MACHINE LEARNING

CIENCIAS DE LA COMPUTACIÓN
M.Sc. Richard F. Fernández Vásquez

¿Por qué es importante realizar un preprocesamiento de datos?

Las principales tareas antes de realizar cualquier modelo de Machine Learning son:

- Estimar valores perdidos
- Identicar outliers y suavizar datos
- Corregir datos incosistentes

Tresp (1995) - el problema de valores faltantes en aprendizaje supervisado usando redes neuronales.

Logro Unidad 1

Al finalizar la unidad, el alumno es capaz de aplicar adecuadamente técnicas de pre procesamiento de datos para posibilitar la implementación de una solución de Machine Learning para un problema del mundo real.

Contenido 3

- Identificación y tratamiento de valores faltantes
- Identificación y tratamiento de valores atípicos
- Transformación de datos
- Caso de Aplicación

Contenido 3

- Identificación y tratamiento de valores faltantes
- Identificación y tratamiento de valores atípicos
- Transformación de datos
- Caso de Aplicación

Valores Faltantes

NA

- Cuando el valor es “not available”

- Si se operan elementos con NA, el resultado será NA

- Se puede omitir los NA al momento de realizar cualquier operación (na.rm = T)

NaN

- Cuando se realizan operaciones que no son números, “not a number”

Inf => Indica un valor “infinito positivo”

-Inf => Indica un valor “infinito negativo”

Valores Faltantes

Según investigaciones el impacto de los valores faltantes son:

1% datos faltantes => trivial

1-5% => manejable

5-15% => requiere métodos sofisticados

Mas del 15% => interpretación perjudicial

Valores Faltantes

Ante la presencia de valores faltantes se puede optar por:

- Eliminar
- Reemplazar
- Mantener

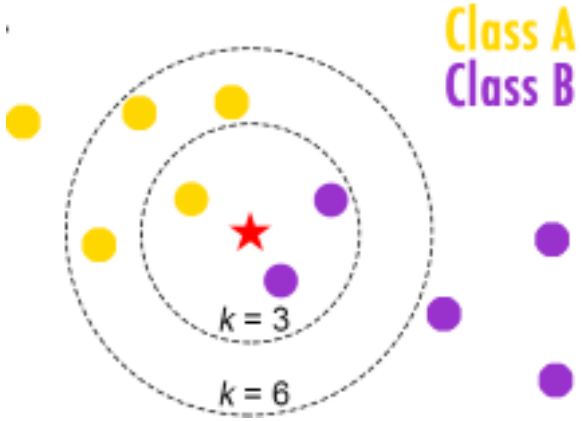
Valores Faltantes

Imputación: Los valores faltantes son reemplazados con valores estimados basados en la información disponible:

- Imputación por la media
- Imputación por la mediana
- Imputación por la moda

Valores Faltantes

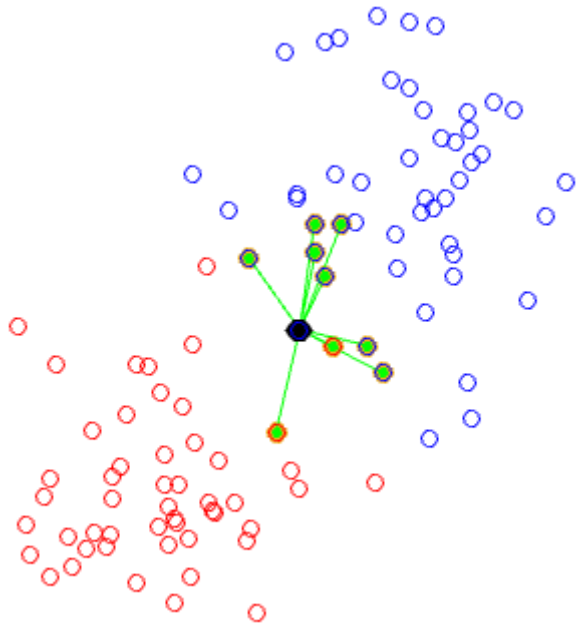
K-vecinos mas cercanos



Si la variable es categórica se reemplaza por la moda de las k-observaciones más cercanas

Valores Faltantes

K-vecinos mas cercanos



Si la variable es numérica se reemplaza por la media de las k-observaciones más cercanas

Valores Faltantes

- No hay diferencia entre usar imputación por la media e imputación por la mediana.

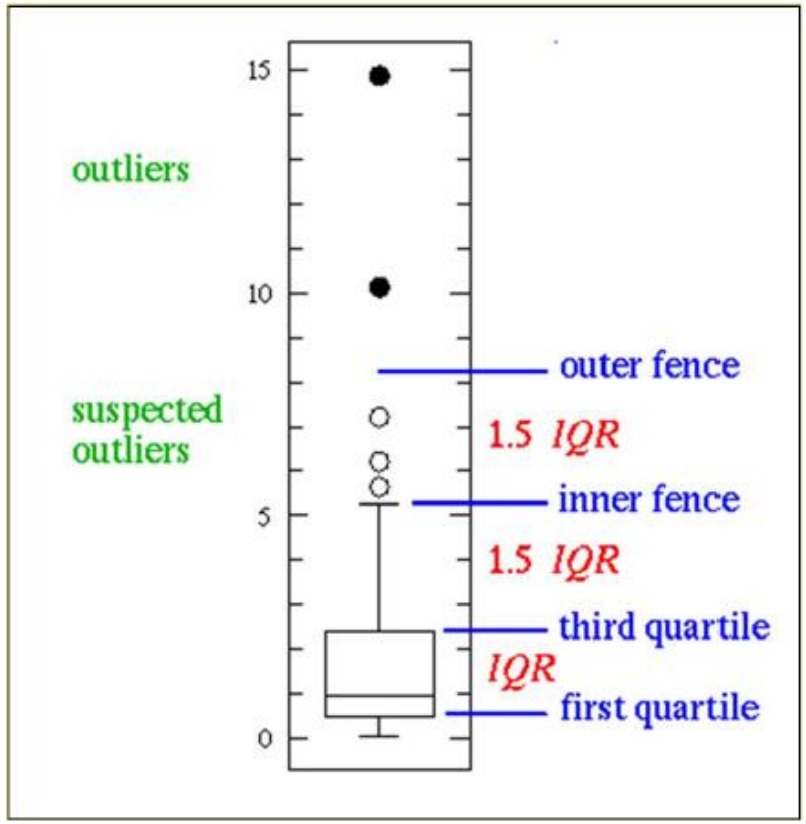
- Para conjuntos de datos con una pequeña cantidad de valores faltantes se observa poca diferencia entre la eliminación de casos y otros métodos de imputación.

Contenido 3

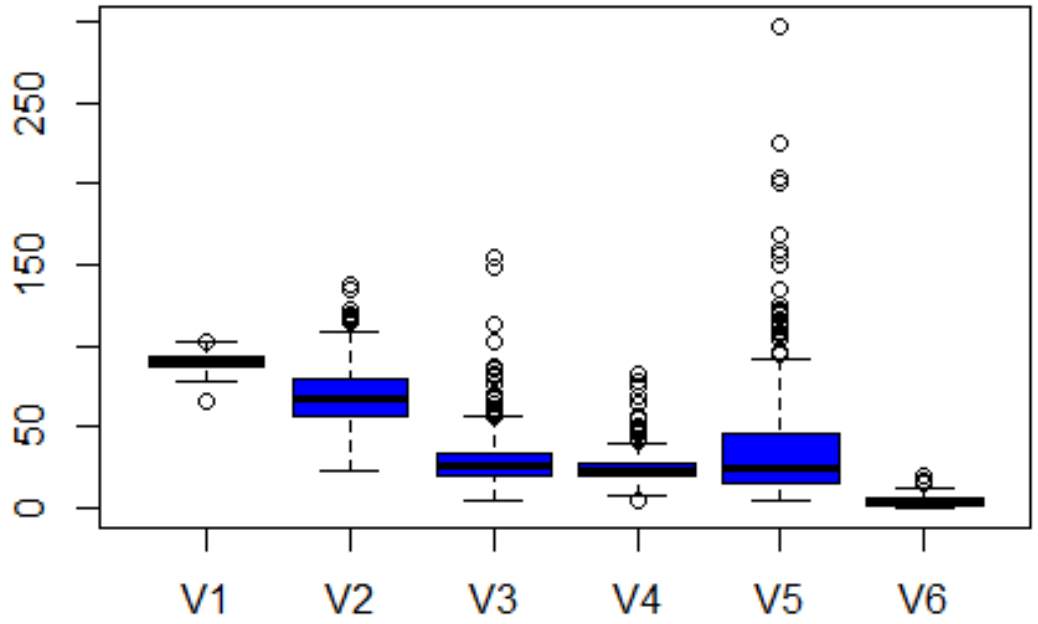
- Identificación y tratamiento de valores faltantes
- Identificación y tratamiento de valores atípicos
- Transformación de datos
- Caso de Aplicación

Outliers

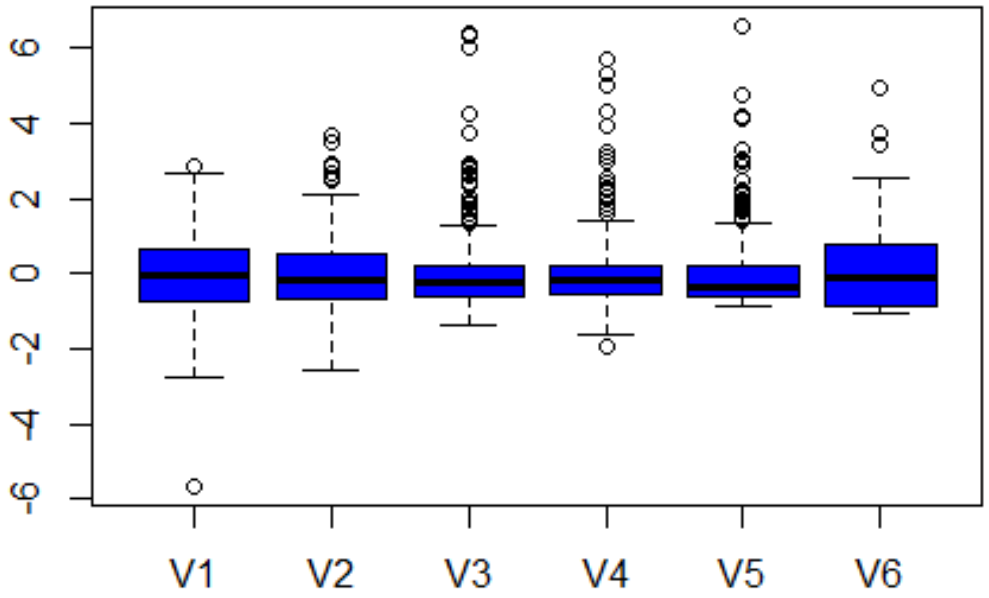
- Considerar outliers valores que $\frac{|x-\overline{x}|}{s} > k$
- donde k es 2 ó 3 si consideramos normalidad.
- Considerando el Boxplot (Tukey, 1977), se considera outlier a los valores que caen fuera de este intervalo. $(Q_1 - 3 \times IQR,\ Q_3 + 3 \times IQR)$



Outliers



Outliers



Prueba de normalidad

Es una prueba para evaluar si una variable se aproxima a una distribución normal.

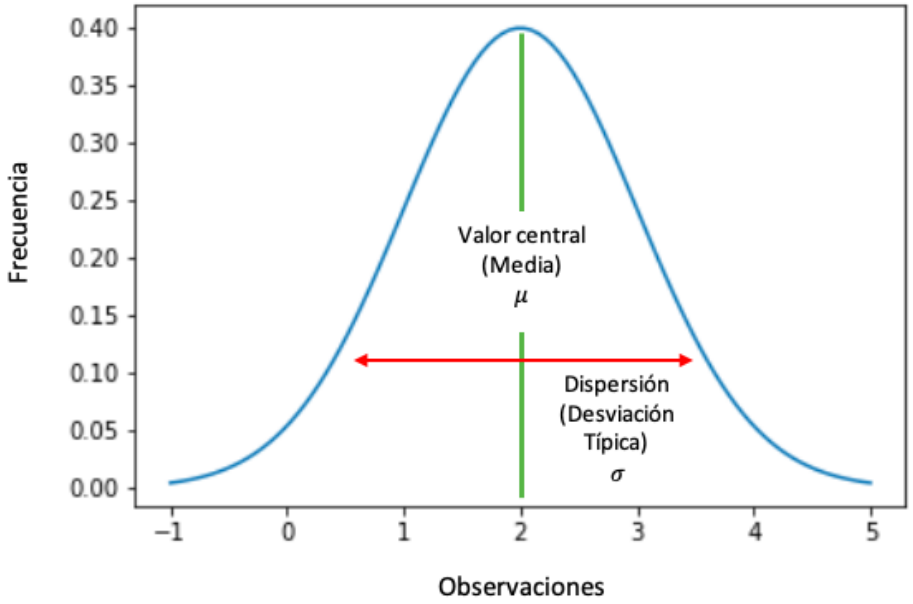
Ho: La variable se aproxima a una distribución normal.

Ha: La variable no se aproxima a una distribución normal.

=> Si el pvalor < 5% => **se rechaza la Ho**, al 95% de confianza se concluye que la variable no se aproxima a una distribución normal.

=> Si el pvalor > 5% => **no se rechaza la Ho**, al 95% de confianza se concluye que la variable se aproxima a una distribución normal.

Prueba de normalidad



α: Nivel de significación, por lo general toma el valor de 5%

1-α: Nivel de confianza, por lo general toma el valor de 95%

Valores atípicos

Si la variable **se aproxima** a una distribución normal:

- Considerar outliers valores que $\frac{|x-\overline{x}|}{s} > k$
- donde k es 2 ó 3 si consideramos normalidad.

Si la variable **NO se aproxima** a una distribución normal:

- Considerando el Boxplot (Tukey, 1977), se considera outlier a los valores que caen fuera de este intervalo. $(Q_1 - 3 \times IQR,\ Q_3 + 3 \times IQR)$

Tratamiento de los valores atípicos

Ante la presencia de valores atípicos se puede optar por:

- Eliminar
- Reemplazar
- Mantener

Tratamiento de los valores atípicos

Reemplazo

Los valores atípicos son reemplazados con valores estimados basados en la información disponible:

- Reemplazo por la media
- Reemplazo por la mediana
- Reemplazo por la moda

Contenido 3

- Identificación y tratamiento de valores faltantes
- Identificación y tratamiento de valores atípicos
- Transformación de datos
- Caso de Aplicación

Transformación de datos

1. Suavizamiento: Remover datos ruidosos

2. Agregación: resumen, construcción de cubos de datos

3. Normalización
   - Normalización min-max
   - Normalización z-score
   - Normalización por escalamiento decimal

4. Construcción de Atributos
   - Nuevos atributos construidos basados en los anteriormente especificados.

Transformación de datos

Normalización

- Normalización z-score
- Normalización min-max
- Normalización por escalamiento decimal
- Normalización Sigmoidal
- Normalización Softmax

Transformación de datos

Normalización

Normalización z-score

$$Z = \frac{X - \mu}{\sigma}$$

Transformación de datos

Normalización

Normalización min-max

Los valores son transformado en forma lineal a un rango pre-especicado [a,b].

$$V' = \frac{(V - X_{min})\,(b - a)}{X_{max} - X_{min}} + a$$

Transformación de datos

Normalización

Normalización por escalamiento decimal

La normalización se realiza moviendo el punto decimal de los valores. Esta normalización transforma los datos al rango [-1,1]

$$V' = \frac{V}{10^j}$$

donde j es el entero mas pequeño tal que $max\left(|V'|\right) < 1$

Transformación de datos

Normalización

Normalización Sigmoidal

Se realiza una transformación no lineal de los datos para llevarlos al rango [-1,1].
Se usa cuando se tienen datos anómalos.

$$V' = \frac{1 - e^{-a}}{1 + e^{-1}}$$

donde $a = \frac{V - \overline{x}}{s}$

Transformación de datos

Normalización

Normalización Softmax

Esta transformación lleva los valores al rango [0,1].

$$V' = \frac{1}{1 + e^{-a}}$$

donde $a = \frac{V - \overline{x}}{s}$

CIERRE

Tratamiento de valores atípicos

¿Qué es estandarizar una variable?

¿Qué es un outlier?

CONSULTAS

pcsirife@upc.edu.pe