

Fastbook 10

**Estadística Aplicada
al Marketing**

Clustering y series temporales



10. Clustering y series temporales

En el fastbook anterior nos adentramos en las técnicas de reducción de dimensionalidad dentro del aprendizaje no supervisado. Más concretamente en las dos técnicas principales: el análisis de componentes principales o PCA y en el análisis factorial o FA.

- Empezamos conociendo el **significado de dimensionalidad** y cuáles son los principales problemas a los que nos enfrentamos cuando estamos en un espacio de dimensión alta.
- Continuamos con la técnica más extendida de la reducción de dimensionalidad: el **análisis de componentes principales**. El PCA es una rotación de las variables originales para encontrar nuevas direcciones que reduzcan el número de variables. Las nuevas componentes son combinación lineal de las originales y maximizan la varianza explicada.
- Por último, nos familiarizamos con el **análisis factorial**. Es una técnica muy común en el ámbito social que tiene como objetivo medir variables latentes (no observables) a través de las variables originales. El FA está basado en la correlación entre las variables originales con las que se crean las variables latentes o factores.

En este último fastbook de la asignatura conoceremos el significado de clustering y su principal técnica, k-means. Continuaremos familiarizándonos con las series temporales y ARIMA.

Autora: Patricia Martín González

[Clustering](#)

[Series temporales](#)

[Resumen](#)

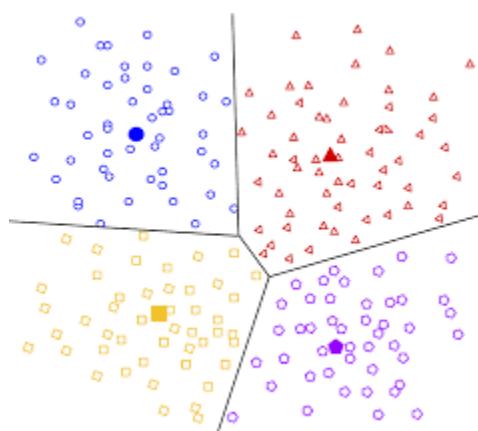
Clustering

X Edix Educación

Dentro del aprendizaje no supervisado (aquel en el que no existe variable objetiva), la pata que nos queda por estudiar es el análisis clúster (o clustering), también conocido como segmentación.



El clustering consiste en agrupar a los individuos (cada una de las observaciones de la base de datos) en grupos que sean homogéneos dentro de ellos, y heterogéneos respecto a otros grupos. De forma simple: el clustering segmenta en k grupos a los individuos de una población.



Estos grupos son llamados clústeres. Se busca que los individuos dentro de cada clúster sean parecidos, y a su vez estos clústeres sean diferentes entre ellos.

Al ser una técnica de aprendizaje no supervisado y no paramétrica, resulta muy importante elegir correctamente las variables que se usarán para hacer la segmentación, ya que serán las que determinen la robustez del modelo.

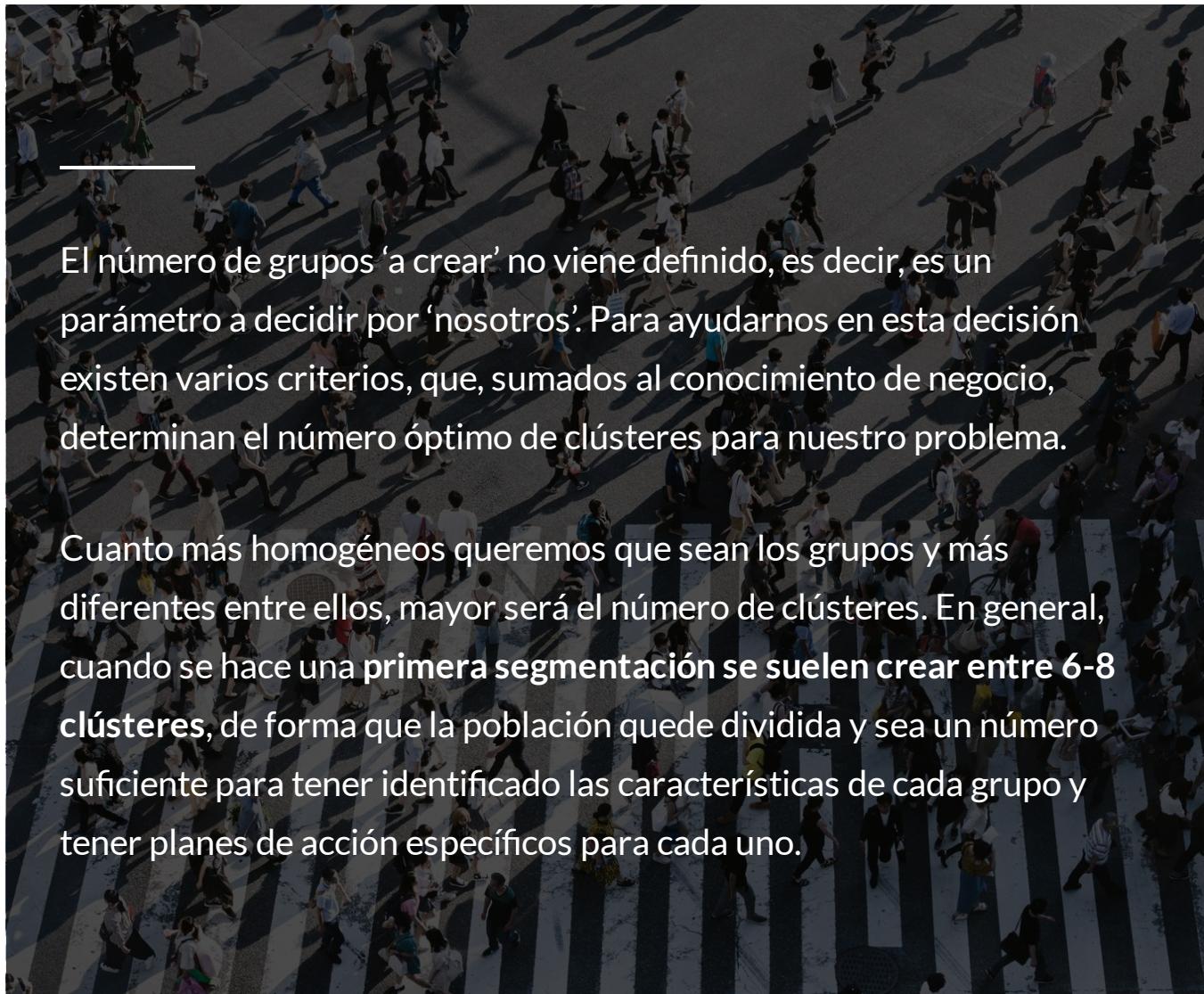
La forma de medir cuánto de parecidos o diferentes son dos individuos para saber a qué clúster asignarlos va a depender del tipo de variables con las que se haga la segmentación.

Si son variables numéricas (cuantitativas) suele usarse la distancia, pero si tenemos variables cualitativas que no se pueden expresar mediante números (por ejemplo, los colores, formas geométricas, nivel de estudios) o que la distancia no tiene sentido (para las variables binarias) se suele aplicar o definir una métrica de similitud o divergencia.

	Distancia más extendida	Medidas de similitud
¿Cuáles destacan?	Euclídea	Índice binario y el Tanimoto
¿Qué hacen?	Es la más extendida, aunque existen más opciones en función de la necesidad o lo que se quiera ponderar, como, por ejemplo, la distancia de Mahalanobis, de Chebyshev, Manhattan, etc.	Tanimoto (mide el número de unos que hay entre los individuos).

La asignación de un individuo a un clúster puede hacerse de varias formas, aunque generalmente se asigna cada individuo al clúster que tenga el centroide (punto central de cada clúster) más cercano.

La **distancia entre clústeres** puede medirse de varias formas, aunque generalmente se usa la distancia entre los centroides. Otra alternativa es medir la distancia entre los individuos más cercanos de cada clúster.



El número de grupos ‘a crear’ no viene definido, es decir, es un parámetro a decidir por ‘nosotros’. Para ayudarnos en esta decisión existen varios criterios, que, sumados al conocimiento de negocio, determinan el número óptimo de clústeres para nuestro problema.

Cuanto más homogéneos queremos que sean los grupos y más diferentes entre ellos, mayor será el número de clústeres. En general, cuando se hace una **primera segmentación** se suelen crear entre **6-8 clústeres**, de forma que la población quede dividida y sea un número suficiente para tener identificado las características de cada grupo y tener planes de acción específicos para cada uno.

Primer paso

El primer paso en clustering es seleccionar las variables que se usarán en la modelización, a continuación, crear los segmentos, pero no menos importantes son los pasos siguientes.

Profiling

Una vez que se tienen los grupos creados y los usuarios asignados a los clústeres, es necesario describir y poner un nombre a cada uno de los grupos. Para ello, nos basamos en métricas de las variables con las que se ha creado el clúster, aunque también se suele utilizar variables adicionales que no han sido utilizadas en la modelización. Una buena práctica, una vez definidos y nombrados todos los grupos, es sacar un perfil prototípico de cada uno de los grupos (generalmente el individuo más cercano al centroide) para tener claro cómo sería un usuario de cada uno.

Plan de acción

Una vez hemos completado el paso anterior, es muy importante decidir las acciones a llevar a cabo para cada uno de los grupos, para que la segmentación tenga sentido y no sea un mero ejercicio que se ‘quede guardado en una caja’. La finalidad de las segmentaciones es crear grupos de la población para conocerlos mejor y decidir acciones sobre cada uno de ellos con la finalidad de mejorar su satisfacción, personalizar las comunicaciones, servicio de atención al cliente, continuidad, etc.

El clustering o segmentación lo podemos encontrar en casi todos los campos:

Banca

En **banca** se suele dividir a los clientes para cubrir las necesidades de los clientes de la mejor manera posible. Por ejemplo: clientes más digitales, clientes de zonas rurales, clientes con necesidad de oficina, etc. Otra segmentación podría ser según la relación con el banco: clientes muy activos, clientes con productos contratados, clientes con riesgo de fuga, etc.

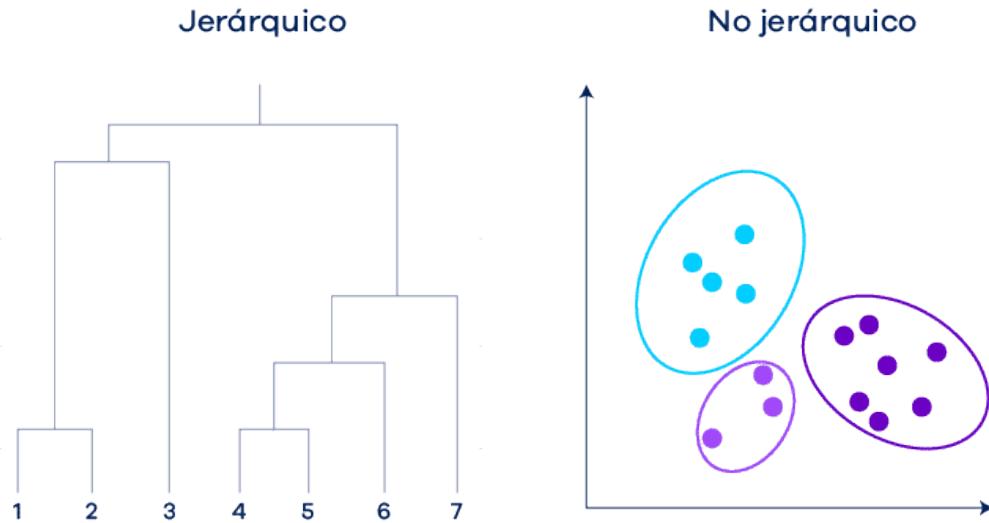
Retail

En **retail** se tiene clasificado a cada cliente en el grupo adecuado para saber qué campañas publicitarias son más efectivas, o qué grupo de productos son los más interesantes, etc. Por ejemplo, una segmentación de los clientes de una empresa de bricolaje podría ser: clientes con piscina, clientes con jardín, clientes ocasionales, clientes con reforma, etc.

Telecomunicaciones

En las compañías de telecomunicaciones se divide a los clientes para saber las necesidades que tienen, cómo tenerlos contentos y retenerlos en la compañía, clasificaciones para atención telefónica, etc. Por ejemplo: clientes mayores que buscan comodidad, clientes que buscan el producto más económico, clientes que viven en casas de alquiler, clientes con todos los productos, etc.

Las técnicas de clustering suelen dividirse en dos grandes grupos: **clustering jerárquico** y **clustering no jerárquico**.



En los **métodos jerárquicos** los individuos no se asignan a un clúster de una sola vez, sino que se van haciendo **particiones sucesivas** a ‘distintos niveles de agregación o agrupamiento’.

Se pueden subdividir en aglomerativos (aquellos que comienzan con tantos clústeres como individuos hay) o de división (se empieza con un único clúster y se va dividiendo).

El resultado es un **dendograma** (imagen superior de la izquierda) en el que se puede ver los niveles de agrupación y la forma en la que se unen los individuos.

Los **métodos no jerárquicos** son los más extendidos. En esta tipología de modelos el número de clústeres está definido y cada individuo es asignado generalmente a un único clúster. Existen numerosas técnicas: DBSCAN, PAM, K-means, K-modes, k-prototypes, etc. De entre todos ellos destaca el **algoritmo de K-means**.

K-means

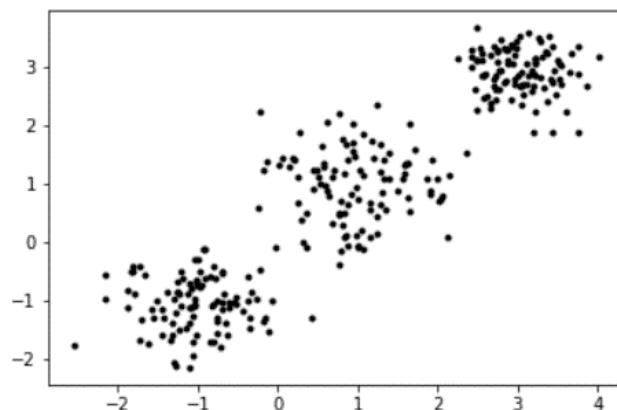
K-means es el algoritmo más extendido y usado de clustering gracias a su fácil implementación a la par que los buenos resultados que se consiguen. De forma intuitiva, k-means asigna cada individuo al clúster más cercano, concretamente al centroide (centro de los clústeres) más cercano. El número de clústeres se establece previamente.

K-means es un algoritmo iterativo en el que se van actualizando los centroides y los clústeres en cada iteración.

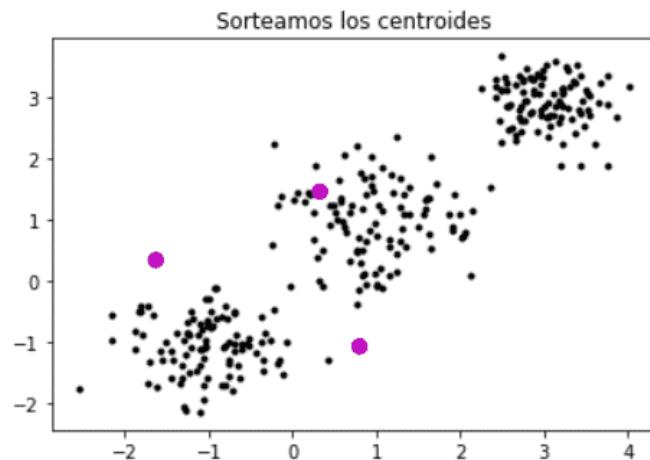
En la primera iteración (como los clústeres no han sido definidos) se escogen los **centroides de manera aleatoria** en el espacio muestral, y a partir de la siguiente iteración se **asigna cada punto al clúster más cercano**. El algoritmo itera hasta que se alcance el número máximo de iteraciones (parámetro que se suele dar como input del modelo) o no haya cambios en los clústeres.

Veamos un ejemplo:

Supongamos que tenemos los siguientes puntos en un espacio 2D y queremos crear 3 clústeres.

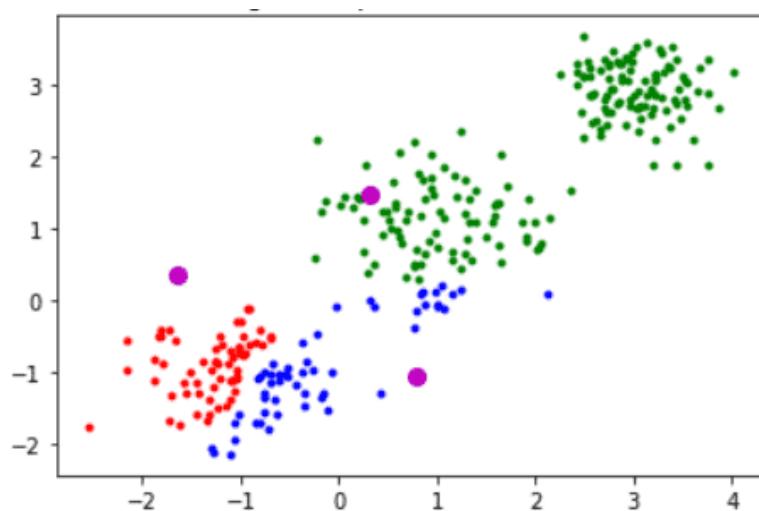


Como no tenemos centroides, los ponemos aleatoriamente dentro del espacio muestral (puntos morados).

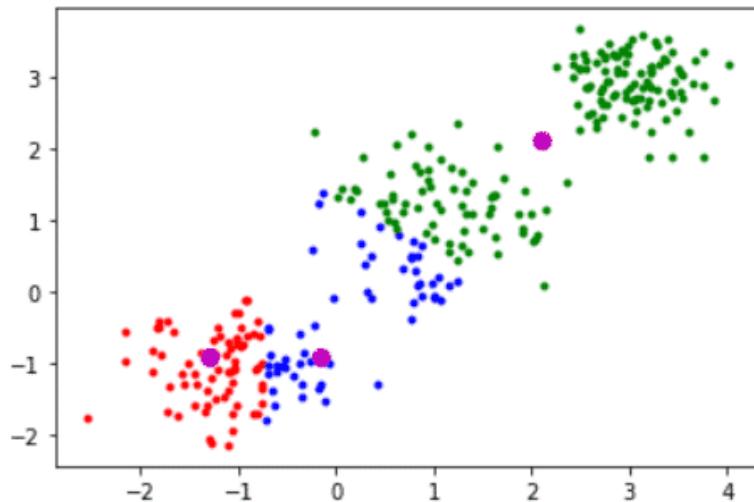


Iteración 1

Empezamos con la iteración 1. Asignamos los puntos al centroide más cercano.

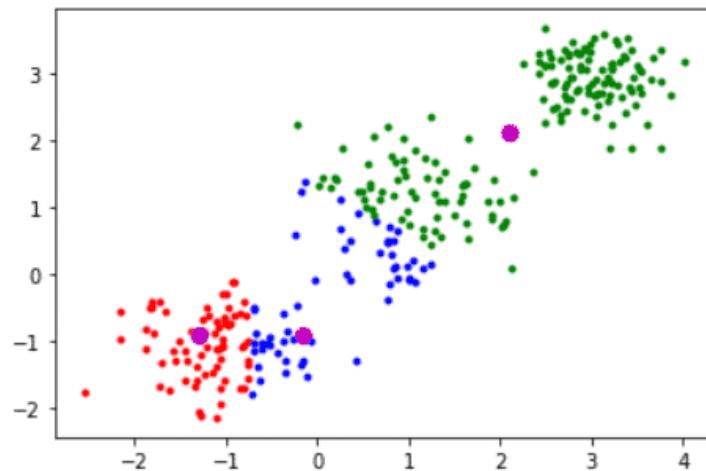


Y recalculamos los centroides.

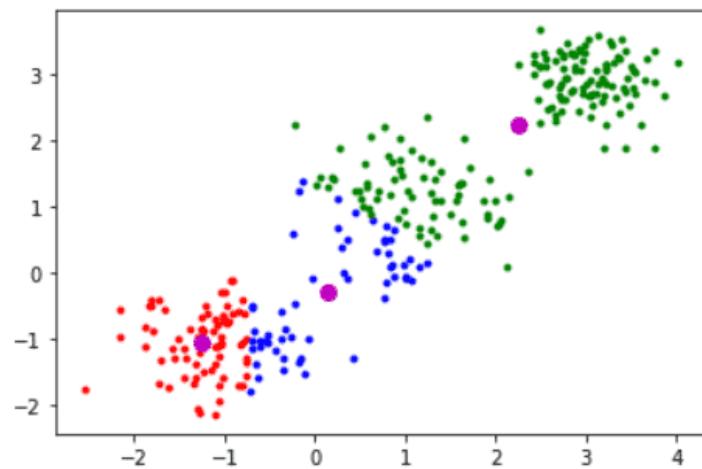


Iteración 2

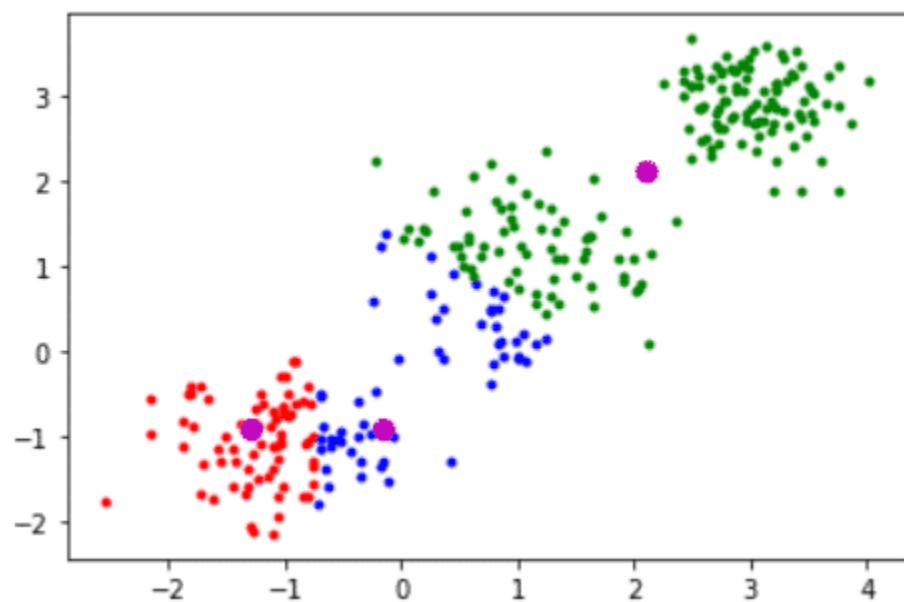
Pasamos a la iteración 2: reasignamos los puntos al clúster con el centroide más cercano y recalculamos centroides.



Reasignamos los puntos a los clústeres.



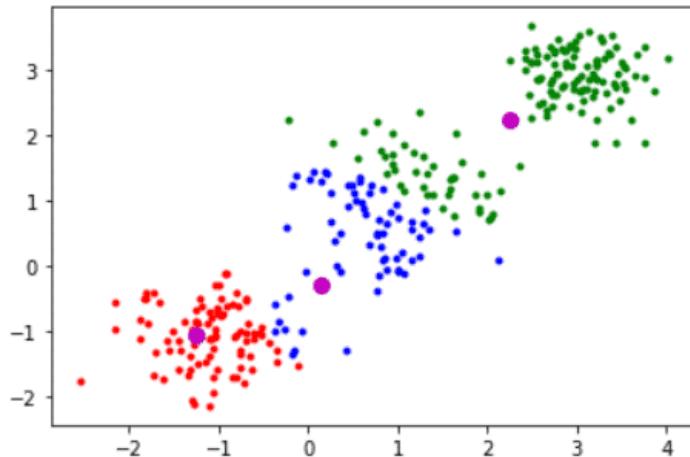
Recalculamos centroides.



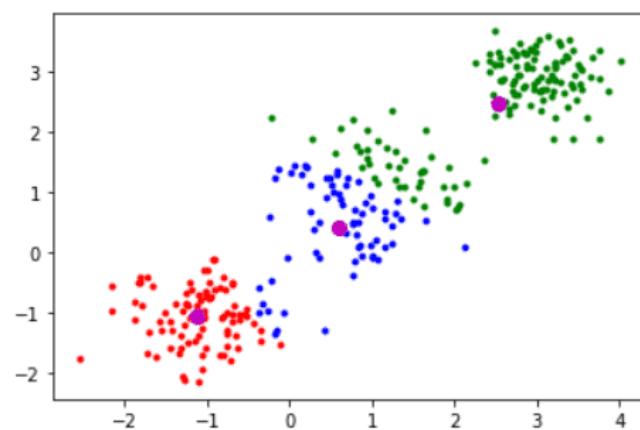
Iteración 3

Reasignamos los puntos a los clústeres con centroides más cercanos, y recalculamos centroides.

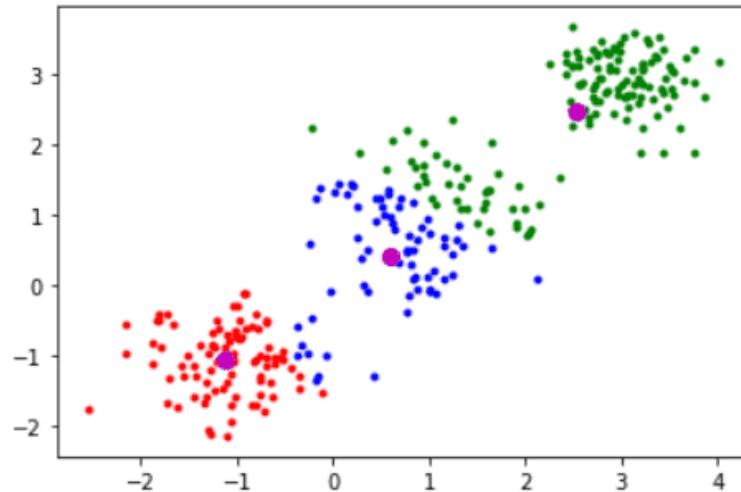
Reasignamos los puntos a los clústeres.



Reasignamos los puntos a los clústeres.

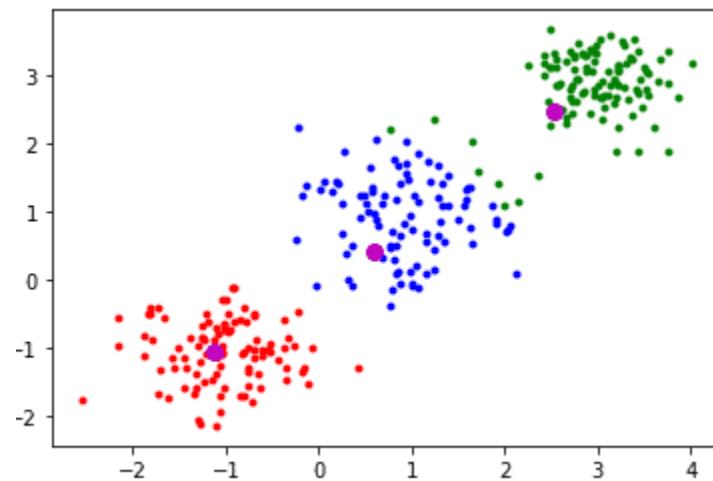


Reasignamos los puntos a los clústeres.

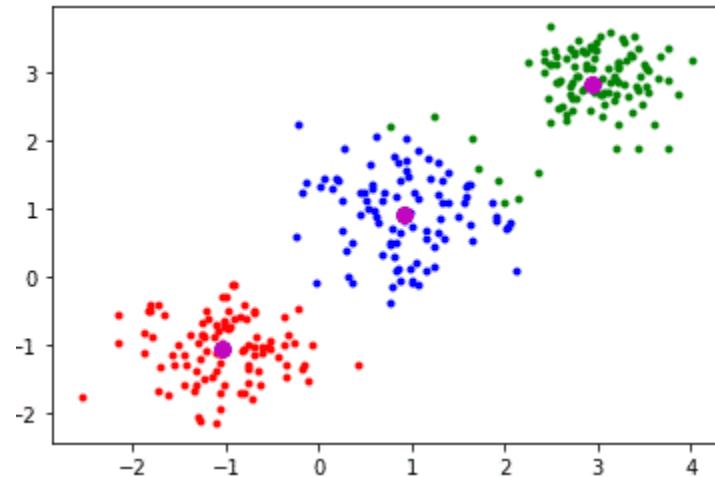


Iteración 4

Reasignamos los puntos a los clústeres.



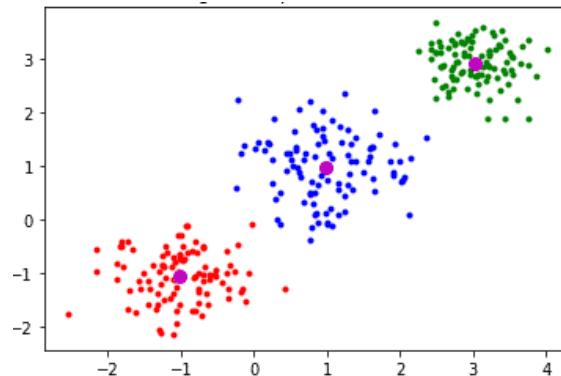
Recalculamos centroides.



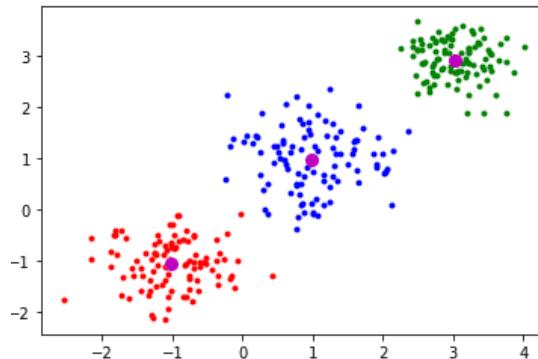
Iteración 8

Seguimos iterando hasta la iteración 8 en la que los clústeres y, por tanto, los centroides, no se modifican y habremos encontrado la solución a nuestro problema.

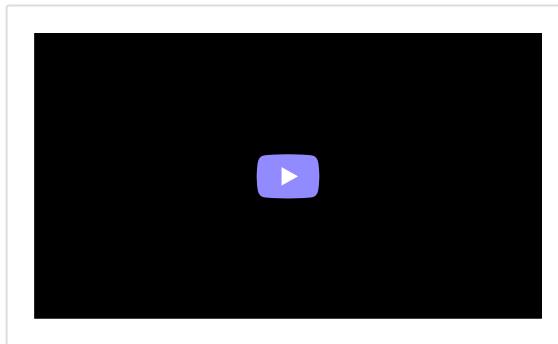
Reasignamos los puntos a los clústeres.



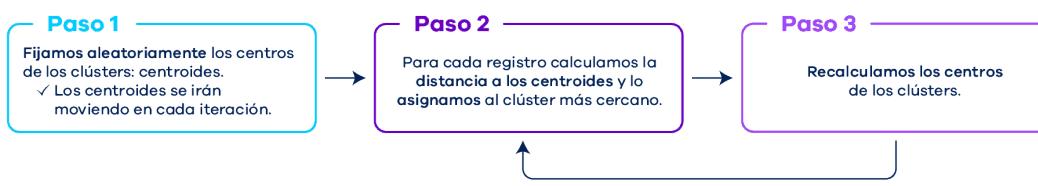
Recalculamos centroides.



Si quieras ver otro ejemplo completo, te animo a que veas de forma dinámica cómo se actualizan los clústeres y la asignación de cada punto en este vídeo.



El algoritmo de K-means podría resumirse como:



Como ves, es un algoritmo muy sencillo basado en distancias que calcula una segmentación con un alto potencial.

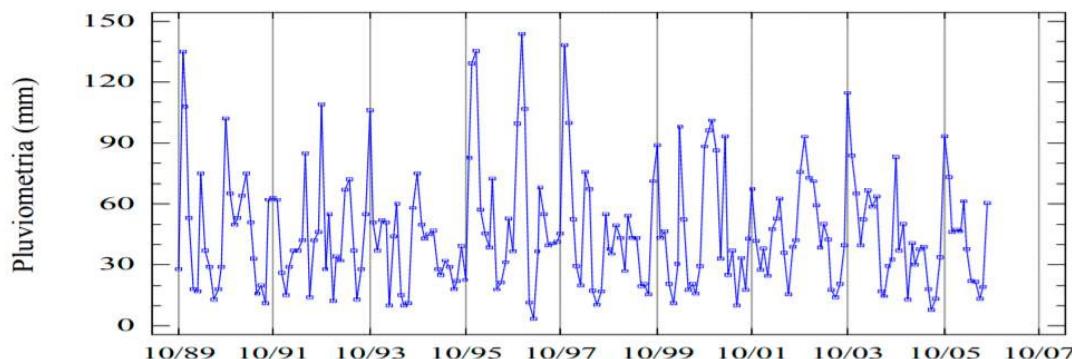
Series temporales

X Edix Educación

Una serie temporal se define como una colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones suelen recogerse en instantes de tiempo equiespaciados.

En las series temporales interesa estudiar los cambios de la variable a lo largo del tiempo y predecir los valores que va a tomar en el futuro.

En el análisis de series temporales resulta de vital importancia graficar la serie. Generalmente (por no decir siempre), suelen representarse en un gráfico bidimensional: en el que el eje X se representa la evolución temporal, y el eje Y las fluctuaciones de la variable a estudiar. En el gráfico inferior encontrarás un ejemplo de una serie temporal: la evolución de los mm promedio de lluvia al mes de la península desde octubre 1989.

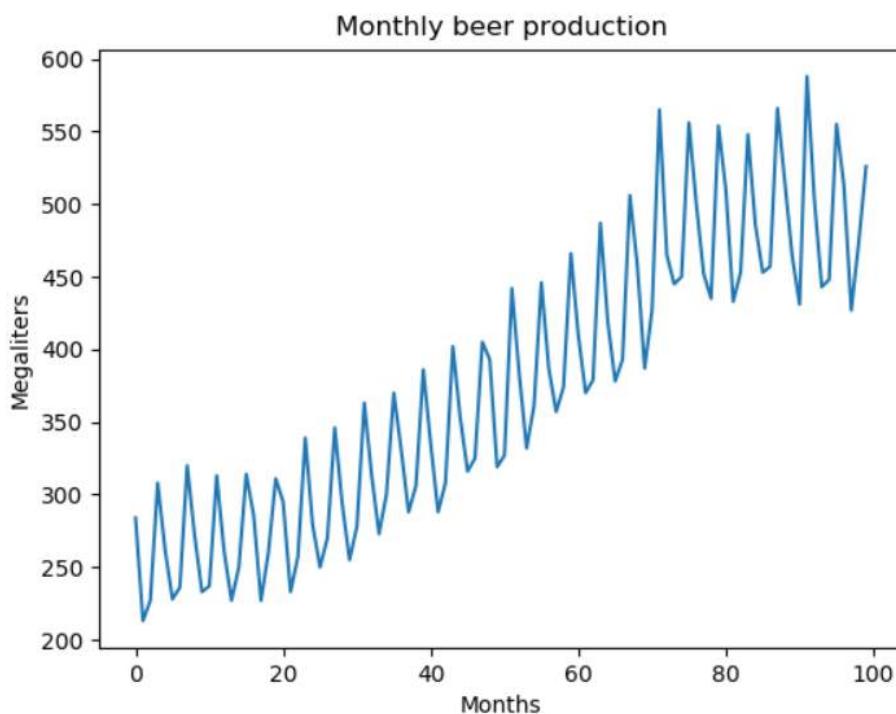


La unidad de la variable temporal es ‘libre’, es decir, puede ser milisegundos, días, años, etc. Dependiendo de la frecuencia de medición se escogerá la unidad apropiada. Los modelos de series temporales no se ven afectados por la frecuencia, ya que lo importante es que sean puntos secuenciales en el tiempo y no la separación entre cada punto.

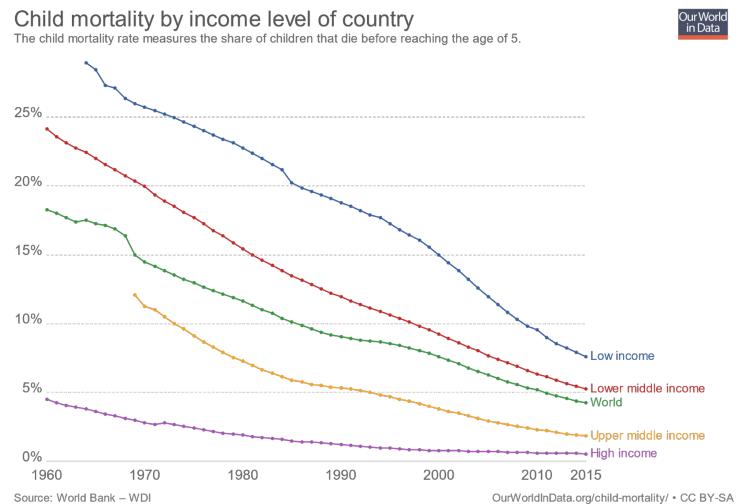
Las **series temporales se caracterizan por ser autoexplicativas**, es decir, cada instante de tiempo se expresa como combinación lineal de los instantes anteriores (no necesariamente de los inmediatamente anteriores), los residuos o la serie diferenciada ($X_{t-1}-X_{t-2}$).

Las series temporales forman parte de ‘nuestro día a día’ y podemos encontrarlas en casi todos los campos. ¿Se te ocurre alguno? Aquí te muestro algunos **ejemplos**:

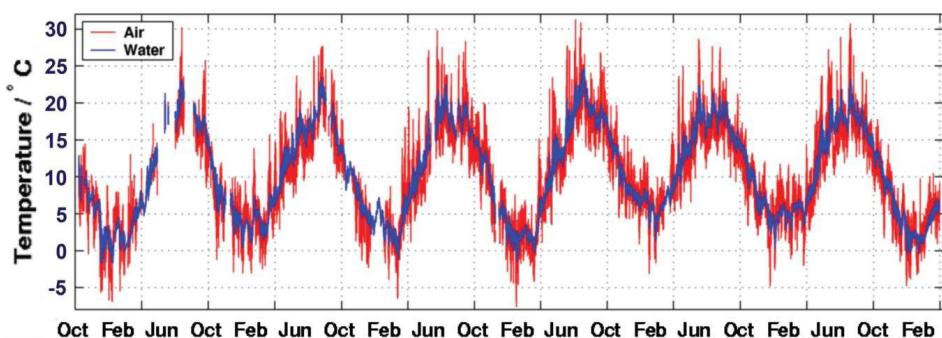
En **economía y marketing** podemos encontrar numerosas series temporales, como, por ejemplo, el precio de los alquileres, los indicadores macroeconómicos (PIB, IPC, desempleo), la demanda de alimentos, los beneficios de una entidad bancaria, etc. En el gráfico inferior puedes ver la evolución temporal de la producción de cerveza a nivel mensual.



En **demografía** podemos encontrar la serie del número de habitantes, la tasa de natalidad, tasa de dependencia, etc. En el gráfico inferior puedes ver la evolución temporal de la tasa de mortalidad infantil anual por tipo de país.

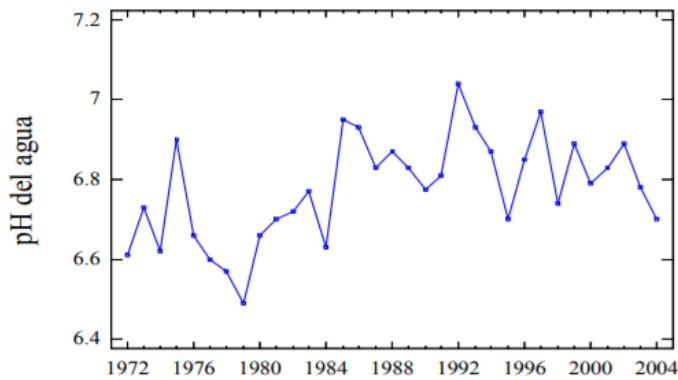


En **meteorología y medio ambiente** podemos encontrar numerosos ejemplos: evolución horaria o semanal o mensual de la contaminación, temperatura mínima media y máxima mensual, vertido de residuos tóxicos a un río, precipitaciones diarias, etc. En el gráfico inferior puedes observar el comportamiento de la temperatura del aire y del agua a nivel semanal.



No todas las series temporales son objeto de estudio, pues hay algunas que son totalmente aleatorias y no pueden modelizarse, como, por ejemplo, el pH anual del agua (gráfico inferior), en el que no podemos sacar ningún patrón de comportamiento y, por tanto, no es una serie de este tipo de estudio.

Rio Santa Cruz (Washington, USA)



Dentro de las **series temporales** podemos diferenciar entre **serie estacionaria** y **serie no estacionaria**:

Serie estacionaria

Diremos que es una serie estacionaria si la media y la variabilidad se mantienen constantes a lo largo del tiempo.

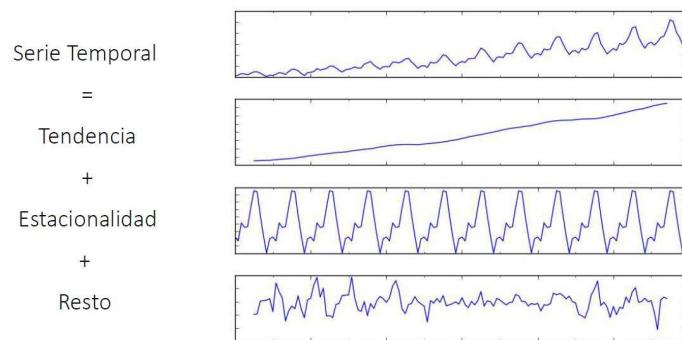
Serie no estacionaria

Se define una serie no estacionaria cuando la media o variabilidad no sean constantes en el tiempo. Esto trae una serie de características:

- Pueden tener tendencia (los datos presentan un comportamiento creciente o decreciente).
- Puede tener efectos estacionales, es decir, que el comportamiento de la serie sea parecido en diferentes momentos equiespaciados de la serie.
- Pueden mostrar cambios de varianza.

Generalmente, las series temporales de la vida real son no estacionarias, y tienen tendencias, cambios de varianza y efectos estacionales (te reto a que intentes encontrar estos comportamientos en los ejemplos anteriores).

Las series temporales usualmente las descomponemos en tres partes de forma simple:



Tendencia

Cambio a largo plazo que se produce en relación al nivel medio. Dicho de forma simple: cambio a largo plazo de la media. La tendencia se identifica con un movimiento suave de la serie a largo plazo.

Efecto estacional

Muchas series temporales presentan cierta periodicidad (anual, mensual...). Por ejemplo, el paro en España generalmente aumenta en invierno y disminuye en verano. Estos tipos de efectos son fáciles de entender y se pueden medir explícitamente o incluso se pueden eliminar del conjunto de los datos desestacionalizando la serie original.

Resto o componente aleatoria

Una vez identificadas las componentes anteriores y después de haberlas eliminado, persisten unos valores que son aleatorios. Se pretende estudiar qué tipo de comportamiento aleatorio presentan estos residuos, utilizando algún tipo de modelo probabilístico que los describa.

Las series temporales, generalmente, se podrían escribir como:

$$X_t = T_t + E_T + I_t$$

Donde X_t es la serie temporal a estudiar, T_t es la componente estacional e I_t es la componente aleatoria.

ARIMA

El ARIMA es la **técnica de análisis de series temporales más extendida y usada**. Es un modelo **autorregresivo** (depende de sus observaciones pasadas) que modeliza el comportamiento de una variable a lo largo del tiempo como combinación lineal de su comportamiento en el pasado. La principal ventaja es que proporciona predicciones óptimas a corto y medio plazo.

El **acrónimo ARIMA significa modelo autorregresivo integrado de media móvil (*autoregresive integrated moving average*)**. Se descompone en tres partes denominadas componentes: **AR-I-MA**. Cada componente modela un comportamiento distinto de la serie, y tiene un parámetro asociado (p, d, q):

AR	<p>La parte AR es la componente autorregresiva y estudia el comportamiento histórico de la serie. Esta componente recoge la influencia de los instantes anteriores sobre el instante actual. Sería de la forma: $X_t = \varphi_{t-1}X_{t-1} + \varphi_{t-2}X_{t-2} + \dots + \varphi_{t-p}X_{t-p}$. El parámetro asociado p indica el grado de la componente autorregresiva más lejana que influye sobre el instante actual. Cada φ_k es independiente, es decir, que el parámetro $p=52$ no significa que todos los anteriores tengan que existir, sino que al menos X_{t-52} influye sobre la serie y, por tanto, su coeficiente $\varphi_{t-52} \neq 0$.</p>
I	<p>La parte I es la componente de regulación estacionalaria, con ella eliminamos la componente estacional. Garantiza la estacionariedad de la serie, es decir que su media y varianza no cambie a lo largo del tiempo. El parámetro asociado d indica el número de diferenciaciones que se aplican a la serie. Por ejemplo, sea X_t la serie original, si $d=1$, la nueva serie diferenciada será $Y_t = X_t - X_{t-1}$.</p>
MA	<p>La parte MA refleja la componente media móvil, es decir, recoge la influencia de los residuos (error de la predicción) sobre la serie original. Se encarga de corregir los errores que comete el modelo pasado. Sería de la forma: $X_t = \theta_{t-1}\varepsilon_{t-1} + \theta_{t-2}\varepsilon_{t-2} + \dots + \theta_{t-q}\varepsilon_{t-q}$ donde ε_{t-1} es el error de predicción de la serie* en el instante $t-1$. El parámetro asociado q indica el grado del residuo más lejano que influye sobre el instante actual. Cada θ_k es independiente, es decir, que el parámetro $q=52$ no significa que todos los anteriores tengan que existir, sino que al menos ε_{t-52} influye sobre la serie y, por tanto, su coeficiente $\theta_{t-52} \neq 0$.</p> <p>* $\varepsilon_{t-1} = X_{t-1} - \hat{X}_{t-1}$ siendo \hat{X}_{t-1} la predicción de la serie para $t-1$.</p>

De forma resumida, una serie que sea de la forma ARIMA (p,d,q) queda formulada de la forma:

$$Y_t^{(d)} = C + \underbrace{\varphi_{t-1} X_{t-1}^{(d)} + \dots + \varphi_{t-p} X_{t-p}^{(d)}}_{\text{Comp. Autorregresiva}} + \theta_{t-1} \varepsilon_{t-1}^{(d)} + \dots + \theta_{t-q} \varepsilon_{t-q}^{(d)} + \varepsilon_t^{(d)}$$

Comp. Autorregresiva *Comp. de Media Móvil*

Como vemos, se trata de un modelo de regresión lineal múltiple, donde la variable dependiente es la propia serie X_t (diferenciada o no) y las variables independientes son valores de la serie X_{t-i} y de los residuos ε_{t-j} hasta unos órdenes p y q, respectivamente.

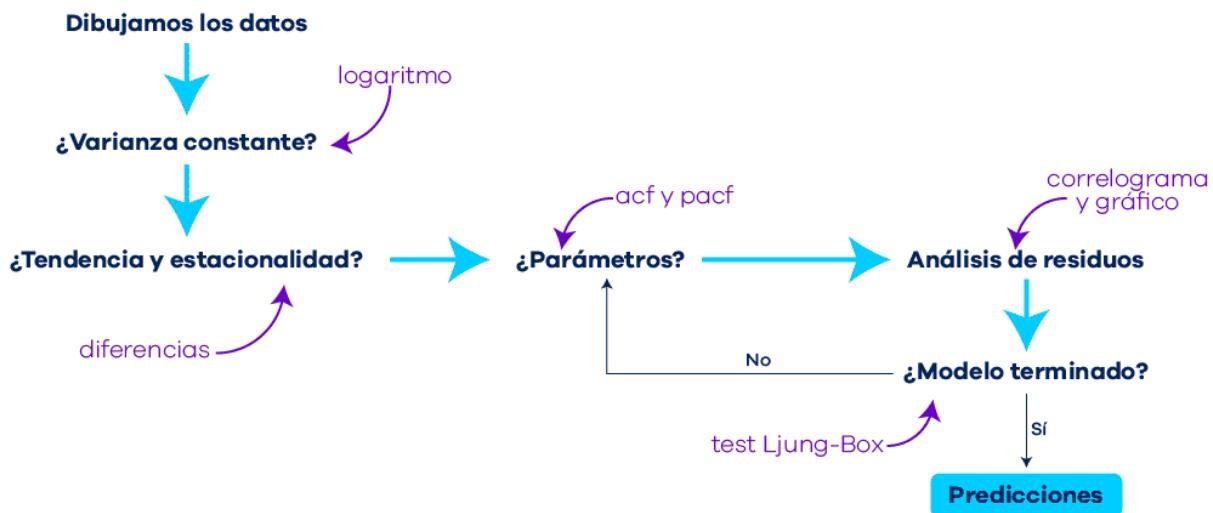
En los modelos ARIMA no es necesario que los tres parámetros sean distintos de 0. Por ejemplo, en un modelo semanal la serie podría ser ARIMA(52,0,1), que significaría que el instante t-52

(un año antes) y el residuo anterior tienen influencia sobre el instante t.

El ajuste de las series temporales no es inmediato y puede hacerse de varias formas. Para determinar el orden de los parámetros p y q nos apoyamos en los gráficos ACF y PACF. Estos gráficos de barras muestran la influencia de los retardos de la serie y de los residuos sobre el instante actual. Con ACF medimos la parte MA y con PACF medimos la parte AR.

Los órdenes de los parámetros se van calculando de **forma iterativa**, es decir, se ajustan los órdenes de p, d, y q con el gráfico, se vuelve a graficar con estos parámetros aplicados, se vuelve a ajustar la serie, etc. El modelo se dirá que está ajustado cuando ya no queden patrones de comportamiento que modelizar y la parte restante sea ruido blanco (aleatorio).

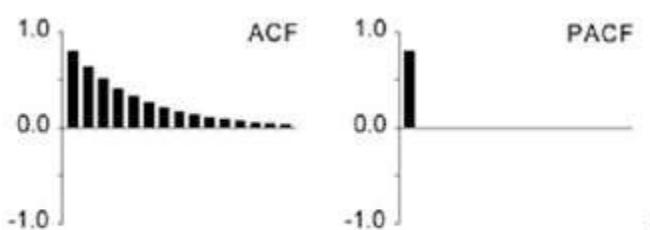
De forma esquemática, un modelo ARIMA sigue los siguientes pasos:



Veamos algunos **ejemplos**.

Ejemplo 1

El siguiente modelo vemos que tiene una barra claramente destacada en el gráfico PACF, por lo tanto, estaríamos ante una serie AR(1) y quedaría modelizado de la forma: $AR(1): X_t = 0,8X_{t-1} + \varepsilon_t$

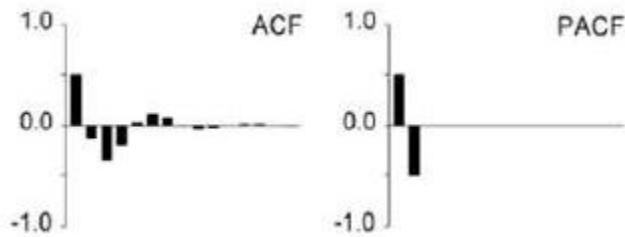


$$AR(1): Y_t = 0.8Y_{t-1} + A_t.$$

Ejemplo 2

En el ejemplo siguiente vemos que hay dos barras claramente destacadas en PACF, por lo que el modelo sería:

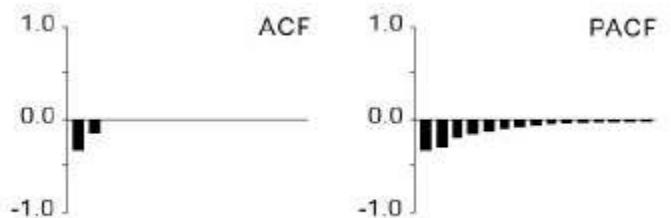
$$AR(2): X_t = 0,76X_{t-1} - 0,5X_{t-2} + \varepsilon_t$$



Ejemplo 3

A continuación, tenemos un ejemplo de un modelo MA(2), con dos barras destacadas sobre el resto en el gráfico ACF. El modelo, por tanto, quedaría de la forma:

$$MA(2): X_t = -0,6\varepsilon_{t-1} - 0,2\varepsilon_{t-2} + \varepsilon_t$$



En las siguientes asignaturas profundizarás en los modelos de series temporales donde te familiarizarás con el cálculo de cada parámetro, como medir la bondad de un modelo, hipótesis para aplicarlos, etc.

Resumen

X Edix Educación

Con este último fastbook de la asignatura hemos cubierto dos de las técnicas más usadas dentro en el área de analytics.

Hemos empezado conociendo significado de **clustering** y dónde se ubica dentro de la estructura de los algoritmos de machine learning. Descubrimos los pasos a seguir para cualquier proyecto de segmentación, así como los puntos importantes de cada fase (selección de variables, distancia, algoritmo a aplicar y número de clústeres). Hemos aprendido la diferenciación entre algoritmos jerárquicos y no jerárquicos, destacando entre todos ellos el K-means. Este **algoritmo iterativo** es el más extendido y está basada en la distancia a los **centroides**.

Cerramos la asignatura con el módulo de **series temporales**. Nos hemos familiarizado con el concepto de serie temporal, las partes básicas en las que se puede descomponer y los múltiples campos en los que podemos encontrar este tipo de modelos. Para finalizar, hemos conocido los principios básicos de los modelos ARIMA (el más extendido de las series temporales), descubriendo el significado de cada una de las patas, los parámetros asociados y los gráficos en los que apoyarnos para ajustar un modelo.

¡Enhorabuena! Fastbook superado

edix

Creamos Digital Workers