ITESO, Universidad
Jesuita de Guadalajara

Departamento de Matemáticas y
Física

Ingeniería de características

Tratamiento de datos faltantes

Dr. Carlos Augusto Arellano Muro

1. Visualización de datos faltantes

Mapa de calor: Muestra los datos en un color de acuerdo a su valor, si no hay dato, el color se queda en blanco.

Mtriz de co–ocurrencia: Describe la frecuencia de los datos de interés.

La Figura 1 es una representación visual de los datos ordenados que se muestran en la Tabla 9.

Figura 1: Ejemplo de mapa de color.

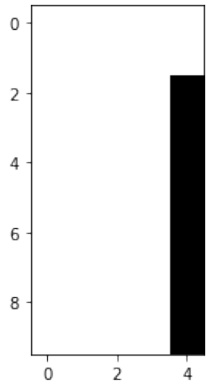

Tabla 1: Muestra de datos faltantes.

	D1	D2	D3	D4	D5
0	8.0	6.0	1.0	9.0	6.0
1	10.0	7.0	3.0	0.0	NaN
2	8.0	11.0	2.0	8.0	1.0
3	1.0	9.0	0.0	9.0	NaN
4	8.0	3.0	11.0	0.0	NaN

PYTHON
MATPLOTLIB.PYPLOT.IMSHOW(X)
SEABORN.HEATMAP(X)
PANDAS.CROSSTAB(INDICE,COLUMNA)

2. Eliminación de observaciones

Cuando sea deseable utilizar modelos que sean intolerantes a los datos faltantes, los valores faltantes deben extraerse de los datos. El enfoque más simple para tratar con valores perdidos es eliminar predictores completos y/o muestras que contienen valores perdidos. Sin embargo, se deben considerar cuidadosamente varios aspectos de los datos antes de adoptar este enfoque. Por ejemplo, los valores perdidos se podrían eliminar borrando todos los predictores que contienen al menos un valor perdido. De manera similar, los valores perdidos se podrían eliminar borrando todas las muestras con valores perdidos. Ninguno de estos enfoques será apropiado para todos los datos.

Una consideración importante es el valor intrínseco de las muestras en comparación con los predictores. Cuando sea difícil obtener muestras o cuando los datos contengan una pequeña cantidad de éstas, no es conveniente

eliminarlas de los datos. En general, las muestras son más críticas que los predictores y se debe dar mayor prioridad a mantener tantas como sea posible.

PYTHON

```
PANDAS.DATAFRAME.DROPNA()  # POR RENGLONES
PANDAS.DATAFRAME.DROPNA(AXIS=1)  # POR COLUMNAS
```

3. Sustitución de datos faltantes

La moda se usa para imputar predictores cualitativos y el promedio o la mediana se usa para imputar predictores cuantitativos.

3.1. Sustitución por media y mediana

- Ambos valores son representativos de la muestra.
- No cambia su valor central.
  - En presencia de valores atípicos, se recomienda la mediana.
- Con valores que no son estrictamente aleatorios, resultan en un sesgo de inconsistencia.
- Alteran el valor de la varianza.
- Alteran la relación con otras variables.

Por ejemplo, en la Tabla 2, la media aritmética de la columna D3 es 4.5 (recordando que los datos mostrados solo representan una muestra):

Tabla 2: Muestra de datos faltantes.

	D1	D2	D3	D4
0	8.0	6.0	1.0	9.0
1	24.0	7.0	NaN	0.0
2	8.0	11.0	2.0	8.0
3	1.0	9.0	NaN	9.0
4	8.0	3.0	11.0	0.0

Tabla 3: Datos sustituidos por la media aritmética.

	D1	D2	D3	D4
0	8.0	6.0	1.0	9.0
1	24.0	7.0	4.5	0.0
2	8.0	11.0	2.0	8.0
3	1.0	9.0	4.5	9.0
4	8.0	3.0	11.0	0.0

3.2. Sustitución por moda y frecuencia

- Es un valor representativo de la muestra. Tampoco altera la moda de los datos.
- Se usa el valor que más se repite para reemplazar el valor perdido.

En este caso, se sabe que el valor que más se repite en la columna D3 es 5.0, por lo que se sustituyen los NaN por este valor:

Tabla 4: Muestra de datos faltantes.

	D1	D2	D3	D4
0	8.0	6.0	1.0	9.0
1	24.0	7.0	NaN	0.0
2	8.0	11.0	2.0	8.0
3	1.0	9.0	NaN	9.0
4	8.0	3.0	11.0	0.0

Tabla 5: Datos sustituidos por la moda

	D1	D2	D3	D4
0	8.0	6.0	1.0	9.0
1	24.0	7.0	5.0	0.0
2	8.0	11.0	2.0	8.0
3	1.0	9.0	5.0	9.0
4	8.0	3.0	11.0	0.0

3.3. Sustitución aleatoria

- Se usa un valor de la muestra escogido al azar.
- Es solo para muestras obtenidas completamente al azar.
- La sustitución puede ser un elemento al azar para todos los valores faltantes o, idealmente, un valor al azar por cada elemento faltante en cada variable.

Tabla 6: Muestra de datos faltantes.

	D1	D2	D3	D4
0	8.0	6.0	1.0	9.0
1	24.0	7.0	NaN	0.0
2	8.0	11.0	2.0	8.0
3	1.0	9.0	NaN	9.0
4	8.0	3.0	11.0	0.0

Tabla 7: Muestra de datos faltantes.

	D1	D2	D3	D4
0	8.0	6.0	1.0	9.0
1	24.0	7.0	4.0	0.0
2	8.0	11.0	2.0	8.0
3	1.0	9.0	3.0	9.0
4	8.0	3.0	11.0	0.0

PYTHON

```
FIT_TRANSFORM(X)  SKLEARN.IMPUTE.SIMPLEIMPUTER(STRATEGY='MEAN')#'MEDIAN','MOST_FREQUENT'
```

3.4. Criterio de selección

El primer criterio para discriminar el método de imputación es el tipo de variable, si ésta es categórica, se prefiere la sustitución por moda; si es numérica, el siguiente criterio a evaluar es su distribución, en una distribución uniforme, el método a emplear debe ser la sustitución aleatoria; en cambio, si la distribución es semejante a la normal, es decir, simétrica y curtósis mayor a -1.0, el siguiente criterio es, si la variable muestra datos atípicos, se aplica la sustitución por mediana, ya que esta medida de tendencia central no varía en la presencia de tales valores; de lo contrario, se usa la sustitución por media aritmética como lo muestra el diagrama de la Figura2.

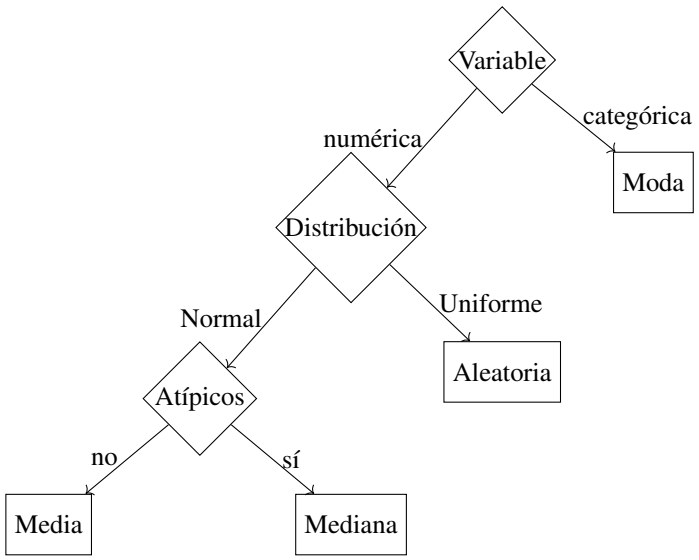

Figura 2: Diagrama de la toma de decisiones para la elección del método de imputación.

Sin embargo, si en la sustitución aleatoria, la elección del elemento a imputar se escoge con una distribución con la misma forma (sesgo y curtósis), dispersión y tendencia central, se puede aplicar para cualquier variable, numérica o categórica.

4. Valores extremos

- Los valores extremos o aberrantes influyen directamente a la distribución de la muestra.
- Se usa el rango o recorrido de tres veces la distancia entre los cuartiles Q_3 y Q_1.
- Otros criterios implican suponer una distribución para la variable (comúnmente Normal) y calcular qué probabilidad existe de encontrar dicho valor.

En el siguiente ejemplo, en la columna D1, los cuartiles correspondientes Q_1, Q_2 y Q_3 son 6, 8 y 9 respectivamente, por lo que el Rango intercuartílico es $R_i = 3$. Entonces $3R_i = 9$, este valor se suma a Q_3 resultando $Q_3 + 3R_i = 18$, así que, cualquier valor superior a 18 se considera extremo en esta columna, que es el caso del 24. Éste y los valores faltantes se sustituyen por la mediana Q_2 respectivos para cada columna:

Tabla 8: Muestra de datos faltantes.

	D1	D2	D3	D4
0	8.0	6.0	1.0	9.0
1	24.0	7.0	NaN	0.0
2	8.0	11.0	2.0	8.0
3	1.0	9.0	NaN	9.0
4	8.0	3.0	11.0	0.0

Tabla 9: Muestra de datos faltantes.

	D1	D2	D3	D4
0	8.0	6.0	1.0	9.0
1	8.0	7.0	4.0	0.0
2	8.0	11.0	2.0	8.0
3	1.0	9.0	4.0	9.0
4	8.0	3.0	11.0	0.0

Referencias

[1] Kuhn M. Johnson K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models. Chapman and Hall/CRC Press, pp. 187–204.

[2] Galli, S. (2022). Python feature engineering cookbook: over 70 recipes for creating, engineering, and transforming features to build machine learning models. Packt Publishing Ltd. pp. 45–91.

[3] `https://www.uv.es/webgid/Descriptiva/23_valores_faltantes.html`

[4] https://www.datasource.ai/es/data-science-articles/todo-sobre-el-manejo-de-datos-faltantes