SEMANA 3 · HISTORIA 24

Series de tiempo (Parte 2)

Grabación

Presentación

El equipo docente subió contenido extra de la meeting ✨

Contenido extra

Grabación

> *"Es difícil hacer predicciones, especialmente sobre el futuro." –Niels Bohr o Yogi Berra.*

¡Gracias Sabina Bercovich por la creación de los contenidos de esta bitácora!

En esta bitácora estaremos profundizando en el procesamiento de las series de tiempo: vamos a estar analizando cómo identificar de manera más formal sus componentes, para poder encontrar distintas formas de predecir su posible comportamiento futuro.

Predicción en series de tiempo

Una de las funciones más buscadas a la hora de trabajar con series de tiempo es la posibilidad de *predecir el futuro*. Los sets de datos temporales nos permiten analizar los distintos comportamientos que tienen nuestros datos, y cuanto mejor sea nuestra comprensión del pasado, más probabilidad tenemos de predecir correctamente el futuro.

Por lo pronto suena todo muy lógico y sencillo, pero estamos asumiendo un supuesto muy importante: que **el futuro se va a comportar como el pasado**. Es importante que sea consciente de que estamos presuponiendo esto a la hora de trabajar con predicciones en series de tiempo. El futuro es incierto y podemos aproximarnos a lo que puede llegar a suceder, ¡pero siempre vamos a tener un componente de sorpresa!

Una forma que tenemos de entender cuán parecida es la serie respecto a sí misma (en distintos momentos del tiempo) es estudiando la autocorrelación. Este concepto es bastante parecido al de correlación que vimos a lo largo de la cursada. El objetivo de este indicador es comprender cómo es la relación entre una misma variable, medida en distintos momentos del tiempo.

Este indicador generalmente se estudia entre distintas ventanas de tiempo, para poder entender cada cuánto es que nuestro proceso "se repite". En el fácil de entender, por ejemplo, con la temperatura promedio, donde nuestros datos tienen aproximadamente un ciclo de 12 meses.

Partiendo de este concepto de la autocorrelación de los valores, o en otras palabras, que el futuro se parece al pasado, vamos a contarte sobre algunos modelos de predicción. Aprenderás desde el enfoque de la estadística clásica, para comprender un poco mejor el marco conceptual. En el fondo, verás que estos modelos se reducen a un problema de regresión, que ya estuvimos trabajando desde el aprendizaje supervisado.

Finalmente vamos a presentar una librería especializada para trabajar con series de tiempo llamada **Prophet**.

Modelos *clásicos*

Aquí te contaremos sobre tres modelos (o procesos) que fueron durante mucho tiempo el "estado del arte" en series de tiempo y, en muchas áreas siguen siéndolo. Por lo que es importante conocer sus principales características. ¡No te preocupes si no logras entender todos los detalles! Aprenderás sobre:

1. Procesos Autoregresivos (AR)
2. Procesos de Media Móvil (MA)
3. Procesos ARMA y ARIMA, que son una combinación de los dos primeros

1. **Procesos Autorregresivos AR(p)**

Son procesos en los que el valor de Y observado en el tiempo t depende de sus p valores anteriores (t-1, t-2, ..., t-p).

Es decir que podría expresarse:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

Podríamos entender estos procesos como el grado en que el valor de hoy depende del mismo valor, pero del pasado. El valor de alpha (α) representa el componente independiente de nuestro valor en t, mientras que los valores de beta (β) corresponden a cuánto peso le otorgamos a nuestros valores pasados para componer el valor al momento t. Como siempre, incluimos a e como el término de error de medición de nuestro proceso. Dentro de las componentes mencionadas en la bitácora anterior, este término se corresponde con el ruido blanco.

2. **Procesos de media móvil MA(q)**

Los procesos de media móvil se conforman como una combinación de variables aleatorias no correlacionadas (ruido blanco), que son ni más ni menos que los términos de error:

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

Así como en los procesos autoregresivos miramos los valores del pasado y los ponderamos, en los procesos de media móvil buscamos incluir en nuestro valor en el momento *t* los distintos valores de ruido blanco que existieron a lo largo del tiempo.

3. **Procesos ARMA y ARIMA**

Tenemos entonces dos modelos: AR, que ve los valores del pasado y los pondera, y MA, que incluye los procesos de ruido blanco.

La suma de estos dos modelos se expresa como ARMA(p, q). Está compuesto por una combinación lineal de los valores pasados junto con los términos de error, ambos ponderados, para explicar el valor de nuestra variable en el momento *t*.

En la bitácora anterior mencionamos el concepto de "procesos no estacionarios" como aquellos que tienen tendencia no nula. Para trabajar mejor con estos modelos, es importante generar alguna modificación a nuestra serie para lograr que sea estacionaria (podemos pensar en esto como una normalización).

Cuando el proceso no es estacionario, diferenciamos la serie. Es decir, creamos una nueva variable:

Zt = Yt - Yt-1

El objetivo de esta diferenciación es disminuir el efecto de la tendencia en la serie. Al trabajar con esta nueva variable Zt, y reemplazando nuestros valores de Yt, Yt-1, etc., por sus respectivos procesos ARMA, obtenemos nuestro modelo ARIMA (p, q, d).

Este proceso es el más comúnmente utilizado en la estadística clásica, así que vamos a detenernos un poco a observar cómo se compone:

- Por un lado tenemos nuestro proceso AR(p) cómo una regresión lineal de los *p* valores anteriores a *t*.
- Por otro lado, tenemos nuestro proceso MA(q) como una regresión lineal de los *q* términos de error anteriores a *t*
- Por último, tenemos la resta de estos dos mismos componentes, pero calculados para Yt -d, con el objetivo de diferenciar nuestra serie y volverla estacionaria.

Si esto te parece complejo, ¡no te asustes! Lo importante de destacar que este modelo **no es más que una combinación lineal de los valores anteriores y el proceso de ruido blanco.**

Para pensar: ¿Se te ocurre algún modelo de machine learning que nos ayude a calcular combinaciones lineales?

Modelos de Machine Learning

Tal como aprendiste con los modelos de estadística anteriores, podemos entender a las series -simplificando mucho- como combinaciones lineales de los sucesos pasados y ruido blanco. Estos modelos podrían ser fácilmente programables y realizables dentro del mundo de Machine Learning.

Un detalle muy importante a la hora de trabajar con modelos es prestarle especial atención a qué parte de nuestros datos usamos como train y cuál como test. En el caso de las series de tiempo y el problema de predicción a futuro, nuestra porción de testeo va a ser siempre los datos más actuales.

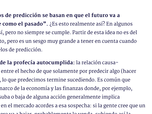

Fig. 1: Diferencias entre sliding window y forward chaining cross validation

Marco de regresión de aprendizaje automático de series temporales. Fuente: Towards Data Science

La elección del método de validación pasa entonces a tener otro componente a tener en cuenta, que es básicamente el tiempo. Existen distintas estrategias que se pueden utilizar para validar nuestros modelos, siempre teniendo en cuenta que las particiones en este caso van a ser versantas *de tiempo* consecutivas, y ya no valores aleatorios.

Modelos de regresión

Los modelos de regresión que vimos a lo largo de la cursada sirven perfectamente a la hora de trabajar con series de tiempo. Por ejemplo, podríamos utilizar los modelos de regresión lineal para aproximar las tendencias.

De todas maneras, estos modelos no están pensados concretamente para comprender el vínculo temporal que existe entre los datos, por ejemplo los patrones de estacionalidad. Es justamente por esto que te mostraremos una librería determinada, pensada específicamente para trabajar con series de tiempo: Prophet.

Para pensar: ¿Podríamos utilizar datos auxiliares para ayudarnos a entender nuestras series de tiempo? Por ejemplo, ¿se podría estudiar la variación del dólar para comprender el aumento de precios? Al trabajar con series de tiempo podemos aprovechar otros datos, notas periodísticas, información de expertos, o incluso la historia misma, para entender un poco mejor el contexto temporal que existe por fuera de nuestros datos específicos. En el capítulo 1 del Python Data Science Handbook puedes encontrar un ejemplo de este estilo, el cual recomendamos fuertemente que leas.

Prophet

Esta librería desarrollada por Facebook, busca estudiar los componentes de nuestras series temporales. Incluye dentro de sus algoritmos modelos de predicción, de detección de anomalías, análisis de estacionalidad para distintas granularidades temporales, estudios de tendencia, entre muchas otras opciones.

Tiene una funcionalidad muy interesante para graficar las distintas partes de nuestra serie de tiempo, incluyendo una predicción con intervalos de confianza:

Podemos ver en este gráfico la tendencia, junto con su predicción, y la estacionalidad anual y semanal.

La documentación de la librería está muy bien explicada y resulta muy clara su aplicación, y la puedes encontrar acá. La instalación requiere que tengamos ciertas librerías previamente instaladas, por lo que recomendamos mucho armar un ambiente propio a la hora de trabajar con Prophet.

¡Atención!

Hay algunos puntos importantes a remarcar cuando estamos trabajando con series de tiempo.

1. **"Los modelos de predicción se basan en que el futuro va a comportarse como el pasado".** ¿Es esto realmente así? En algunos casos tal vez sí, pero no siempre se cumple. Partir de esta idea no es del todo incorrecto, pero es un sesgo muy grande a tener en cuenta cuando usemos modelos de predicción.
2. **La paradoja de la profecía autocumplida:** la relación causa-consecuencia entre el hecho de que solamente por predecir algo (hacer una profecía), lo que predecimos termine sucediendo. Es común que suceda en el marco de la economía y las finanzas donde, por ejemplo, predecir una suba o baja de alguna acción generalmente implica movimientos en el mercado acordes a esa suspecha: si la gente cree que un activo financiero va a bajar, probablemente lo venda, subiendo así la oferta y generando que el precio baje efectivamente.

Empieza a trabajar en el notebook

Para el encuentro, mira las secciones 1 y 2.1. El resto del notebook lo trabajaremos en clase.

DS_Bitácora_42_Series_de_Tiempo

¡Prepárate para el próximo encuentro!

Profundiza
Te invitamos a conocer más sobre el tema de esta bitácora.

Herramientas
Programas necesarios para facilitar tu experiencia.

Proyecto
Consejos para comenzar a trabajar sobre el proyecto del bloque.

Contenido extra de la meeting

Grabación
Código de acceso para acceder: Acamica22*