Series temporales y minería de flujo de datos: Trabajo autónomo I - Series Temporales (Parte teoría)

Máster en ciencia de datos e ingeniería de computadores - UGR

10-4-2018

Mª Cristina Heredia Gómez

crstnheredia@correo.ugr.es

Índice

Preprocesamiento

Como los datasets no contenían muchos valores perdidos (5 el diario y 37 el mensual), y la temperatura normalmente cambia de forma gradual, se ha optado por calcular los valores perdidos mediante interpolación linear de los datos conocidos que lo rodean. Es decir, sean el punto (x_0, y_0) y el punto (x_1, y_1) conocidos, y (x, y) un punto entremedias de los dos anteriores, para $x \in [x_0, x_1]$ podemos calcular y como:

$$y = \frac{y_0 \cdot (x_1 - x) + y_1 \cdot (x - x_0)}{x_1 - x_0}$$

Análisis de tendencia y estacionalidad

Puesto que ninguna de las dos series presentaba tendencia, nos centramos en el análisis de la estacionalidad, para lo que principalmente he usado el ACF, PACF y el test ADF.

ACF

También conocido como correlograma, muestra los coeficientes de autocorrelación en un gráfico, es decir, la relación lineal que existe entre los lags de la serie temporal. Sea una serie temporal T, se definen los sucesivos coeficientes de correlación como:

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$$

donde y_t denota la serie en el tiempo t.

Las líneas azules horizontales en un gráfico ACF denotan los límites a partir de los cuales las correlaciones son significativamente lejanas a cero. Estos límites de calculan como $\pm 2/\sqrt{|T|}$ donde $|T|$ denota la longitud de la serie. Cuando en torno al 95 % de los lags que aparecen representados en el gráfico ACF quedan entre los límites, la serie es ruido blanco, es decir, no hay autocorrelaciones. Por el contrario si en torno al 5 % de los lags sobresalen de los límites, podemos afirmar que la serie no es ruido blanco.

Detección de tendencia y estacionalidad

Este gráfico es especialmente útil para detectar si la serie tiene componentes no estacionarias como la tendencia o la estacionalidad. Si una serie tiene tendencia, el gráfico ACF mostrará valores positivos que decrecen lentamente a 0, a medida que los lags aumentan. Esto se debe a que los lags más pequeños mostrarán correlaciones altas debido a que las observaciones están muy cercanas entre sí en el tiempo. Para detectar la estacionalidad, observaremos si en la gráfica hay patrones que se repitan con una frecuencia dada, generalmente dando lugar a una especie de ondas u ondulaciones.

PACF

Muestra las autocorrelaciones parciales, es decir, al igual que ACF mide la relación entre y_t e y_{t-k}, pero eliminando los efectos de los lags entre medias de ambos $1, 2, 3, 4, ..., k$. De esta forma podemos obtener si y_t e y_{t-k} están realmente correladas de forma directa, es decir, si realmente existe alguna información en y_{t-k} que sirva para predecir y_t y no de forma indirecta, es decir, que estén correladas porque ambas están conectadas con otra intermedia.
Sea α_k el coeficiente de autocorrelación k-ésimo, entonces α_k se puede calcular como la estimación de ϕ_k en $AR(k)$, es decir, cada correlación parcial se calcula como el último coeficiente de un modelo autoregresivo.

Detección de patrones ARIMA

El PACF junto con el ACF puede ser usado para detectar si los datos se pueden modelar con un ARIMA(p,d,0) o un ARIMA(0,d,q). En concreto:

- ARIMA(p,d,0): ACF presenta aspecto exponencial decreciente o sinusoidal, y en el PACF hay únicamente un pico que sobresale en el lag p
- ARIMA(0,d,q): PACF presenta aspecto exponencial decreciente o sinusoidal y en el ACF hay únicamente un pico que sobresale en el lag q

Puesto que en los datos de la predicción diaria y en los de la predicción mensual se daban patrones similares a los de un ARIMA(p,d,0) en el ACF y PACF, se ajustaron diversos modelos ARIMA cambiando los parámetros p y d en cada caso.

Test ADF

El test de Dickey-Fuller aumentado, conocido como test ADF, trata de determinar si existen o no raíces unitarias en una serie temporal, es decir, si existen componentes no estacionarias que puedan cambiar con el tiempo. La H_0 es que existe una raíz unitaria en la serie. Es un modelo AR(1):

$$y_t = \rho y_{t-1} + u_t$$

donde y_t es la variable a predecir, t es el tiempo, u_t el error y ρ un coeficiente. El modelo de regresión se describe como:

$$\nabla y_t = (\rho - 1) y_{t-1} + u_t$$

Donde ∇ es el operador de primera diferencia. Si ρ toma valor 1, entonces existe una raíz unitaria, por lo que el test nos devolverá un p-valor > 0.05 que no nos permitirá rechazar la hipótesis nula H_0. Si por el contrario obtenemos un p-valor < 0.05, entonces podemos rechazar la H_0, concluyendo que la serie es estacionaria.

A lo largo de esta práctica se ha aplicado varias veces este test antes y después de diferenciar las series, para comprobar su estacionariedad.

Estacionariedad

En ambas series (diaria y mensual) se ha asumido una descomposición aditiva:

$$y_t = T_t + S_t + R_t$$

donde T_t representa la tendencia, S_t representa la estacionalidad y R_t es ruido, que puede ser ruido blanco o no.
En ninguna de ellas se observa tendencia, pero sí se ha detectado estacionalidad, si bien en la mensual bastó con quitar los 12 primeros valores que se repiten a la serie de entrenamiento para eliminar la estacionalidad, con la serie para predicciones diarias esto no fue suficiente para eliminar la estacionalidad y hubo que diferenciar una vez para eliminarla.

Diferenciación: Random walk model

Para diferenciar se ha utilizado un modelo sencillo que se conoce como random walk model, que calcula la diferencia de dos series como el cambio entre dos observaciones consecutivas:

$$\nabla y_t = y_t - y_{t-1}.$$

Donde la serie resultante tendrá $|T| - 1$ valores.
Finalmente para comprobar si las series eran estacionarias antes de modelar, se usaron las técnicas de ACF, PACF y test de Dickey-Fuller descritos en la sección anterior.

Modelado de series temporales

En esta práctica y para cada una de las dos predicciones (diaria y mensual) se han ajustado distintos modelos ARIMA(p,d,0), ya que eran los patrones que se observaban en los ACF, PACF respectivos, tal y como se menciona en la subsección *Detección de patrones ARIMA*, pero vamos a introducir brevemente los modelos autoregresivos y los de medias móviles, dado que ARIMA es una combinación de diferenciación con estos dos.

Modelos autoregresivos

Predicen la variable de interés a partir de combinaciones lineales de valores anteriores de esa misma variable. Un modelo autoregresivo se define como:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

donde ε_t es ruido blanco y p denota el orden del modelo autoregresivo.

Medias móviles

Similar a los modelos autoregresivos, pero en lugar de usar valores anteriores de la variable de interés para realizar las predicciones, medias móviles usa los errores de predicción anteriores:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

donde ε_t es ruido blanco y q denota el orden del modelo, tal que cada valor de y_t puede considerarse una media móvil ponderada de los últimos errores de predicción.
Tanto en los modelos autoregresivos como en medias móviles, una varianza en ε_t supondrá un cambio en la escala de la serie, pero no modificará su patrón.

ARIMA

Los modelos ARIMA(p,d,q) no estacionarios combinan diferenciación con modelos autoregresivos y de medias móviles. El modelo se describe como:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

donde y'_t es la serie diferenciada más de una vez, los predictores de la parte derecha denotan lags de y_t y errores, p denota el orden de la parte autoregresiva, d denota el grado de diferenciación y q denota el orden de la parte de medias móviles. La ecuación anterior puede escribirse como:

$$\underset{\substack{\uparrow \\ AR(\mathrm{p})}}{(1 - \phi_1 B - \cdots - \phi_p B^p)} \quad \underset{\substack{\uparrow \\ \mathrm{d}differences}}{(1 - B)^d y_t} \quad = \quad c + \underset{\substack{\uparrow \\ MA(\mathrm{q})}}{(1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t} \tag{1}$$

Donde se puede observar mejor la descomposición en parte autoregresiva, diferencial y medias móviles.

Validación del modelo

Tras ajustar los modelos ARIMA, se aplicaron distintos test sobre los residuos del mismo para validar el modelo. En concreto se aplicaron los test de Box-Pierce para comprobar la aleatoriedad de los residuos, y los test de Jarque Bera y Shapiro, para comprobar la normalidad de los residuos.

Box-Pierce

Trata de medir la aleatoriedad de los residuos del modelo comprobando si las autocorrelaciones son diferentes de cero. H_0 se formula como: los datos se distribuyen de forma independiente. El test viene dado por:

$$Q_{BP} = n \sum_{k=1}^{h} \hat{\rho}_k^2$$

donde n es el tamaño de los datos, $\hat{\rho}_k^2$ es la autocorrelación en el lag k y h es el número de lags que se están testeando. Si se obtiene un p-valor > 0.05 indica que no se puede rechazar H_0 y que por tanto, los residuos son aleatorios.

Jarque Bera

Comprueba la normalidad de los residuos comprobando si los datos presentan la asimetría y curtosis típicas de una distribución normal. El test estadístico se define como:

$$JB = \frac{n-k+1}{6}(S^2 + \frac{1}{4}(C-3)^2)$$

donde n es el número de datos, S es la asimetría de la muestra, C la curtosis y k el número de regresores. Si el p-valor > 0.05 podemos asumir que los errores son normales.

Shapiro Wilk

Comprueba si los datos, en este caso, los residuos, siguien una distribución normal. H_0 es que la muestra provenga de una distribución normal. El test estadístico se define como:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

donde $x_{(i)}$ es el dato en la posición i-ésima en los datos, $\bar{x}$ es la media de los datos, y a_i son variables que se calculan a partir de valores medios del estadístico ordenado y de la matriz de covarianza.
De nuevo, si obtenemos un p-valor < 0.05 rechazamos la H_0, pero si el p-valor es > 0.05 no podemos rechazar H_0 y por tanto podemos decir que los datos son normales.