ITESO, Universidad
Jesuita de Guadalajara

Tratamiento de datos faltantes

Dr. Gaddiel Desirena López

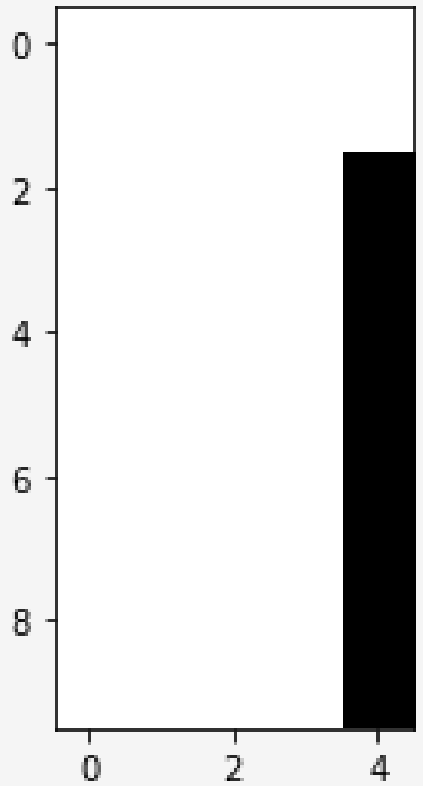
Primavera 2024

Contenido

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.



Eliminación de observaciones

- El enfoque más simple es eliminar predictores completos y/o muestras que contienen valores perdidos.
- Los valores perdidos se podrían eliminar borrando todos los predictores que contienen al menos un valor perdido.
- Los valores perdidos se podrían eliminar borrando todas las muestras con valores perdidos.
- Cuando sea difícil obtener muestras o cuando los datos contengan una pequeña cantidad de éstas, no es conveniente eliminarlas de los datos.
- Las muestras son más críticas que los predictores y se debe dar mayor prioridad a mantener tantas como sea posible.

Sustitución por media y mediana

- Ambos valores son representativos de la muestra.
- No cambia su valor central.
- 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.

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.

Sustitución aleatoria

- Se usa un valor de la muestra escogido al azar.
- Es solo para muestras obtenidas completamente al azar.

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.