que necesitamos son **datos**, y podrían tener la siguiente forma:

ID_cliente	Género
000001	F
000002	M
000003	M
...	...
999999	F

Como vemos, los datos así ¿de qué nos sirven? Pues de muy poco.

Para que nos sean de utilidad hay que transformarlos, por lo tanto, ¿qué tal si calculamos la proporción de hombres y mujeres? Supongamos que tenemos un 57% mujeres y un 43% hombres. Estos números ya son información, y un analista podría inferir de estos datos que nuestro producto o servicio tiene una capacidad de atracción mayor sobre cliente mujer que sobre cliente hombre.

Podríamos quedarnos aquí, con la información sin más, pero estaríamos cometiendo errores en muchos casos.

- ¿Qué pasa si la población nacional en la que se encuentran nuestros clientes es también de 57% mujeres y un 43% hombres? En este caso, nuestra anterior conclusión era falsa pues nuestro cliente o servicio no tiene mayor capacidad de atraer clientes mujeres que hombres, simple y llanamente tenemos la proporción habitual de hombres y mujeres.

- Por el contrario, ¿y si la proporción nacional es de 50% mujeres y 50% hombres? En este caso, la conclusión inicial sí sería correcta. Como vemos, siempre hay que llegar a esta fase en la que comparamos con otros elementos, contemplamos posibles consecuencias de hacer A o B acciones, dar contexto a la información con otras fuentes de datos...

De esta forma, hemos convertido la información en conocimiento: nuestro producto atrae/no atrae a un género más que al otro (a ambas conclusiones podemos catalogarlas como ‘conocimiento’).

Ejemplo 2. Vamos a ver qué ocurre ahora con una segmentación, que es el análisis que nos interesa en esta asignatura.

Supongamos que queremos hacer una segmentación tradicional. Podríamos partir de los siguientes datos:

ID cliente	Género	Edad	País	Código Postal	...	Cliente Premium
000001	F	31	España	28850	...	No
000002	M	25	España	28010	...	No
000003	M	56	Portugal	110608	...	Si
...
999999	F	44	Francia	06240	...	No

Nuestro objetivo es segmentar a nuestros clientes en grupos de clientes similares entre sí. Pero ¿cómo convertimos estos datos en información que nos diga a qué segmento pertenece cada cliente? La respuesta es: usando modelos de segmentación.

Recordemos el siguiente diagrama:



El input serán nuestros datos y el output será una nueva tabla con información de a qué segmento asignar cada cliente:



Ahora que tenemos a qué segmento pertenece cada cliente. ¿Habríamos terminado?

Para aquellos que crean que hemos terminado lanzo las siguientes preguntas: ¿qué es el segmento 1?, ¿y el segmento 2? Es más, ¿qué hace que un cliente esté en un segmento cualquiera?

Con el output de un modelo de segmentación tenemos información, pero es necesario convertir esta en conocimiento, y solo lo conseguiremos si respondemos a las anteriores preguntas.

Lo que podemos hacer es un **análisis descriptivo de cada segmento**: analizando proporciones de hombres y mujeres, la edad media, los niveles de estudios medios, etc., podemos **crear un perfil de los clientes en cada segmento** y así darles un sentido.

De esta manera, descubriremos que el segmento 1 son jóvenes (<30), solteros, con nivel de estudios altos, por ejemplo, y que el segmento 2 son personas de mediana edad (~50), mayoritariamente de Madrid y casados, por ejemplo, etc.

En función de cómo de ricas sean las variables que hayamos empleado en la segmentación, descubriremos perfiles igual de interesantes.

Es común denominar cada segmento con un nombre que resuma el perfil. En este ejemplo podríamos decir que el segmento 1 son los *Millennials Estudiosos*, mientras que el segmento 2 podría llamarse los *Baby Boomers Madrileños* o algo similar (para esta tarea suele requerirse el apoyo creativo del departamento de Marketing).

Con este análisis obtendremos un conocimiento profundo sobre qué perfil de clientes componen cada segmento con el valor que esto supone a una compañía para una cantidad incontable de acciones, siendo la más representativa la personalización de campañas que veremos a continuación.

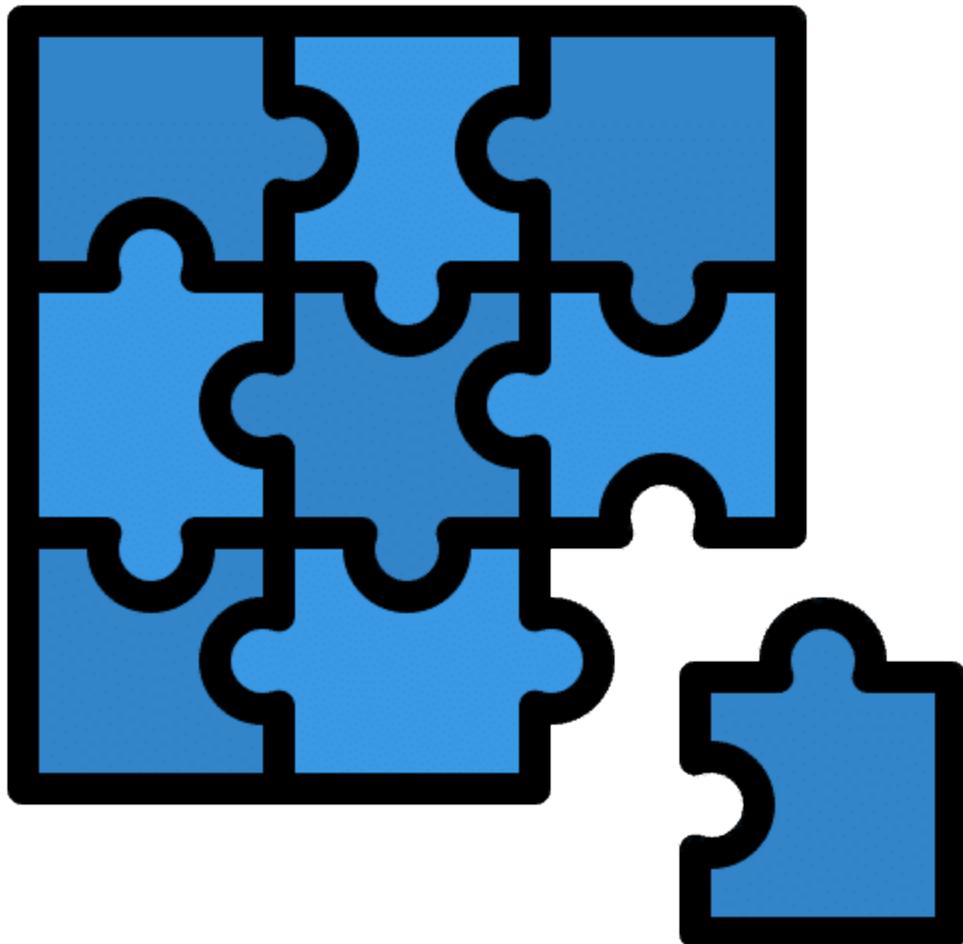
Campañas

Una vez tenemos asignados a los clientes por segmentos es posible **afinar muchísimo con los mensajes y las creatividades que generamos desde un departamento de Marketing**.

Nadie estaría de acuerdo en lanzar los mismos mensajes a un Millennial que a un Baby Boomer. Incluso entre los Baby Boomers casados y solteros podríamos diferenciar matices que den lugar a mensajes diferentes. Podríamos seguir así y cruzar la *Edad* con la *Provincia*, el *Género* con la *Edad*, el *Género* con el *Estado civil*... y acabaríamos con una infinidad de posibilidades.

Sería muy poco realista el generar mensajes para cada una de estas audiencias. Por eso es tan útil el análisis de segmentación: **no solo nos permite segmentar a nuestros clientes, sino que segmenta a nuestros clientes en N grupos de la forma más óptima posible.**

Si nuestro departamento de Marketing solo puede soportar 9 audiencias, hay modelos de segmentación que segmentarán a nuestros clientes en el número de segmentos que les indiquemos, en este caso sería 9, con la optimización en los resultados de campañas correspondiente.



Ir más allá de los clientes

En muchas ocasiones acabamos encontrando a **uno o varios grupos minoritarios de clientes 'muy raros'**. Con 'raro' queremos decir que son un segmento muy, muy, muy diferente a cualquier otro segmento que hayamos generado.

Pensemos en clientes que tienen unos niveles de actividad cien veces más altos que el resto de clientes, con niveles de estudios de doctorado, de un país extranjero y del sector laboral de los *Videojuegos*. Claramente este es un segmento de **clientes raros** ('frikis', que dirían algunos).

Cuando segmentamos clientes no hay más que añadir, este segmento lo podemos definir como un segmento de clientes **outliers**, con el valor que tiene este descubrimiento para el negocio.

Esta es una utilidad nueva que ganamos al usar modelos de segmentación, el encontrar segmentos de elementos raros.

Esto es precisamente lo que motiva esta sección.

Hasta ahora solo hemos hablado de segmentar clientes, que es el principal objetivo de esta asignatura, pero **¿por qué sólo segmentar clientes?**

Al igual que los modelos de segmentación nos ayudan a segmentar clientes, nos sirven también para segmentar cualquier otro grupo de elementos. Pensemos que **podemos ir siempre más allá de nuestros clientes**.

Por ejemplo, imaginemos que somos un banco y tenemos transacciones. Asociada a cada transacción hay mucha información que podríamos usar, por lo que tendríamos una tabla como la siguiente:

ID_transaccion	Cantidad	Divisa	País origen	...	País destino
000000001	20.00	€	España	...	España
000000002	34.99	€	España	...	España
000000003	200,000.00	\$	Portugal	...	Japón
...
999999999	150.00	€	Francia	...	España

Usando un modelo de segmentación podemos segmentar las transacciones en grupos de transacciones parecidas entre sí. Este es un análisis muy común en el que suelen generarse segmentos de transacciones ‘raras’. Fijémonos en la tercera transacción de la tabla anterior:

ID_transaccion	Cantidad	Divisa	País origen	...	País destino
000000001	20.00	€	España	...	España
000000002	34.99	€	España	...	España
000000003	200,000.00	\$	Portugal	...	Japón
...
999999999	150.00	€	Francia	...	España

Una transacción de 200,000 \$ de una cuenta de Portugal a otra de Japón sin duda merece la pena ser investigada. Este tipo de transacciones suele ser de origen fraudulento, por lo que a los bancos les interesa estar al tanto, como es normal.

De esta forma, además de segmentar, la tercera gran utilidad que tienen los modelos de segmentación es la **detección de anomalías o outliers**, simplemente por medio de la generación de un segmento de elementos muy raros.

Lección 3 de 5

Modelos de segmentación

 Edix Educación

Llegados a este punto, debemos tener claro qué es una segmentación y qué utilidades tiene.

El último ingrediente que nos falta son los propios **modelos de machine learning para hacer una segmentación**, que es lo que cubriremos en esta sección.

En este curso centraremos nuestros esfuerzos en conocer tres modelos: **clustering jerárquico**, **K-Means** y **DBSCAN**.

Clustering jerárquico

Existen dos variantes del **clustering jerárquico**: el **clustering aglomerativo** y el **clustering divisivo**.

1

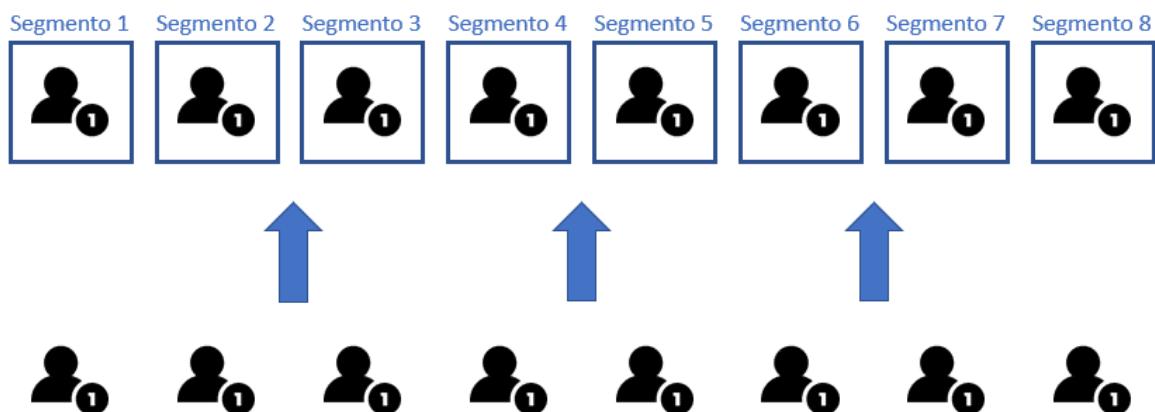
Clustering aglomerativo

Entonces, supongamos que tenemos 1M clientes y queremos hacer una segmentación tradicional de los mismos. Es decir, tenemos la siguiente tabla de variables:

ID_cliente	Género	Edad	País	Código Postal	...	Cliente Premium
000001	F	31	España	28850	...	No
000002	M	25	España	28010	...	No
000003	M	56	Portugal	110608	...	Si
...
999999	F	44	Francia	06240	...	No

Cuando le damos esta tabla a un **clustering jerárquico aglomerativo**, ¿qué es lo que hace con la información para poder generar segmentos? El algoritmo consta de los siguientes pasos:

Paso 1. En primer lugar, el modelo coge a todos los clientes de nuestra población objetivo y considera cada uno como un **segmento (clúster) independiente**, es decir, si tenemos 5 millones de clientes, vamos a empezar generando 5 millones de clústeres, como ilustra la siguiente figura con 8 clientes de ejemplo:

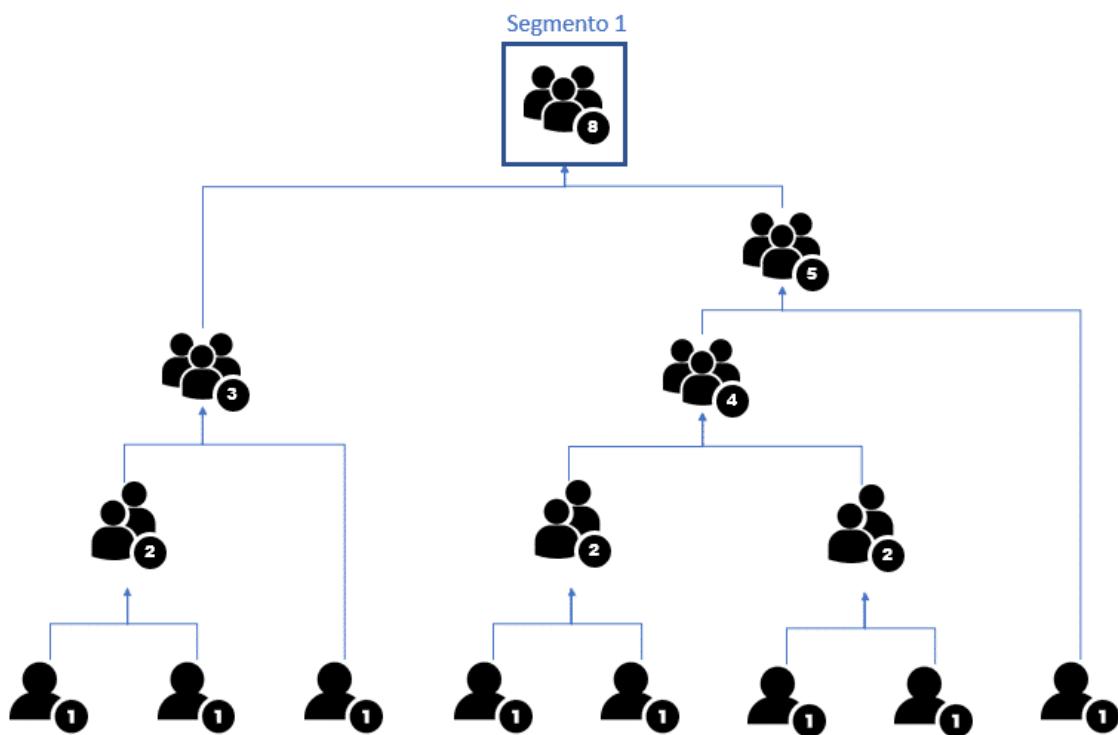


Paso 2. De entre todos los segmentos, el modelo **busca los dos clientes más ‘parecidos’ entre sí y los agrupa (aglomera) en un único segmento**. De esta manera, en el ejemplo pasamos de tener 8 segmentos a tener 7. Supongamos que los clientes más parecidos entre sí son el 1 y el 2:



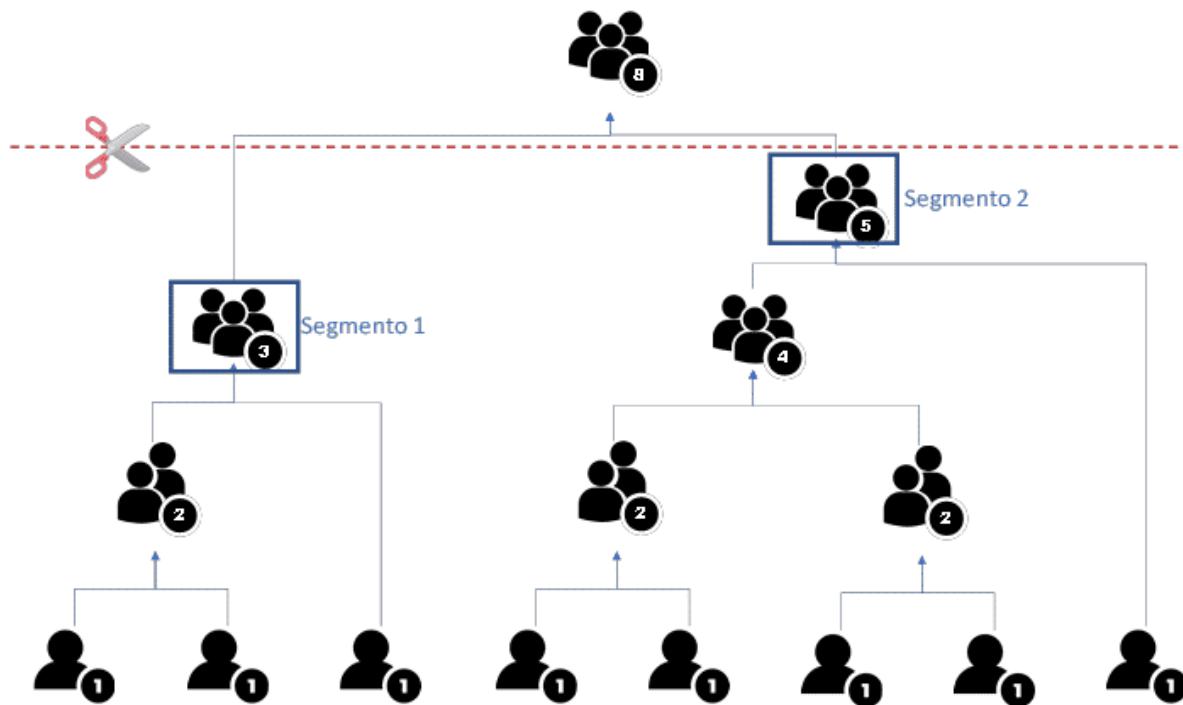
Antes hemos entrecerrado ‘parecidos’ porque todavía no hemos aclarado cómo medir lo parecidos o distintos que son dos clientes. Este aspecto es importante y lo veremos al final del fastbook, pero, por ahora, limitémonos a los que nos dice nuestra intuición.

Paso 3. Consiste en **repetir el paso 2 hasta que se agrupen todos los clientes en un único segmento**, es decir, hasta que hayamos aglomerado a todos ellos, quedando una estructura llamada **dendrograma** como la siguiente:

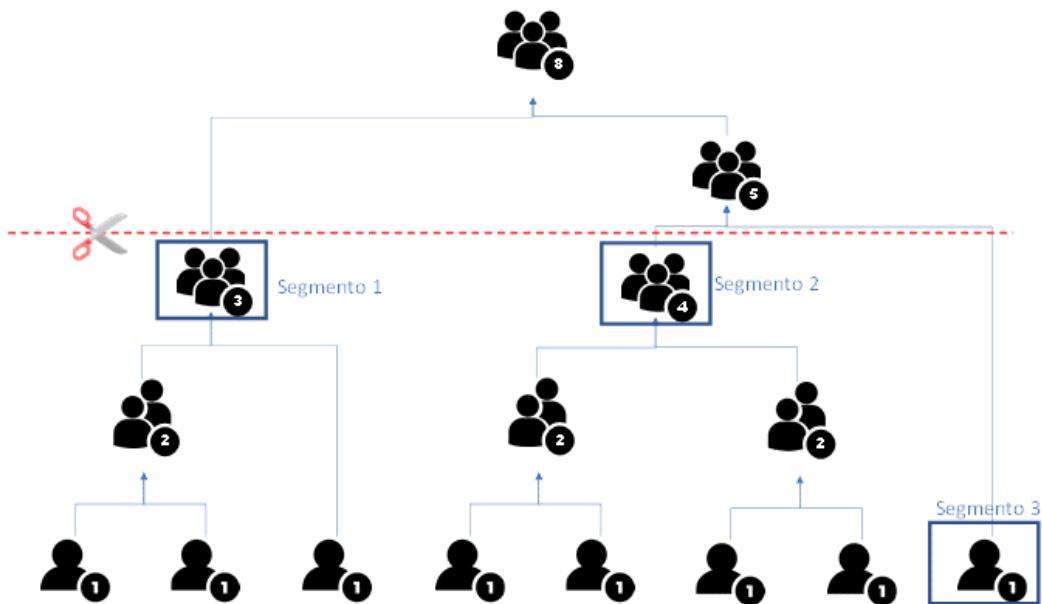


Paso 4. El último paso consiste en seleccionar cuántos segmentos queremos generar en total. Este es un input que damos nosotros al modelo. Aprovechando el dendrograma que hemos construido al terminar el paso 3, es tan sencillo como ‘cortar’ a la altura que queramos.

Si queremos generar dos segmentos, cortaremos de la siguiente forma generando un segmento con 5 clientes y otro con 3:



Si queremos generar tres segmentos, cortaremos de la siguiente forma:



Viendo el procedimiento del algoritmo, nos queda más claro el porqué del nombre de aglomerativo.

2

Clustering divisivo

El clustering jerárquico **divisivo**, como veremos a continuación, es la versión inversa al aglomerativo. En lugar de ir agrupando segmentos por parecido hasta generar un segmento que los contenga a todos, vamos a hacerlo al revés. Los pasos son los siguientes:

Paso 1. El modelo de segmentación jerárquico **divisivo** empieza cogiendo a todos los clientes y agrupándolos en un único segmento:

Segmento 1



Paso 2. Este paso consiste en **dividir** a todos los clientes en dos grupos de segmentos que sean lo más diferentes entre sí, pero que los clientes que contengan sean lo más parecidos entre ellos posible:

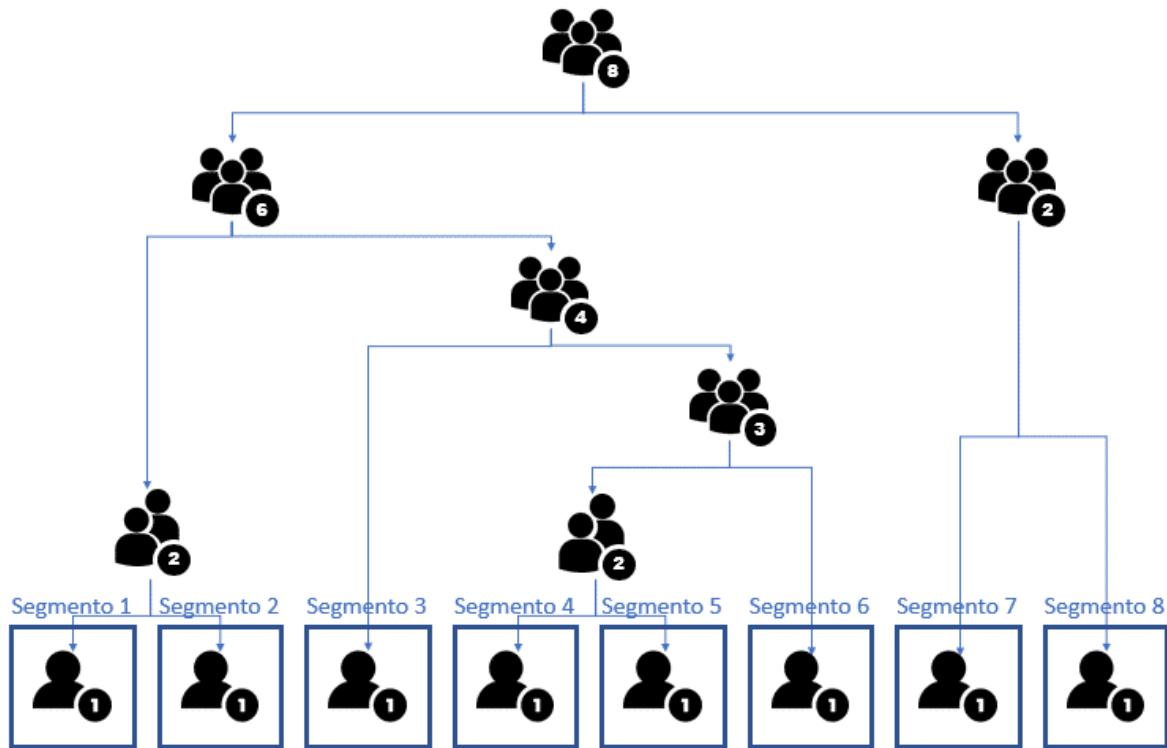


Segmento 1



Segmento 2

Paso 3. Consiste en repetir el paso 2 hasta que todos los clientes pertenecen a segmentos únicos, generando nuevamente un dendrograma como el que ilustra la siguiente figura:



Paso 4. En el último paso decidimos cuántos segmentos queremos generar para cortar el dendrograma por el nivel correspondiente, como hemos visto en la versión aglomerativa.

Es importante tener en cuenta que las versiones aglomerativa y divisiva del clustering jerárquico no son equivalentes, sino que cada una genera unos segmentos distintos. Cómo de distintos lo marcará la estructura que tengan los datos que demos al modelo como input.

Si te estás preguntando cuál es la mejor de las dos versiones, la respuesta es ‘ninguno es mejor que el otro’. En ocasiones, será la versión aglomerativa la que genere unos segmentos que nos gusten más, y, otras veces, será la versión divisiva nuestra preferida.

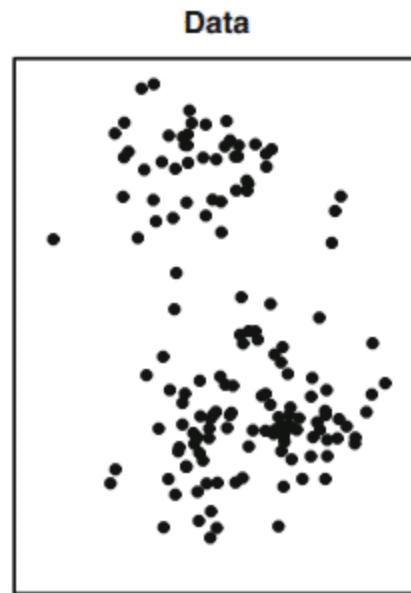
Por esta razón solemos ajustar distintos tipos de modelos, de la misma familia y de distintas familias, ya que, en la práctica, es complicado asegurar qué modelo dará mejores resultados.

K-Means

Este es probablemente el modelo de segmentación más conocido, más estudiado y sobre el que más se habrá escrito en la literatura. Tiene un punto de sofisticación y elegancia mayor a lo que hemos visto con los modelos de segmentación, pero sigue siendo fácil de entender.

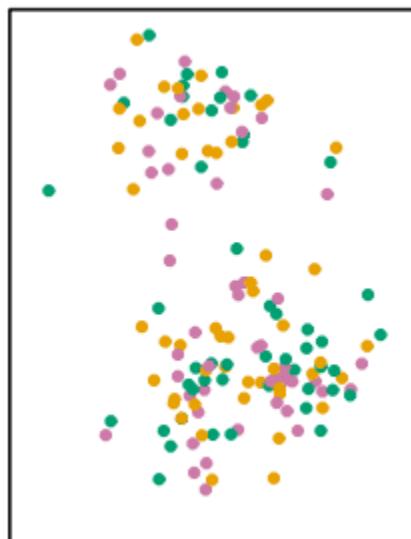
La estrategia que sigue K-Means para generar **K segmentos** consiste en diferentes pasos que ilustraremos apoyándonos en una de las referencias que todo analista de datos debe tener en su biblioteca: [An Introduction to Statistical Learning](#). Por suerte para todos, los profesores autores de la Universidad de Stanford tienen público un ejemplar en formato PDF al que podemos acceder en el link anterior.

Supongamos que tenemos los siguientes datos para segmentar, donde cada punto es un cliente:



Alguno seguramente está tentado a coger el lápiz y segmentar a mano, pero vamos a emplear K-Means a ver qué ocurre.

Paso 1. Supongamos que queremos generar tres segmentos. Este es un input que tenemos que dar al modelo K-Means al igual que a los clustering jerárquicos. Una vez dado, el primer paso que da el modelo es asignar de manera aleatoria cada punto (cliente) a uno de los tres segmentos que vamos a generar. De esta forma tenemos clasificados a los clientes usando tres colores con la siguiente distribución:



A partir de aquí, el modelo realizará varias **iteraciones** donde se harán **2 subpasos**:

1

Calcular centroides (el punto medio de todas las variables) de cada segmento.

2

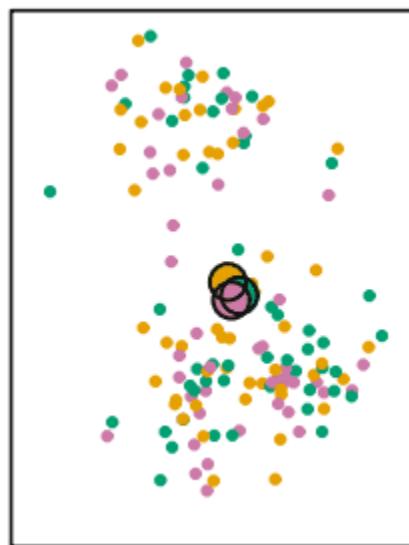
Reasignar cada punto al grupo del centroide más cercano.

En nuestro ejemplo tendríamos:

Iteración 1

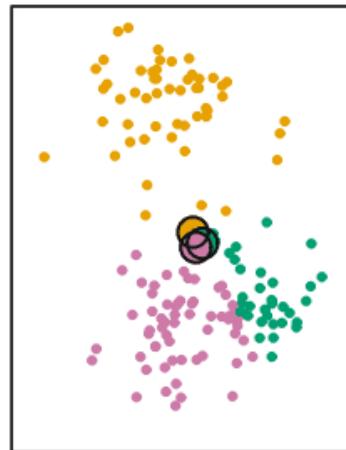
Subpaso 1

Calculamos los centroides, que están representados como bolas grandes en la siguiente imagen:



Subpaso 2

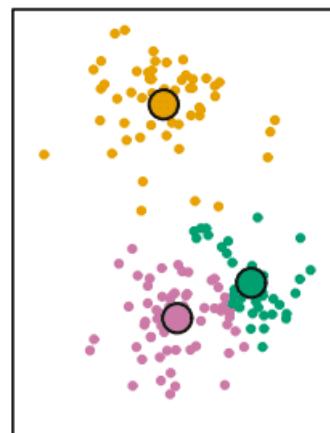
Asignamos de nuevo cada punto al centroide más cercano que tenga:



Iteración 2

Subpaso 1

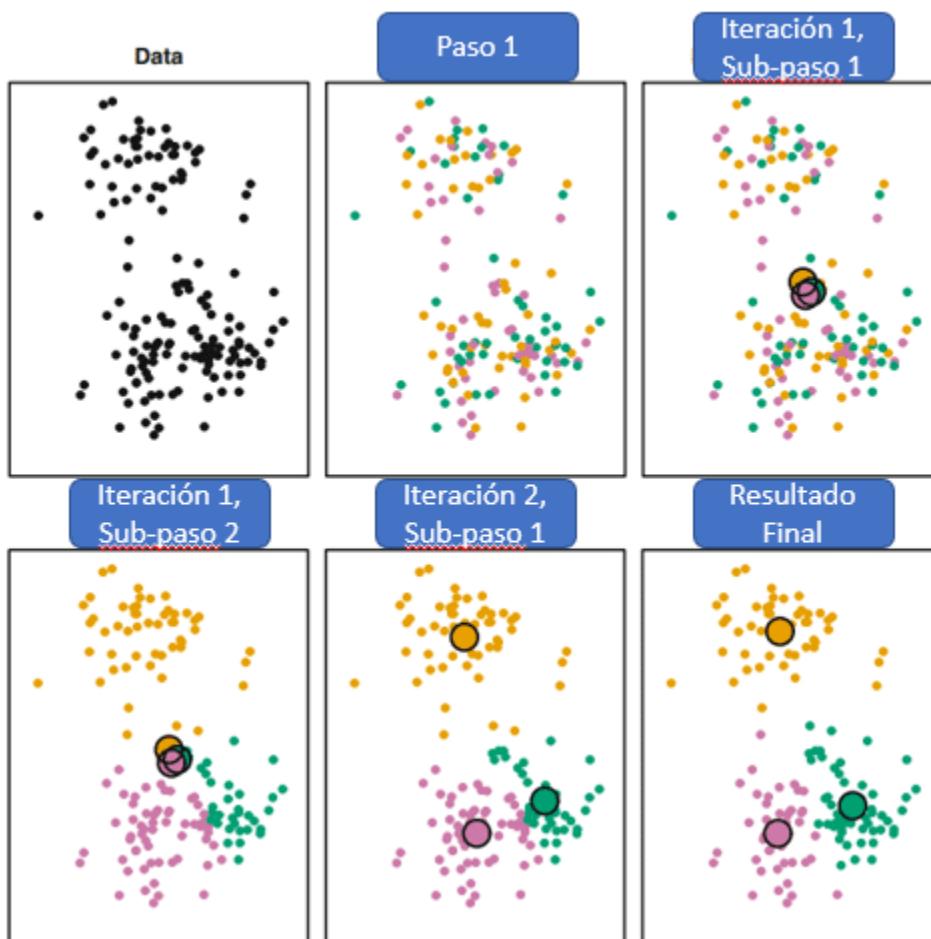
Si damos una segunda iteración para recalcular los centroides, tenemos el resultado final:



De esta manera K-Means realiza una segmentación. En este ejemplo solo hemos empleado dos variables para poder entender visualmente cada paso que realiza el algoritmo. En nuestros ejemplos reales tendremos muchas más de dos variables, pero el **modelo realizará los mismos pasos** sea cual sea el número de variables que demos de input para segmentar.

Por otro lado, este ejemplo nos ha permitido llegar a una solución final en dos iteraciones. En la práctica pueden requerirse más iteraciones, pero esto lo decidirá el modelo.

Para que los pasos del algoritmo puedan seguirse quizás un poco mejor, cerramos esta sección con todos los pasos que hemos visto agrupados en una imagen:



Reflexionemos

Como hemos visto, el objetivo es claro: dada una serie de puntos (clientes de los que tenemos variables con información) queremos segmentarlos en **K segmentos**. Pero las estrategias para conseguir este objetivo son muy diferentes de unos modelos a otros.

Mientras que los clustering jerárquicos tratan de ir emparejando o dividiendo grupos de clientes parecidos (como una hormiguita, paso a paso), K-Means tiene una estrategia más óptima y elegante aprovechando los centroides para ir asignando los clientes al centroide que tengan más cercano.

Cada modelo pertenece a una familia de modelos concreta, por ejemplo, **los modelos aglomerativos y divisivos** pertenecen a los clusterings jerárquicos y K-Means pertenece a los clusterings particionales.

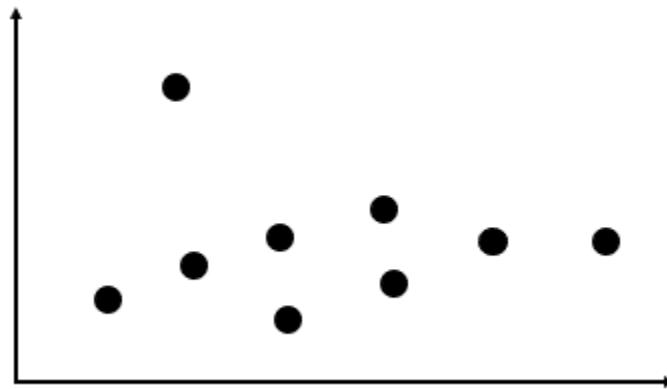
DBSCAN

Cada modelo cuenta uno con sus puntos fuertes y sus puntos débiles, por eso en ciencia de datos es importante contar con un abanico de modelos lo más extenso posible.

Con este objetivo vamos a ver un modelo de una familia nueva, los denominados **clusterings basados en densidad**. Este modelo es **DBSCAN** (por sus siglas, Density-Based Spatial clustering of Applications with Noise). El nombre puede asustar un poco, y es cierto que nos va a exigir un poco más de esfuerzo que los anteriores modelos para entender cómo funciona, pero, al igual que antes, veremos paso a paso el algoritmo para ganar intuición sobre cómo funciona.

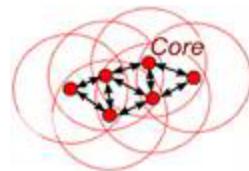
Antes de ver los pasos que componen el algoritmo, necesitamos definir tres conceptos: **clientes core** (núcleo), **clientes border** (frontera) y **clientes noise** (atípicos), que son los ingredientes que usará DBSCAN para trabajar.

Supongamos que tenemos dos variables y nuestros clientes se distribuyen de la siguiente manera:



DBSCAN diferencia tres tipos de clientes (es decir, puntos, siendo más generales):

- **Clientes core:** son aquellos clientes que tienen varios clientes similares próximos a ellos. Es decir, no son ‘raros’, hay otros clientes parecidos a ellos y componen el núcleo de lo que será un segmento.



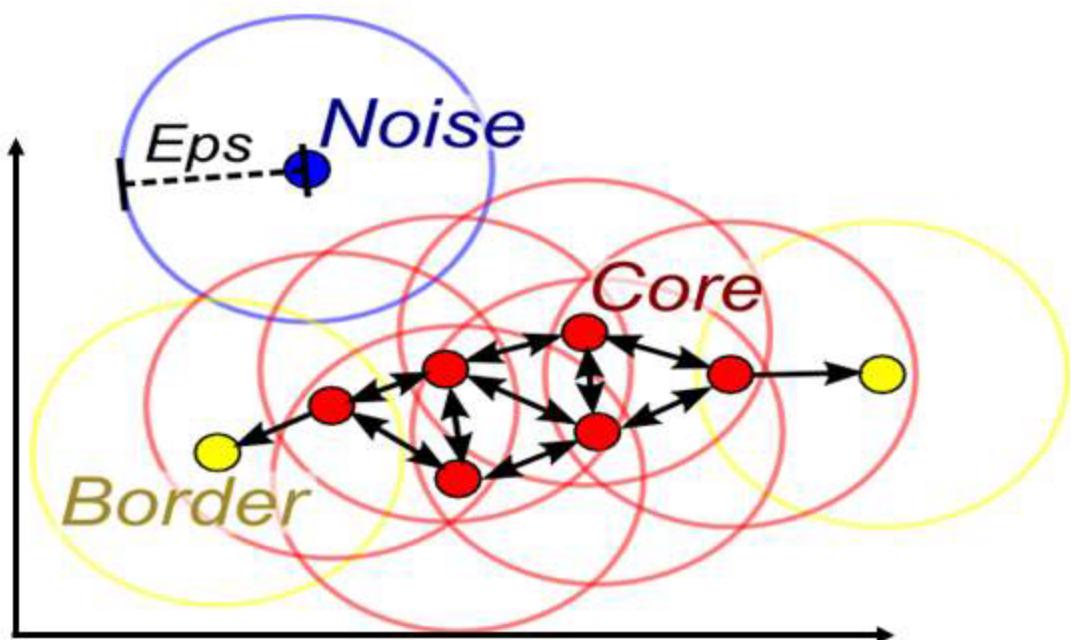
- **Clientes frontera:** son aquellos clientes que están en el borde de un grupo de clientes cercanos entre sí. Es decir, son clientes que tienen a algún cliente cercano, por lo que tampoco es un cliente ‘raro’ por sus características e irá asignado a un segmento en concreto, pero bien podría estar cerca de otro segmento. Está en la frontera de un segmento propiamente hablando.



- **Clientes atípicos:** son aquellos clientes que no tienen a nadie parecido a ellos, es decir, próximos a ellos. Estos son los que podríamos denominar clientes 'raros'.



En el anterior ejemplo DBSCAN clasificaría los puntos de la siguiente manera:



Como vemos, el cliente azul no tiene a ningún cliente próximo a él (a una distancia inferior a Eps). Los puntos rojos y amarillos están conectados unos con otros con clientes cercanos. Podemos ir dando saltos de tamaño inferior a Eps y llegar desde cualquier punto a cualquier punto, solo que diferenciamos los clientes que tienen varios clientes cerca (los rojos, clientes core) de los clientes que tienen solo un cliente cerca (los amarillos, clientes border).

De esta manera, sólo debemos decirle a DBSCAN dos cosas (hiperparámetros, formalmente):

1

Eps: epsilon, que es la distancia mínima que tienen que estar dos clientes para considerarse próximos (parecidos).

2

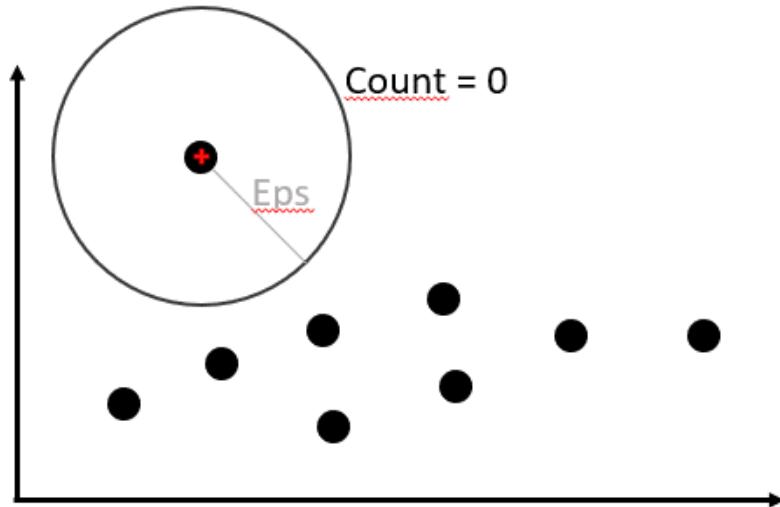
El número de clientes mínimo que tienes que tener cerca para clasificarte como cliente atípico o como cliente Core o Frontera. En el anterior ejemplo este valor es 1, es decir, si tienes al menos un cliente parecido a ti, entonces serás clasificado como un cliente Core o Frontera, pero, si no es así, entonces serás clasificado como cliente atípico.

Los valores más adecuados para estos hiperparámetros los aprenderemos a base de probar distintos valores en cada ejemplo que tengamos. No hay forma de saber *a priori* cuáles pueden ser valores adecuados sin un conocimiento experto.

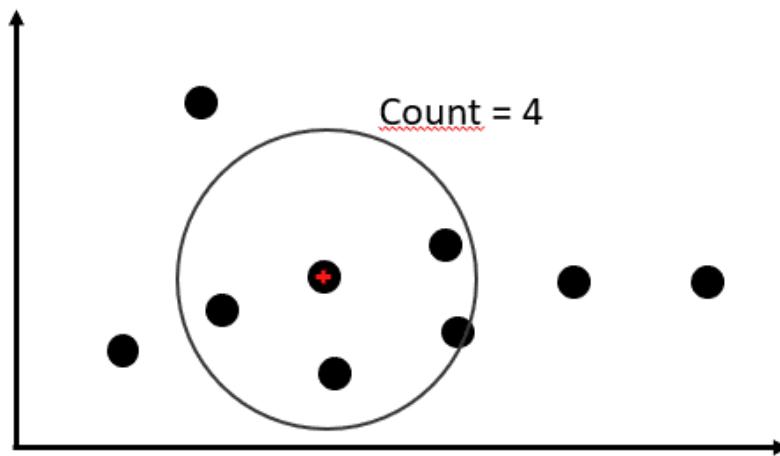
Con estos conceptos claros podemos definir los pasos que sigue DBSCAN para segmentar a nuestros clientes:

Paso 1

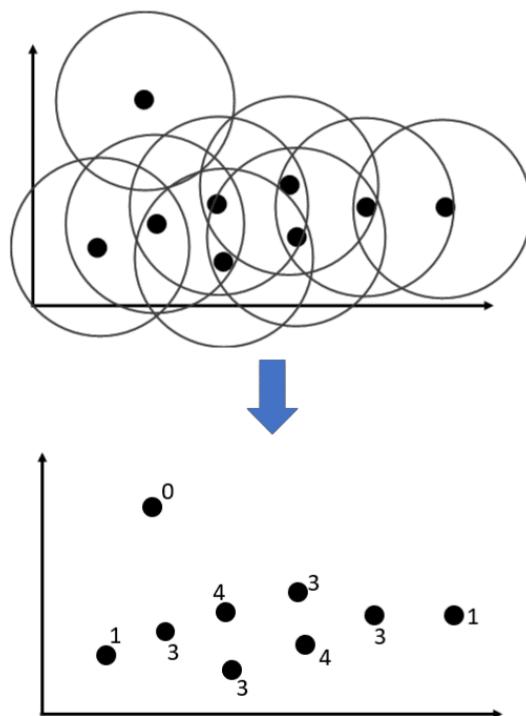
Para cada cliente contamos cuántos clientes hay a una distancia inferior a Eps. Por ejemplo, el cliente que está más arriba tiene 0 clientes a una distancia inferior a Eps:



Mientras que el siguiente cliente tiene a 4 clientes cerca de él:



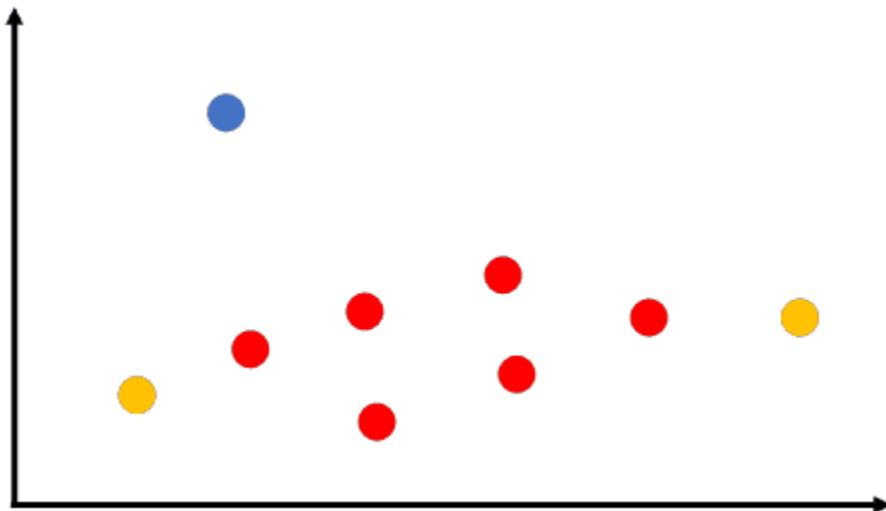
De esta manera DBSCAN va cliente a cliente haciendo el mismo ejercicio:



Paso 2

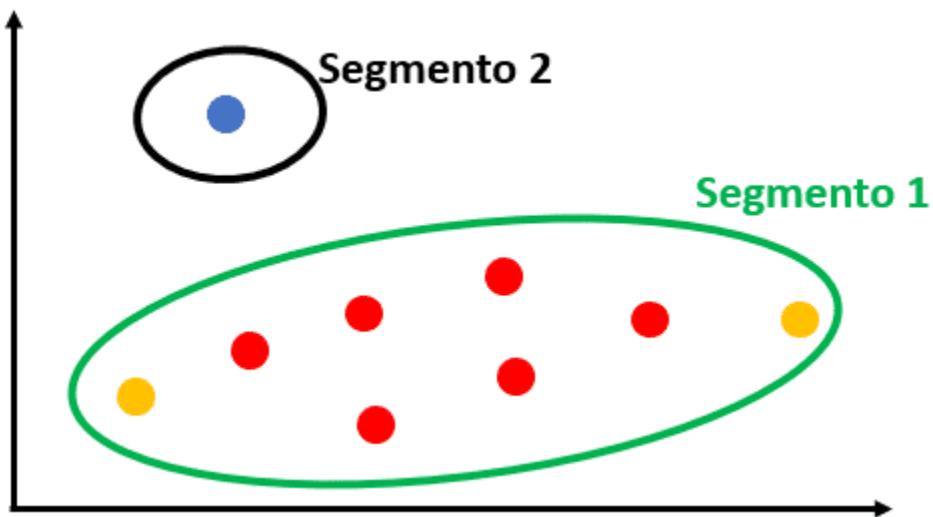
Decidimos el **mínimo de clientes que tiene que haber cerca es 1** para ser core o frontera, y tenemos la clasificación que vimos antes donde:

- **Azul**: clientes atípicos, que tienen menos de 1 cliente cerca.
- **Rojo**: clientes core, que tienen más de 1 cliente cerca.
- **Amarillo**: clientes frontera, que tienen a 1 cliente cerca.

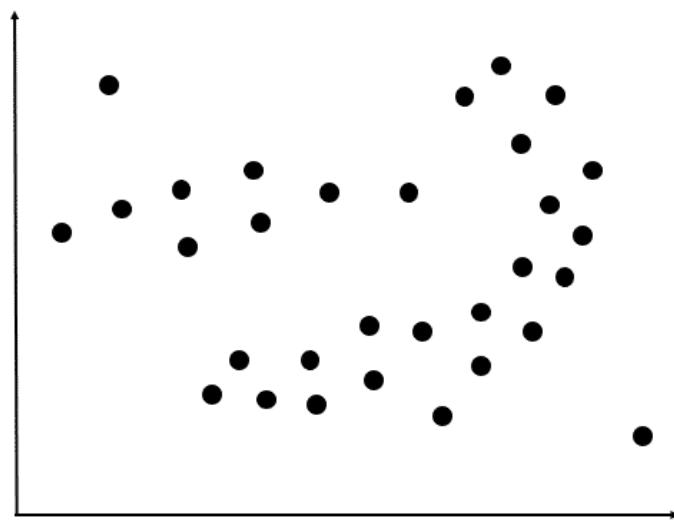


Paso 3

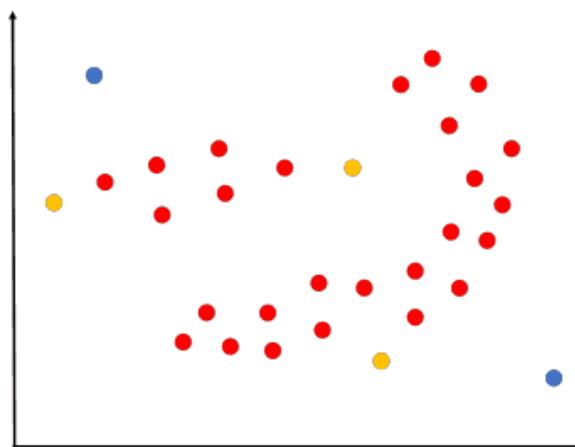
Y, por último, DBSCAN agrupa en un mismo segmento a todos los clientes que pueden conectarse entre sí dando pasos inferiores a Eps: de esta forma tenemos un primer segmento de clientes y un segundo segmento con un cliente atípico.



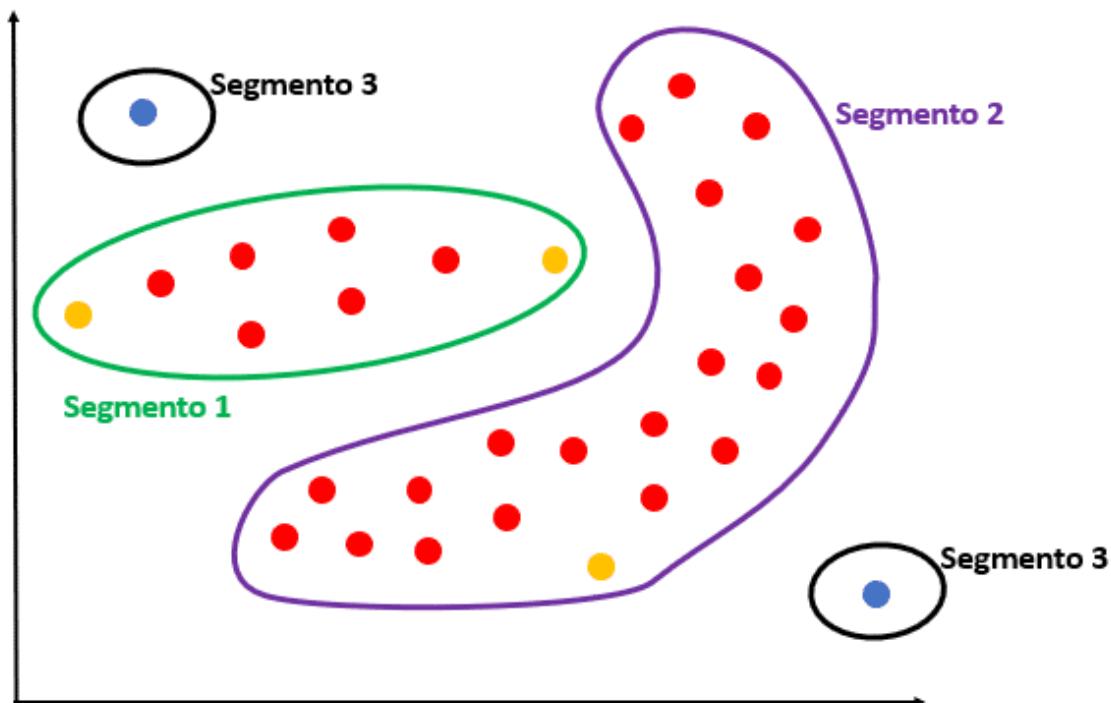
Este ejemplo es muy básico y excepto un cliente, el resto se agrupan en un mismo segmento, lo cual es muy poco interesante, pero esto es debido a la sencillez del ejemplo por propósitos didácticos. Si tuviéramos la siguiente nube de puntos algo más compleja:



DBSCAN haría la siguiente clasificación:



De esta manera, DBSCAN agruparía los puntos azules en un mismo clúster de clientes atípicos (esta una de las principales puntos fuertes de usar DBSCAN, ya que tiene una capacidad muy alta de detectar outliers) y diferenciaría dos segmentos de puntos que se conectan densamente unos con otros:



Con todo esto ya tenemos tres tipos de modelos con lo que poder llevar a cabo una segmentación de clientes. En la siguiente sección hablaremos de dos conceptos técnicos importantes para terminar de cerrar nuestra teoría de análisis clúster.

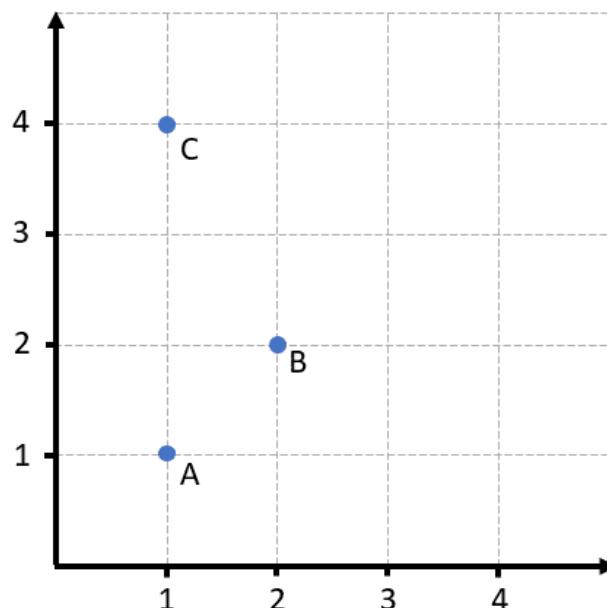
Conceptos técnicos

X Edix Educación

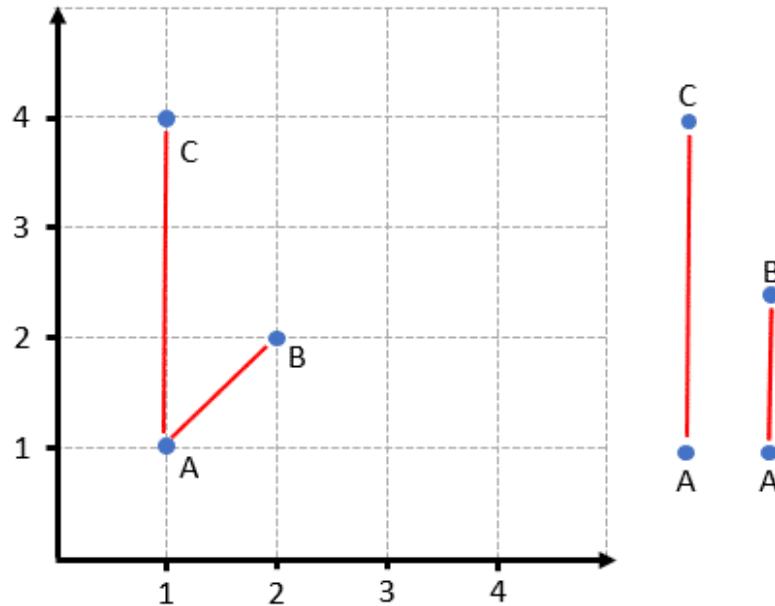
Distancias

En todo momento hemos hablado de si un cliente es ‘parecido’ a otro. Aunque en la vida real tenemos una intuición clara de qué son dos cosas parecidas o diferentes, los ordenadores no entienden nada de esto. Es necesario, por lo tanto, que definamos matemáticamente una manera de medir cómo de parecidos o distintos son dos elementos cualesquiera o dos clientes para nuestra asignatura, y para ello están las denominadas *distancias*.

Si tuviéramos los puntos A con coordenadas (1, 1), B con coordenadas (2, 2) y C con coordenadas (1, 4) ¿Qué punto está más cerca de A? ¿B o C? Si pintamos los puntos, la respuesta sería muy fácil:



Pero si tuvieras que explicar por qué, ¿qué dirías? Probablemente cogerías lápiz y papel, pintarías una línea desde B hasta A y desde C hasta A y me dirías que claramente la línea de B a A es la más corta.



Con esto nuestro profesor de educación plástica del colegio estaría muy contento, pero el de matemáticas no tanto. Lo que estamos midiendo con cada línea roja es la distancia entre dos puntos, pero concretamente estamos utilizando la **distancia Euclídea**, que tiene la siguiente fórmula para medir distancias desde un punto (x_1, y_1) a un punto (x_2, y_2) :

$$D((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Con esta fórmulas sabemos cuánto miden cada una de las líneas rojas, y no nos hace falta ni dibujar ni hacer ninguna comparativa visual, puesto que podemos decir con certeza que:

La distancia de A a B es de $\sqrt{(1 - 2)^2 + (1 - 2)^2} = \sqrt{(-1)^2 + (-1)^2} = \sqrt{1 + 1} = \sqrt{2} \approx 1.41$

La distancia de A a C es de $\sqrt{(1 - 1)^2 + (1 - 4)^2} = \sqrt{(0)^2 + (-3)^2} = \sqrt{0 + 9} = \sqrt{9} = 3$

Por tanto $D(A, C) = 3 > 1.41 = D(A, B)$, es decir, $D(A, C) > D(A, B)$, como escribiría un matemático formalmente.

Esto mismo es lo que hacen los modelos de clustering cuando tienen que medir cómo de parecidos (o distintos) son dos clientes entre sí, solo que en vez de usar dos variables como en el ejemplo anterior, usan todas ellas y, por defecto, suelen emplear la **distancia Euclídea** que acabamos de ver.

Existen muchos otros tipos de métricas, no solo existe la Euclídea, y cada una tiene sus propiedades. No obstante, el entendimiento de métricas queda fuera de alcance de este curso donde simplemente debemos tener clara la intuición de distancia para saber cómo están trabajando los ordenadores cuando empleamos estos algoritmos.

Métricas de calidad

Antes hemos comentamos la gran dificultad que hay *a priori* para saber cuál de los modelos de segmentación que conocemos va a ofrecernos los mejores resultados. Habrá ocasiones que por los datos que empleemos sea un K-Means el que nos genere unos buenos segmentos, en otras será DBSCAN o algunos de los jerárquicos, pero esto no es posible saberlo sin un conocimiento experto del dominio en el que usamos los datos y de las propiedades que tienen los modelos que empleemos.

Esto no es un problema grave, simplemente tendremos que probar distintos modelos. Pero hay un aspecto que es importante, podríamos decir: "venga vale, acepto que no puedo saber cuál va a ser el mejor modelo, pero, una vez que he utilizado varios modelos, ¿cómo puedo comparar los resultados y elegir el mejor? Porque no puedo ponerme a comparar cliente a cliente o segmento a segmento cómo se compone".

Pues bien, para poder comparar la calidad de los resultados, es decir, la calidad de los segmentos que nos genera cada modelo, existen las denominadas métricas de calidad.

No te sorprenderá saber que existen muchas métricas distintas, cada una con sus propiedades (como siempre), pero una de las más conocidas se denomina **Silhouette**.

Podemos encontrar una definición exacta, matemáticamente hablando, en la [wiki](#) (reto para los valientes), pero no nos tienen que despistar todas esas fórmulas y definiciones que unos buenos matemáticos han definido para mantener las matemáticas como la ciencia más robusta que hay.

El Silhouette es un número entre -1 y 1. Cuanto más cerca esté de -1, peor calidad tendrán los segmentos que hemos generado, y al contrario, cuanto más cerca esté de 1, más calidad tendrán dichos segmentos.

$$-1 < \text{Silhouette} < +1$$

En los dos siguientes fastbooks veremos ejemplos de segmentaciones donde aprenderemos a calcular el Silhouette después de haber aplicado un modelo junto al resto de conceptos y técnicas que hemos visto en estos dos primeros fastbooks sobre segmentación de clientes.

Conclusiones

 Edix Educación

En este fastbook hemos visto qué utilidades tiene el hacer una segmentación de nuestros clientes, siendo la base de todas ellas la obtención de conocimiento sobre nuestros clientes. Esto los podemos aprovechar para personalizar y optimizar nuestras campañas de marketing o para detectar anomalías.

Después de toda la teoría general sobre segmentación, hemos estudiado y entendido cómo funcionan en detalle tres modelos: **clustering jerárquico, K-Means y DBSCAN**.

Finalmente hemos visto el concepto de *distancias*, que usan los modelos para medir lo parecidos que son dos clientes, y el concepto de *métrica*, para comparar de forma ágil y objetiva los segmentos que nos ofrecen varios modelos. Cerramos con ellos este segundo fastbook con toda la teoría necesaria para hacer una segmentación de clientes.

¡Enhорabuena! Fastbook superado

edix

Creamos Digital Workers