
Universidad
Tecmilenio®

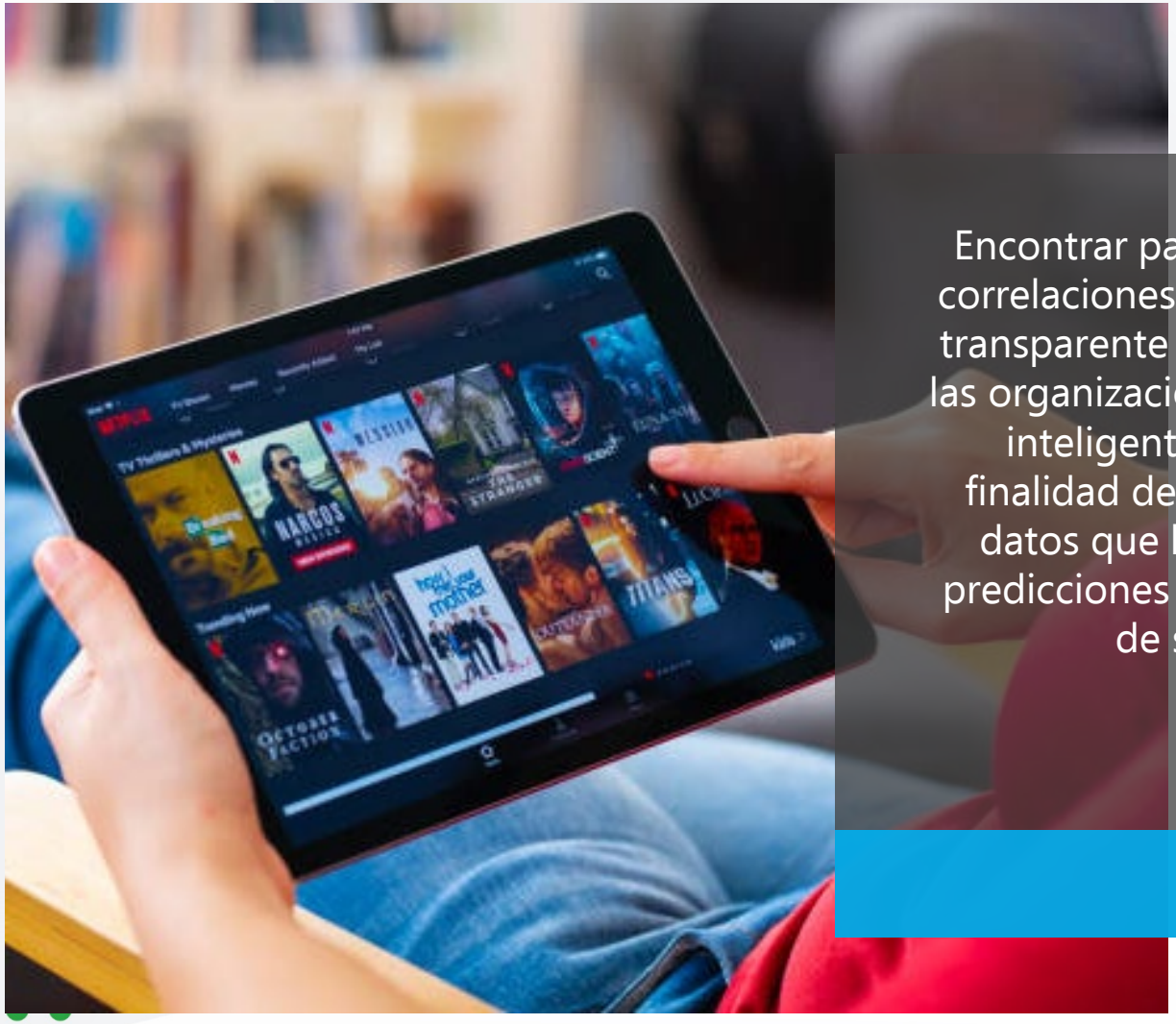
Universidad Tecmilenio

Ingeniería de datos masivos

Correlaciones en Big Data

Semana 8

Introducción



Encontrar patrones de consumo y correlaciones otorga una impresión transparente del mercado para que las organizaciones tomen decisiones inteligentes y sólidas, con la finalidad de seleccionar aquellos datos que les ayuden a realizar predicciones sobre las preferencias de sus usuarios.

Explicación

Crear modelos de análisis de datos implica entender las variables registradas en los datos que logran describir la situación de análisis de la vida real y cómo se relacionan entre sí. La correlación sirve para medir esa relación y es el análisis básico más importante que se hace para identificar patrones, seleccionar variables para los modelos, hacer predicciones e identificar causas (Bhan, 2018).

Los modelos con frecuencia involucran una gran cantidad de variables que están correlacionadas. Por lo tanto, aprender a medir esa correlación y el grado de dependencia lineal entre las variables ayuda a entender los datos, seleccionar variables e investigar si esa correlación es una relación de causalidad o no.

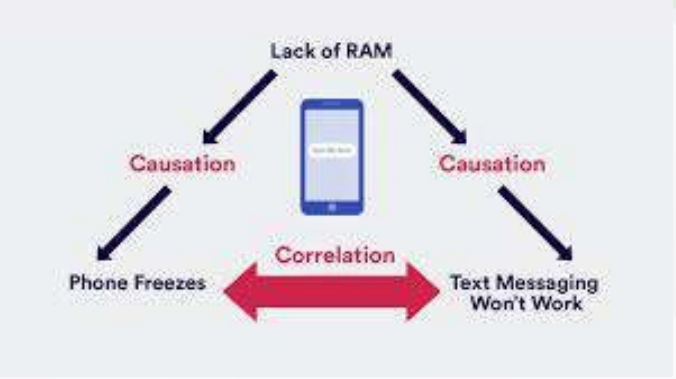
Para saber más sobre cómo Netflix emplea la ciencia de datos, revisa el siguiente video:

Stanford University School of Engineering. (2016, 9 de mayo). *Caitlin Smallwood: How data science helps Netflix please customers* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=xOriwLUcTAg

Explicación

Causalidad y **correlación** no son lo mismo, puesto que la relación de causalidad de las variables se da por la naturaleza de la situación del análisis de la vida real, conocerla permite entender la situación, hacer predicciones, explicar el pasado e intervenir para cambiar los resultados. Por lo tanto, comprender si las inferencias que se hacen con los datos son correctas o incorrectas es particularmente importante para la ciencia de datos.

Ejemplo de causalidad contra correlación, donde la falta de memoria RAM de un teléfono inteligente ocasiona que este se congele y no se puedan enviar mensajes de texto, existiendo una correlación entre ambos sucesos.



Fuente: Valdellon, L. (2019). *Correlation vs Causation: Definition, Differences, and Examples*. Recuperado de https://clevertap.com/blog/correlation-vs- causation/

Medir la causalidad implica hacer experimentación y análisis para descubrir relaciones de causa-efecto que con frecuencia no están en su totalidad registradas en los datos.

Explicación

La **correlación** es la medida normalizada de asociación de la covarianza entre dos variables cuantitativas, es decir, para determinar si dos variables están correlacionadas se identifica si existe covarianza. Asimismo, la correlación mide la fuerza y la dirección de la relación lineal entre dos variables.

- ✓ El signo **(+ / -)** determina la dirección de la relación.
- ✓ El valor **[-1, 1]** determina la fuerza de la relación.
- ✓ Un valor de correlación de **+ 0.5** significa que una variable se mueve en la misma dirección que la otra la mitad de las veces.
- ✓ Un valor de correlación de **0** significa que las variables no están correlacionadas.

El valor de la correlación se puede clasificar de la siguiente manera:

1. **Correlación no significativa < 0.1**
2. **0.1 < Correlación baja <= 0.3**
3. **0.3 < Correlación media <= 0.5**
4. **Correlación alta > 0.5**

De acuerdo con su signo se puede clasificar de la siguiente manera:

Correlación positiva

Si aumenta una variable, aumenta la otra.

Correlación negativa

Si aumenta una variable, disminuye la otra.

Correlación cero

Si cambia una variable, no hay garantía con la otra.

Explicación

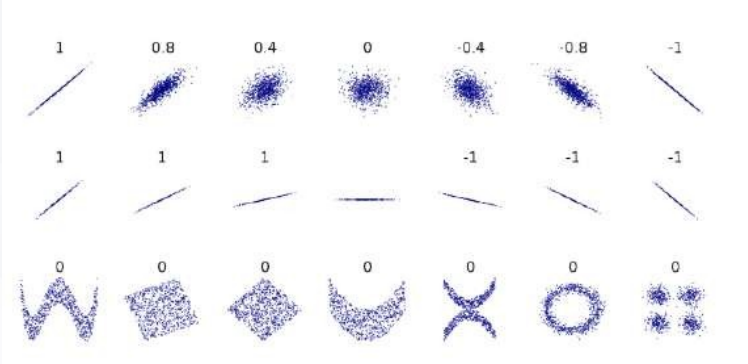
La correlación se puede utilizar para identificar:

- Las variables que son más predictivas o determinantes para considerarlas al hacer un modelo de predicción.
- Las variables para hacer análisis de causalidad.
- Multicolinealidad, que es la correlación en un conjunto de variables que afecta los valores de coeficientes de una regresión, generando modelos menos precisos.

Para conocer ejemplos de correlaciones de datos en Python, mira el siguiente video:

Dirección de Producción de Contenidos. (2020, 3 de agosto). *Tema 9. Correlaciones de datos en Python* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=BMHEi6CGGmk&feature=youtu.be

Visualmente las correlaciones se pueden ver con **gráficos de dispersión** en los que se aprecia la desviación de los puntos con respecto al modelo lineal.
La siguiente imagen muestra los gráficos de dispersión de parejas de variables aleatorias (etiquetadas con las correlaciones de Pearson asociadas):



Fuente: Kom, Y. (2018). *Pearson's Correlation, Linear Regression, And Why 'Beta' Grossly Underestimates Portfolio Sensitivity To Market Returns*. Recuperado de https://towardsdatascience.com/the-black-swans-in-your-market-neutral-portfolios-part-i-7521683a7317

Cierre

La importancia de la correlación radica en que es uno de los conceptos básicos dentro del análisis de datos, permitiendo al profesional precisar tendencias, realizar predicciones y descubrir las causas de ciertos fenómenos, ya que no se debe perder de vista que el valor real de los datos es cuando estos son comprensibles, por ejemplo, mediante las herramientas de análisis de datos se están encontrando correlaciones de gran trascendencia (BSA, 2015).

Bibliografía

- Bhan, S. (2018). *Correlation Coefficient In Data Science.* Recuperado de https://stepupanalytics.com/correlation- coefficient-in-data-science/
- Dirección de Producción de Contenidos. (2020, 3 de agosto). *Tema 9. Correlaciones de datos en Python* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=BM HEi6CGGmk&feature=youtu.be
- Kom, Y. (2018). *Pearson's Correlation, Linear Regression, And Why 'Beta' Grossly Underestimates Portfolio Sensitivity To Market Returns.* Recuperado de https://towardsdatascience.com/the- black-swans-in-your-market-neutral-portfolios-part-i-7521683a7317
- BSA. (2015*). ¿Por qué son importantes los datos?* Recuperado de https://data.bsa.org/wp-content/uploads/2015/10/BSADataStudy
- _es.pdf
- Stanford University School of Engineering. (2016, 9 de mayo). *Caitlin Smallwood: How data science helps Netflix please customers* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=xOri wLUcTAg
- Valdellon, L. (2019). *Correlation vs Causation: Definition, Differences, and Examples.* Recuperado de https://clevertap.com/blog/correlation- vs- causation/

Universidad Tecmilenio

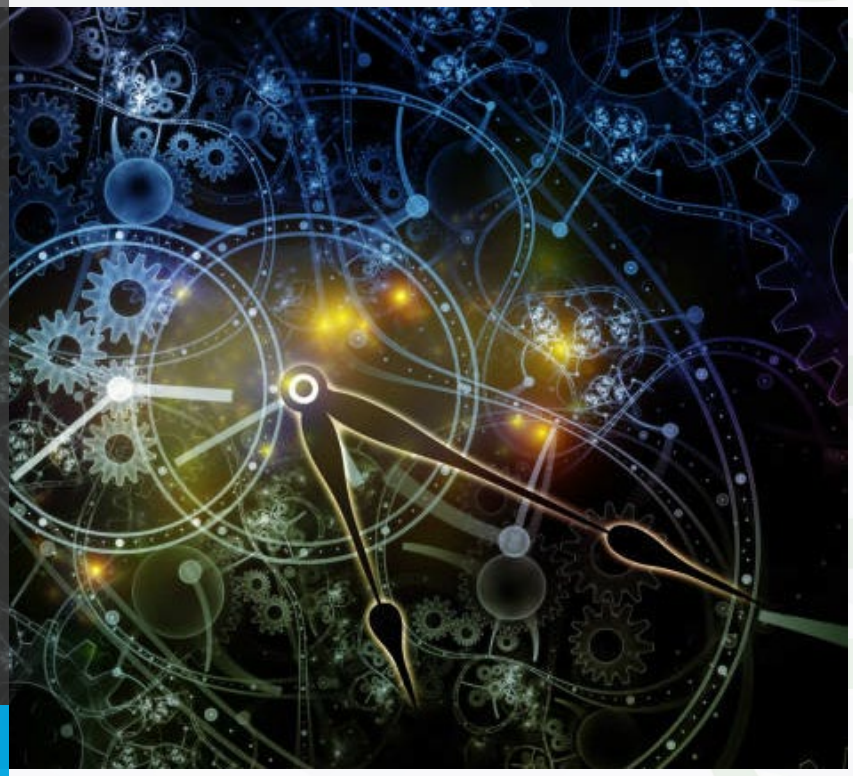
Ingeniería de datos masivos

Análisis de series temporales para Big Data

Semana 8

Introducción

El análisis de series de tiempo se usa para identificar patrones basados en intervalos de tiempo de los datos, permitiendo crear modelos para pronosticar un valor o un comportamiento futuro en un intervalo de tiempo (Clower, 2020). En las empresas se emplea este análisis frecuentemente debido a que sus métricas de negocios como ventas, precios y presupuestos están asociadas a fechas determinadas.



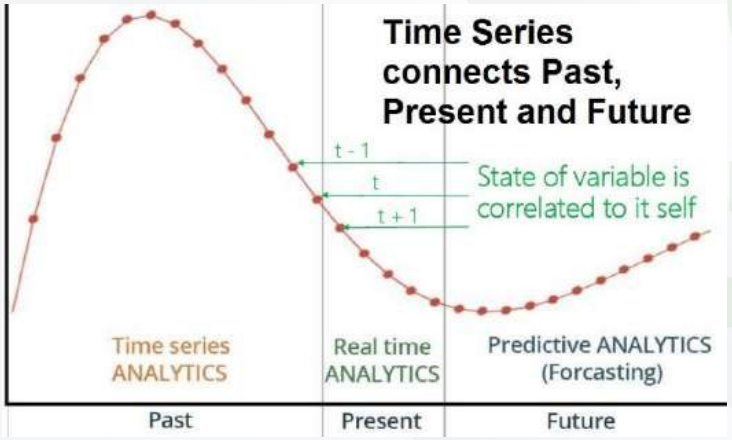
Explicación

Universidad Tecmilenio

Análisis de series de tiempo

Es una secuencia ordenada de valores de una variable a intervalos de tiempo igualmente espaciados, por ejemplo, las ventas mensuales o el número de pasajeros diarios en los últimos 10 años.

Asimismo, este análisis (Singh, 2019) consiste en comprender los factores determinantes y la estructura detrás de los datos observados con el fin de elegir un modelo que permita pronosticar, mejorando la toma de decisiones.



Fuente: Fenjiro, Y. (2019). *Time Series for Business: A general introduction*. Recuperado de https://medium.com/@fenjiro/time-series-for-business-a- general-introduction-50968346e660

El análisis de series de tiempo se utiliza cuando lo único que se sabe es el registro en el tiempo del evento a analizar, por lo que no es un análisis causal en el que se sabe acerca de los determinantes de la variable que se está analizando; consiste en realizar un modelo que describa la serie de tiempo analizada, implicando descomponer la serie temporal en sus componentes (Ogayo, 2020).

Explicación

Componentes del análisis de series de tiempo:

Tendencia	Estacionalidad	Ciclos	Aleatorio
Describe el movimiento de aumento o disminución de la serie. Indica si los valores de observación están aumentando o disminuyendo a lo largo del tiempo.	Describe la fluctuación fija y periódica de la serie a lo largo del tiempo.	Describe una fluctuación periódica de la serie a lo largo del tiempo, pero que no es fija como en el caso de un componente de estacionalidad.	Variabilidad que el modelo no puede explicar y consiste en el resto irregular que queda después de la extracción de todos los componentes que no se describen con ninguna de las tres clasificaciones anteriores.

Los datos de series de tiempo a menudo requieren limpieza, escala e incluso transformación. Por lo tanto, para tener una secuencia ordenada de valores de una variable a intervalos de tiempo igualmente espaciados se requiere lo siguiente:

- Hacer intervalos de tiempo para el análisis.
- Identificar y manejar valores atípicos.
- Identificar faltantes que deben ser imputados.

Explicación

El proceso que se sigue para hacer el análisis es el siguiente (Brid, 2018):

- ✓ Se grafican los datos.
- ✓ Se verifica si la serie es estacionaria (en caso de no serlo se manipula para hacerla estacionaria).
- ✓ Se identifica un modelo para hacer el pronóstico.
- ✓ Se construye el modelo.
- ✓ Se hacen predicciones con el modelo.

Una serie de tiempo es **estacionaria** si sus propiedades estadísticas de media y varianza permanecen constantes en el tiempo.

Por lo tanto, es importante determinar si una serie de tiempo es estacionaria, ya que la mayoría de los modelos funcionan bajo el supuesto de que la serie es de este tipo.

Las técnicas para estimar o modelar la tendencia para luego eliminarla de la series son las siguientes (Khandelwal, 2019):

Para conocer un ejemplo de análisis de series de tiempo en Python, mira el siguiente video:

Dirección de Producción de Contenidos. (2020, 3 de agosto). *Tema10. Análisis de serie de tiempo en Python* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=ZxQMMjKhBBU&feature=youtu.be

Explicación

Las técnicas para estimar o modelar la estacionalidad para luego eliminarla de la serie son las siguientes (Hyndman y Athanasopoulos, 2018):

Diferenciación

Tomando la diferencia con un retraso de tiempo, utilizando lo que se conocen como *lags*.

Descomposición

Identificando cada componente de la serie por separado.

Explicación

El modelo **ARIMA** es el más utilizado para hacer análisis de series de tiempo porque combina los tres tipos de análisis que se realizan en series de tiempo.

De esta forma, este modelo es una ecuación lineal (Prabhakaran, 2019) que se configura con tres parámetros (p, q y d).

p es el número de términos AR (auto- regresivos), los cuales son las variables anteriores a la variable dependiente.

q es el número de términos MA (promedio móvil), los cuales son errores de pronóstico rezagados en la ecuación de predicción.

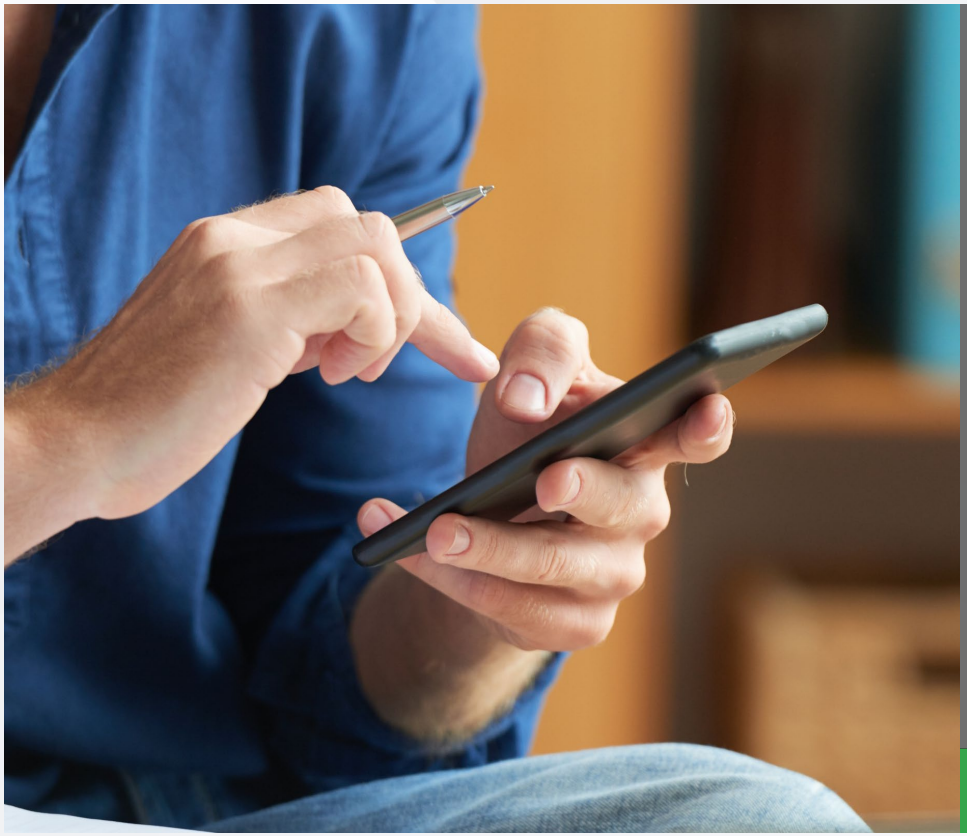
d es el número de diferencias (d) no estacionales a utilizar.

Para determinar los valores de p y q se utilizan las siguientes funciones:

Función de autocorrelación (ACF): es una medida de la correlación entre la serie de tiempo con una versión retrasada de sí misma.

Función de autocorrelación parcial (PACF): mide la correlación entre la serie de tiempo y la versión retrasada de sí misma después de eliminar las variaciones ya explicadas por las comparaciones intermedias.

Cierre



Una serie de tiempo es una colección de cantidades que se encajan en intervalos regulares en el tiempo de forma cronológica.
Mientras que el análisis de series de tiempo es una técnica estadística que trabaja justamente con los datos provenientes de estas series de tiempo, el cual consta de los siguientes componentes: tendencia, estacionalidad, ciclos y aleatoriedad. Por otro lado, el modelo ARIMA (AutoRegressive Integrated Moving Average) se caracteriza por tres parámetros (p, d y d).

Bibliografía

- Dirección de Producción de Contenidos. (2020, 3 de agosto). Tema 10. Análisis de serie de tiempo en Python [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=Z xQMMjKhBBU&feature=youtu.be
- Fenjiro, Y. (2019). *Time Series for Business: A general introduction.* Recuperado de https://medium.com/@fenjiro/time- series-for-business-a-general- introduction-50968346e660
- Ogayo, P. (2020). *Time Series Analysis For Beginners.* Recuperado de https://towardsdatascience.com/time- series-analysis-for-beginners- 8a200552e332
- Singh, K. (2019). *Beginner's Guide for Time-Series Forecasting.* Recuperado de https://dimensionless.in/beginners- guide-for-time-series-forecasting/
- Brid, R. (2018). *Introduction to Time Series.* Recuperado de https://medium.com/greyatom/time- series-b6ef79c27d31
- Khandelwal, R. (2019). *Step by Step Time Series Analysis.* Recuperado de https://medium.com/datadriveninvestor/step-by-step-time-series-analysis- d2f117554d7e
- Hyndman, R., y Athanasopoulos, G. (2018). *Forecasting: principles and practice (2nd edition). Australia: OTexts.* Recuperado de https://otexts.com/fpp2/intro.html
- Prabhakaran, S. (2019). *ARIMA Model – Complete Guide to Time Series Forecasting in Python.* Recuperado de https://www.machinelearningplus.com/t ime-series/arima-model-time-series- forecasting-python/