

Contenido

[**6ta Clase – 15/07/2021**](#_heading=h.gjdgxs) **4**

[Aprendizaje supervisado vs No supervisado](#_heading=h.30j0zll) 4

[*Problemas Aprendizaje No supervisado*](#_heading=h.1fob9te) *5*

[Aprendizaje no supervisado: Clustering](#_heading=h.3znysh7) 5

[*Reducción de la dimensionalidad*](#_heading=h.2et92p0) *6*

[*Resumen*](#_heading=h.tyjcwt) *7*

[*Clústering: concepto y ejemplos*](#_heading=h.3dy6vkm) *7*

[*Clústering concepto de distancia*](#_heading=h.1t3h5sf) *8*

[Medidas de proximidad](#_heading=h.4d34og8) 8

[*Variedad de Medidas de proximidad*](#_heading=h.2s8eyo1) *9*

[*Distancia euclidiana*](#_heading=h.17dp8vu) *9*

[Desventajas](#_heading=h.3rdcrjn) 9

[Casos de uso](#_heading=h.26in1rg) 10

[*Similitud de coseno o separación angular*](#_heading=h.lnxbz9) *12*

[Desventajas](#_heading=h.35nkun2) 12

[Casos de uso](#_heading=h.1ksv4uv) 12

[*Distancia de Hamming*](#_heading=h.44sinio) *12*

[Desventajas](#_heading=h.2jxsxqh) 12

[Casos de uso](#_heading=h.z337ya) 13

[*Distancia de Manhattan*](#_heading=h.3j2qqm3) *13*

[Desventajas](#_heading=h.1y810tw) 13

[Casos de uso](#_heading=h.4i7ojhp) 14

[*Distancia de Chebyshev*](#_heading=h.2xcytpi) *14*

[Desventajas](#_heading=h.1ci93xb) 15

[Casos de uso](#_heading=h.3whwml4) 15

[*Distancia de Minkowski*](#_heading=h.2bn6wsx) *16*

[Desventajas](#_heading=h.qsh70q) 16

[Casos de uso](#_heading=h.3as4poj) 16

[*Índice de Jaccard*](#_heading=h.1pxezwc) *16*

[Desventajas](#_heading=h.49x2ik5) 17

[Casos de uso](#_heading=h.2p2csry) 17

[*Distancia de Haversine*](#_heading=h.147n2zr) *18*

[Desventajas](#_heading=h.3o7alnk) 18

[Casos de uso](#_heading=h.23ckvvd) 18

[*Índice de Sørensen-Di*](#_heading=h.ihv636) *19*

[Desventajas](#_heading=h.32hioqz) 19

[Casos de uso](#_heading=h.1hmsyys) 19

[Distintos tipos de Clústering](#_heading=h.41mghml) 20

[*Clasificación de los algoritmos de clústering*](#_heading=h.2grqrue) *20*

[K-means](#_heading=h.vx1227) 22

[Como elegir el K](#_heading=h.3fwokq0) 24

[Notebook](#_heading=h.1v1yuxt) 25

[**7ta Clase – 22/07/2021**](#_heading=h.4f1mdlm) **26**

[Clústering Jerárquico](#_heading=h.2u6wntf) 26

[*Dendrograma*](#_heading=h.19c6y18) *27*

[*Clústering jerárquico divisivo*](#_heading=h.3tbugp1) *29*

[Proceso de clúster jerárquico divisivo: algoritmo DIANA](#_heading=h.28h4qwu) 29

[*Clústering jerárquico aglomerativo*](#_heading=h.nmf14n) *29*

[Proceso de Clústering jerárquico aglomerativo](#_heading=h.37m2jsg) 30

[*¿Cómo se define una medida de disimilitud entre clústeres?*](#_heading=h.1mrcu09) *31*

[Enlace único (Single)](#_heading=h.46r0co2) 31

[Enlace completo (complete)](#_heading=h.2lwamvv) 31

[Vinculación promedio (average)](#_heading=h.111kx3o) 31

[Enlace Pabellón (Ward):](#_heading=h.3l18frh) 32

[Enlace centroide (centroid)](#_heading=h.206ipza) 32

[*Características de Clústering Jerárquico*](#_heading=h.4k668n3) *32*

[*Noteboook*](#_heading=h.2zbgiuw) *33*

[*Coeficiente cofenético*](#_heading=h.1egqt2p) *33*

[Partition Clústering](#_heading=h.3ygebqi) 34

[*El método PAM (Partition around Medoids)*](#_heading=h.2dlolyb) *34*

[Algoritmo](#_heading=h.sqyw64) 35

[Ejemplo](#_heading=h.dinleb8qvhtn) 36

[Clústering basado en densidad](#_heading=h.3cqmetx) 41

[*El método DBSCAN*](#_heading=h.1rvwp1q) *41*

[El algoritmo DBSCAN](#_heading=h.4bvk7pj) 42

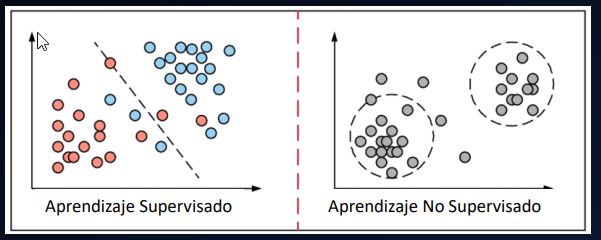
[Notebook](#_heading=h.2r0uhxc) 43

[Clústering basado en densidad](#_heading=h.o10z6p7ftu1l) 45

# 6ta Clase – 15/07/2021

## Aprendizaje supervisado vs No supervisado

* La principal diferencia entre los dos tipos es que el **aprendizaje supervisado** tenemos un **conocimiento previo** de cuáles deberían ser los valores de salida para nuestras muestras. Por lo tanto, el objetivo del aprendizaje supervisado **es aprender una función** que, **dada una muestra de datos y resultados deseados, se aproxima mejor a la relación entre entrada y salida observable en los datos**.
* El **aprendizaje no supervisado**, por otro lado, **no tiene etiquetas**, y por lo tanto su objetivo **es inferir la estructura natural presente dentro de un conjunto de puntos de datos**.



* Las tareas más comunes dentro del **aprendizaje no supervisado** son:
  + **agrupación clustering**
  + **el aprendizaje de representaciones**
  + **la estimación de densidad.**
* En todos estos casos, deseamos conocer la estructura inherente de nuestros datos sin utilizar etiquetas proporcionadas explícitamente. Algunos algoritmos comunes incluyen **agrupamiento de k-medias**, análisis de componentes principales y codificadores automáticos (autoencoders).

*(Agrupar en k-medias ayuda a reducir el problema en menos dimensiones, por ej. Analizo y saco un promedio ponderado de varias variables convirtiéndolas en una sola)*

* Dado que no se proporcionan etiquetas, no existe una forma específica de comparar el rendimiento del modelo en la mayoría de los métodos de aprendizaje No supervisados.

### Problemas Aprendizaje No supervisado

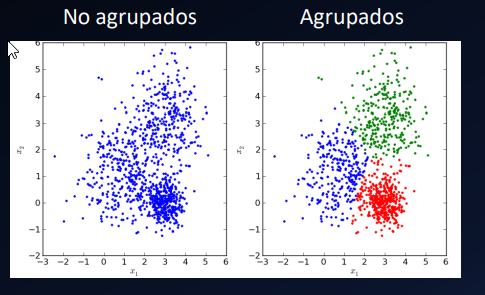
* El aprendizaje no supervisado es más difícil en comparación con las tareas de aprendizaje supervisado.
* *¿Cómo sabemos si los resultados son significativos dado que* ***no hay etiquetas de respuesta disponibles****?* 
  + Necesitamos que un experto mire los resultados (evaluación externa)
  + Definir una función objetiva sobre la agrupación en clústeres (evaluación interna)
* *¿Por qué se necesita el aprendizaje no supervisado a pesar de estos problemas?* 
  + Anotar grandes conjuntos de datos es muy costoso y, por lo tanto, solo podemos etiquetar algunos ejemplos manualmente. Ejemplo: reconocimiento de voz
  + Es muy costoso en tiempo y dinero armar la clase (a mano) cuando no la tenemos y no se puede construir en modo automatizado.
    - A veces podemos targetearlos a todos
  + Pregunta: no se podría aplicar supervisado cuando el es muy costoso?, Cuando se aplicaría en un caso concreto aprendizaje supervisado o no supervisado?
  + Puede haber casos en los que no sepamos en cuántas / en que clases se dividen los datos. Ej. Minería de datos.
  + Es posible que deseemos utilizar la agrupación en clústeres para obtener una idea de la estructura de los datos antes de diseñar un clasificador(algoritmo supervisado de clasificación)
    - Puedo usar un no supervisado para identificar grupos y después cuando obtengo la clase, puedo llevarlo al ámbito de los supervisados.

## Aprendizaje no supervisado: Clustering

* El aprendizaje no supervisado es muy útil en el análisis exploratorio porque puede identificar automáticamente la estructura de los datos.
* Los métodos de agrupación en clústeres no supervisados son un excelente punto de partida. En situaciones en las que es imposible o poco práctico para un ser humano proponer tendencias en los datos, el aprendizaje no supervisado puede proporcionar conocimientos iniciales que luego pueden usarse para probar hipótesis individuales.

*(Ejemplo: tenía 10.000 variables, de las cuales generé dos variables nuevas y a partir de ellas género la agrupación. Otro ejemplo: la edad de la persona que compro una casa, no sirve para predecir el precio de una casa, por lo tanto la puedo quitar del dataset.)*

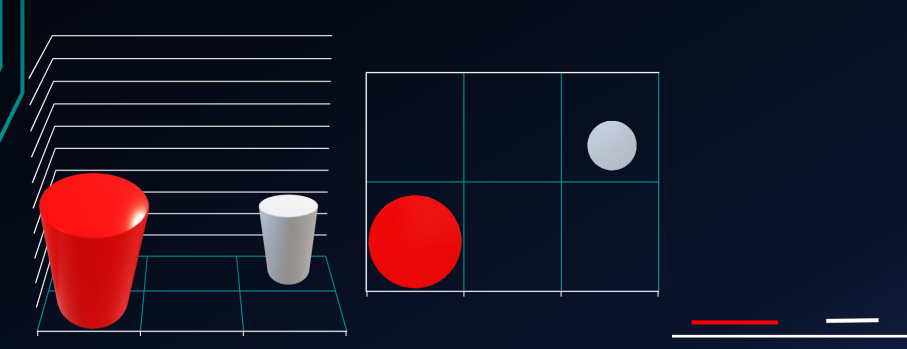
*Método no supervisado, el proceso ignora la variable respuesta que indica a qué grupo pertenece realmente cada observación (si es que existe tal variable). Esta característica diferencia al clústering de las técnicas supervisadas, que emplean un set de entrenamiento en el que se conoce la verdadera clasificación.*



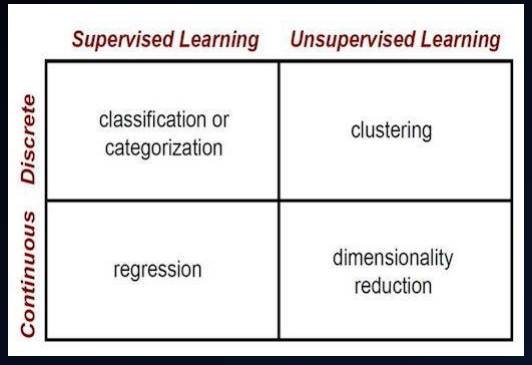
### Reducción de la dimensionalidad

* La **reducción de dimensionalidad** se refiere a los métodos utilizados para representar datos utilizando menos variables.
* La idea es representar nuestros datos utilizando las características “latentes” que interrelacionan nuestras variables iniciales.
* Esta estructura latente que se presenta en forma rala, a menudo se representa con muchas menos variables de las que teníamos al principio, y por lo tanto puede hacer que el procesamiento de datos sea mucho menos intensivo y puede eliminar las funciones redundantes.

*(Siempre es bueno reducir la dimensionalidad. La que más pesa son las columnas más que las filas. Esto conviene tanto para simplificar el modelo como para que el tiempo de procesamiento sea menor)*



### Resumen



Siempre es bueno reducir la dimensionalidad. La que más pesa son las columnas, más que las filas. Esto conviene tanto para simplificar el modelo como para que el tiempo de procesamiento sea menor.

La reducción de la dimensionalidad es solo para variables continuas? ¿Por qué no podría aplicar reducción de dimensionalidad con variables discretas? RTa. **Mario**: En general se aplica a variables continuas, no funciona bien con las variables discretas. Esto es algo generalizado.

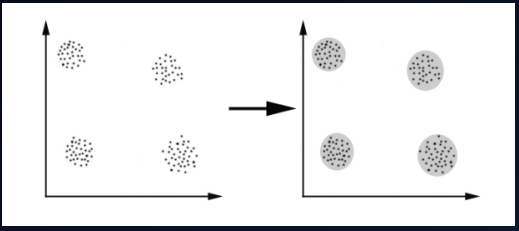
### Clústering: concepto y ejemplos

* La agrupación se refiere a un amplio conjunto de técnicas para encontrar subgrupos o agrupaciones en un conjunto de datos. Esto nos ayuda a dividir las observaciones en grupos distintos para que cada grupo contenga observaciones similares entre sí.
* Por ejemplo, en el escenario del cáncer de mama, los grupos podrían representar el grado del tumor. También es muy útil en marketing para la segmentación del mercado con el fin de identificar un grupo de personas que serían más receptivas a un determinado tipo de producto.

*(Si aplico un no supervisado a un dataset de estos de juguete sacándole la clase, seguramente me va a detectar la clase)*

* Una definición de clústering es "el proceso de organizar objetos en grupos cuyos miembros son similares de alguna manera". Por lo tanto, un grupo es una colección de objetos que son "similares" entre ellos y son "diferentes" a los objetos que pertenecen a otros grupos.

*(Maximiza la similitud de las observaciones dentro del grupo asignado y la diferencia con el resto de los grupos. Los objetos dentro de un grupo son muy similares y los objetos de grupos distintos son muy distintos).*



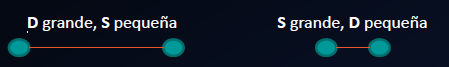
### Clústering concepto de distancia

* *Todos los métodos de clústering tienen una cosa en común, para poder llevar a cabo las agrupaciones necesitan definir y cuantificar la similitud entre las observaciones.*
* *El término distancia se emplea dentro del contexto del clústering como cuantificación de la similitud o diferencia entre observaciones. Si se representan las observaciones en un espacio p dimensional, siendo p el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán, de ahí que se emplee el término distancia.*
* *La característica que hace del clústering un método adaptable a escenarios muy diversos es que puede emplear cualquier tipo de distancia, lo que permite al investigador escoger la más adecuada para el estudio en cuestión. A continuación, se describen algunas de las más utilizada*
* Dado un conjunto de puntos, con una noción de distancia entre puntos, agrupando los puntos en cierto número de grupos, de modo que:
  + Las distancias internas (dentro del grupo) deben ser pequeñas, es decir, los miembros de los grupos son cercanos / similares entre sí.
  + Las distancias externas (entre los grupo) deben ser grandes, es decir, los miembros de diferentes grupos son diferentes.



## Medidas de proximidad

* Para la agrupación en clústeres, necesitamos definir una medida de proximidad para dos puntos de datos. La proximidad aquí significa cuán similares/diferentes son las muestras entre sí.
* **Medida de similitud** ( , 𝑥𝑘 ) es grande si 𝑥𝑖 y 𝑥𝑘 son similares entre sí.
* **Medida de disimilitud** (o distancia) D(𝑥𝑖 , 𝑥𝑘 ) es pequeña 𝑥𝑖 y 𝑥𝑘 son similares entre sí.



* Los métodos de clústering presuponen la existencia de relaciones entre n objetos que se representan con una matriz de tamaño n x n y que esas relaciones se indican utilizando una medida de similitud.

*(Las distancias deben tener ciertas condiciones: no pueden ser negativas, etc. Hay cuestiones*

*Matemáticas que hay que tener en cuenta.)*

* Las matrices de disimilitud (D) deben cumplir con varias propiedades:
* La disimilitud de objetos idénticos es cero, D(a,a) = 0.
* NO Negativas. D sólo puede ser mayor o igual que cero, D(a,b) ≥ 0.
* Los objetos no idénticos pueden ser distinguibles o no. Si a ≠ b entonces D(a,b) ≥ 0. Si a = b, D(a,b) = 0. Esto quiere decir que si no son idénticos, puedo ser o no iguales?, un ejemplo? RTA **Mario**: Si, la distancia entre ellos podría ser igual, similar o distinta.
* Simétricas. Las medidas de disimilitud son simétricas: D(a, b) = D(b, a)
* Desigualdad triangular: D(a,b) ≤ D(a,c) + D(b,c)

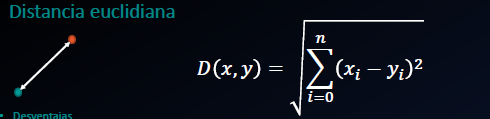
### Variedad de Medidas de proximidad



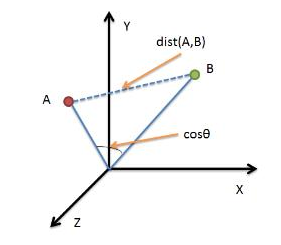
### Distancia euclidiana

#### Desventajas

* Aunque es una medida de distancia común, las distancias calculadas pueden estar sesgadas según las unidades de las entidades. Normalmente, es necesario normalizar los datos antes de utilizar esta medida de distancia.
  + *Es sensible a los datos extremos*
* Además, a medida que aumenta la dimensionalidad de sus datos, la distancia euclidiana se vuelve menos útil porque el espacio de dimensiones superiores no actúa como, intuitivamente, esperaríamos

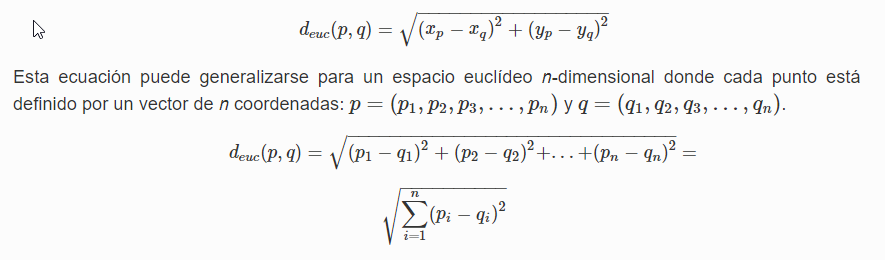


*(Diremos que una distancia es euclidiana cuando pueda encontrarse un espacio vectorial de dimensión igual o inferior a la dimensión del espacio de las variables en el que podamos representar a los individuos por puntos cuya distancia euclídea ordinaria coincida con la distancia utilizada.)*



#### Casos de uso

* La distancia euclidiana funciona muy bien cuando tiene datos de baja dimensión y es importante medir la magnitud de los vectores. Los métodos como kNN y HDBSCAN muestran excelentes resultados listos para usar si se aplica la distancia euclidiana en datos de baja dimensión.
* Aunque se han desarrollado muchas otras medidas para tener en cuenta las desventajas de la distancia euclidiana, sigue siendo una de las medidas de distancia más utilizadas por buenas razones. Es increíblemente intuitiva de usar, fácil de implementar y muestra excelentes resultados en muchos casos de uso.
* La distancia euclídea entre dos puntos p y q se define como la longitud del segmento que une ambos puntos. En coordenadas cartesianas, la distancia euclídea se calcula empleando el teorema de Pitágoras. Por ejemplo, en un espacio de dos dimensiones en el que cada punto está definido por las coordenadas (x,y)
* (x,y), la distancia euclídea entre p y q viene dada por la ecuación:

**

* Una forma de dar mayor peso a aquellas observaciones que están más alejadas es emplear la distancia euclídea al cuadrado. En el caso del clústering, donde se busca agrupar observaciones que minimicen la distancia, esto se traduce en una mayor influencia de aquellas observaciones que están más distantes.
* La siguiente imagen muestra el perfil de dos observaciones definidas por 10 variables (espacio con 10 dimensiones).
* *Ejemplo en R*

***library****(ggplot2)*

*observacion\_a <-* ***c****(4, 4.5, 4, 7.5, 7, 6, 5, 5.5, 5, 6)*

*observacion\_b <-* ***c****(4, 4.5, 4, 7.5, 7, 6, 5, 5.5, 5, 6)* ***+*** *4*

*datos <-* ***data.frame****(observacion =* ***rep****(****c****("a", "b"), each = 10),*

*valor =* ***c****(observacion\_a, observacion\_b),*

*predictor = 1****:****10)*

***ggplot****(data = datos,* ***aes****(x =* ***as.factor****(predictor), y = valor,*

*colour = observacion))* ***+***

***geom\_path****(****aes****(group = observacion))* ***+***

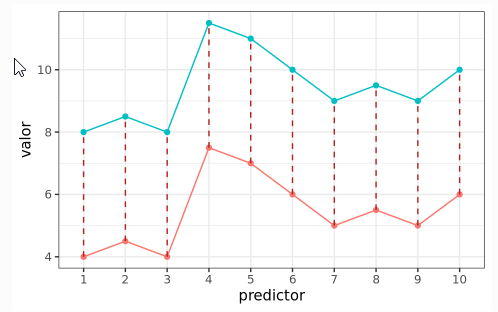
***geom\_point****()* ***+***

***geom\_line****(****aes****(group = predictor), colour = "firebrick", linetype = "dashed")* ***+***

***labs****(x = "predictor")* ***+***

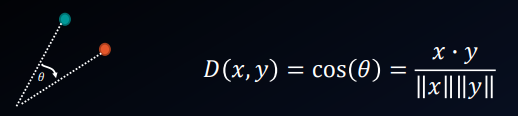
***theme\_bw****()* ***+***

***theme****(legend.position = "none")*

**

* La distancia euclídea entre las dos observaciones equivale a la raíz cuadrada de la suma de las longitudes de los segmentos rojos que unen cada par de puntos. Tiene en cuenta por lo tanto el desplazamiento individual de cada una de las variables.

### Similitud de coseno o separación angular



La similitud de coseno se ha utilizado a menudo como una forma de contrarrestar el problema de la distancia euclidiana con alta dimensionalidad. La similitud del coseno es simplemente el coseno del ángulo entre dos vectores. Dos vectores con exactamente la misma orientación tienen una similitud de coseno de 1, mientras que dos vectores diametralmente opuestos entre sí tienen una similitud de -1.

#### Desventajas

Una de las principales **desventajas** de la similitud de coseno es que no se tiene en cuenta la magnitud de los vectores, simplemente su dirección. ¿A que se refiere con magnitud? ¿Se refiere a cuán largo es el vector? RTA: **Mario**: Se refiere a la distancia un del vector el cual está entre dos puntos

#### Casos de uso

* Usamos la similitud de coseno a menudo cuando tenemos datos de alta dimensión y cuando la magnitud de los vectores no es importante.
* Para los análisis de texto, esta medida se usa con bastante frecuencia cuando los datos están representados por recuentos de palabras. Por ejemplo, cuando una palabra aparece con más frecuencia en un documento que en otro, esto **no significa necesariamente** que un documento esté más relacionado con esa palabra.

### Distancia de Hamming



La distancia de Hamming es el número de valores que son diferentes entre dos vectores. Normalmente se utiliza para comparar dos cadenas binarias de igual longitud.

#### Desventajas

Como era de esperar, la distancia de Hamming es difícil de usar cuando dos vectores no tienen la misma longitud. Además, no tiene en cuenta el valor real siempre que sean diferentes o iguales. (No entiendo la frase) Por lo tanto, no se recomienda utilizar esta medida de distancia cuando la **magnitud** es una medida importante. Rta **Mario**: No mide cuán distinto es, sino que es distinto. Pone más énfasis en lo que se parece que en lo que no se parece.

La distancia de Hamming entre dos cadenas de igual longitud es el número de posiciones en las que los símbolos correspondientes son diferentes. En otras palabras, es el número de sustituciones necesarias para transformar una cadena en otra.

**Example**

If

s1 = 'cats'

s2 = 'dogs'

then

d = 3

because three substitutions are required to transform s1 into s2.

cats => dats (substitute 'd' for 'c')

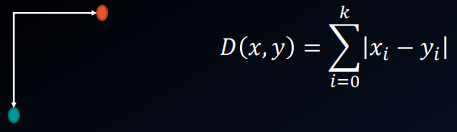
dats => dots (substitute 'o' for 'a')

dots => dogs (substitute 'g' for 't')

#### Casos de uso

Los casos de uso típicos incluyen la corrección/detección de errores cuando los datos se transmiten a través de redes informáticas. Además, también puede utilizar la distancia de Hamming para medir la distancia entre variables categóricas.

### Distancia de Manhattan



La distancia de Manhattan, a menudo llamada distancia de taxi o distancia de bloque de ciudad, calcula la distancia entre vectores. La distancia de Manhattan se refiere a la distancia entre dos vectores si solo pudieran moverse en ángulos rectos. No hay ningún movimiento diagonal involucrado en el cálculo de la distancia.

No entiendo la fórmula

Rta **Mario**: Es la diferencia entre componentes: a en x menos b en y, esto en dos dimensiones.

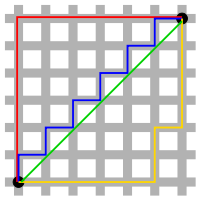
#### Desventajas

* Aunque la distancia de Manhattan parece funcionar bien para datos de alta dimensión, es una medida que es algo menos intuitiva que la distancia euclidiana.
* Además, es más probable que dé un valor de distancia más alto que la distancia euclidiana, ya que no es el camino más corto posible.

#### C*a*sos de uso

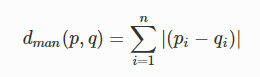
Cuando su conjunto de datos tiene atributos discretos y/o binarios, Manhattan parece funcionar bastante bien ya que toma en cuenta las rutas que efectivamente se podrían tomar para conectar distintos valores de una variable.

En una ciudad como Nueva York, formada en gran medida por una matriz de edificios, la distancia entre dos puntos nunca es la distancia "en línea recta". Más bien viene determinada por los bordes de los bloques que hay que rodear. Consideremos el siguiente diagrama en el que hay dos puntos (en negro) cuya distancia queremos calcular:

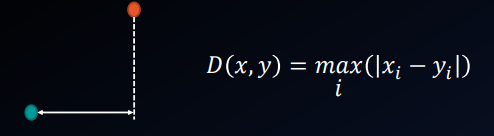


Aun cuando la distancia "ordinaria" sea la correspondiente a la línea verde, si los recuadros blancos son edificios, estaremos obligados a bordearlos. En estas condiciones, tanto el camino rojo como el azul y el amarillo son equivalentes y representan la distancia Manhattan mínima entre los puntos.

La distancia de Manhattan define la distancia entre dos puntos p y q como la sumatoria de las diferencias absolutas entre cada dimensión. Esta medida se ve menos afectada por outliers (es más robusta) que la distancia euclídea debido a que no eleva al cuadrado las diferencias.

**

### Distancia de Chebyshev



La distancia de Chebyshev se define como la mayor diferencia entre dos vectores a lo largo de cualquier dimensión de coordenadas. En otras palabras, es simplemente la distancia máxima a lo largo de un eje.

Es la distancia a una de los ejes del otro punto

#### Desventajas

Chebyshev se usa generalmente en casos de uso muy específicos y por esta razón, se sugiere usarlo solo cuando esté absolutamente seguro de que se adapta a su caso de uso.

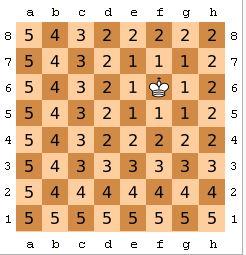
#### Casos de uso

Se puede usar para extraer el número mínimo de movimientos que necesita una pieza de ajedrez para ir de una casilla a otra. Además, puede ser una medida útil en juegos que permiten un movimiento de 8 direcciones sin restricciones.

La distancia de Chebysev es la mayor de las diferencias entre sus dimensiones. Por ejemplo, si las posiciones x (en un plano) de dos puntos son 2 y 5 (esto es, si la diferencia entre ellas es 3) y las posiciones y son -3 y 4 (esto es, si la diferencia entre ellas es 7), la distancia de Chebyshev es de 7 (valor máximo de 3 y 7).Distancia de Chebyshev

Esta distancia viene dada por la siguiente expresión:

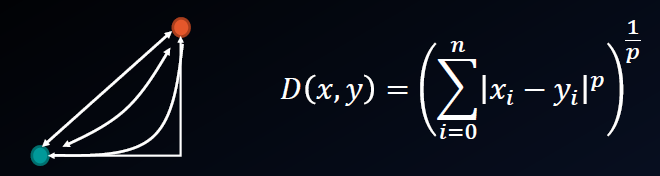
Esta distancia también se denomina distancia del tablero de ajedrez, pues representa el número mínimo de movimientos que el rey necesita para moverse entre dos casillas (pues el rey en el ajedrez puede moverse horizontalmente, verticalmente y en diagonal):



¿Alguna vez has jugado al ajedrez? El rey puede moverse a cualquiera de las 8 casillas adyacentes dando un paso (Figura 1.11). ¿Cuántos pasos necesita el rey para caminar desde la cuadrícula (x1, y1) a la cuadrícula (x2, y2)? Inténtalo tú mismo. Encontrará que el número mínimo de pasos es siempre max (| x2-x1 |, | y2-y1 |) pasos. Existe un método de medición de distancia similar llamado distancia de Chebyshev (norma L∞).

La distancia de Chebyshov entre dos casillas del tablero de ajedrez coincide con el mínimo número de movimientos necesarios para que la ficha del rey se desplace entre ellas. Esto es debido a que el rey se puede mover tanto en diagonal como en horizontal o vertical. En el tablero figuran las distancias de Chebyshov desde la casilla f6.

### Distancia de Minkowski



* La distancia de Minkowski es una medida un poco más compleja que la mayoría y puede utilizarse en un espacio donde las distancias se pueden representar como un vector que tiene una longitud.
* Lo más interesante de esta medida de distancia es el uso del parámetro **p**. Podemos usar este parámetro para manipular las métricas de distancia para que se parezcan mucho a otras métricas de distancia:
  + p = 1 (Distancia de Manhattan)
  + p = 2 (Distancia euclidiana)
  + p = ∞ (Distancia de Chebyshev)

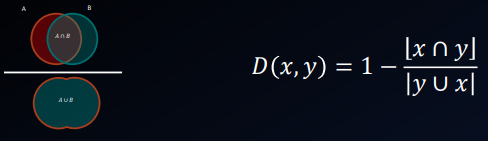
#### Desventajas

Minkowski tiene las mismas desventajas que las medidas de distancia que representan. Además, el parámetro **p** en realidad puede ser problemático para trabajar, debido a que encontrar el valor correcto puede ser bastante ineficiente.

#### Casos de uso

La ventaja de **p** es la posibilidad de iterar sobre él y encontrar la medida de distancia que funcione mejor para su caso de uso.

### Índice de Jaccard



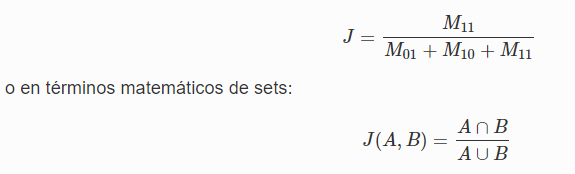
* El índice de Jaccard es una métrica que se utiliza para calcular la similitud y diversidad de conjuntos.
* Es el tamaño de la intersección dividido por el tamaño de la unión de los conjuntos de muestras.
* En la práctica, es el número total de entidades similares entre conjuntos dividido por el número total de entidades.
* Por ejemplo, si dos conjuntos tienen 1 entidad en común y hay 5 entidades diferentes en total, entonces el índice de Jaccard sería 1/5 = 0,2.

#### Desventajas

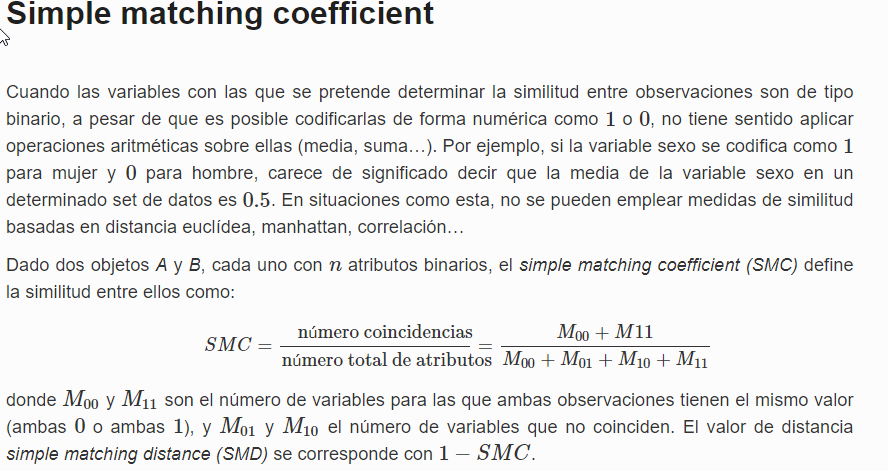
* Una gran desventaja del índice Jaccard es que está muy influenciado por el tamaño de los datos.

#### Casos de uso

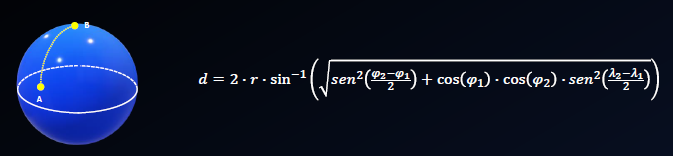
* El índice Jaccard se utiliza a menudo en aplicaciones donde se utilizan datos binarios o binarizados. De manera similar, se puede usar en el análisis de similitud de texto para medir cuánta superposición de palabras existe entre documentos.
* El índice Jaccard o coeficiente de correlación Jaccard es similar al simple matching coefficient (SMC). La diferencia radica en que el SMC tiene el término M00
* M00 en el numerador y denominador, mientras que el índice de Jaccard no. Esto significa que SMC considera como coincidencias tanto si el atributo está presente en ambos sets como si el atributo no está en ninguno de los sets, mientras que Jaccard solo cuenta como coincidencias cuando el atributo está presente en ambos sets.



* *La distancia de Jaccard (1−J) supera a la simple matching distance en aquellas situaciones en las que la coincidencia de ausencia no aporta información. Para ilustrar este hecho, supóngase que se quiere cuantificar la similitud entre dos clientes de un supermercado en base a los artículos comprados. Es de esperar que cada cliente solo adquiera unos pocos artículos de los muchos disponibles, por lo que el número de artículos no comprados por ninguno (M00). (M00) será muy alto. Como la distancia de Jaccard ignora las coincidencias de tipo M00. M00, el grado de similitud dependerá únicamente de las coincidencias entre los artículos comprados.*



### Distancia de Haversine



La distancia de Haversine es la distancia entre dos puntos en una esfera dadas sus longitudes y latitudes. Es muy similar a la distancia euclidiana en que calcula la línea más corta entre dos puntos.

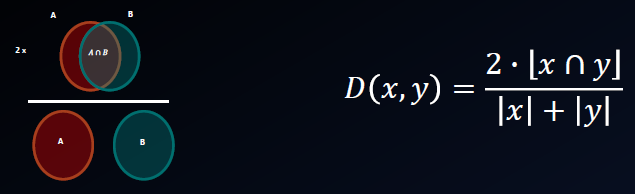
#### Desventajas

Una desventaja de esta medida de distancia es que se supone que los puntos se encuentran en una esfera. En la práctica, esto rara vez ocurre.

#### Casos de uso

Como era de esperar, la distancia de Haversine se utiliza a menudo en la navegación. Por ejemplo, puede usarlo para calcular la distancia entre dos países cuando vuela entre ellos.

### Índice de Sørensen-Di



El índice de Sørensen-Dice es muy similar al índice de Jaccard. Sin embargo es un poco más intuitivo porque el porcentaje de superposición entre dos conjuntos varía entre 0 y 1.

#### Desventajas

Al igual que el índice de Jaccard, pesa cada elemento de forma inversamente proporcional al tamaño del conjunto relevante en lugar de tratarlos por igual. No entiendo la frase.

#### Casos de uso

Los casos de uso son similares, si no iguales, a los del índice Jaccard. En general está asociado a tareas de segmentación de imágenes o análisis de similitud de texto.

Para un mismo algoritmo, podes usar distintas formas de medir las distancias para ver como agrupa los datos.

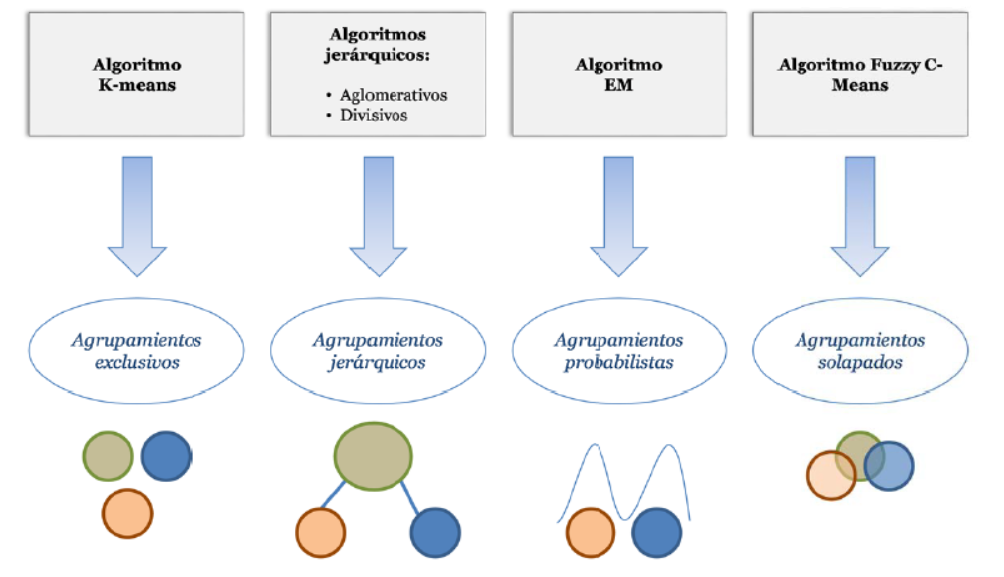
Me gustaría ver un ejemplo práctico para saber en qué tipo de Dataset se puede aplicar cada una de estas medidas de proximidad, para comprender cuándo aplicar cada una.

Vimos el notebook: 6ta Clase - Distancias

[https://www.cienciadedatos.net/documentos/37\_clústering\_y\_heatmaps#Introducci%C3%B3n](https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps#Introducci%C3%B3n)

## Distintos tipos de Clústering

Dada la utilidad del clústering en disciplinas muy distintas (genómica, marketing…), se han desarrollado multitud de variantes y adaptaciones de sus métodos y algoritmos. Pueden diferenciarse tres grupos principales:

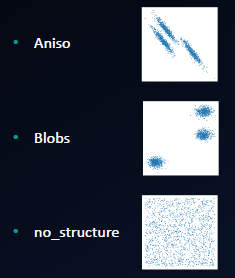
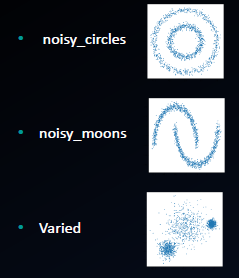


* **Partitioning Clústering:** Este tipo de algoritmos requieren que el usuario especifique de antemano el número de clústers que se van a crear (K-means, K-medoids, CLARA).
* **Hierarchical Clústering:** Este tipo de algoritmos no requieren que el usuario especifique de antemano el número de clústers. (agglomerative clústering, divisive clústering).
* **Agrupamientos Probabilistas:** Los clústeres se generan mediante algún método probabilístico. El ejemplo más representativo sea posiblemente el algoritmo Expectación-Maximización (EM).
* **Métodos que combinan o modifican los anteriores** (hierarchical K-means, fuzzy clústering, model based clústering y density based clústering).

### Clasificación de los algoritmos de clústering

* Agrupación exclusiva (K-mean)
* Clústeres superpuestos (K-mean fuzzy)
* Agrupación jerárquica (Hierarchical clústering)
* Agrupación probabilística (Mixture of Gaussians)
* En el primer caso, los datos se agrupan de forma mutuamente exclusiva, de modo que si un determinado punto de datos pertenece a un grupo definido, no podría incluirse en otro grupo.
* Por el contrario, el segundo tipo, el agrupamiento superpuesto, utiliza conjuntos difusos para agrupar datos, de modo que cada punto puede pertenecer a dos o más grupos con diferentes grados de pertenencia.
* Un algoritmo de agrupamiento jerárquico se basa en la unión entre los dos grupos más cercanos. La condición inicial se realiza configurando cada punto de datos como un grupo. Después de algunas iteraciones, llega a los grupos finales deseados.

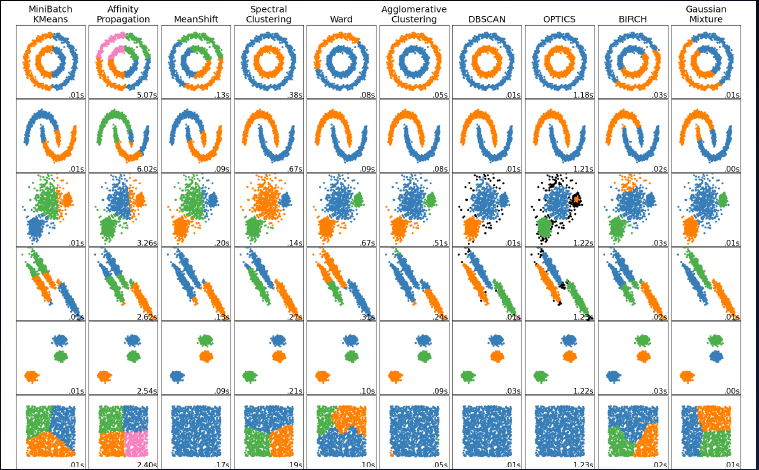
Puntos definidos por dos variables que se distribuyen de las siguientes formas:



Cuanto más variables tengo, más observaciones tenemos que tener para ver la variabilidad.

Siempre hay que practicar un método para reducir variables y luego la clasterización.

La reducción de variables aplica tanto para supervisados y no supervisados.



Cada gráfico muestra cómo clasifica cada algoritmo a cada Dataset

El último dataset no tiene estructura. A este dataset sin estructura lo clasifica bien los algoritmos a los que no tenés que decirle el k

X: Tipos de algoritmos para hacer Clústering. Y: Cada dataset

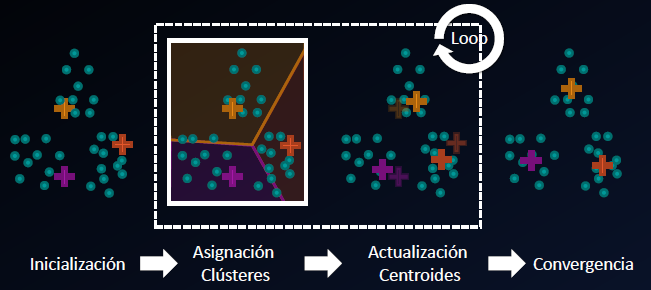
Esto se puede ver en dos o tres dimensiones, luego para más dimensiones, no me puedo imaginar nada.

Hay métricas para saber si funciona o no. Es más fácil darse cuenta que no funciona. Por esto es importante disminuir la dimensional dad.

No voy a poder ver la estructura intrínseca de los datos. Tengo primero que ver aplicando distintos algoritmos o alguien que conozca o ya haya hecho algo. Es RECOMENDABLE aplicar lo que ya otros están aplicando, ver lo que otros analistas de Ciencias de datos han aplicado y ver cual tiene relación con mi problema. Siempre conviene leer Papers, foros, etc. O sumarse a alguna clase donde expliquen algo específico del problema.

#### K-means

* El algoritmo de agrupación en clúster más común y simple que existe es el agrupamiento de K-Means.
* Este algoritmo sigue un procedimiento simple y fácil **y requiere que se le indique a priori un cierto número de grupos (supongamos k grupos) que se piensan existen en conjunto de Datos**.
* La idea central es definir k centros, uno para cada grupo. Estos centroides deben colocarse de manera inteligente debido a que su ubicación puede dar lugar a resultados diferentes.
* Por lo tanto, la mejor opción es colocarlos lo más lejos posible entre sí.
* El siguiente paso es tomar cada punto que pertenece a un conjunto de datos dado y asociarlo al centroide más cercano. Cuando no queda ningún punto pendiente, se completa el primer paso y se realiza un agrupamiento preliminar.
* En este punto, necesitamos volver a calcular nuevos k centroides y realizar una nueva asociación entre los datos del conjunto analizado y los nuevos centroides más cercanos.
* Se ha generado un bucle. Como resultado de este ciclo, podemos notar que los k centroides cambian su ubicación paso a paso hasta que no se realizan más cambios. En otras palabras, los centroides ya no se mueven.
* Como resultado de este ciclo, los k centroides cambian su ubicación paso a paso hasta que los centroides no se mueven.



1. Define tres centroide y calcula todas las distancias a todos los puntos
2. Hace lo mismo cambiando los centroides
3. Los hace hasta que encuentra un valor del error cuadrático que converja. Si el error cuadrático es mayor, cambia la ubicación de los puntos y vuelve a calcular. Lo hace hasta que error deja de cambiar. Aquí es cuando el error converge.

Este algoritmo tiene como objetivo minimizar una función objetivo, en este caso una función de error cuadrático. La función objetivo



𝑘=número de clústeres;

𝑛=número de muestras,

𝑥𝑖=muestra𝑖,

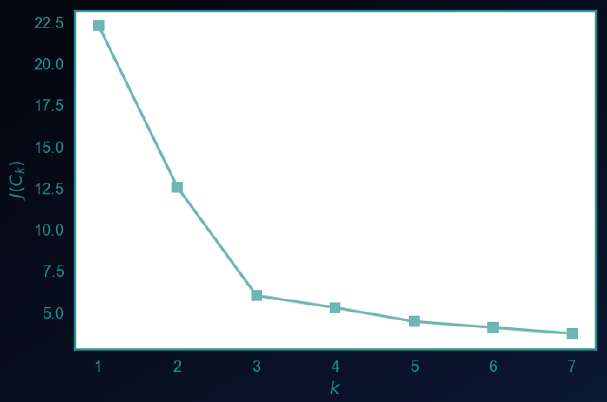
𝑐𝑗=centroide clúster**𝑗**

Es una medida de distancia elegida entre un punto de datos 𝑥𝑖 y el centroide 𝑐𝑗.

#### Como elegir el K

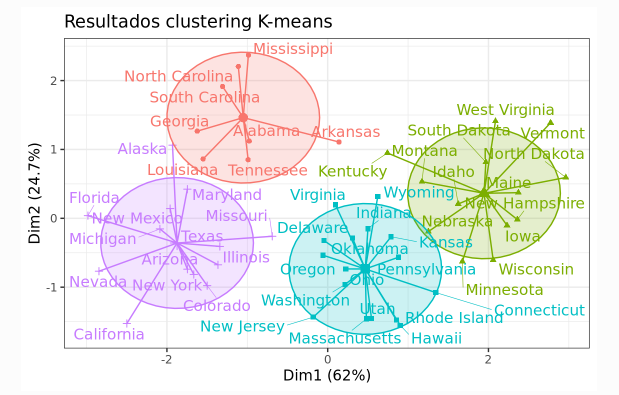
* Los resultados dependen del valor de k y no existe una forma óptima de describir el mejor "k".
* Este último problema es particularmente problemático, ya que a menudo no tenemos forma de saber cuántos clústeres existen.
* Desafortunadamente, no existe una solución teórica general para encontrar el número óptimo de conglomerados para un conjunto de datos dado.
* Un enfoque que se suele usar es comparar los resultados de múltiples corridas con diferentes k y elegir la mejor de acuerdo con un criterio dado, pero debemos tener cuidado porque aumentar k da como resultado valores de función de error más pequeños, pero también aumenta el riesgo de sobreajuste (overfitting).
* Por lo general, cuando se trabaja con k-medias, optimizamos la suma de las distancias al cuadrado entre las observaciones y sus centroides. Pero hay un problema: el óptimo se alcanza cuando el número de centroides es igual al número de observaciones y por lo tanto cada observación será un grupo separado.
* Para evitar ese caso, deberíamos elegir una serie de conglomerados después de los cuales una función J (C) disminuye con menor rapidez
* Para evitar ese caso, deberíamos elegir un k para el cual la función J disminuye con menor rapidez.

 (esta es la función del error)



La regla del codo o del quiebre. Eso es ojímetro puro. No hay un modo analítico

[https://www.cienciadedatos.net/documentos/37\_clústering\_y\_heatmaps#Introducci%C3%B3n](https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps#Introducci%C3%B3n)



<https://interactivechaos.com/es/manual/tutorial-de-machine-learning/gestion-de-valores-nulos>

## Notebook

6ta Clase - Credit Alemania

6ta Clase - KMeans

6ta Clase - Distancias

6ta Clase - Dist\_Clúster

6ta Clase - Mall

Tengo que ver los notebook de esta clase

# 7ta Clase – 22/07/2021

## Clústering Jerárquico

El objetivo de la agrupación en clústeres jerárquica es crear una jerarquía de clústeres.

Esto puede ser útil cuando se desea flexibilidad en la cantidad de clústeres que se desea agrupar.

No le decimos el número de clúster que tiene que haber. (Ejemplo: Amazon, estudiantes que estudian cierta carrera o en caso de animales: vertebrados / invertebrados y así hasta llegar a la mínima expresión en la estructura, donde vea que los grupos tienen sentido y son distintos entre sí).

Tiene que existir una jerarquía entre los datos. Si existe orden jerárquico es un buen método, de lo contrario no es aconsejable.

Por ejemplo, imagina agrupar artículos en Amazon. Deberán existir algunas categorías amplias, pero a medida que ingrese a categorías de compras más específicas, se necesitará niveles cada vez mayores de granularidad.

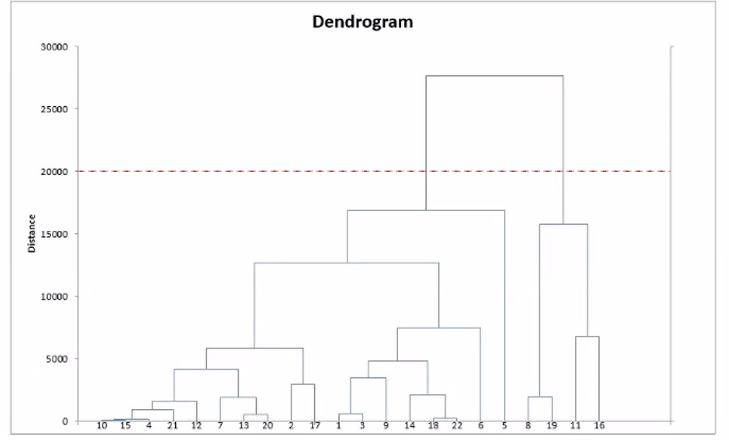
Clústeres jerárquicos es una alternativa a los métodos de partitioning clústering que no requiere que se pre-especifique el número de clúster. Los métodos que engloba el clúster jerárquico se subdivide en dos tipos dependiendo de la estrategia seguida para crear los grupos:

* Agglomerative clústering (bottom-up): el agrupamiento se inicia en la base del árbol, donde cada observación forma un clúster individual. Los clúster se van combinado a medida que la estructura crece hasta converger en una única “rama” central.
* Divisive clústering (top-down): es la estrategia opuesta al agglomerative clústering, se inicia con todas las observaciones contenidas en un mismo clúster y se suceden divisiones hasta que cada observación forma un clúster individual.

En ambos casos, los resultados pueden representarse de forma muy intuitiva en una estructura de árbol llamada dendrograma.

### Dendrograma

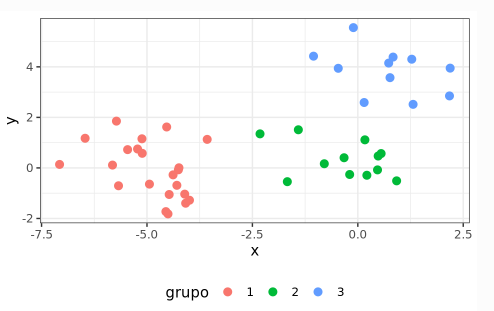
Un dendrograma se utiliza comúnmente para estudiar los Clústering jerárquicos antes de decidir el número de clústeres que resulta más apropiado para el conjunto de datos. La distancia a la que se combinan dos grupos se denomina distancia dendrograma. La distancia del dendrograma es una medida de si dos o más grupos están separados o pueden combinarse para formar un grupo.

.

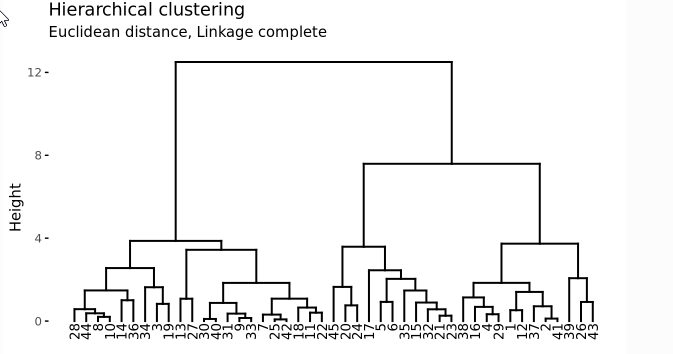
* Luego defino una línea de corte para armar los clúster.
* El gran desafío es qué nombre le pongo a cada clúster, el título se construye en base a las variables que hayamos medido.

Para ilustrar cómo se interpreta un dendrograma, se simula un set de datos y se somete a un proceso de clúster jerárquico.

Supóngase que se dispone de 45 observaciones en un espacio de dos dimensiones, que pertenecen a 3 grupos. Aunque se ha coloreado de forma distinta cada uno de los grupos para facilitar la comprensión de la idea, se va a suponer que se desconoce esta información y que se desea aplicar el método de clúster jerárquico, para intentar reconocer los grupos.



Al aplicar clúster jerárquico, empleando por ejemplo: la distancia euclídea y linkage complete (estos conceptos se ven más abajo), se obtiene el siguiente dendrograma. Como los datos se han simulado en aproximadamente la misma escala, no es necesario estandarizarlos, de no ser así, sí se tendrían que estandarizar.



En la base del dendrograma, cada observación forma una terminación individual conocida como hoja o leaf del árbol. A medida que se asciende por la estructura, pares de hojas se fusionan formando las primeras ramas. Estas uniones (nodos) se corresponden con los pares de observaciones más similares. También ocurre que ramas se fusionan con otras ramas o con hojas. Cuanto más temprana (más próxima a la base del dendrograma) ocurre una fusión, mayor es la similitud. Esto significa que, para cualquier par de observaciones, se puede identificar el punto del árbol en el que las ramas que contienen dichas observaciones se fusionan. La altura a la que esto ocurre (eje vertical) indica cómo de similares/diferentes son las dos observaciones. Los dendrogramas, por lo tanto, se deben interpretar únicamente en base al eje vertical y no por las posiciones que ocupan las observaciones en el eje horizontal, esto último es simplemente por estética y puede variar de un programa a otro. Por ejemplo, la observación 10 es la más similar a la 8 ya que es la primera fusión que recibe la observación 8 (y viceversa). Podría resultar tentador decir que la observación 26, situada inmediatamente a la izquierda de la 8, es la siguiente más similar, sin embargo, las observaciones 28 y 37 son más similares a la 8 a pesar de que se encuentran más alejadas en el eje horizontal. Del mismo modo, no es correcto decir que la observación 28 es más similar a la observación 8 de lo que lo es la 37 por el hecho de que está más próxima en el eje horizontal. Prestando atención a la altura en que las respectivas ramas se unen, la única conclusión válida es que la similitud entre los pares 8-28 y 8-37 es la misma. No me queda claro el tema de las similitudes. Rta **Mario**: esto es así, ya que si estamos clasificando un grupo de cuadrúpedos de África por ejemplo y quedan clasificados como en el dendrograma anterior, no podemos asegurar que el 8 es más similar al 44 que al 37.

**Verificar el árbol resultante**

Una vez creado el dendrograma, hay que evaluar hasta qué punto su estructura refleja las distancias originales entre observaciones. Una forma de hacerlo es empleando el coeficiente de correlación entre las distancias cophenetic del dendrograma (altura de los nodos) y la matriz de distancias original (¿Esto sería como la matriz de correlación como la que hicimos para el ejemplo de las casas?) Rta. **Mario**: Si. Cuanto más cercano es el valor a 1, mejor refleja el dendrograma la verdadera similitud entre las observaciones. Valores superiores a 0.75 suelen considerarse como buenos. Esta medida puede emplearse como criterio de ayuda para escoger entre los distintos métodos de linkage. En R, la función cophenetic() calcula las distancias cophenetic de un clúster jerárquico.

### Clústering jerárquico divisivo

La agrupación jerárquica divisiva funciona de manera opuesta. En lugar de comenzar con n grupos (en el caso de n observaciones), comenzamos con un solo grupo y asignamos todos los puntos a ese grupo. En cada iteración, dividimos el punto más lejano del clúster y repetimos este proceso hasta que cada clúster solo contenga un único punto



#### Proceso de clúster jerárquico divisivo: algoritmo DIANA

El algoritmo más conocido de clúster jerárquico divisivo es DIANA (DIvisive ANAlysis Clústering). Este algoritmo se inicia con un único clúster que contiene todas las observaciones, a continuación, se van sucediendo divisiones hasta que cada observación forma un clúster independiente. En cada iteración, se selecciona el clúster con mayor diámetro, entendiendo por diámetro de un clúster la mayor de las diferencias entre dos de sus observaciones. Una vez seleccionado el clúster, se identifica la observación más dispar, que es aquella con mayor distancia promedio respecto al resto de observaciones que forman el clúster, esta observación inicia el nuevo clúster. Se reasignan las observaciones en función de si están más próximas al nuevo clúster o al resto de la partición, dividiendo así el clúster seleccionado en dos nuevos clústeres.

A diferencia del clúster aglomerativo, en el que hay que elegir un tipo de distancia y un método de linkage, en el clúster divisivo solo hay que elegir la distancia, no hay linkage.

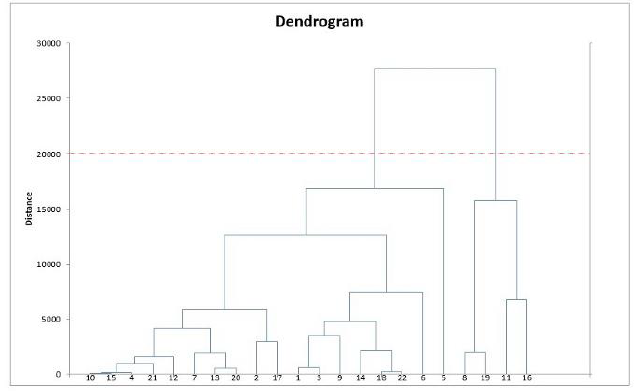
### Clústering jerárquico aglomerativo

Asignamos cada punto a un grupo individual en esta técnica. Luego, en cada iteración, fusionamos el par de clústeres más cercanos y repetimos este paso hasta que solo quede un solo clúster



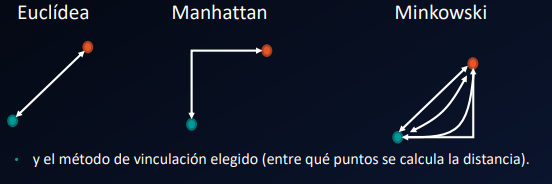
#### Proceso de Clústering jerárquico aglomerativo

1. Comience con N grupos, uno para cada punto de datos
2. Combine los dos grupos más cercanos entre sí .Ahora tienes grupos N-1
3. Vuelva a calcular las distancias entre los grupos. Hay varias formas de hacer esto, por ejemplo considerar la distancia entre dos grupos como la distancia promedio entre todos sus miembros respectivos.
4. Repita los pasos 2 y 3 hasta que obtenga un grupo de N puntos de datos.
5. Elija varios grupos y dibuje una línea horizontal en el dendrograma



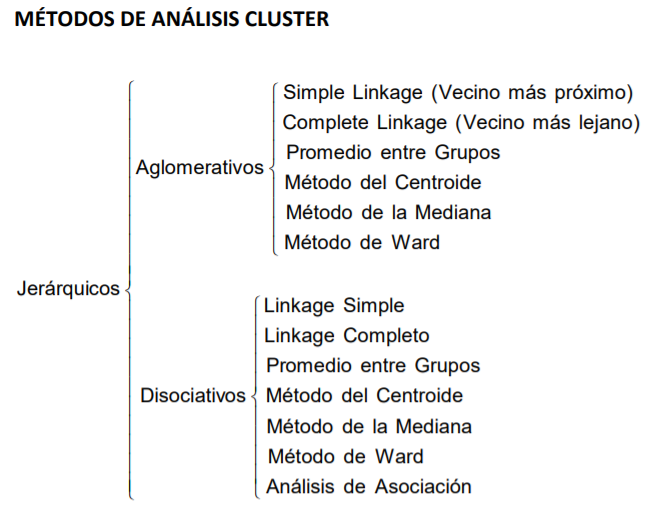
El resultado de este algoritmo se ve afectado por dos cosas principales:

* La métrica de afinidad elegida (cómo se calcula la distancia entre puntos)



* El método de vinculación elegido (entre qué puntos se calcula la distancia)

### ¿Cómo se define una medida de disimilitud entre clústeres?



[https://www.estadistica.net/Master-Econometria/Analisis\_Clúster.pdf](https://www.estadistica.net/Master-Econometria/Analisis_Cluster.pdf)

Los tipos más comunes de vinculación son:

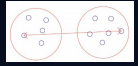
#### Enlace único (Single)

Es la distancia más corta entre un par de observaciones en dos grupos distintos. A veces puede producir agrupaciones en las que las observaciones de diferentes agrupaciones están más juntas que las observaciones dentro de sus propias agrupaciones. Estos grupos pueden parecer dispersos.



#### Enlace completo (complete)

Se mide la distancia entre el par de observaciones más lejano en dos grupos. Este método suele producir agrupaciones más estrechas que el enlace simple, pero estas agrupaciones estrechas pueden terminar muy juntas.



#### Vinculación promedio (average)

La distancia entre cada par de observaciones en cada grupo se suma y se divide por el número de pares para obtener una distancia promedio entre grupos.



#### Enlace Pabellón (Ward):

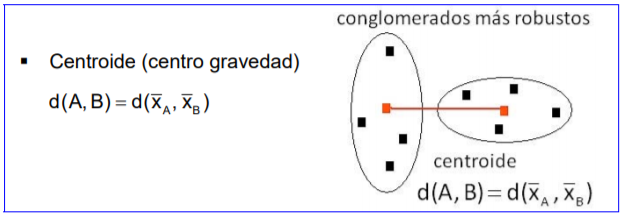
Utiliza el método de análisis de varianza para determinar la distancia entre los clústeres.

Método de Ward (método de mínima varianza): Cuando se unen dos conglomerados, con independencia del método utilizado, la varianza aumenta. El método de Ward une los casos buscando minimizar la varianza dentro de cada grupo. Para ello se calcula, en primer lugar, la media de todas las variables en cada conglomerado. A continuación, se calcula la distancia entre cada caso y la media del conglomerado, sumando después las distancias entre todos los casos. Posteriormente se agrupan los conglomerados que generan menos aumentos en la suma de las distancias dentro de cada conglomerado. Este procedimiento crea grupos homogéneos y con tamaños similares



#### Enlace centroide (centroid)

La distancia entre los centroides de dos grupos. A medida que los centroides se mueven con nuevas observaciones, es posible que los grupos más pequeños sean más similares al nuevo grupo más grande que a sus grupos individuales.



### Características de Clústering Jerárquico

* Los beneficios de la agrupación jerárquica, en comparación con otros métodos de agrupación, es que no es necesario especificar el número de agrupaciones.
* Además, el algoritmo **no es tan sensible a la métrica de distancia**, y por lo tanto los resultados no deberían verse afectados por la elección de la métrica de afinidad
* Este algoritmo es especialmente útil cuando los datos subyacentes tienen una estructura jerárquica o eso es lo que esperamos de los datos, a partir de los cuales podemos ver cómo los clústeres finalmente se agrupan para formar cada vez menos clústeres (o al revés)
* Sin embargo, esto tiene el costo de una menor eficiencia debido a una mayor complejidad temporal. Asimismo, no existe una métrica de rendimiento para la agrupación jerárquica que indique qué número de K se ajusta mejor a los datos. ¿A qué se refiere con complejidad temporal?. Rta **Mario**: Al tiempo de procesamiento
* Una vez que se define un clúster, durante la iteración no se romperá y solo continuará formando grupos más grandes
* Por lo tanto, en general, aunque se utiliza menos que otros algoritmos por razones de complejidad y optimización, este algoritmo es útil cuando se espera que los datos posean una estructura jerárquica intrínseca.
* Los no supervisados no son fáciles de saber si están bien o no, porque no hay métricas que digan si está funcionando bien o no.

### Noteboook

Notebook: 7ta Clase - Delitos\_Londres (Clústering delitos en Londres)

* Estos grupos de delitos, identificados por agrupaciones jerárquicas, no son la cantidad de delitos sino el tipo de delito en Londres. Por ejemplo, ¿hay áreas en Londres donde, independientemente de la cantidad de delitos, hay un alto porcentaje de una cierta cantidad de delitos como Robo o Entraderas?
* Si se identifican estas áreas, entonces, en teoría, las intervenciones específicas pueden estar dirigidas a atacar tipos específicos de delitos en lugar de distribuir estos recursos por todo Londres.

**Dudas:**

En la parte del código donde hace el primer mapa:

Rta Mario: esto es algo típico de Python que hay que hacer para que funcione.

Me gustaría entenderlo un poco más.

# planchamos los subplot para poder iterar sobre ellos

'''ROMI: Que es planchar los ejes?

#Creamos un subeje en la figura

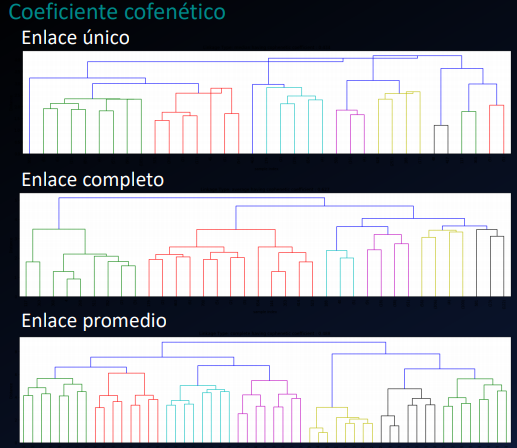
'''Romi: a que se refiere?

En conclusión: ¿Cuál sería la mejor división de clúster?

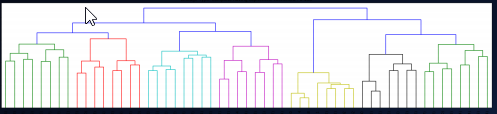
Rta **Mario**: A este estudio habría que acompañarlo de otras métricas que podrían resaltar y ayudar en el análisis: Dinero que cuesta, percepción de la gente, tipos de policías, nuevo escuadrón, etc..

### Coeficiente cofenético

No es una métrica en sí, porque no estas midiendo en las mismas condiciones. Es una especie de hiperparametrización.



* La elección de cómo medir la distancia entre clúster impacta directamente en cómo se forman los clústeres
* Imagine dos grupos, A y B con puntos A con puntos A1 , A₂ y A₃ en el grupo A y puntos B₁, B₂ y B₃ en el grupo B
* Para que estos dos grupos A y B estén bien separados, los puntos A₁, A₂, A₃ y puntos B₁, B₂ y B₃ también deben estar alejados entre sí.
* El coeficiente Cofenético es una medida de la correlación entre la distancia de puntos en el dendrograma
* Se calculan la distancia euclidiana entre los puntos. Luego calcula la distancia del dendrograma en la que se combinan los grupos A y B. Si la distancia entre los puntos aumenta con la distancia del dendrograma entre los grupos A y B, entonces el coeficiente cofenético está más cerca de 1
* El coeficiente cofenético interesa que sea lo mas elevado posible, siendo siempre menor o igual que 1.



## Partition Clústering

### El método PAM (Partition around Medoids)

* PAM es un método de clústering por partición, como k-medias
* La diferencia con este último algoritmo es que PAM busca “**medoides**” o prototipos, esto es, objetos representativos del clúster que tienen una distancia mínima al resto de los miembros de su clúster
* En el método de PAM el objetivo es la minimización de la suma de disimilitudes entre los miembros del clúster y su medoide. Esta diferencia en cuanto a la función a optimizar hace que PAM tienda a ser más robusto que kmedias en conjuntos de datos con outlier
* K-medoids es un método de clústering muy similar a K-means en cuanto a que ambos agrupan las observaciones en K clústers, donde K es un valor preestablecido por el analista. La diferencia es que, en K-medoids, cada clúster está representado por una observación presente en el clúster (medoid), mientras que en K-means cada clúster está representado por su centroide, que se corresponde con el promedio de todas las observaciones del clúster pero con ninguna en particular
* Una definición más exacta del término medoid es: elemento dentro de un clúster cuya distancia (diferencia) promedio entre él y todos los demás elementos del mismo clúster es lo menor posible. Se corresponde con el elemento más central del clúster y por lo tanto puede considerarse como el más representativo. El hecho de utilizar medoids en lugar de centroides hace de K-medoids un método más robusto que K-means, viéndose menos afectado por outliers o ruido. A modo de idea intuitiva puede considerarse como la analogía entre media y mediana.
* El algoritmo más empleado para aplicar K-medoids se conoce como PAM (Partitioning Around Medoids) y sigue los siguientes pasos:

#### Algoritmo

**El algoritmo tiene dos etapas:**

Etapa de construcción (Build phase)

1. Elegir k objetos al azar (o utilizar los seleccionados por el usuario) y utilizarlos como medoides
2. Calcular la matriz de disimilitud –en principio, con cualquier medida-, o utilizar la matriz que se pasó como argumento para calcular una matriz de distancia euclídea. ¿No entiendo esto? Rta **Mario**: es la matriz de correlación.
3. Asignar cada objeto (observación) a su medoide más cercano. Para cada clúster calcular la sumatoria de las disimilitudes entre miembros y medoides; calcular la sumatoria para todos los clústeres

Etapa de intercambio (Swap phase)

1. Para cada clúster intercambiar el medoide con alguno de sus otros miembros y determinar si este cambio baja la suma de las disimilitudes del grupo. Si esto ocurre, remplazar el mediode

Si se modificó el mediode de al menos un grupo volver a 3; si no, terminar

A diferencia del algoritmo K-means, en el que se minimiza la suma total de cuadrados intra-clúster (suma de las distancias al cuadrado de cada observación respecto a su centroide), el algoritmo PAM minimiza la suma de las diferencias de cada observación respecto a su medoide.

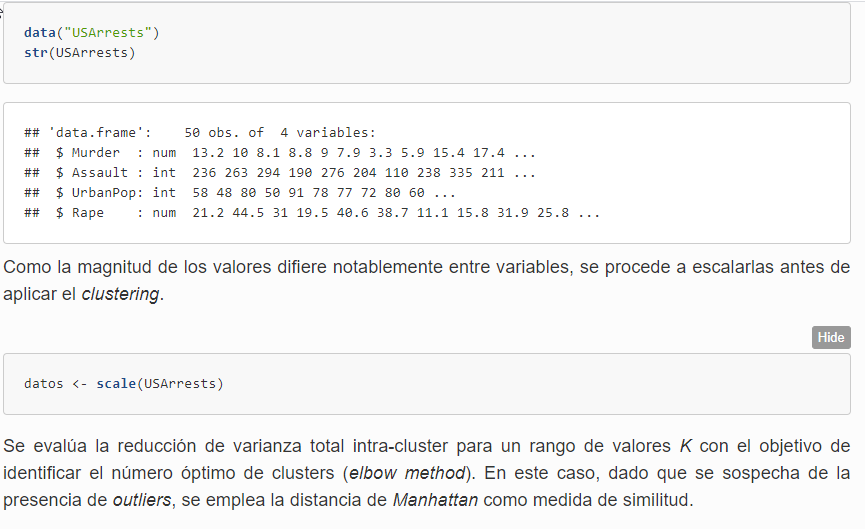
Por lo general, el método de K-medoide se utiliza cuando se conoce o se sospecha de la presencia de outliers. Si esto ocurre, es recomendable utilizar como medida de similitud la distancia de Manhattan, ya que es menos sensible a outliers que la euclídea.

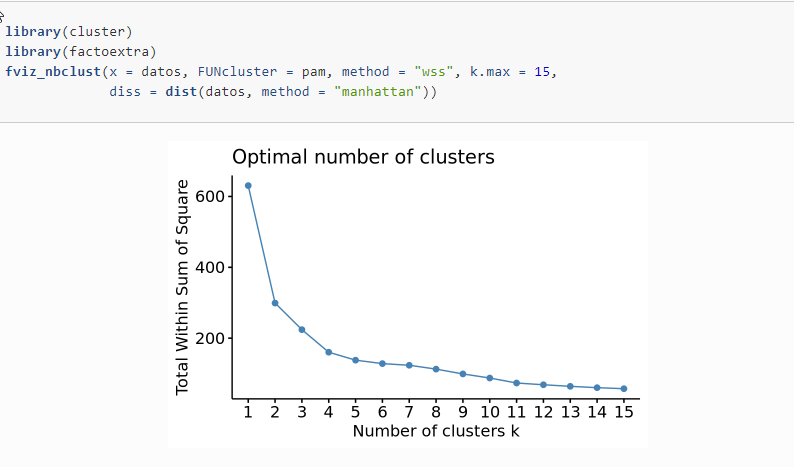
* K-medoide es un método de clustering más robusto que K-means, por lo es más adecuado cuando el set de datos contiene outliers o ruido.
* Al igual que K-means, necesita que se especifique de antemano el número de clústeres que se van a crear. Esto puede ser complicado de determinar si no se dispone de información adicional sobre los datos. Muchas de las estrategias empleadas en K-means para identificar el número óptimo, pueden aplicarse en K-medoide.
* Para sets de datos grandes necesita muchos recursos computacionales. En tal situación se recomienda aplicar el método CLARA. ¿Cuál es el método Clara?

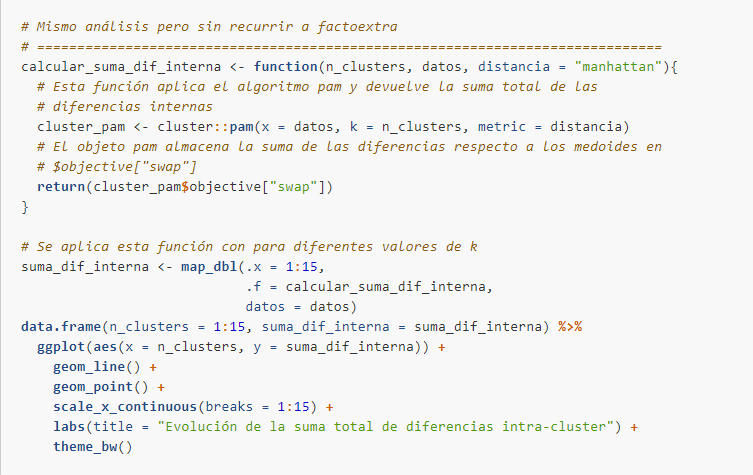
##### Ejemplo

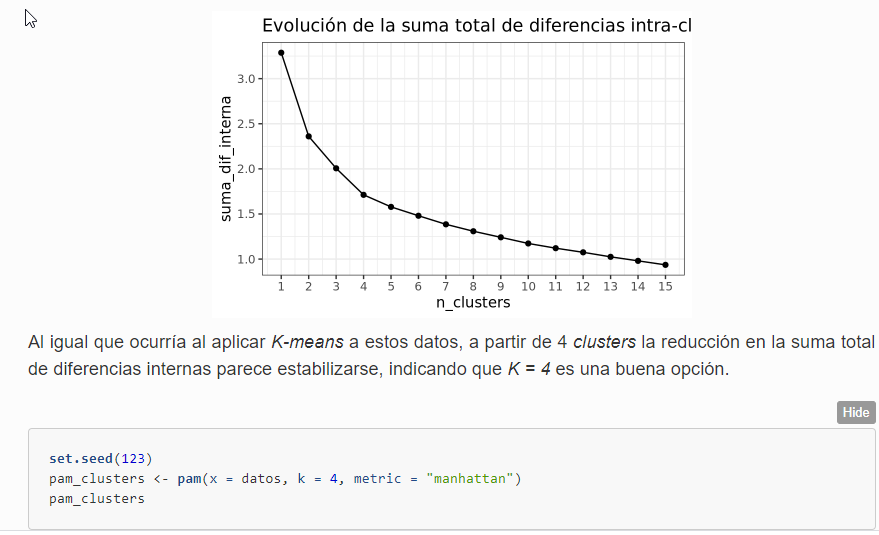
El set de datos USArrests contiene información sobre el número de delitos (asaltos, asesinatos y secuestros) junto con el porcentaje de población urbana para cada uno de los 50 estados de USA. Se pretende estudiar si existe una agrupación subyacente de los estados mediante clustering. Dado que se sospecha de la presencia de outliers se recurre a K-medoide.

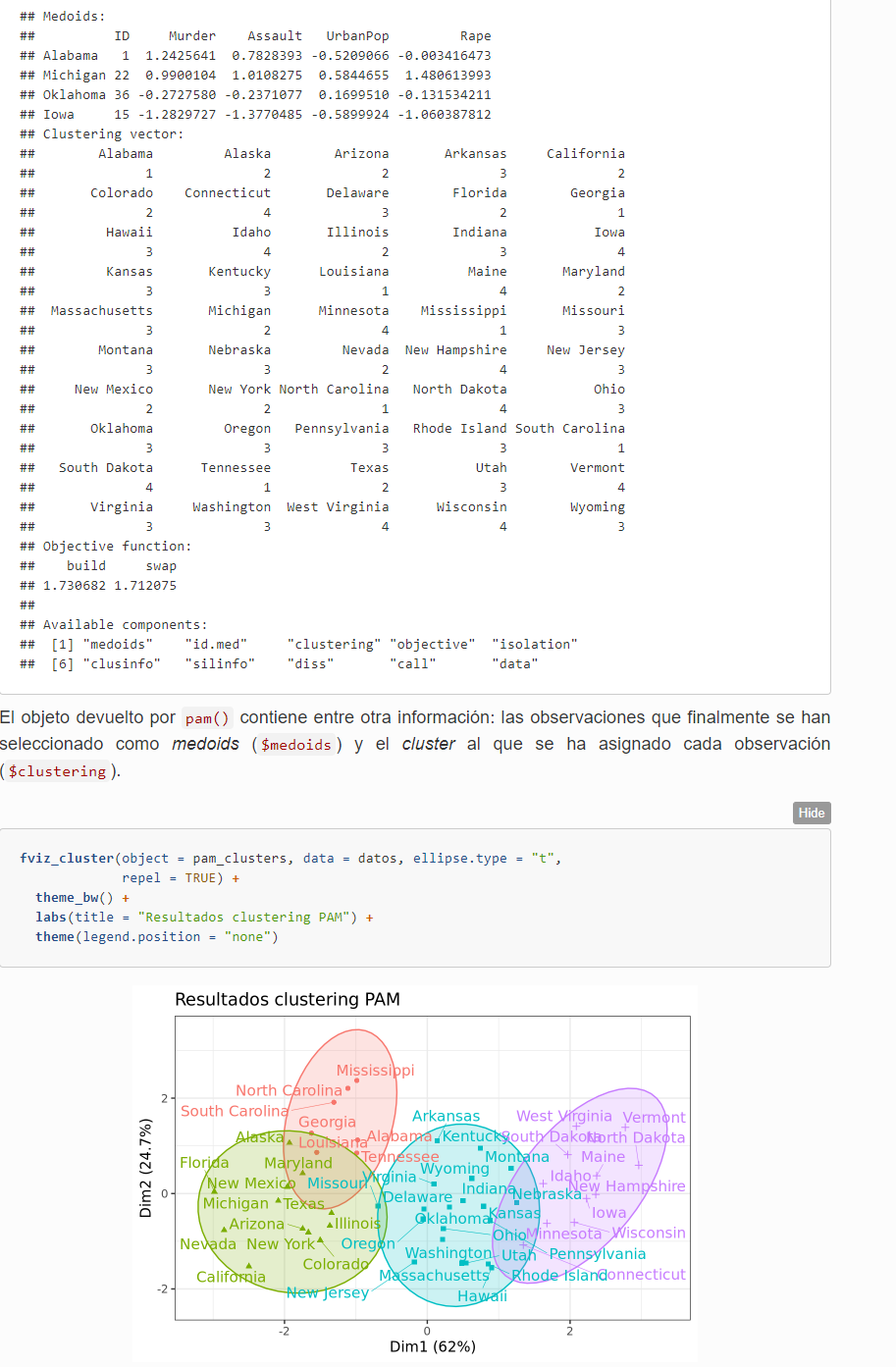
El proceso a seguir en R para aplicar el método de K-medoide es igual al seguido en K-means (ejemplo 2), pero en este caso empleando la función pam() del paquete clúster.

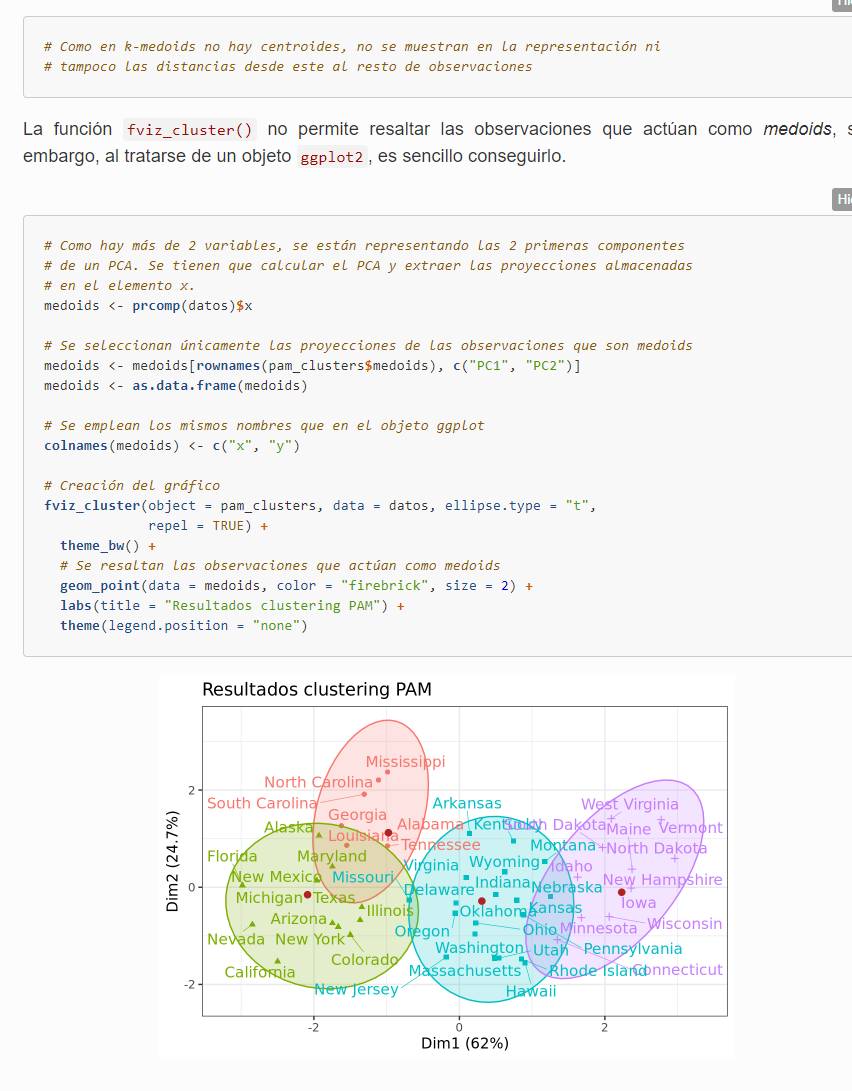
**

**

**

**

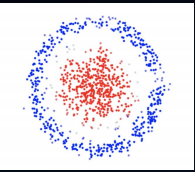
**

**

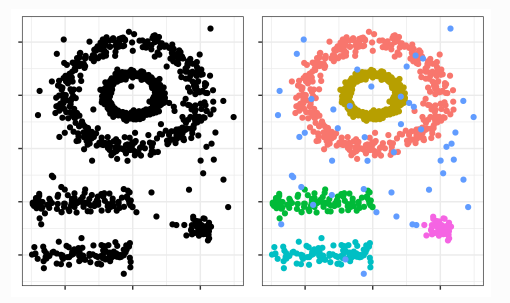
## Clústering basado en densidad

### El método DBSCAN

Hay situaciones, donde existen grupos bien definidos, pero los métodos de clustering como K-medias o PAM, que buscan clúster esféricos, fallan.



* Para solucionar esto se diseñaron los métodos de clustering basados en densidad. Estos métodos consideran que un clúster es una región de alta densidad de puntos rodeados por regiones de baja densidad. La variante más conocida de estos métodos es DBSCAN, que además tiene la particularidad de ser un método de clustering no completo. Esto significa que aquellos objetos que caigan en regiones de baja densidad se consideran ruido y no se asignan a ningún grupo, se descartan
* Es importante tener en cuenta que DBSCAN puede tener problemas para encontrar clústeres cuando las densidades de cada grupo son muy diferentes entre sí
* Density-based spatial clústering of applications with noise (DBSCAN) una forma de identificar clústeres siguiendo el modo intuitivo en el que lo hace el cerebro humano, identificando regiones con alta densidad de observaciones separadas por regiones de baja densidad
* Véase la siguiente representación bidimensional de los datos multishape del paquete factoextra



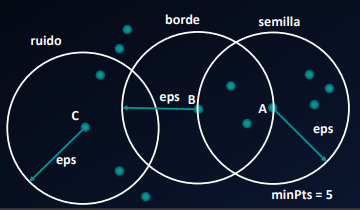
* El cerebro humano identifica fácilmente 5 agrupaciones y algunas observaciones aisladas (ruido). Véanse ahora los clúster que se obtienen si se aplica, por ejemplo, K-means clustering.

[*https://www.cienciadedatos.net/documentos/37\_clústering\_y\_heatmaps#Density\_based\_clústering\_(DBSCAN)*](https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps#Density_based_clustering_(DBSCAN))

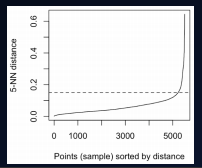
#### El algoritmo DBSCAN

1. Determinar el número de objetos que tiene un objeto dado dentro de un radio eps, incluyendo dicho objeto. Si la cantidad de objetos dentro de ese radio excede un número umbral, llamado minPts, ese punto se llama “semilla”
2. Un objeto que no es semilla, porque su cantidad de vecinos no supera minPts, pero cae dentro del área de otro objeto semilla, se denomina “borde”
3. Un objeto que no es ni borde, ni semilla, es un objeto “ruido”
4. Eliminar los objetos ruido
5. Establecer un borde o límite entre todos los objetos semilla que están dentro de una distancia eps entre sí
6. Crear un clúster a partir de cada grupo de objetos “semillas” conectados del punto anterior
7. Asignar cada punto borde a uno de los clúster de sus objetos semilla asociados

Ver video animado: [https://www.digitalvidya.com/blog/the-top-5-clústering-algorithms-data-scientists-should-know/](https://www.digitalvidya.com/blog/the-top-5-clustering-algorithms-data-scientists-should-know/)



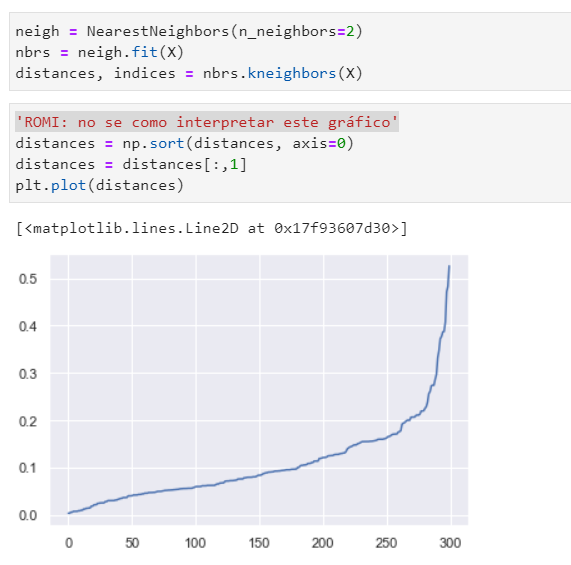
* Un punto crítico del algoritmo es la selección de Eps y minPts
* No existe una forma automática de determinar el valor MinPts para DBSCAN. En última instancia, el valor de MinPts debe establecerse utilizando el conocimiento del dominio y la familiaridad con el conjunto de datos
* Algunos criterios que pueden usarse:
  + Cuanto mayor sea el conjunto de datos, mayor debe ser el valor de MinPts
  + Si el conjunto de datos es más ruidoso, elija un valor mayor de MinPts
  + Generalmente, MinPts debe ser mayor o igual que la dimensionalidad del conjunto de datos
  + Para datos bidimensionales, use el valor predeterminado de DBSCAN de MinPts = 4 (Ester et al., 1996).Si sus datos tienen más de 2 dimensiones, elija MinPts = 2 \* dim, donde dim = las dimensiones de su conjunto de datos (Sander et al., 1998).
* Una forma empírica de hacer esto es determinar la distancia k-dist al k vecino más cercano para todos los objetos y graficar esas distancias rankeadas
* Esta técnica calcula la distancia promedio entre cada punto y sus k vecinos más cercanos, donde k = el valor MinPts que seleccionó. Las k-distancias medias se trazan luego en orden ascendente en un gráfico de k-distancias. El valor óptimo para ε en el punto de máxima curvatura (es decir, donde el gráfico tiene la mayor pendiente). No entiendo cómo interpretar este gráfico. Rta **Mario**: Abajo son los puntos ordenados por la distancia al 5to punto más cercano. En base a esto defino el epsilon
* No define zonas, sino que define cuán cerca están los puntos unos con otros.



#### Notebook

Notebook: 7ta Clase - DDBSCAN\_1: Es un ejemplo del algoritmo: DDBSCAN con datos generados.

Dudas:



7ta Clase - DDBSCAN\_2: Es un ejemplo del algoritmo: DDBSCAN con datos de ingresos y score para cliente de un banco.

# 8va Clase - 29/07/2021

## Métricas Clustering

* Hasta ahora exploramos los principales tipos de métricas que se utilizan para evaluar los modelos de aprendizaje automático:
  + Clasificación
  + Regresión
* Medir correctamente la performance de los algoritmos de clustering es fundamental.
* Esto es particularmente relevante debido a que a menudo los clústeres se inspeccionan manual y cualitativamente para determinar si los resultados son significativos.
* Tres factores importantes mediante los cuales se puede evaluar la agrupación son:
  + (a) Tendencia a la agrupación
  + (b) Número de clústeres, k
  + (c) Calidad de la agrupación

### Tendencia a la agrupación test de Hopkins

* Ver detalle de cómo se hace
* Antes de evaluar la performance o directamente antes de aplicar un algoritmo de clustering, es muy importante asegurarse de que el conjunto de datos en el que estamos trabajando tenga tendencia a agruparse y no contenga puntos distribuidos uniformemente.
* Si los datos no contienen tendencia al agrupamiento, entonces los agrupamientos identificados por cualquier algoritmo de agrupamiento pueden ser irrelevantes (especialmente en aquellos casos donde si o si define clústeres).
* Se puede utilizar el test de Hopkins, una prueba estadística de aleatoriedad espacial de una variable, para medir la probabilidad de puntos de datos generados por la distribución uniforme de datos.
* Hipótesis nula (H0): los puntos de datos se generaron mediante una distribución uniforme (no existen agrupaciones significativas).
* Hipótesis alternativa (H1): los puntos de datos se generan mediante puntos de datos aleatorios (presencia de clústeres).
* Si H ~ 0.5, la hipótesis nula puede rechazarse y es muy probable que los datos contengan clústeres. Si H está más cerca de 0, entonces el conjunto de datos tiene tendencia a agruparse .
* La prueba de Hopkins está muy influenciada por valores atípicos y por lo tanto es importante normalizarlos.

### Número de clústeres k

* Obtener el número óptimo de clústeres es muy importante en el análisis. Si k es demasiado alto, cada punto comenzará a representar en líneas generales un grupo y si k es demasiado bajo, los puntos de datos se agrupan incorrectamente.
* No hay una respuesta definitiva para encontrar el número correcto de clústeres, porque depende de:
  + (a) la forma de distribución
  + (b) la escala en el conjunto de datos
  + (c) la resolución de clústeres requerida por el usuario.
* Hay dos enfoques principales para encontrar el número óptimo de clústeres:
  + (1) Conocimiento del dominio
  + (2) Enfoque basado en datos
* Conocimiento del dominio: el conocimiento del dominio puede proporcionar algún conocimiento previo sobre cómo encontrar el número de agrupaciones.
* Enfoque basado en datos: si el conocimiento del dominio no está disponible, los métodos matemáticos ayudan a encontrar el número correcto de grupos.

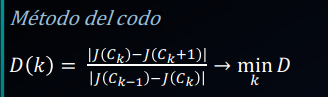
#### Método empírico

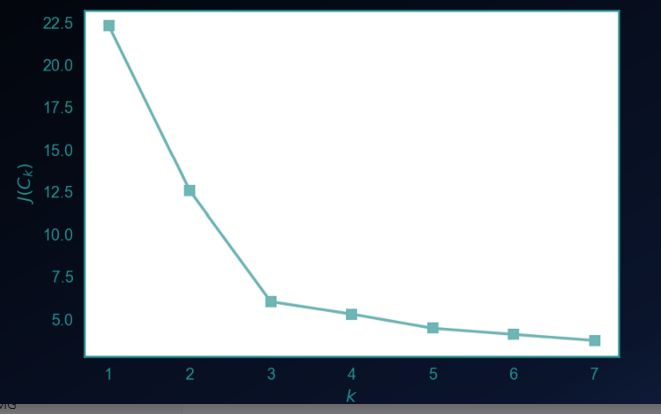
Un método empírico simple para encontrar el número de grupos es , donde N es el número total de datos, de este modo cada grupo contiene puntos.

#### 

#### Método del codo

Buscar más información





### Calidad de la agrupación

* Una vez que se realiza la agrupación en clústeres, se puede cuantificar el rendimiento de la agrupación mediante una serie de métricas. El agrupamiento ideal se caracteriza por una distancia mínima entre los elementos del grupo y una distancia máxima entre grupos.

#### Tipos de Medidas

* Existen principalmente dos tipos de medidas para evaluar el rendimiento de la agrupación.

(i) **Medidas extrínsecas** que requieren etiquetas. Algunos ejemplos son:

* + - Adjusted Rand index
    - Fowlkes-Mallows scores
    - Mutual information based scores
    - Homogeneity
    - Completeness and V-measure

Aquí hay una cierta variable predictora. Este tipo de medidas sirven para verificar lo que ya tenemos, quizás para encontrar algún otro tipo de comportamiento en nuestra variable predictora . Es como hacer una re ingeniería.

Nos va a mostrar un Notebbok de jugadores de fútbol. *Cómo se clasifican los jugadores de la FIFA. Si sos arquero o no. Hay que mirar la clase para saber si está clasificando bien*

(ii) **Medidas intrínsecas** que no requieren etiquetas de verdad fundamental. Algunas de las medidas de rendimiento de la agrupación son:

* + - Silhouette Score (es la más típica)
    - índice Calinski-Harabasz
    - índice Davies-Bouldin
    - etc.

##### Medidas extrínsecas

###### Índice Rand (Rand Index)

Calcula una medida de similitud entre dos clústeres considerando todos los pares de muestras y contando los pares que se asignan en el mismo o en diferentes clústeres en los clústeres predichos y verdaderos. No entiendo la frase



Donde el *RI* puede variar de 0 a 1

###### Índice Rand ajustado (Adjusted Rand Index)

Es idéntico pero se "ajusta” por posibles efectos del azar.

###### Mutual Information

La información mutua es otra métrica que se utiliza a menudo para evaluar el rendimiento de los algoritmos de clustering.

**

Es una medida de la similitud entre dos etiquetas de los mismos datos. Dónde | Ui | es el número de muestras en el conglomerado Ui; y | Vj | es el número de muestras en el conglomerado Vj.

*Solo saber cuales son las medidas extrínsecas que existen. Te dicen si lo clasificaste bien en función a un criterio externo*

##### Medidas intrínsecas

En general aquí se ve las medidas dentro del clustering y entre los clustering

###### Índice de Calinski-Harabasz

* Se conoce como el criterio de la relación de varianza.
* La puntuación se define como la relación entre la dispersión dentro de los clústeres y la dispersión entre los clústeres.



* Donde tr (Bk) es la traza de la matriz de dispersión entre grupos y tr (Wk) es la traza de la matriz de dispersión dentro del clúster.

###### Índice de Davies-Bouldin

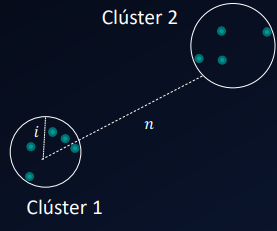
El índice de Davies-Bouldin se define como la medida de similitud promedio de cada clúster con su clúster más similar. La similitud es la relación entre las distancias dentro de los clústeres y las distancias entre los clústeres. De esta manera, los clústeres que estén más separados y menos dispersos conducirán a una mejor puntuación.

###### 

###### Silhouette Score

* El **Silhouette Score** y el **gráfico de Silhouette** se utilizan para medir la distancia de separación entre grupos. Muestra una medida de la cercanía de cada punto de un cluster a los puntos de los clusters vecinos.
* Esta medida tiene un **rango de −1,1** y es una gran herramienta para inspeccionar visualmente las similitudes dentro de los clústeres y las diferencias entre ellos.
* El **Silhouette Score** se calcula utilizando la **distancia media dentro del cluster (*i*)** y la distancia media al cluster más cercano *(n)* para cada muestra.
* El El **Silhouette Score** para una muestra es 
* *n* es la distancia entre cada muestra y el cluster más cercano del que la muestra no forma parte, mientras que *i* es la distancia media dentro de cada grupo
* ¿Cuál sería un buen valor de esta métrica?

*Aplicas tu cluster y te calcula para cada cluster una distancia media dentro de ese cluster y la distancia más próxima al cluster más cercano.*

* Si los cluster están solapados, n e i pueden ser casi iguales.*

###### 

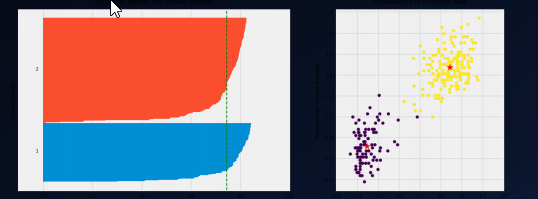
###### Gráficos de Silhouette

* Los gráficos de Silhouette representan la etiqueta del cluster en el eje y, mientras que la Silhouette Score en el eje x.
* El tamaño de las siluetas es proporcional al número de muestras dentro del grupo que se analiza.
* Cuanto más altos sean los Silhouette Score (más cercanos a +1), más alejadas estarán las muestras del clúster de las muestras de los clústeres vecinos.
* Un valor de 0 indica que la muestra está en o muy cerca del límite de decisión entre dos clústeres vecinos.
* Los valores negativos, en cambio, indican que esas muestras podrían haber sido asignadas al grupo incorrecto.
* Al promediar los Silhouette Score, obtenemos un Silhouette Score global que se puede usar para describir la performance global.

*Se hace un gráfico de Silhouette como se muestra abajo. lo que se tiende a calcular el valor promedio -...*

*Cuando hay valores negativos indica que está clasificando para un cluster, pero es muy probable que si da negativo deba pertenecer a otro cluster (tiene más chance de pertenecer a otro cluster)*

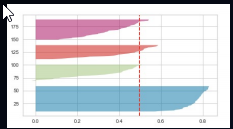
*Si son mayores que cero es probable que sí corresponden al cluster*

**

Cada cada punto amarillo, se represetna en el área roja, se represetna como una línea roja

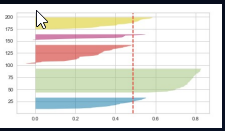
*Te dice cuan bien te clasificó*

* Si todos los **Gráficos de Silhouette** (de cada clúster) superan el valor promedio de Silhouette Score

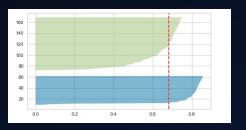


* Si no hay grandes fluctuaciones en el tamaño de los Clústeres. *Aquí hay algunos negativos, significa que hay poquitos puntos que están mal clasificados*

*Esto depende del dominio. El punto rojo indica que de ese cluster el 95% de los puntos, la distancia media dentro del cluster ……..*



* Si el ancho del Gráfico de Silhouette de los clústeres es uniforme. *En base a esto puedo cambiar los hiper parámetros o trabajar un poco más las variables. Esta no es una métrica como la AUC. Yo decido si representa o no algo.*



*Si no tenés mucha idea aplica con distintos k . Hay cluster que a veces dan negativo. Tenés que ir ajustando y quedarte con el peor de los males.*

*Mario sugiere primero agarrar algún sistema de agrupación y luego aplicar* Silhouette. *Esto no es muy distinto a cuando aplicaba la regla del codo.*

###### Notebooks

*8va Clase - Hopkins\_Silhouette.*

*8va Clase - Distinas\_metricas*

*8va Clase - Segmentación\_Clientes*

# 9na Clase - 05/08/2021

## Reducción de la dimensionalidad

* La reducción de dimensionalidad, se refiere a los métodos utilizados para representar datos utilizando menos variables. La idea es representar nuestros datos utilizando las características “latentes” que interrelacionan nuestras variables iniciales. Esta estructura latente que se presenta en forma rala, a menudo se representa con muchas menos variables de las que teníamos al principio, y por lo tanto puede hacer que el procesamiento de datos sea mucho menos intensivo y puede eliminar las funciones redundantes

*Reducción de dimensionalidad: Es proyectar y mirar a los objetos que queremos mirar con menos dimensiones*

*Cada punto en un punto en la dimensionalidad.*

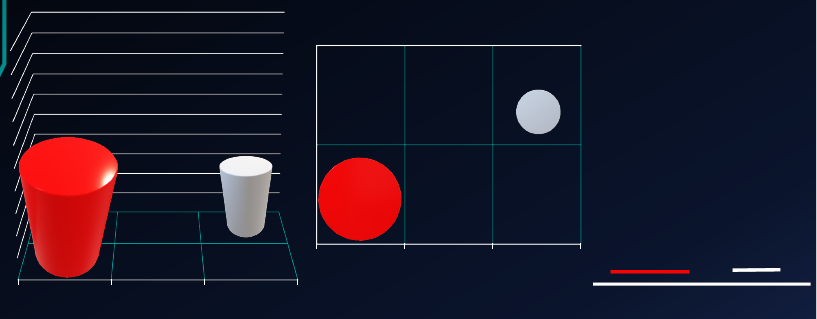
* *Una posibilidad es quedarse con las variables que más identifican al problema.los Árboles me dicen cuales tienen más peso predictivo.*
* *Otra forma es combinando variables: existe un método que me permite determinar como puedo combinar variables o cual es la mejor combinación. Cuando las variables están correlacionadas es un indicador de combinación de variables, generando un nuevo componente. Si la correlación es muy alta (correlación 1) puedo eliminarlas.*

*No tiene sentido aplicar una reducción de dimensionalidad sino no hay correlación de variables*

*Si aplica sobre un nuevo set de datos se debe mirar si las variables que saqué no eran importantes. En tal caso, se debe meter en mi modelo de entrenamiento y volver a procesar.*

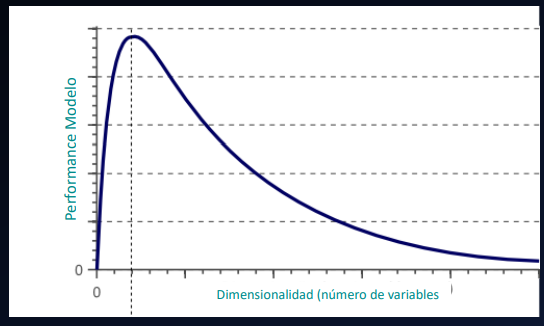
La reducción de dimensiones es un paso previo al entrenamiento del algoritmo

Un modo es tomar todas las variables, aplicar transformaciones para crear nuevas variables.



### La maldición de la dimensionalidad

* La maldición de la dimensionalidad se refiere a todos los problemas que surgen al trabajar con datos que se representan en un gran número de dimensiones.
* Podríamos pensar que al aumentar el número de variables explicativas se mejora la capacidad predictiva de los modelos.
* Lo cual, en general, sería cierto si realmente los predictores fuesen de utilidad para explicar la respuesta. Sin embargo, al aumentar el número de dimensiones se pueden agravar notablemente muchos de los problemas que ya pueden aparecer en dimensiones menores, esto es lo que se conoce como la maldición de la dimensionalidad (curse of dimensionality, Bellman, 1961).
* A medida que aumenta el número de variables, el número de muestras también debe aumentar proporcionalmente. Cuantas más variables tenemos, mayor número de muestras necesitaremos para tener todas las combinaciones de valores de características bien representadas en nuestra muestra.

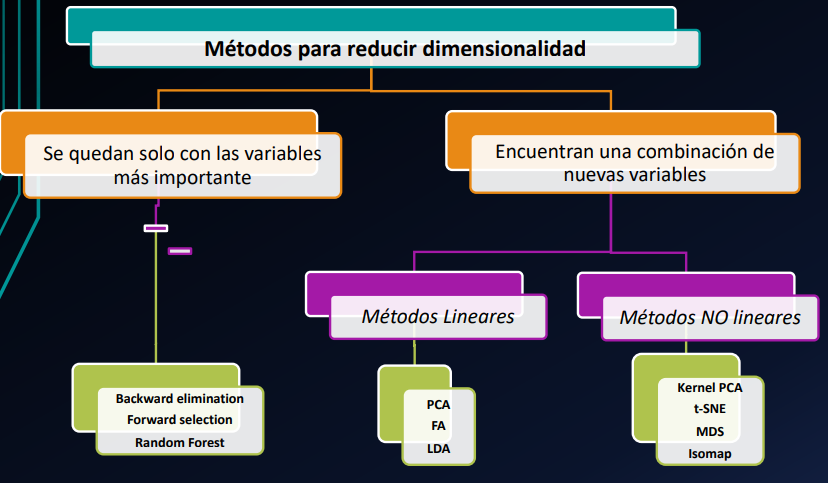


*Tiene que tratar de quedarse lo más cercano a este pico.*

* A medida que aumenta el número de variables, el modelo se vuelve más complejo. Cuanto mayor sea el número de variables, mayores serán las posibilidades de sobreajuste. *Las Imágenes, textos y audios son multidimensionales. Son los que más variables tienen. Aquí está bueno disminuir la dimensionalidad. Ej: MP3: todo lo que no se escucha no vale la pena guardarlo.*
* Un modelo de aprendizaje automático que se entrena con una gran cantidad de variables, se vuelve cada vez más dependiente de los datos en los que fue entrenado y, a su vez, se sobreajusta, y por lo tanto resulta en una performance deficiente con datos reales
* Evitar el sobreajuste es una motivación importante para realizar la reducción de dimensionalidad, pero eso no es todo y la reducción de la dimensionalidad tiene muchas más ventajas:
  + Menos datos singulares significan que mejora la precisión del modelo.
  + Menos datos significa que los algoritmos se entrenan más rápido.
  + Menos datos significa menos espacio de almacenamiento requerido
  + Menos dimensiones permiten el uso de algoritmos no aptos para una gran cantidad de dimensiones
  + Elimina las funciones redundantes y el ruido
  + La reducción de la dimensionalidad evita el problema del sobreajuste
  + La reducción de dimensionalidad es extremadamente útil para la visualización de datos
  + La reducción de dimensionalidad se puede utilizar para la compresión (imágenes, textos, audios, etc.)

### Métodos de reducción de dimensionalidad

* Existen dos tipos de métodos de reducción de dimensionalidad:
  + Un tipo de método **sólo mantiene las variables más importantes** en el conjunto de datos y elimina las variables redundantes. No se aplica ninguna transformación al conjunto de variables.
  + El otro método **encuentra una combinación de nuevas variables**. Se aplica una transformación adecuada al conjunto de variables. El nuevo conjunto de variables contiene valores diferentes en lugar de los valores originales. Este método se puede dividir en métodos lineales y métodos no lineales.
* Hay varios métodos de reducción de dimensionalidad que se pueden utilizar con diferentes tipos de datos para diferentes requisitos.



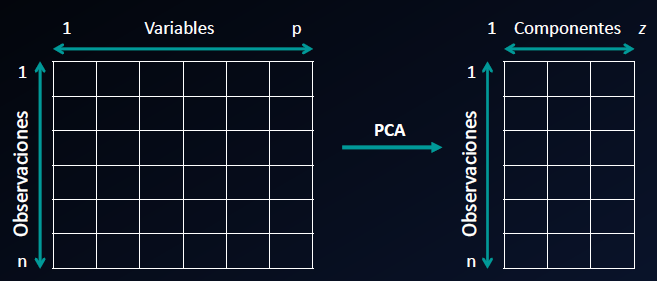
#### Análisis de componentes principales (PCA)

*PCA: análisis de componentes principales: combinan linealmente las variables*

* PCA es una técnica de reducción de dimensionalidad lineal. Transforma un conjunto de variables correlacionadas (p) en un número menor de variables k (k <p) no correlacionadas denominadas componentes principales, al tiempo que conserva la mayor cantidad posible de variación en el conjunto de datos original.
* El concepto principal detrás del PCA es considerar la correlación entre variables. Si la correlación es muy alta entre un subconjunto de las características, PCA intentará combinar las características altamente correlacionadas y representar estos datos con un número menor de variables linealmente no correlacionadas.
* Con estos componentes, es casi siempre posible reconstruir las variables originales. El algoritmo PCA intenta activamente minimizar el error de reconstrucción durante su búsqueda de los componentes óptimos. RTA Mario: a veces no es fácil hacer el camino inverso desde los componentes generados volver as variable originales
* Al reducir la dimensionalidad de los datos, PCA reducirá el tamaño de los datos mejorando el rendimiento de los algoritmos de aprendizaje automático.
* **PCA es una técnica no supervisada**, lo que significa que no utiliza la información del vector objetivo y, en cambio, sólo considera la matriz de características.
* Permite encontrar variables que están midiendo lo mismo, que están correlacionadas y entonces las combina para formar un nuevo componente

##### 

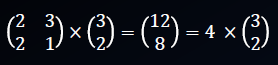
* Supóngase que existe una muestra con 𝑛individuos cada uno con 𝑝variables 𝑋1,𝑋2,…,𝑋𝑝 es decir, el espacio muestral tiene 𝑝dimensiones PCA permite encontrar un número de factores subyacentes 𝑧<𝑝que explican aproximadamente lo mismo que las 𝑝variables originales Donde antes se necesitaban 𝑝valores para caracterizar a cada individuo, ahora bastan 𝑧valores Cada una de estas nuevas variables recibe el nombre de componente principal



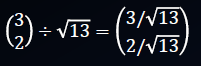
* El método de PCA permite por lo tanto " la información aportada por múltiples variables en solo unas pocas componentes.
* Aun así, no hay que olvidar que sigue siendo necesario disponer del valor de las variables originales para calcular las componentes
* Dos de las principales aplicaciones del PCA son la visualización y el preprocesado de predictores previo ajuste de modelos supervisados
* Haces proyecciones en otros ejes que te permiten realizar otro análisis.

##### Álgebra lineal - Eigenvectors (autovectores)

* Los eigenvectors son un caso particular de multiplicación entre una matriz y un vector. Obsérvese la siguiente multiplicación



* Un autovector es una matriz \* un vector que sea igual a una constante \* el vector
* 3/2: autovector
* 4: autovalor
* Si encuentro otro autovector, será perpendicular a 3/2
* El vector resultante de la multiplicación es un múltiplo entero del vector original. Los eigenvectors de una matriz son todos aquellos vectores que, al multiplicarlos por dicha matriz, resultan en el mismo vector o en un múltiplo entero del mismo. Los eigenvectors tienen una serie de propiedades matemáticas específicas
  + Los eigenvectors solo existen para matrices cuadradas y no para todas. En el caso de que una matriz n x n tenga eigenvectors el número de ellos es n
  + Si se escala un eigenvector antes de multiplicarlo por la matriz, se obtiene un múltiplo del mismo eigenvector Esto se debe a que, si se escala un vector multiplicándolo por cierta cantidad, lo único que se consigue es cambiar su longitud pero la dirección es la misma
  + Todos los eigenvectors de una matriz son perpendiculares ( entre ellos, independientemente de las dimensiones que tengan
* CORRELACIÓN: RANGO, PERIMETRO Y AREA: mayor radio, mayor período, mayor área .. hay correlación aunque no sea proporcionalmente directa.
* Puede que sea INVERSAMENTE CORRELACIONADO, es una correlación. (ver el valor si es cercana a 1 o 0).
* Dada la propiedad de que multiplicar un eigenvector solo cambia su longitud pero no su naturaleza de eigenvector, es frecuente escalarlos de tal forma que su longitud sea 1. De este modo se consigue que todos ellos estén estandarizados A continuación se muestra un ejemplo:
* El eigenvector  tiene una longitud de . Si se divide cada dimensión entre la longitud del vector, se obtiene el eigenvector estandarizado con longitud 1

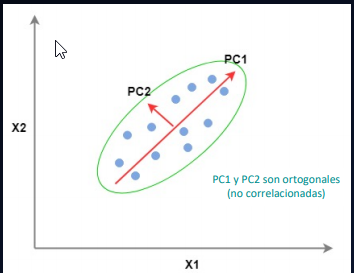


* Cuando se multiplica una matriz por alguno de sus eigenvectors se obtiene un múltiplo del vector original, es decir, el resultado es ese mismo vector multiplicado por un número. Al valor por el que se multiplica el eigenvector resultante se le conoce como eigenvalue A todo eigenvector le corresponde un eigenvalue y viceversa



* En el método PCA, cada una de las componentes se corresponde con un eigenvector y el orden de componente se establece por orden decreciente de eigenvalue Así pues, la primera componente es el eigenvector con el eigenvalue más alto

##### ¿Se requiere algún procesamiento antes de aplicar PCA?

* PCA no se puede aplicar al conjunto de datos con valores nulos. Por lo tanto, se deben tratar los valores nulos antes de continuar con PCA. Hay diferentes formas de tratar los valores nulos, como descartar las variables e imputar los datos faltantes utilizando la media o la mediana.
* No debemos aplicar PCA en el conjunto de datos que tiene atributos en diferentes escalas. Necesitamos estandarizar las variables antes de aplicar PCA.
* PCA se basa en principios de álgebra lineal y, por lo tanto, se vuelve ineficaz cuando las relaciones son de naturaleza no lineal.
* No entiendo el gráfico

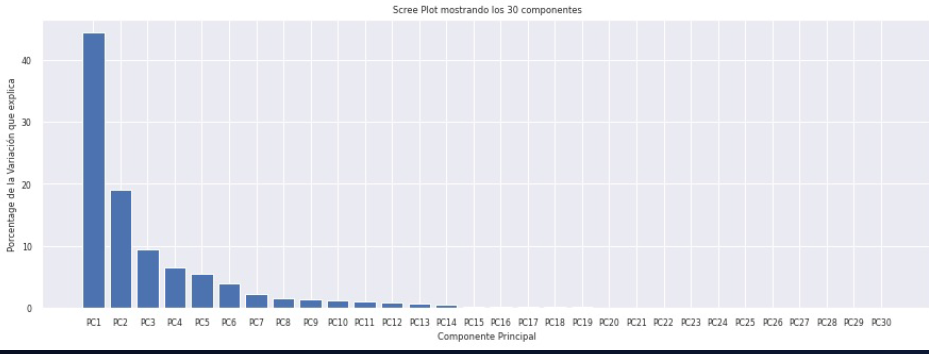




* *Buscar proyecciones de los datos que me expliquen la mayor variabilidad del set . (Ver autovectores)*
* *Si tengo 2 variables voy a poder encontrar una relación que me pesque la mayor variabilidad de los datos que otra. En ese caso, me voy a quedar con esa*
* *PC1, PC2: lo grafico en otros ejes. Según me quede graficada voy a ver cómo proyecta el resto de las variables. Ya no tengo más las variables originales, ahora tengo el componente 1 y el componente 2.*
* Hay 2 autovectores perpendiculares entre sí. Los cuales permiten explicar mejor lo que está pasando con los puntos dentro del gráfico. Esto es en lugar de usar x e y.
* Cada punto tiene un valor en x1 y en x2
* Por ejemplo, los puntos a y b tienen más diferencia en el componente PC1 que en el componente PC2. Mirando el dataset, veo que PC1 pesca más variabilidad que PC2, por lo cual puedo sacar PC2.

##### Pasos para realizar PCA

1. Estandarizar los datos
2. Calcule la matriz de covarianza: La covarianza es una medida de cómo una variable varía con respecto a la otra, en términos de la propia variación de sus valores con respecto a la media: esto podrá dar un valor positivo, negativo o cero; en este último caso ambas variables no varían una con respecto a la otra, es decir, son independientes. (Esta es la matriz de correlación que ya vimos?)
3. Calcule los autovectores y los autovalores de la matriz de covarianza: Una propiedad de los autovectores de una matriz es que estos son perpendiculares (ortogonales ) entre sí. Lo importante es que uno puede expresar sus datos en términos de autovectores
4. Clasifique los autovectores por sus autovalores correspondientes y obtenga los componentes principales con sus vectores de columna correspondientes a los k superiores. *(ordenarlos)*



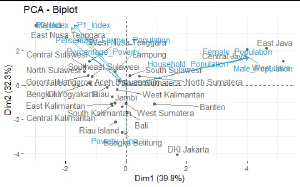
Genera 30 componentes principales, hay que ver cómo quedarme con menos variables, reduciendo el espacio (ej de 30 variables, me quedo con 3 variables).

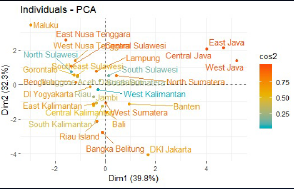
Pregunta Romina: podría pasar que cierta variable que están correlacionadas entre sí, no necesariamente estén relacionada con lo que queremos predecir? Rta **Mario**: me ayuda a explicar relaciones entre variables y me aumenta la posibilidad de predicción.

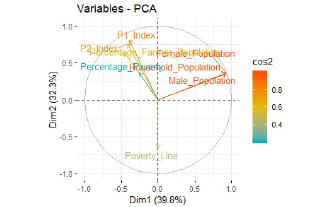
*NO SIEMPRE PUEDES REDUCIR A 2 COMPONENTES. Y ESO PUEDE PASAR. CUANDO NO ESTÁN CORRELACIONADAS LAS VARIABLES.*

#### 

#### Bi Plot







*Muestra las variables que yo medí y las componentes. El componente me dice el % de algo, y puedo verlo visualmente.. trato de entender la correlación de la nueva variable que estoy creando.*

*Siempre se trata de explicar lo que estamos mostrando, analizando..*

*Ej. en un análisis de laboratorio y al analizar las variables veo que el colesterol y los triglicéridos estaban relacionados y eso se notaba visualmente.*

***PCA puedo combinar esto con Clustering.***

***Reducción de dimensionalidad se aplica en mayor medida sobre los no supervisados.***

#### Notebooks

* 9na Clase - Cancer\_Mama\_PCA
* 9na Clase - PCA\_arrest\_USA
* 9na Clase - PCA\_Kmeans