

Fastbook 07

Analítica de Cliente & Predictive Analytics

Teoría general de series temporales (II)



07. Teoría general de series temporales (II)

En el fastbook 06 vimos lo natural que es encontrar series temporales en cualquier negocio. Todo se mide, todo se quiere monitorizar, pero, como las personas no somos muy buenas procesando mucha información, tratamos de resumir eventos y procesos en simples números (KPIs) y, al recopilar varios de estos valores a lo largo del tiempo, aparecen las series temporales.

Ahora que ya hemos cubierto los aspectos fundamentales de las series temporales, es el momento de dar el salto y ver los primeros modelos. Este será el objetivo de este fastbook, pero, antes, es importante responder a las siguientes tres preguntas para situarnos:

- ¿Qué utilidades tienen las series temporales más allá de describir lo que ha ocurrido?
- ¿Qué serie o series temporales queremos modelizar?
- ¿Es necesario reajustar las series temporales antes de modelizar?

Una vez las hayamos respondido, pasaremos a la segunda sección, donde estudiaremos en profundidad los siguientes modelos:

- Regresión lineal.
- Descomposición.
- Suavizado exponencial.

Conocer estos tres modelos es esencial para poder entender modelos más complejos de series temporales, como veremos en el fastbook 09.

Y, por último, cerraremos este fastbook comentando conceptos técnicos que aparecerán cuando ajustemos modelos de series temporales.

Autor: Miguel Ángel Fernández

-  **Utilidades, objetivo y reajustes**
-  **Modelos de series temporales**
-  **Métricas de calidad**
-  **Resumen y conclusiones**

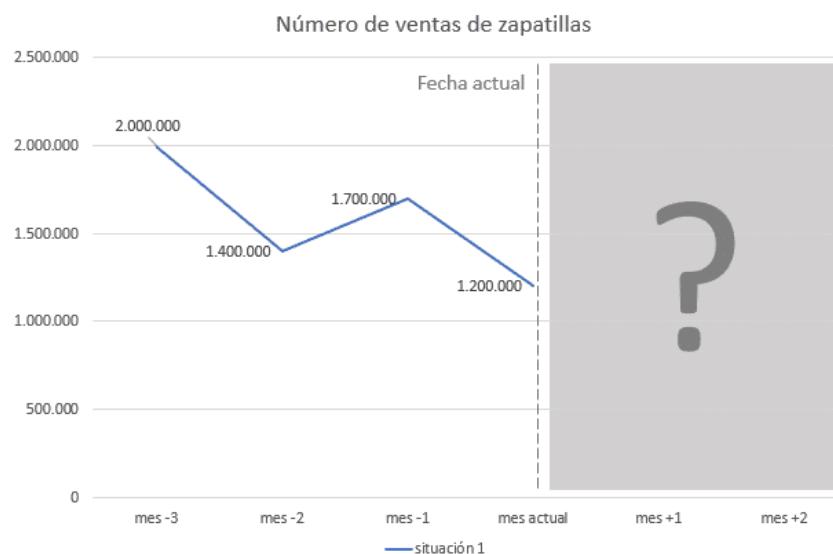
Utilidades, objetivo y reajustes

X Edix Educación

En el fastbook 06 nos hemos lanzado directos a analizar series temporales y ver aspectos importantes de ellas como outliers, tendencias, estacionalidades, etc. Aunque, de momento, la única utilidad que hemos mostrado de las series temporales es la descripción a pasado del suceso que mida dicha serie. Pero ¿qué ocurrirá en el futuro?

Cuando tenemos una serie temporal es natural hacerse preguntas sobre lo que ocurrió en el pasado, pero también lo es querer conocer cómo va a seguir evolucionando una serie en el futuro. Analizar el pasado tiene un gran valor para hacer retrospectiva de acciones que realizó la compañía o campañas de marketing y entender qué impacto tuvieron en el negocio, pero conocer qué ocurrirá en el futuro nos permitirá anticipar acciones y planificar mejor.

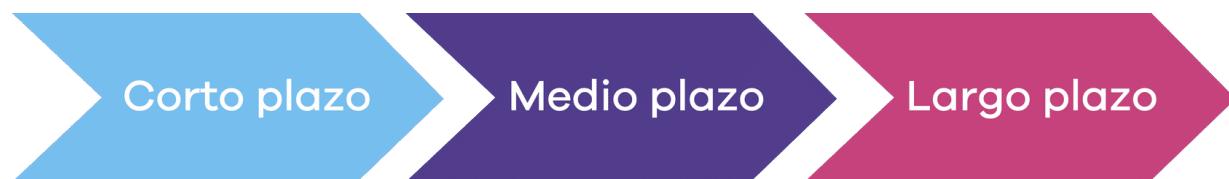
Si recordamos el ejemplo de las ventas de zapatillas, querríamos conocer qué valor va a tomar la serie temporal azul (es decir, qué ventas tendremos) en los meses ‘mes+1’ y ‘mes+2’.



Responder qué ocurrirá en el futuro nos proporciona las grandes utilidades que tienen las series temporales.

Utilidades

En las siguientes secciones veremos los modelos de series temporales que usaremos para predecir qué valores va a tomar una serie temporal en el futuro. Pero, antes de ello, vamos a distinguir qué tipos de predicciones vamos a querer hacer y entender la utilidad más común de cada una de ellas. Dependiendo del horizonte al que vayamos a predecir, tenemos:



- **Predicciones a corto plazo:** se utilizan para sucesos que requieren una respuesta muy temprana. Por ejemplo: para programar el transporte de mercancías, la producción y los procesos de los empleados de una compañía.
- **Predicciones a medio plazo:** se utilizan para anticipar los recursos que va a necesitar una compañía. Por ejemplo: predicciones de demanda para planificar la contratación de personal, la compra de materiales que se necesitarán o la adquisición de maquinaria.
- **Predicciones a largo plazo:** se utilizan para la planificación estratégica de la compañía.

Toda compañía que pretenda llegar a lo más alto en la práctica de data science necesita los siguientes elementos:

- 1 Ganar experiencia identificando problemas de predicción de series temporales.
- 2 Conocer un amplio número de modelos.
- 3 Seleccionar modelos apropiados para cada problema.
- 4 Evaluar y refinar estos modelos a lo largo del tiempo.

Puede ser muy tentador lanzarnos directamente a hacer predicciones a corto, medio o largo plazo. Una vez que tenemos los datos accesibles y conocemos un par de modelos de series temporales, queremos ponernos manos a la obra y empezar a ajustar modelos y lanzar predicciones. No obstante, es muy importante lo siguiente:

- **Elegir correctamente la serie temporal a modelizar:** en la siguiente sección veremos que surgen varias preguntas sobre qué serie temporal elegir y que, normalmente, no es una pregunta sencilla de responder.
- **Transformar la serie temporal para que el patrón sea lo más sencillo posible:** este paso es importante para simplificar lo máximo que podamos el patrón de nuestra serie y que las predicciones de nuestros modelos sean lo más precisas posibles.

¿Qué serie temporal modelizamos?

Para que podamos sacar el máximo partido a la utilidad que tienen las series temporales es muy importante elegir correctamente qué serie o series temporales queremos modelizar para obtener predicciones. Hay muchos detalles que tenemos que decidir al comienzo de un proyecto de series temporales.

Por ejemplo, en el caso de la venta de zapatillas, sencillamente podríamos estar interesados en predecir el número de zapatillas que vamos a vender para anticiparnos a grandes volúmenes de demanda y poder planificar mejor el stock disponible en las tiendas físicas. Pero:

- ¿Predecimos las ventas con granularidad semanal o mensual?
- ¿Predecimos la serie de ventas totales? ¿O mejor predecimos cada uno de los modelos que hay? ¿O agrupamos las ventas por zapatillas de hombres, mujeres y niños?
- ¿Con qué horizonte predecimos? ¿Predecimos a un mes vista, seis meses, cinco años...?

No es sencillo responder a cada una de estas preguntas. Debemos entender bien nuestro negocio para poder responder correctamente. Pero, para ganar algo de intuición, vamos a analizar cada una de ellas y ver en líneas generales qué debemos tener en cuenta:

- **Para determinar la granularidad más adecuada** suele ser útil visualizar la serie temporal con distintas granularidades y elegir aquella con la mayor granularidad, pero que tenga un patrón sencillo. Por ejemplo, una serie semanal tiene más detalle que una mensual, pero, si el patrón es mucho más complejo en la semanal que en la mensual, puede ser mejor elegir la serie mensual.

- **Para determinar si usamos la serie total o agrupamos por producto, por región o por género** es esencial entender quiénes van a ser las personas en la compañía que van a hacer uso de las predicciones que vayamos a hacer. Si las predicciones nos las ha pedido el CEO y sabemos que el negocio tiene líneas de productos para hombres, mujeres y niños será conveniente agrupar las ventas por estos tres grupos para que nuestros resultados estén alineados con el resto de la compañía.
- **Para determinar el horizonte** simplemente debemos entender la antelación que se necesita para poder emprender acciones según lo que nos informe la serie. Por ejemplo, si queremos estimar la demanda de zapatillas que va a tener cada tienda de nuestra compañía pero el inventario y la programación de envíos de mercancía se hace solo cada final de mes, será conveniente predecir con uno o dos meses de horizonte. De este modo, podremos anticiparnos a picos de demanda y tendremos tiempo suficiente para reorganizar el envío de mercancía a cada tienda.

Reajustes

Por último, antes de ponernos a modelizar la serie que hayamos decidido, es muy conveniente realizar un adecuado tratamiento de los datos. Lo primero es calcular correctamente la serie temporal, mientras que lo segundo que vamos a discutir en esta sección son los reajustes más comunes que aplicar a las series temporales para que nuestros resultados sean óptimos.

Los reajustes son las transformaciones que vamos a realizar sobre la serie original que tenemos antes de empezar a modelizar para que la modelización sea más sencilla y las predicciones más precisas.

Cuando veamos el primer ejemplo quedará más claro. Los reajustes que vamos a realizar a las series podemos clasificarlos en las siguientes cuatro categorías:

Reajustes por calendario



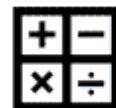
Reajustes por población



Reajustes por efectos económicos



Reajustes por transformaciones matemáticas



1

Reajustes por calendario

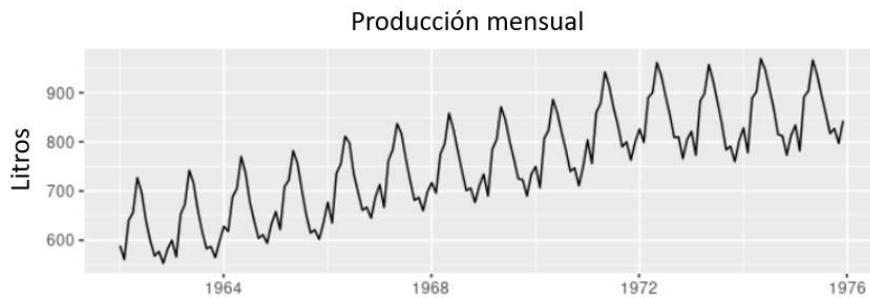
Hay efectos que tiene el propio calendario que pueden provocar pequeñas variaciones en los datos y es conveniente corregir con un reajuste para modelizar de forma más sencilla.

Por ejemplo, supongamos que estamos modelizando una serie de venta mensual de un producto cualquiera. Cada mes obtenemos el número de ventas totales que hemos tenido, pero ¿qué pasa en los meses que tienen 30 días frente a los que tienen 31?

Claramente, un día menos de tiendas abiertas afectará a las ventas totales provocando que los meses de 30 días tengan de forma natural menos ventas de media que los meses de 31 días (mejor ni hablamos del mes de febrero).

En este reajuste dividiríamos las ventas por el número de días que tiene el mes: en vez de modelizar las ventas totales de cada mes, modelizamos las ventas medias de cada día del mes.

Los reajustes por efectos del calendario parecen una finura, que queremos ser muy puristas o exactos, pero el efecto que tiene al simplificar la serie sorprende (y mucho). La siguiente serie se basa en los litros de leche producidos al mes por una granja:



Si dividimos cada mes por el número de días que tiene la serie, se convierte en la leche media diaria, y tiene la siguiente forma:



Como podemos comprobar, este simple reajuste nos proporciona una serie temporal con un patrón mucho más sencillo. Esto hará que los modelos de series temporales generen predicciones mucho más precisas y que los intervalos de confianza sean más estrechos.

2

Reajustes por población

En la mayoría de las situaciones, nuestras series temporales se ven afectadas por efectos poblacionales. Pensemos de nuevo en el ejemplo anterior de la producción de leche. La cantidad total de leche que produce una granja está estrechamente relacionada con el número de vacas que tenga. Si de un mes a otro se compran cien vacas más, la producción de leche crecerá proporcionalmente por cada vaca.

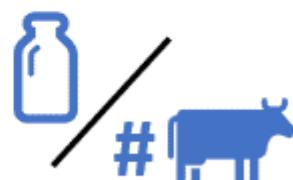
Si pensamos en Netflix y nos interesa la serie temporal del número de horas de video que consumen los usuarios en la plataforma, encontraremos que este KPI está estrechamente relacionado con el número de clientes que tiene nuestra compañía.

Al igual que con los reajustes por calendario, el tratamiento más adecuado cuando nuestras series se ven afectadas por efectos poblacionales es dividir la serie por el número de individuos que afecten a la serie. Para el ejemplo de las vacas, sería convertir los litros diarios que produce la granja en el número medio diario de litros que produce cada vaca:

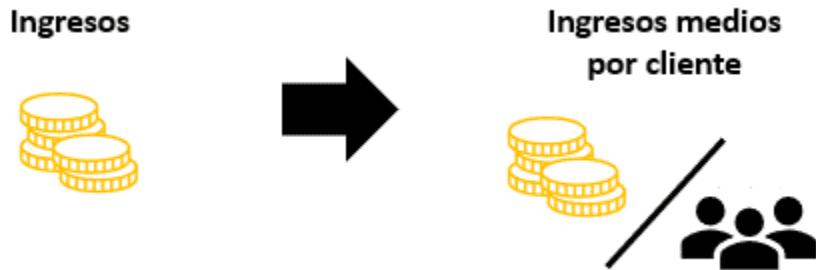
Litros de leche



**Litros de leche
medios por vaca**



Y en el ejemplo de Netflix, la serie de ingresos totales deberíamos convertirla en ingresos medios por cliente:



3

Reajustes por inflación

Hay series temporales que se ven afectadas por efectos económicos. El más común es la inflación. Un euro hoy no vale lo mismo que un euro mañana y lo mismo les ocurre a todas las monedas (el dólar, la libra...).

Un ejemplo de serie temporal que se ve afectada por el efecto de la inflación, con el que es probable que todos estemos familiarizados, es el precio de la vivienda. Si me compro una casa por 200.000€ en el 2021, su valor no es el mismo que el que tenía hace veinte años, debido al efecto de la inflación.

Por esta razón, las series financieras suelen reajustarse para tener en cuenta el efecto de la inflación. Al igual que los anteriores reajustes, la forma de tratar estas series es dividir por el IPC.

4

Reajustes matemáticos

Los anteriores reajustes se hacen para corregir efectos de la realidad sobre nuestra serie temporal. Los reajustes matemáticos serán aquellos reajustes que hacemos a las series aplicando transformaciones por el simple hecho de corregir la forma que tienen y hacerlas más sencillas.

La serie final que obtenemos no tiene por qué tener un sentido exacto, simplemente tendrá una forma más sencilla para poder modelizar después. Como tal, cualquier función matemática que apliquemos sobre nuestra serie es un reajuste matemático. Los más comunes son las siguientes:

- Logaritmo:

$$y = \log(x)$$

- Exponencial:

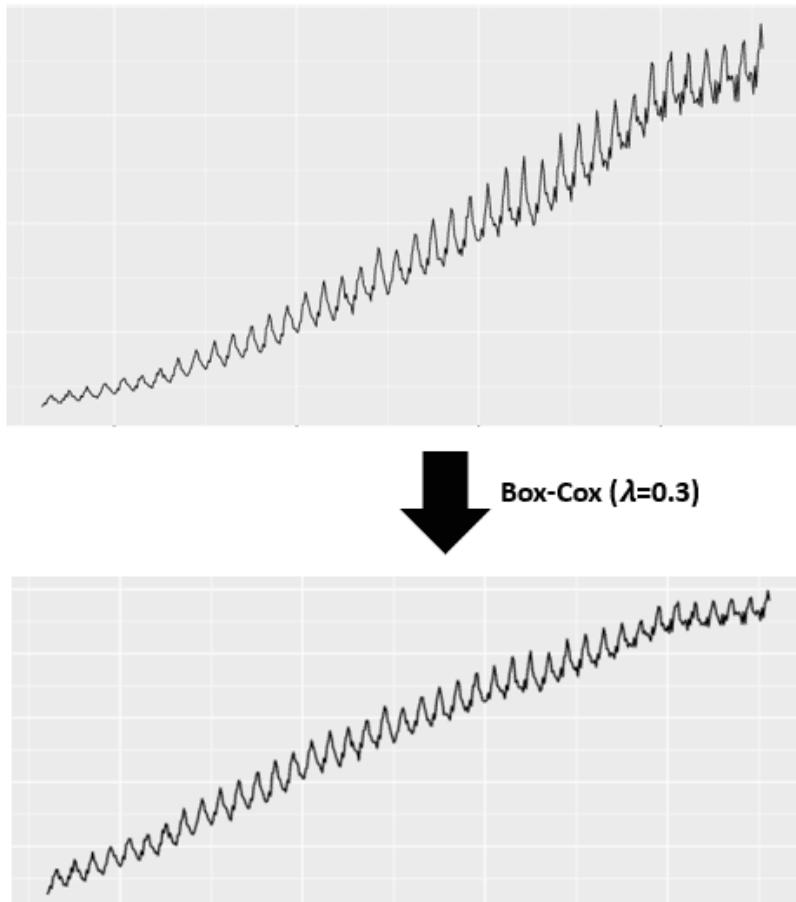
$$y = e^x$$

- Box-Cox:

$$y = \begin{cases} \log(x) & \text{si } \lambda = 0 \\ \frac{(x^\lambda - 1)}{\lambda} & \text{si } \lambda \neq 0 \end{cases}$$

Pensemos en estas transformaciones como fórmulas matemáticas con propiedades para simplificar el patrón que tienen nuestras series temporales.

Con estas fórmulas podemos transformar series como en el siguiente ejemplo:



Antes de aplicar modelos sobre nuestras series temporales es muy recomendable sentarse y pensar qué reajustes de los cuatro que hemos visto puede ser adecuado aplicar para simplificar el patrón de la serie.

Modelos de series temporales

X Edix Educación

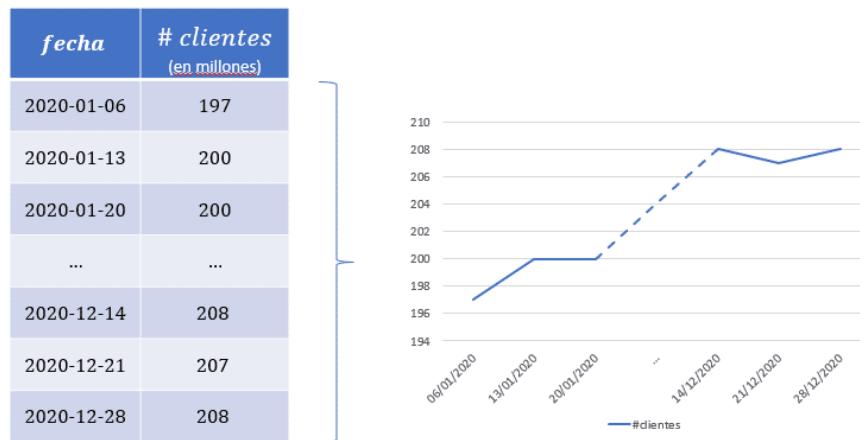
Llegó el momento. A partir de esta sección, la asignatura pasa de centrarse en los datos a los modelos que podemos aplicar sobre los mismos. El objetivo lo tenemos claro: **predecir el valor que va a tomar una serie temporal cualquiera**, y para ello contamos con una caja de herramientas repleta de modelos que vamos a conocer de uno en uno, tal y como hemos hecho con los modelos de segmentación.

En este fastbook vamos a cubrir los primeros modelos de series temporales para ir ganando intuición sobre cómo trabajar con este tipo de modelos. Ya después, en el fastbook 09, estudiaremos modelos más complejos, pero empecemos por la clásica regresión lineal.

Regresión lineal

El modelo de regresión lineal es probablemente el modelo más conocido de toda la literatura de la ciencia de datos. En este curso ya hemos visto la regresión lineal como protagonista de otras asignaturas y, en esta asignatura, también vamos a ver que es posible utilizar la regresión lineal para modelizar series temporales.

Supongamos que somos Netflix y que tenemos interés en modelizar el número de clientes que hay suscritos actualmente a nuestro servicio cada semana del 2020. La información tendría la siguiente forma en una tabla:



La tengamos en una tabla o la pintemos en un gráfico, la serie temporal del número de clientes de Netflix de 2020 será siempre el mismo concepto. La notación más empleada para referirnos a una serie temporal cualquiera es la siguiente: Y_t , donde la letra Y representa en nuestro caso el número de clientes de Netflix y t representa el índice de la semana en la que tomamos dicha medición.

Es decir, Y_0 es el número de clientes en el instante 0 (primera semana del histórico, 197 millones), Y_1 es el número de clientes en el instante 1 (segunda semana del año, 200 millones), y así sucesivamente hasta Y_{51} , que es el número de clientes en el instante 51 (última semana del año 2020, 208 millones).

fecha	# clientes (en millones)	
2020-01-06	197	$\leftarrow Y_0$
2020-01-13	200	$\leftarrow Y_2$
2020-01-20	200	$\leftarrow Y_3$
...
2020-12-14	208	$\leftarrow Y_{49}$
2020-12-21	207	$\leftarrow Y_{50}$
2020-12-28	208	$\leftarrow Y_{51}$

Y_t

La situación de la que partimos en todo modelo es que no solo conocemos la serie temporal que nos interesa Y_t sino que tenemos más información, otras variables que también han sido medidas a lo largo del tiempo y nos pueden ayudar a explicar los valores que toma nuestra serie Y_t . Por ejemplo, Netflix puede conocer el número de followers que Twitter cada semana ($X_{1,t}$), o la inversión en publicidad que realiza cada semana ($X_{2,t}$).

Cada una de estas variables puede explicar una parte de las variaciones que toma la serie Y_t . Si el número de followers crece ($\uparrow X_{1,t}$) esto debería afectar de forma positiva al número de suscriptores de Netflix (Y_t). De igual manera, en las semanas que Netflix haga publicidad ($X_{2,t}$ tome valores positivos) deberíamos esperar que el número de suscriptores también crezca ($\uparrow Y_t$).

Todas las variables que disponemos y que pueden explicar la serie Y_t las vamos a ir denotando como $X_{1,t}, X_{2,t}, \dots, X_{n,t}$. Lo que nos falta por formular del modelo de regresión simplemente es cómo podemos relacionar exactamente qué valores toma Y_t con los valores que toma cada una de las X . Y como el nombre indica, la relación que suponemos es lineal:

$$Y_t \sim \beta_1 \cdot X_{1,t} + \beta_2 \cdot X_{2,t} + \dots + \beta_n \cdot X_{n,t}$$

Lo que tratamos de hacer en un modelo de regresión es encontrar los mejores $\beta_1, \beta_2, \dots, \beta_n$ para que, al combinar las variables $X_{1,t}, X_{2,t}, \dots, X_{n,t}$, tengamos los valores más parecidos a Y_t . Como no siempre podemos explicar el 100% de la variable Y_t a partir de las variables que conocemos $X_{1,t}, X_{2,t}, \dots, X_{n,t}$ solemos meter un término de error ε_t en la fórmula para que la relación sea exacta:

$$Y_t = \beta_1 \cdot X_{1,t} + \beta_2 \cdot X_{2,t} + \dots + \beta_n \cdot X_{n,t} + \varepsilon_t$$

Toda esta nomenclatura no debe asustarnos, en el fondo solo estamos ponderando unas variables ($X_{i,t}$) con unos coeficientes (β_i) para obtener una serie lo más parecida a la serie temporal que queremos modelizar (Y_t).

Entonces, ¿cómo funciona el modelo de regresión? Pues muy sencillo: nosotros trataremos de elegir variables, $X_{1,t}, X_{2,t}, \dots, X_{n,t}$, que creemos que pueden tener una fuerte relación con la serie temporal que queremos modelizar y_t y el modelo de regresión determinará la mejor manera que podemos combinar las variables para obtener un resultado lo más parecido a y_t , es decir, determinará los mejores $\beta_1, \beta_2, \dots, \beta_n$.



En el fastbook 08 veremos ejemplos de regresión lineal y cómo utilizar correctamente este modelo para hacer predicciones sobre series temporales.

Descomposición

El segundo método que vamos a estudiar para la modelización de series temporales se llama *descomposición*. El objetivo que tiene esta técnica es el separar de una serie temporal los distintos patrones que la conforman.

Empleando de nuevo notación matemática, nosotros vamos a estar interesados en una serie temporal Y_t . Y lo que estamos asumiendo en el modelo de descomposición es que Y_t es el resultado de combinar distintos patrones, es decir:

$$Y_t = S_t + T_t + R_t$$

Donde:

- Y_t : serie temporal a modelizar
- S_t : patrón estacional
- T_t : componente de tendencia + ciclo
- R_t : el error (ruido aleatorio)

Las técnicas de descomposición tratan de obtener S_t , T_t y R_t a partir de Y_t .



Para que veamos más claro el resultado de los modelos de descomposición es el siguiente:



En el fastbook 08 veremos varios modelos de descomposición en la práctica y cómo podemos utilizarlos para realizar predicciones a futuro.

Suavizado exponencial

En las series temporales, el valor que toma la serie en un instante t suele ser parecido a valor que toma la serie en instantes próximos, por ejemplo, en $t-1$. Por esta razón, hay veces que, si queremos predecir qué valor va a tomar una serie mañana (Y_{t+1}), una buena predicción suele ser el valor que haya tenido hoy (Y_t).

A partir de esta idea (valores cercanos en el tiempo suelen tomar valores parecidos) es de donde surge el modelo de suavizado exponencial. Si usamos notación matemática, podemos modelizar una serie temporal Y_t usando el valor que ha tomado en el instante anterior con un constante de corrección que denotaremos α . Es decir, modelizamos Y_t como:

$$Y_t = \alpha \cdot Y_{t-1} + \varepsilon_t$$

Con este modelo, podríamos determinar el mejor valor de α y predecir el valor que la serie va a tomar en el instante siguiente como:

Pero ¿qué pasa con los valores que ha tomado la serie antes del instante anterior?, ¿qué hacemos con los valores que ha tomado nuestra serie en Y_{t-1} , Y_{t-2} , Y_{t-3} ... pues también podríamos tenerlos en cuenta para modelizar Y_t .

Podríamos proponer una nueva fórmula como la siguiente:

$$Y_t = \alpha_1 \cdot Y_{t-1} + \alpha_2 \cdot Y_{t-2} + \alpha_3 \cdot Y_{t-3} + \varepsilon_t$$

Con esta fórmula tendríamos que determinar los mejores α_1 , α_2 y α_3 .

Este modelo es perfectamente válido. El problema es que, si seguimos añadiendo términos Y_{t-n} , hay que determinar un parámetro α_n y se añade complejidad al modelo.

Por otra parte, que varios términos Y_{t-n} se tengan en cuenta está bien, pero el término Y_{t-1} debería tener un peso (α_1) superior al peso de Y_{t-2} (α_2) por estar más próximo al valor actual de la serie. A su vez, el peso de Y_{t-2} (α_2) debería ser superior al peso de Y_{t-3} (α_3), y así sucesivamente.

Dicho de forma más breve, los valores más recientes de la serie deben tener unos α 's más altos (más peso) que los valores más antiguos de la serie.

El modelo de suavizado exponencial incluye las dos propiedades que acabamos de discutir con la siguiente formulación:

$$Y_t = \alpha \cdot Y_{t-1} + \alpha \cdot (\alpha - 1) \cdot Y_{t-2} + \alpha \cdot (\alpha - 1)^2 \cdot Y_{t-3} + \dots + \alpha \cdot (\alpha - 1)^{n-1} \cdot Y_n + \varepsilon_t$$

Donde:

$$0 \leq \alpha \leq 1$$

De esta forma tenemos las siguientes propiedades en el modelo de suavizado exponencial:

1. Solo tenemos que determinar un único parámetro en modelo: α .
2. Podemos tener en cuenta todas las observaciones a pasado que nos interesen, n es arbitrario en el modelo, por lo que podemos incluir en él los valores: $Y_{t-1}, Y_{t-2}, Y_{t-3}, \dots, Y_{t-n}$.
3. Como el parámetro α está entre 0 y 1, el peso de cada Y_{t-k} es mayor que Y_{t-k-1} , ya que $\alpha \cdot (\alpha - 1)^{k-1} \leq \alpha \cdot (\alpha - 1)^{k-2}$ si y solo si $(\alpha - 1) \leq 1$, es decir $\alpha \leq 1$.

Con esta formulación, tenemos descrito el último de los modelos de series temporales que vamos a ver. En fastbook 08 veremos ejemplos de cómo determinar el valor de α y cómo lanzar predicciones de la serie Y_t utilizando software en lugar de la notación matemática que tanto dolor de cabeza nos levanta.

Métricas de calidad

X Edix Educación

Con la sección anterior ya hemos tenido un primer contacto teórico con los modelos de series temporales. Cada uno de los modelos que hemos visto (*regresión, descomposición y suavizado*), propone una fórmula distinta con la que modelizar una serie temporal Y_t sobre la que estemos interesados.

Si recapitulamos, cada uno de los modelos propone:

- El modelo de regresión propone utilizar varias series temporales que nos puedan informar sobre Y_t y usar la siguiente fórmula:

$$Y_t = \beta_1 \cdot X_{1,t} + \beta_2 \cdot X_{2,t} + \dots + \beta_n \cdot X_{n,t} + \varepsilon_t$$

Y trata de determinar los mejores: $\beta_1, \beta_2, \dots, \beta_n$.

- Los modelos de descomposición, como hemos visto, proponen la siguiente fórmula:

$$Y_t = S_t + T_t + R_t$$

Y estos modelos tratan de calcular las series S_t, T_t y R_t .

- Y por último, el modelo de suavizado exponencial propone la siguiente fórmula para modelizar Y_t :

$$Y_t = \alpha \cdot Y_{t-1} + \alpha \cdot (\alpha - 1) Y_{t-2} + \cdots + \alpha (\alpha - 1)^{n-1} Y_n + \varepsilon_t$$

Donde solo hay que elegir un n adecuado y el modelo tratará de determinar el mejor valor de α .

Entonces, cada uno de los modelos trata de modelizar la serie temporal Y_t con una fórmula, y con ella es posible realizar predicciones a futuro de la serie. Por ejemplo, el modelo de regresión puede utilizarse para predecir el valor que va a tomar la serie en el instante Y_{t+1} con la siguiente fórmula:

$$\hat{Y}_{t+1} = \hat{\beta}_1 \cdot X_{1,t+1} + \hat{\beta}_2 \cdot X_{2,t+1} + \dots + \hat{\beta}_n \cdot X_{n,t+1}$$

Para referirnos a las predicciones de Y_t en h instantes utilizamos la notación \hat{Y}_{t+h} y simplemente sustituimos $t + h$ en lugar de t en la fórmula de regresión. De igual manera, los mejores valores de $\beta_1, \beta_2, \dots, \beta_n$ que estime el modelo de regresión los denotamos como $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n$.

En esencia, podemos utilizar cada uno de los modelos de series temporales para lanzar predicciones a horizonte $t + h$, como veremos en el fastbook 08 con ejemplos.

La pregunta que respondemos en esta sección es ‘¿cómo medimos lo buenas que son estas predicciones?’. Pues de la misma forma que cuando estudiamos los modelos de segmentación: utilizando métricas que midan la calidad de las predicciones.

Existen muchas métricas que miden cómo de buenas son las predicciones de un modelo de series temporales, pero, sin duda, la más famosa, útil e intuitiva es la llamada MAPE, que estudiaremos a continuación como métrica para evaluar las predicciones de nuestros modelos.

MAPE

Cuando trabajamos con modelos de series temporales, uno de los aspectos más importantes es determinar el horizonte al que queremos predecir: ¿queremos predecir lo que va a ocurrir dentro de un instante o a más largo plazo?

Esta elección suele establecerse por sentido común y criterios de negocio.

Si nuestra serie de ventas es mensual y queremos anticiparnos con al menos un mes de antelación a lo que vaya a ocurrir, tendremos que lanzar predicciones a horizonte 1 o 2.

Independientemente del horizonte que se elija, la serie temporal que tenemos suele dividirse en dos partes:

- Una parte con los datos que van a ver nuestros modelos, los datos con los que se va a entrenar y a ajustar los parámetros.

- Y otra parte sobre la que se lanzarán predicciones y se comparará con los valores reales para medir la calidad predictiva del modelo.

Gráficamente, si cada punto es una unidad temporal, solemos dividir nuestra serie de la siguiente forma:

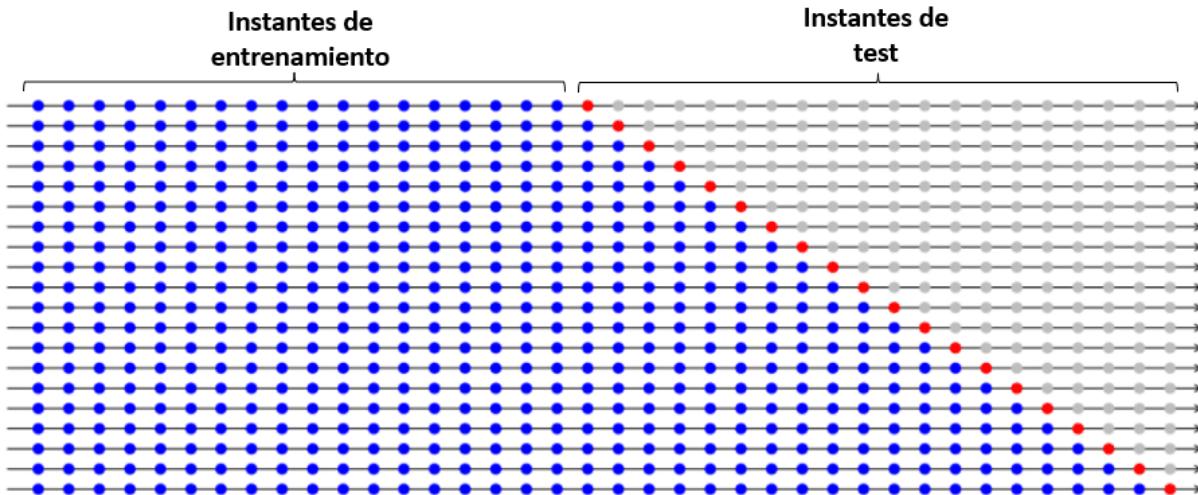


La idea entonces es muy sencilla:

- 1 Usaremos los datos de la serie Y_t en los instantes azules para entrenar los modelos de series temporales.
- 2 Usaremos los datos de la serie para ir lanzando predicciones al horizonte que consideremos.
- 3 Compararemos las predicciones que hagamos en los instantes de test (instantes rojos) con los valores reales de la serie.

El punto 1 es sencillo, simplemente tenemos que ‘recortar’ la serie Y_t y quedarnos con los valores que tome en los instantes azules. Estos valores se los pasamos al modelo y este determinará los parámetros que necesite dicho modelo.

Una vez hecho esto, iremos lanzando predicciones a horizonte h y lanzaremos predicciones sobre todos los instantes de test. Si, por ejemplo, queremos predecir a horizonte 1 ($h=1$), usaremos los siguientes instantes azules para lanzar predicciones a horizonte 1:



Con este procedimiento tendremos una predicción a horizonte 1 sobre cada instante de test. En el ejemplo anterior hemos reservado 20 instantes de test (instantes sobre los que tenemos el valor real de la serie) y sobre cada uno de ellos tenemos una predicción, es decir, tenemos 20 valores predichos.

Los valores reales de la serie los denotamos como:

$$Y_{t+1}, Y_{t+2}, \dots, Y_{t+19} \text{ y } Y_{t+20}$$

Y las predicciones que tengamos las denotamos como:

$$\hat{Y}_{t+1}, \hat{Y}_{t+2}, \dots, \hat{Y}_{t+19} \text{ y } \hat{Y}_{t+20}$$

Con esta notación podemos medir cómo de cerca han estado las predicciones de los valores reales usando el MAPE, que tiene la siguiente fórmula:

$$\begin{aligned}
 MAPE &= \frac{1}{20} \sum_{h=1}^{20} \left| \frac{Y_{t+h} - \hat{Y}_{t+h}}{Y_{t+h}} \right| = \\
 &= \left| \frac{Y_{t+1} - \hat{Y}_{t+1}}{Y_{t+1}} \right| + \left| \frac{Y_{t+2} - \hat{Y}_{t+2}}{Y_{t+2}} \right| + \dots + \left| \frac{Y_{t+20} - \hat{Y}_{t+20}}{Y_{t+20}} \right|.
 \end{aligned}$$

Es decir, para cada predicción (desde $t + 1$ hasta $t + 20$) calculamos la diferencia porcentual y en valor absoluto que hay entre el valor real y_{t+h} y la predicción que hace nuestro modelo \hat{Y}_{t+h} , esto es: $\left| \frac{Y_{t+h} - \hat{Y}_{t+h}}{Y_{t+h}} \right|$. Y, para tener una métrica global de todas las predicciones, simplemente calculamos la media de estos valores.

En resumen, cuando hayamos entrenado un modelo de series temporales podremos lanzar predicciones y obtener el MAPE, que será un valor que medirá el error porcentual medio que tienen las predicciones de nuestro modelo.

Si nuestro modelo tiene un MAPE con valor 0.02, quiere decir que las predicciones de nuestro modelo se desvían de media un 2% (hacia arriba o hacia abajo) de los valores reales que tiene la serie temporal que estemos modelizando.

Resumen y conclusiones

X Edix Educación

En este fastbook ya hemos entrado de lleno a estudiar los primeros modelos de series temporales. Concretamente hemos visto las fórmulas que proponen los modelos de **regresión lineal, descomposición y suavizado exponencial** para la modelización de una serie temporal Y_t cualquiera.

La notación matemática y las fórmulas que hemos visto no deben ni intimidarnos ni asustarnos. Simplemente debemos entender cada modelo como un algoritmo que recibirá información, determinará unos parámetros para tener una fórmula fija y que podrá emplearse para **lanzar predicciones de qué valores va a tomar la serie en un horizonte h que determinemos**.

Lo importante será entender **cómo podemos modelizar y lanzar predicciones de una serie temporal usando software (R)** y medir cómo de precisas son las predicciones de dichos modelos.

Cualquier aspecto o duda que se tenga relativa a cómo utilizar cada uno de estos modelos en la práctica se podrán clarificar en el siguiente fastbook por medio de ejemplos, que es, realmente, la mejor manera de aprender ciencia de datos.

¡Enhорabuena! Fastbook superado

edix

Creamos Digital Workers