UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Imputación de Datos

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 21) 23.marzo.2022

Imputación de Datos

Son técnicas (estadísticas) para rellenar o completar datos faltantes en una base de datos. La idea es que estas técnicas permiten hacer el relleno de una forma apropiada.

- Imputación usando resúmenes estadísticos (de una variable)
- Imputación usando resúmenes estadísticos multivariados
- Imputación usando regresión
- Otras técnicas más avanzadas

Resúmenes estadísticos

Rellenar con valores 0: Típicamente completamos los datos nulos con 0. Esto no es la mejor estrategia, pero al menos llena la base de datos. De alguna forma, la presencia de 0 indica un dato faltante.

Ventajas:

- Muy simple.

Desventajas:

- Altera la distribución de la variable o columna.
- Puede confundir los datos 0 faltantes, con datos 0 correctos.

Obs: Lo mismo vale al sustituir datos faltantes por un valor constante.

Resúmenes estadísticos

Rellenar con la media: Los datos faltantes se reemplazan por la media de la columna. Aquí se reemplaza por un valor constante pero al menos se presenva la media de cada columna

Ventajas:

- Simple. Preserva la media de la distribución.

Desventajas:

- Altera la distribución de la variable o columna.

Obs: Se puede también sustituir por la mediana, o la moda (el dato más frecuente).

Resúmenes estadísticos

Rellenar con la media por grupo: En ocasiones, la media no es un buen descriptor para todos los datos faltantes. Es posible que en la base de datos haya una columna o variable categórica.
Si la variable que deseamos reemplazar se comporta distinto en cada una los grupos de esta variable categórica, es convieniente usar la media de cada grupo en lugar de la media global.

Ventajas:

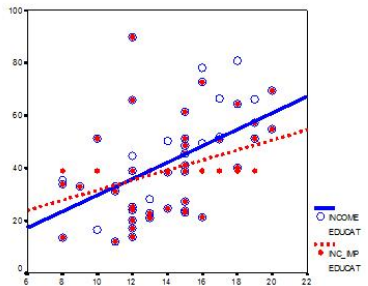
- Simple.
- Preserva la media de cada grupo.
- Aproxima mejor el comportamiento de los datos faltantes en función de su grupo.

Desventajas:

- Sólo funciona si existe alguna variable categórica para agrupar.

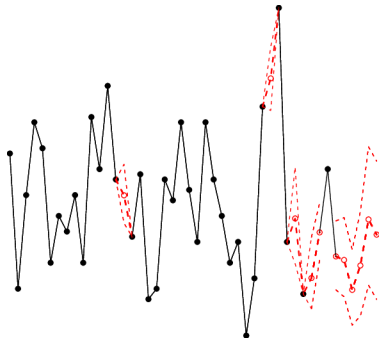
Regresión lineal

En ocasiones, podríamos basar nuestra decisión del valor de los datos faltantes en función de los valores de una o más variables numéricas (continuas).
En ese caso, se suele utilizar una regresión lineal para decidir el valor del dato faltante.

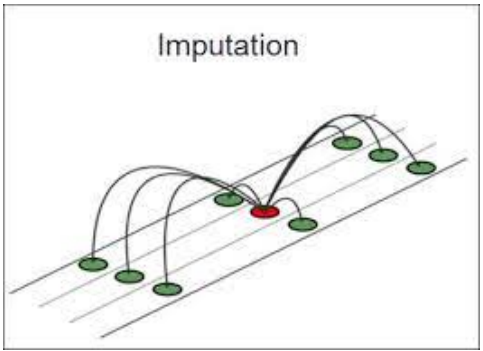


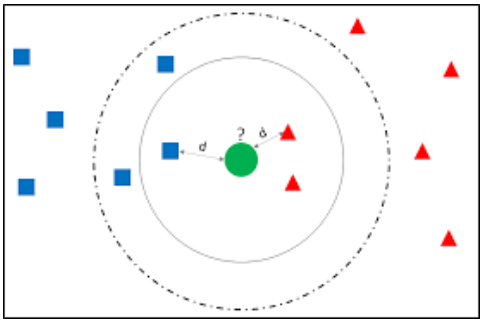
Otros métodos avanzados

Imputación usando KNN (*K nearest neighbours*). Busca los *k* datos más cercanos al registro donde está el dato faltante. Reemplaza el datos faltante por el promeido de los valores encontrados en los *k* vecinos más cercanos.



KNN Imputation





KNN Imputation

El algoritmo es el siguiente: Trabamos a los datos como si fuesen vectores $\mathbf{x}_i \in \mathbb{R}^d$, donde d es la dimensión o número de columnas (variables).

Algoritmo (Imputación KNN):

1. Establecer el parámetro k de número de vecinos.

2. Para cada dato $\mathbf{x}_i$ (con algíún campo faltante), localizamos los k vecinos más cercanos a $\mathbf{x}_i$.

3.
   - Si el campo faltante es categórico, lo reemplazamos con la clase más frecuente entre los k vecinos más cercanos.
   - Si el campo faltante es continuo, lo reemplazamos con el promedio de los k vecinos más cercanos.