Pre-procesamiento para Ciencias de los Datos.

Luis Alexander Calvo Valverde
Material base: M. Sc. Saúl Calderón Ramírez.

19 de febrero de 2020

Resumen

Este material está basado en el libro *Análisis de señales*, de Pablo Irarrazabal[2], y el libro de Procesamiento digital de imágenes, de Rafael González, [1].

1 Pre-procesamiento

El preprocesamiento consiste en aplicar a un conjunto de datos, representado matricialmente $X \in \mathbb{R}^{n \times m}$, una transformación:

$$X_p = T\left(X\right)$$

la cual **modifique la información en cada observación** o muestra $\overrightarrow{x}_i \in \mathbb{R}^n$, y/o **modifique la cantidad de muestras** m con el fin de mejorar el desempeño de etapas posteriores. Es por ello que entonces $X_p \in \mathbb{R}^{n \times m'}$. Observe que las técnicas de preprocesamiento preservan la dimensionalidad original de las muestras $\overrightarrow{x}_i \in \mathbb{R}^n$.

2 Preprocesamiento de registros

Muchas veces los datos a utilizar para resolver distintas preguntas en las ciencias de los datos, se conocen como **registros, tuplas** o **vectores de atributos**, los cuales hasta ahora hemos representado como arreglos $\overrightarrow{x} \in \mathbb{R}^n$:

$$\overrightarrow{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

donde entonces, cada componente, para un registro, es denominado como un **atributo** (también a veces conocido como una dimensión, característica o variable). Un atributo x_i puede tomar un valor discreto o continuo. La distribución

de un atributo o variable se le conoce como distribución univariable, y de dos atributos, bivariable, etc.

Los siguientes son algunos tipos básicos de atributos:

1. **Atributos numéricos** $x_i \in \mathbb{N}$ o $x_i \in \mathbb{R}$: Los atributos numéricos son atributos que toman una serie de valores discretos o continuos. Las operaciones matemáticas están definidas en estos atributos, tal como la moda, media y mediana y las **distancias** (ℓ_p por ejemplo). Por ejemplo atributos como *temperatura, humedad* o *fecha* pueden tomar infinidad de valores, por lo que son entonces **continuos**. Atributos numéricos **discretos** son aquellos que toman valores naturales, como por ejemplo *cantidad_clientes*.

2. **Atributos ordinales** $x_i \in \mathbb{N}$: Corresponden a atributos con un conjunto de valores los cuales pueden ser ordenados de forma significativa, aunque las magnitudes de tales valores no sea conocida.

   (a) Por ejemplo atributos como *satisfaccion_usuario* el cual puede tomar los valores *muy_satisfecho, poco_satisfecho*, o *insatisfecho*, a los cuales se les pueden asignar los valores 3,2 y 1 respectivamente.

   (b) Los atributos ordinales pueden derivarse a partir de un atributo numérico, partiendo el dominio de tal atributo en intervalos, como por ejemplo el atributo *temperatura* puede tomar los valores 0-15, 16-25, 25-50.

   (c) Para los atributos ordinales se puede calcular la moda (valor más frecuente), o mediana, con significado matemático, no así la media.

3. **Atributos binarios** $x_i \in \{0, 1\}$: Un atributo binario es un atributo cualitativo que puede tomar únicamente dos valores (Booleano), usualmente codificados como 0 o 1. Un ejemplo de atributo binario es *identificacion_nacional* el cual es 0 si la identificación es extranjera, y 1 de lo contrario. Existen dos tipos de atributos binarios:

   (a) **Atributos binarios simétricos**: son atributos cuyos valores presentan usualmente igual probabilidad. Por ejemplo el atributo de *genero* puede ser *hombre* o *mujer* con igual probabilidad usualmente.

   (b) **Atributos binarios asimétricos**: se refiere así a los atributos cuyos valores pueden presentar probabilidades distintas. Por ejemplo, la variable *hiv* puede tomar los valores $negativo = 0$ o $positivo = 1$, con lo que es usual que el valor de *negativo* sea más frecuente, con lo cual se usa asignarle el valor numérico de 0 al valor de mayor frecuencia.

4. **Atributos nominales o categóricos** $x_i \in \mathbb{N}$: son atributos cualitativos que se refieren a símbolos o nombres de elementos. Cada valor representa algún tipo de categoría, código o estado, con valores sin ningún tipo de relación de precedencia. También se les conoce como enumeraciones.

Valor de *nacionalidad*	*variable_dummy1*	*variable_dummy2*	*variable_dummy3*
1: Armenio	1	0	0
2: Ucraniano	0	1	0
3: Bielorruso	0	0	1
4: Ruso (variable de ref.)	0	0	0

Cuadro 1: Ejemplo de registros con atributo faltante.

(a) Por ejemplo atributos como la *contextura_cuerpo* puede presentar distintos valores: *delgado, grueso, muy delgado,* etc. Otro ejemplo de atributo categórico el *estado_civil* el cual puede tomar los valores de *soltero*, *casado* o *divorciado.*

Observe que aunque los distintos valores que puede tomar una variable categórica son numéricos (para la variable *estado_civil* se puede asignar $soltero = 0$, $casado = 1$, etc.), al n**o existir relación de precedencia entre los valores, estos no pueden operarse matemáticamente**. Por ejemplo, no tiene sentido calcular la media o la mediana de atributos categóricos. Además, para calcular la distancia entre dos vectores de atributos nominales $\overrightarrow{v}, \overrightarrow{w} \in \mathbb{R}^n$, se deben utilizar otras distancias como la **tasa de diferencias**:

$$d\left(\overrightarrow{v}, \overrightarrow{w}\right) = \frac{n-p}{n}$$

donde p se refiere a la cantidad de atributos nonimales distintos.

Es por ello que el manejo y representación de atributos nominales debe ser distinto, como se detallará más adelante.

2.1 Codificación de variables categóricas

Las variables categóricas al no ser posible manejarse directamente de forma numérica, deben traducirse a una representación numérica. Tómese el siguiente ejemplo en el cual se toma una variable nominal *nacionalidad,* la cual puede tomar tres valores: ruso, ucraniano, armenio y bielorruso, en este caso entonces con $n = 4$ niveles. A continuación se presenta el enfoque de **codificación tonta o dummy**

La codificación tonta *o dummy* crea una variable dicotómica o binaria por los $n-1$ valores, y para el otro valor restante, conocido como el valor de referencia, el cual se codifica con todas las otras variables dicotómicas en cero. La tabla 1 muestra la codificación tonta para la variable categórica o nominal *nacionalidad.*

El valor de referencia puede escogerse usando un criterio estadístico de las pruebas, definiendo el valor de referencia como el modelo base, con ausencia de aplicación de un *tratamiento.*

Ejemplo: Para entender mejor el efecto de la codificación *dummy,* tomese el siguiente problema: estimar el nivel de felicidad usando dos variables:

Nombre	Edad	Razon_social
Alejandro Bertinelli	15	CUIT
Luana Grifoi	_	CUIL
Betina Roca	12	CUIT

Cuadro 2: Ejemplo de registros con atributo faltante.

estimar la felicidad de una persona $y \in [0 - 100]$ usando los siguientes predictores o variables de entrada:

- Salario neto $x_1 \in \mathbb{R}$, variable numérica, con valores positivos en dólares.
- Religión $x_2 \in \mathbb{N}$: variable categórica o binaria, con dos niveles: ateo y religioso. Lo modelaremos con una sola variable *dummy* x_2

El modelo de regresión lineal a construir viene entonces dada por:

$$y\,(\overrightarrow{x}) = w_0 + w_1 x_1 + w_2 x_2$$

Observe que el modelo puede simplificarse de la siguiente forma:

$$y\,(\overrightarrow{x}) = \begin{cases} \beta_0 + w_1 x_1 & \text{con } x_2 = 1 \text{ y } \beta_0 = w_0 + w_2 \\ w_0 + w_1 x_1 & \text{con } x_2 = 0 \end{cases}$$

lo cual indica que la regresión realiza la estimación de dos modelos distintos para las muestras con variable categórica $x_2 = 0$ y $x_2 = 1$, correspondientes a dos líneas paralelas. Un modelo indicará el efecto del tratamiento *religión*, y el otro indicará la ausencia de este.

2.2 Manejo de valores faltantes

Muchas veces es posible que en un conjunto de registros como por ejemplo los ejemplificados en la siguiente tabla:

Para lidiar con los registros o tuplas faltantes, existen distintas técnicas:

1. **Ignorar las tuplas incompletas:** En circunstancias donde el porcentaje de tuplas incompletas es alto, es poco recomendable desecharlas. Los atributos incompletos pudieron ser de valor..
2. **Llenar los valores faltantes con constantes**: Por ejemplo, en atributos numéricos asignar la constante *Inf*. El uso de tales constantes puede confundir a muchos algoritmos de aprendizaje automático, cuando se calculan distancias, histogramas, etc.
3. **Usar la media, mediana o moda para rellenar tales valores**: Para atributos con valores de distribuciones no *torcidas* como la Gaussiana, se recomienda usar la media, mientras para datos con distribuciones *torcidas*

como la distribución gamma, se recomienda usar la moda. El problema de utilizar el primer momento estadístico es la alteración de la desviación estándar de los datos.

   (a) Una variante de este enfoque consiste en realizar un ***clustering*** de los datos, y asignar la media, moda o mediana del clúster correspondiente a cada registro con valor faltante.

4. **Usar el valor más probable**: realizando inferencia bayesiana o regresión, se puede estimar el valor más probable dados los valores del resto del registro, y también, del resto de los registros.

5. **Algoritmos más complejos como MICE (Multivariate imputation by chained equations)**, el cual se basa en la idea de generar múltiples muestras de una regresión múltiple secuencial, tomando en cuenta que una muestra es generada de un fenómeno aleatorio, por lo que es una mejor aproximación generar múltiples muestras.

2.3 Transformaciones

Muchas veces, los datos originales pueden necesitar de funciones complejas para realizar ya sea una regresión o clasificación de los mismos, o incluso encajar alguna distribución de probabilidad. Por ejemplo, en la Figura 1, se muestra un conjunto de datos, para el cual es necesario realizar una regresión. Es fácil notar que un modelo lineal no logrará un error bajo. Sin embargo, dado que los datos parecen tener un comportamiento logarítmico, al aplicar la inversa de tal función a los datos y realizar la regresión en tales datos transformados, el modelo lineal es suficiente para lograr una regresión de error bajo. Tal función exponencial inversa no es más que una transformación

$$T(X)$$

En general, las técnicas de preprocesamiento a estudiar para señales e imágenes, tienen el objetivo de facilitar la extracción de características, clasificación o regresión, tal cual se muestra en la Figura 1.

2.4 Balanceo de datos

Un conjunto de datos desbalanceado, se refiere a un conjunto de datos $X \in \mathbb{R}^{n \times m}$, para el cual existen muchas más muestras de una clase respecto a las demás. Por ejemplo, en un problema de clasificación binaria, si las muestras de la clase C_1 se almacenan en la matriz $X \in \mathbb{R}^{n \times m_1}$, y las muestras de la clase C_2 están almacenadas en $X \in \mathbb{R}^{n \times m_2}$, se dice que el conjunto de datos está desbalanceado si $m_1 \gg m_2$ (o $m_2 \gg m_1$). Al realizar la implementación de un algoritmo de clasificación como mínimos cuadrados, los errores de ambas clases serían igualmente penados, por lo que es probable que el clasificador al que se arribe esté sesgado hacia la clase con mayor cantidad de muestras.

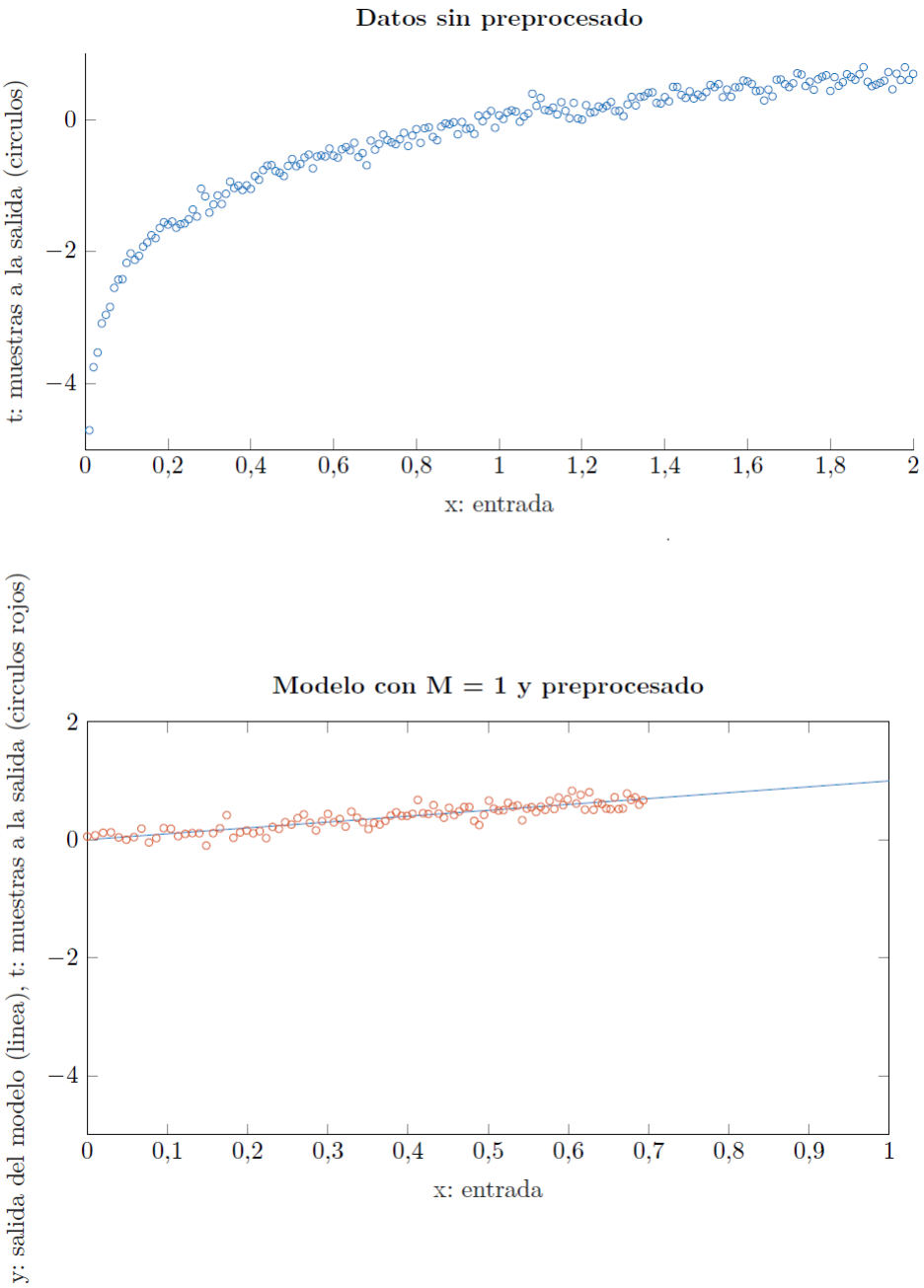

Figura 1: Regresión sin y con datos preprocesados.

Para corregir tal problema existen distintas técnicas posibles a implementar, manipulando el conjunto de datos X (suponga que $m_1 \gg m_2$):

1. **Sobre-muestreo u *oversampling:*** Consiste en generar las muestras faltantes para el conjunto de datos, repitiendo las muestras de la clase con menos muestras (C_2 en este caso), tantas veces sea necesario y de forma aleatoria.

2. **Sub-muestreo o *subsampling***: Elimina las muestras sobrantes de la clase con mayor número de muestras, en este caso m_1.

3. **Generación de nuevas muestras usando aprendizaje generativo**: El **aprendizaje generativo** se enfoca en aprender la distribución de los datos, a diferencia del aprendizaje discriminativo, el cual tiene por objetivo construír un modelo que disminuya el error de clasificación o discriminación.

Otro enfoque, no basado en la generación o eliminacion de nuevas muestras, consiste en modificar la función de pérdida del modelo para *castigar* con mayor fuerza los errores con muestras de las clases con menos muestras. Sin embargo tal enfoque no corresponde al de una etapa de preprocesamiento.

Referencias

[1] Rafael C Gonzalez, Richard Eugene Woods, and Steven L Eddins. *Digital image processing using MATLAB*. Pearson Education India, 2004.

[2] Pablo Irarrázaval. *Análisis de señales*. McGraw-Hill Interamericana, 1999.