Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Tratamiento de datos. Parte 2

1

Profesora: *Dra. Fabiola Ocampo Botello*

Tratamiento de datos

Ejemplo tomado de Larose & Larose (2015:Sección 2.2).

¿Qué errores encuentra en la siguiente tabla?

2

CustomerID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	100048	M	75,000	**C**	M	5000
1002	**J2S7K7**	F	-40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	**6269**	M	50,000	**0**	S	1000
1005	55101	F	99,999	30	D	3000

3

Debido a que algunas veces las bases de datos provienen de diversas fuentes, puede ser que no se verifique que los valores se encuentren en las mismas unidades de medida.

¿Qué hacer con los datos faltantes?

Han, Kamber & Pei (2012) presentan diversas rutinas de limpieza de datos para tratar los valores faltantes, suavizar el ruido al encontrar valores atípicos y corregir inconsistencias.

1. Ignorar la tupla. Esto se hace por ejemplo cuando falta la etiqueta de la clase y lo que se realiza es la clasificación. Este método no es muy efectivo, a menos que a la tupla le falten varios valores.

4

2. Completar manualmente el valor faltante. Este enfoque consume mucho tiempo y puede no ser factible considerando un gran conjunto de datos con muchos datos faltantes.

3. Uso de una constante global para completar los valores faltantes. Se puede utilizar una etiqueta como "Desconocido". Pero, hay que tener cuidado porque el algoritmo de minería de datos puede detectar erróneamente que es un concepto interesante ya que existen muchos datos con ese valor.

4. Uso de una medida de tendencia central para el atributo. Puede ser la media o la mediana.

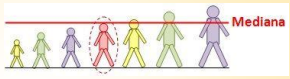


Foto foto de Autor desconocido está bajo licencia CC BY-NC-ND

Foto foto de Autor desconocido está bajo licencia CC BY-NC-ND

5. Usar la media o la mediana para todas las muestras que pertenecen a la misma clase. Por ejemplo, si se clasifica a los clientes de acuerdo al riesgo de crédito, se puede reemplazar el valor faltante con el promedio del valor de ingreso de los clientes de esa categoría.

6. Usar el valor más probable para completar el valor faltante. Este valor se puede determinar mediante una regresión.

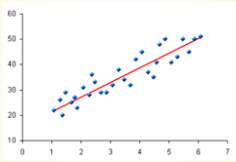


Foto foto de Autor desconocido está bajo licencia CC BY-SA

Además de los anteriores, Larose & Larose (2015) proponen reemplazar los valores faltantes con valores imputados considerando otras características del registro.

Ejemplo que ilustra los tres casos:

1. Reemplazar el valor faltante con una constante definida por el usuario.
2. Reemplazar el valor faltante por la media para valores numéricos.
3. Reemplazar el valor faltante por la moda para valores categóricos.

Ejemplo tomado de Larose & Larose (2015:Sección 3.3)

Tabla original:

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900		71	Europe
3	17.000	302	140	US
4	15.000	400	150	
5	37.700	89	62	Japan

Figura 2.1. Tomada de Larose & Larose (2015:Sección 3.3)

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900		71	Europe
3	17.000	302	140	US
4	15.000	400	150	Missing
5	37.700	89	62	Japan

Figura 2.2. Tomada de Larose & Larose (2015:Sección 3.3)

Se reemplazó el valor faltante por una constante definida por el usuario.

Tabla con los valores reemplazados, para los numéricos se aplicó la media y para los categóricos se aplicó la moda.

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	200.65	71	Europe
3	17.000	302	140	US
4	15.000	400	150	US
5	37.700	89	62	Japan

Figura 2.3. Tomada de Larose & Larose (2015:Sección 3.3)

Debe enfatizarse que reemplazar los valores perdidos es una apuesta, y los beneficios deben sopesarse frente a la posible invalidez de los resultados.

9

Métodos de imputación

J. Nissen, R. Donatello, & B. Van Dusen. (2019) definen **imputación como una técnica ejemplar para manejar los datos faltantes**. La imputación maneja los datos faltantes con valores plausibles, de modo que un investigador pueda analizar el conjunto de datos completo sin preocuparse por los datos faltantes.

Los métodos de imputación se dividen en dos grandes categorías: deterministas y probabilísticos.

Los métodos deterministas incluyen la imputación de la media y el último valor observado.

Los métodos probabilísticos consideran la imputación múltiple (*Multiple Imputation, MI*) y la estimación de la máxima verosimilitud.

10

Larose & Larose (2015) expresan que la pregunta que se plantea en los métodos de imputación es:

¿Cuál sería el valor más probable para este valor perdido considerando todos los demás atributos para un registro en particular?

11

Larose & Larose (2015) presentan diversas formas para iniciar el análisis de los datos. Mencionando las tablas de frecuencias y dos métodos gráficos: histogramas y las gráficas de dispersión.

Los diagramas de caja y bigotes también permiten visualizar datos atípicos.

La tabla de frecuencias permite conocer la cantidad de datos que se encuentran en cada clase.

Ejemplo de tablas de frecuencias presentado en Larose & Larose (2015) para identificar errores de clasificación en las categorías o clases.

Marca	Frecuencia
USA	1
France	1
US	156
Europe	46
Japan	51

Tabla 2.2. Adaptada de Larose & Larose (2015).

12

Los métodos gráficos permiten identificar valores atípicos, los cuales hay que tener presentes debido a que hay técnicas de minería de datos que son susceptibles a estos valores.

Las gráficas de barras se utilizan para variables categóricas, cualitativas, nominales.

Los histogramas se utilizan para representar frecuencias de variables continuas, cuantitativas, en donde cada barra representa la frecuencia de un intervalo de valores.

13

Ejemplo de histograma presentado en Larose & Larose (2015) para identificar valores atípicos mediante un histograma.

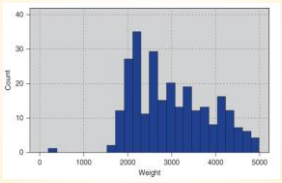


Figura 2.5 tomada de Larose & Larose (2015:Sección 2.5), esta gráfica muestra un histograma de los pesos de los vehículos (ligeramente modificado) de un conjunto de datos.

14

Ejemplo de gráfica de dispersión presentado en Larose & Larose (2015) para identificar valores atípicos.

La gráfica de dispersión permite visualizar valores de dos variables.

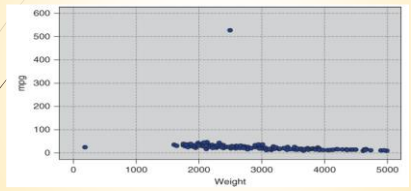


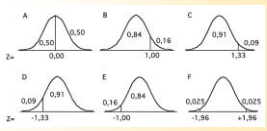
Figura 2.6 tomada de Larose & Larose (2015:Sección 2.5).

Transformación de datos

15

Las variables pueden contener dominios de valores que difieren unos de otros, por lo que para que puedan ser comparables, es necesario **normalizarlos**.

Una de las técnicas de normalización de datos es la conversión de los valores numéricos a desviaciones estándar, puntajes conocidos como puntajes Z, pero la restricción que se tiene es que estos datos deben tener una **distribución normal** (Larose & Larose, 2015).



16

Otra técnica presentada en Larose & Larose (2015) es la llamada normalización Min-Max, la cual analiza qué tan mayor es el valor del campo con respeto al valor mínimo (X) y escala esta diferencia por el rango. La fórmula es:

$$X^*_{mm} = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Suponga que se tienen los siguientes datos:

weightlbs
Statistics

Mean	3005.490
Min	1613
Max	4997
Range	3384
Standard Deviation	852.646

Se desea realizar la normalización de los siguientes datos: 1613, 3384 y 4997.

- Como el valor mínimo es 1613, este tendrá un valor de normalización igual a cero.

$$X^*_{mm} = \frac{X - \min(X)}{\text{range}(X)} = \frac{1613 - 1613}{3384} = 0$$

- El rango medio es el promedio entre el valor máximo y el valor mínimo.

$$\text{Midrange}(X) = \frac{\max(X) + \min(X)}{2} = \frac{4997 + 1613}{2} = 3305 \text{ pounds}$$

Imágenes tomadas de Larose & Larose (2015)

El valor normalizado para vehículos con un peso igual al rango medio es:

$$X^*_{mm} = \frac{X - \min(X)}{\text{range}(X)} = \frac{3305 - 1613}{3384} = 0.5$$

- El vehículo más pesado tiene una normalización de:

$$X^*_{mm} = \frac{X - \min(X)}{\text{range}(X)} = \frac{4497 - 1613}{3384} = 1$$

Nota: El valor en la fórmula anterior debe ser 4997 en vez de 4497.

Imágenes tomadas de Larose & Larose (2015)

La escala decimal

La escala decimal asegura que cada valor normalizado se encuentre entre -1 y 1.

$$X^*_{\text{decimal}} = \frac{X}{10^d}$$

donde d representa el número de dígitos que tiene el valor absoluto del dato. Para los datos que hacen referencia al peso de los automóviles, el valor absoluto más grande en este caso es $d = 4$ dígitos.

La escala decimal para el peso mínimo y máximo son

$$\text{Min}: X^*_{\text{decimal}} = \frac{1613}{10^4} = 0.1613 \quad \text{Max}: X^*_{\text{decimal}} = \frac{4997}{10^4} = 0.4997$$

Imágenes tomadas de Larose & Larose (2015)

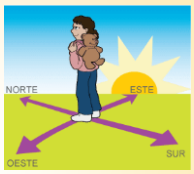
Transformación de variables categóricas en valores numéricos

Variables banderas (flag)

Larose & Larose (2015) establece que algunos métodos de análisis, como los métodos de regresión requieren que los predictores sean valores numéricos, por lo que si el analista desea usar predictores categóricos en la regresión es necesario que recategorice las variables en una o más variables bandera, también conocida como *variable dummy* o variable indicador debido a que sólo puede tomar uno de dos posibles valores: 1 ó 0.

Ejemplo:

Suponga la variable *región* tiene k = 4 posibles valores {north, east, south, west} por lo que es suficiente realizar k-1 transformaciones. En este ejemplo región = *west*.



21

north_flag: If *region* = *north* then *north_flag* = 1; otherwise *north_flag* = 0.

east_flag: If *region* = *east* then *east_flag* = 1; otherwise *east_flag* = 0.

south_flag: If *region* = *south* then *south_flag* = 1; otherwise *south_flag* = 0.

Imagen tomadas de Larose & Larose (2015)

Binning (contenedores) de variables numéricas

Se utiliza para categorizar valores numéricos. Por ejemplo cuando se tiene una serie de datos numéricos y se desea crear categorías de los mismos. Por ejemplo: precio de una casa, sueldo de las personas, edad, etc.

Larose & Larose (2015) establecen cuatro métodos:

22

1. El binning de igual amplitud (*Equal width binning*), divide el predictor numérico en k categorías de igual amplitud, donde k es elegido por el cliente o el analista.

2. El binning con la misma frecuencia de ocurrencia de los elementos (*Equal frequency binning*), categorías que contengan la misma cantidad de elementos, en este caso la frecuencia divide el predictor numérico en k categorías, cada una con k/n registros, donde n es el número total de registros.

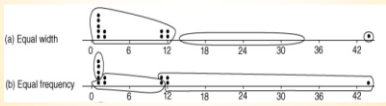


Figura 2.22 tomada de Larose & Larose (2015).

23

3. El binning por agrupamiento (*Binning by clustering*) utiliza un algoritmo de agrupamiento (cluster), como el k-means para calcular automáticamente la partición "óptima".

Figura 2.22 tomada de Larose & Larose (2015).

4. Binning basado en el valor predictivo (*Binning based on predictive value*). Los métodos (1)-(3) ignoran la variable objetivo; el binning basado en el valor predictivo particiona el valor numérico considerando el efecto de cada partición tiene sobre el valor de la variable objetivo.

24

Larose & Larose (2015) sugieren que la reclasificación de variables categóricas que contienen muchas clases se pueden volver a categorizar considerando alguna caracteristica que tengan en común, por ejemplo las entidades federativas de un país podrían categorizarse por regiones o nivel económico.

Larose & Larose (2015) sugiere **remover** variables únicas, es decir aquellas que se presentan en uno de los siguientes casos:

- Variables únicas, aquellas que tienen el mismo valor en todo el conjunto de datos, son constantes.
- Variables que son casi únicas, aquellas en las cuales la frecuencia de una de ellas predomina sobre las demás, por ejemplo suponga que en un grupo de personas el 99.95% son mujeres y el 0.05% son hombres.

Variables que probablemente no deberían eliminarse

Larose & Larose (2015) establecen que algunas variables no deberían eliminarse y que es una práctica común hacerlo. Por ejemplo los siguientes casos:

1. Variables a las que les falta el 90% de los datos.
2. Variables que están fuertemente correlacionadas

Antes de eliminar una variable porque tiene un 90% o más de valores perdidos, hay que tener en cuenta que puede existir un patrón en la falta del dato y, por lo tanto, información útil, que se puede estar descartando. Las variables que contienen 90% de valores perdidos presentan un desafío para cualquier estrategia de imputación de datos faltantes.

Por ejemplo, supongamos que tenemos un campo llamado *donation_dollars* en una base de datos de encuestas autorreportadas. Posiblemente, aquellos que donan mucho estarían inclinados a reportar sus donaciones, mientras que aquellos que no donan mucho pueden estar inclinados a saltarse esta pregunta de la encuesta.

Así, el 10% que informa no es representativo del conjunto. En este caso, puede ser preferible construir una variable de bandera, *donation_flag*, ya que hay un patrón en la falta que puede tener poder predictivo (Larose & Larose, 2015).

Referencias bibliográficas

Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). *Data Mining: concepts and techniques*. Third edition. Morgan Kaufman Series.

Larose, T. Daniel & Larose, D. Chantal. (2015). *Data Mining and Predictive Analytics*. Second Edition. Wiley.

J. Nissen, R. Donatello, & B. Van Dusen. (2019). Missing data and bias in physics education research: A case for using multiple imputation. *Physical Review Physics Education Research*.