Instituto Politécnico Nacional
Centro de Investigación en Computación
Semestre B22

Tarea 02

Raquel Eugenia Meléndez Zamudio

Los valores perdidos consisten en aquellos que para una variable determinada no se tienen algunos valores. En series de tiempo los datos son recolectados bajo diferentes condiciones en el tiempo y existen varios mecanismos que pueden contribuir a la generación de valores perdidos en diferentes períodos. Existen tres tipos de valores perdidos:

- *Missing Completely at Random (MCAR)*: No existe relación entre los valores perdidos y los valores observados. La probabilidad de los valores perdidos es completamente aleatoria y no dependiente de los valores observados.
- *Missing at Random (MAR)*: La ausencia de los datos puede depender de los valores observados.
- *Missing not at Random (MNAR)*: La ausencia de los datos depende de la variable por sí misma.

Para tratar los valores perdidos en las series de tiempo existen dos caminos: eliminar la sección donde aparezca el valor perdido o tratar los valores perdidos mediante imputación; sin embargo, cuando se habla de series de tiempo se hace referencia a este proceso como Interpolación.

Los diferentes tipos de Interpolación para tratar con los valores perdidos de una serie de tiempo, son:

- Interpolación por Media
- Interpolación por Mediana
- Interpolación por Moda
- Interpolación Lineal
- Interpolación Spline

Por otro lado, los outliers están definidos como una observación que difiere mucho de otras observaciones como para suponer que fue generada mediante un mecanismo diferente; en otras palabras, puede decirse que los outliers son observaciones que no siguen el comportamiento esperado.

Las técnicas de detección y tratamiento de outliers pueden dividirse en tres categorías: por el tipo de entrada (series de tiempo univariables o multivariables), el tipo de outlier (puntual, subsecuente, series de tiempo) o la naturaleza del método (univariable o multivariable).

Instituto Politécnico Nacional
Centro de Investigación en Computación
Semestre B22

Los outliers puntuales consisten en un dato que se comporta de forma inusual en un tiempo específico cuando se compara con otros valores de una serie de tiempo (outlier global) o con sus datos vecinos (outlier local).

Los outliers subsecuentes, se refieren a puntos consecutivos en el tiempo cuyo comportamiento es inusual, aunque cada observación individual puede no ser necesariamente un outlier, estos pueden ser globales o locales.

Finalmente, los outliers también pueden ser series de tiempo completas, pero solo pueden ser detectados cuando la información de entrada es una serie de tiempo multivariable.

Existe una gran variedad de técnicas para detección de outliers en series de tiempo, entre ellas podemos encontrar:

- Detección STL (Seasonal-Trend)
- Árboles de Clasificación y Regresión (CART)
- Detección usando predicción
- Detección basada en clustering
- Autoencoders

Así mismo, existen diversos métodos para el tratamiento de estos como lo son los métodos estádisticos.

Por otro lado, dentro de las métricas de desempeño para series de tiempo podemos encontrar:

- Error medio (ME – Mean Error):

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)$$

  Es una métrica simple; sin embargo, se encuentra sesgada debido al efecto de compensación de errores de predicción positivos y negativos que pueden ocultar la imprecisión de la predicción de observaciones correctas. Esta puede mostrar rápidamente la simetría de la distribución de errores.

- Error absoluto medio (MAE – Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f_i|$$

  Utiliza los valores absolutos de los errores en los cálculos, lo que arregla el problema de la cancelación de errores con signos opuestos. Da un promedio de la magnitud absoluto de todos los valores de los errores, sin importar si eran positivos o negativos.

- Error cuadrático medio (MSE – Mean Square Error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2$$

Al igual que MAE, el MSE también arregla el problema de la cancelación de errores positivos y negativos; sin embargo, otorga una mayor penalización en los errores de predicción grandes.

- Raíz del error cuadrático medio (RMSE – Root Mean Square Error)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2}$$

En estadística se conoce como desviación estándar de los errores. Se utiliza comúnmente en la predicción y análisis de las regresiones para verificar resultados experimentales. Tiene la ventaja de tener las mismas unidades que la variable predicha; por lo que, es más fácil de interpretar.

- Error porcentual medio (MPE – Mean Percentage Error)

$$MPE = \frac{1}{n}\sum_{i=1}^{n} 100 * \frac{y_i - f_i}{y_i}$$

Los errores de previsión positivos y negativos pueden compensarse entre sí; por lo que se puede utilizar para medir el sesgo en las previsiones. La desventaja es que no es adecuada para conjuntos de datos que contienen valores observados iguales a cero.

- Error porcentual absoluto medio (MAPE – Mean Absolute Percentage Error)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} 100 * \left|\frac{y_i - f_i}{y_i}\right|$$

Arregla el problema con la compensación de errores y funciona mejor si no hay extremos en los datos.

- Métricas o coeficiente de precisión U de Theil
  Primer coeficiente con valores en un rango entre (0,1).

$$U_1 = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2}}{\sqrt{\frac{1}{T}\sum_{i=1}^{n} y_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_i^2}}$$

Cuanto mayor sea la precisión de la predicción, menor será el valor del coeficiente.

En el segundo coeficiente U_2 se indica cuanto más o menos preciso es un modelo en relación con una predicción trivial.

$$U_2 = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{f_{i+1} - y_{i+1}}{y_i}\right)^2}{\sum_{i=1}^{n}\left(\frac{y_{i+1} - y_{i+1}}{y_i}\right)^2}}$$