Laboratorio de Datos

Preprocesamiento: escalamiento y variables categóricas

Primer Cuatrimestre 2025

Turnos noche

Facultad de Ciencias Exactas y Naturales, UBA

Preprocesamiento

En muchos casos, necesitamos realizar algunas transformaciones previas en los datos antes de ajustar un modelo de regresión (o cualquier modelo cuantitativo). Estas transformaciones pueden ser:

1. Transformar variables categoricas binarias a 0-1.
2. Transformar variables categóricas con varias categorías.
3. Normalizar variables numéricas, escalándolas para que los valores caigan en el intervalo [0,1]
4. Normalizar variables numéricas llevándolas a media 0 y varianza 1.

Escalamiento min-max

Este escalamiento es una transformación lineal. Es decir, se aplica una fórmula de la forma

$$x_{\text{nuevo}} = a \cdot x + b$$

Los valores de a y b se eligen de forma que los valores de la variable transformada se encuentren en el intervalo [0,1].

La fórmula que debemos aplicar es

$$x_{\text{nuevo}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

El escalamiento MinMax lleva a que todas las variables tengan una dispersión similar y por lo tanto podamos comparar el peso de cada variable en el modelo mirando los coeficientes.

Escalamiento min-max

¿Cuáles pueden ser las ventajas de este escalamiento, comparado con solamente dividir con el máximo?

Escalamiento min-max

¿Cuáles pueden ser las ventajas de este escalamiento, comparado con solamente dividir con el máximo?

Ejemplo. Si una variable tiene todos valores entre 390 y 400, si solo dividimos por el máximo quedarán valores entre 0.975 y 1.

Para que esas pequeñas diferencias impacten en el modelo, necesitaremos multiplicar la variable por un coeficiente alto.

Escalamiento min-max

Supongamos que

- x_1 tiene valores en el intervalo $[0, 1]$
- x_2 tiene valores en el intervalo $[0.975, 1]$

En un modelo

$$y = 4x_1 + 4x_2$$

la variable x_1 tiene más peso. Las observaciones con puntaje alto van a ser las que tengan un valor de x_1 alto, las variaciones en x_2 van a tener un aporte muy menor.

Escalamiento min-max

Otro ejemplo: En muchos concursos docentes, la prueba de oposición tiene un puntaje alto (50 o 70 puntos sobre 100), y por lo tanto a priori es la parte más importante del concurso.

Sin embargo en muchos concursos los valores son similares, la mayoría de los concursantes obtiene notas entre 60 y 70. Por lo tanto el peso de la prueba de oposición en el concurso termina siendo mucho menor.

Escalamiento min-max

Pregunta: ¿En qué situación este escalamiento podría resultar inadecuado?

Escalamiento min-max

Pregunta: ¿En qué situación este escalamiento podría resultar inadecuado?

Este escalamiento se ve muy influenciado por la presencia de outliers.

Si tenemos outliers en una variable, llevaremos el outlier a 1, y los demás valores a valores muy pequeños.

Ejemplo: infecciones en hospitales

Variable Number	Variable Name	Description
1	Identification number	1–113
2	Length of stay	Average length of stay of all patients in hospital (in days)
3	Age	Average age of patients (in years)
4	Infection risk	Average estimated probability of acquiring infection in hospital (in percent)
5	Routine culturing ratio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
6	Routine chest X-ray ratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
7	Number of beds	Average number of beds in hospital during study period
8	Medical school affiliation	1 = Yes, 2 = No
9	Region	Geographic region, where: 1 = NE, 2 = NC, 3 = S, 4 = W
10	Average daily census	Average number of patients in hospital per day during study period
11	Number of nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number full time plus one half the number part time)
12	Available facilities and services	Percent of 35 potential facilities and services that are provided by the hospital

Reference: Special Issue, "The SENIC Project," *American Journal of Epidemiology* **111** (1980), pp. 465–653. Data obtained from Robert W. Haley, M.D., Hospital Infections Program, Center for Infectious Diseases, Centers for Disease Control, Atlanta, Georgia 30333.

2. Escalamiento estándar

Otro escalamiento muy común es llevar las variables a media 0 y varianza 1.

Para eso aplicamos la fórmula:

$$\tilde{x} = \frac{x - \bar{x}}{\sigma},$$

donde $\bar{x}$ es la media y σ es el desvío estándar.

Escalamiento estándar

Las principales ventajas son:

1. Menor sensibilidad que min-max (aunque no inmune) a outliers.
2. La variabilidad explicada por cada variable se puede ver directamente comparando coeficientes.

Posibles ventajas de min-max:

1. Es más simple.
2. No introduce valores negativos (que pueden traer problemas si aplicamos logaritmos u otras funciones).
3. En algunos casos, los valores de una variable se mueven en un rango prefijado y solo queremos ajustar el rango, por ejemplo llevar variables 0-100 a variables 0-1.

Codificación de Variables Binarias

- Variables con dos valores posibles, por ejemplo: `sí/no`, `verdadero/falso`.
- Codificación habitual:
  - `sí` $\rightarrow 1$
  - `no` $\rightarrow 0$
- Ejemplo:

Respuesta	Codificada
Sí	1
No	0
Sí	1

Codificación de Variables Categóricas

- Variables con más de dos categorías, por ejemplo: `rojo, verde, azul`.
- Codificación mediante **dummies** (variables indicadoras):
  - Una columna por cada categoría (o $k-1$ para evitar multicolinealidad).
- Ejemplo:

Cliente	Color de auto	Rojo	Verde	Azul
Juan	Rojo	1	0	0
Margarita	Verde	0	1	0
Ana	Azul	0	0	1
Pedro	Rojo	1	0	0

La última columna es redundante, se puede deducir de las anteriores y por lo tanto se puede eliminar. Si el modelo tiene intercept, habría dependencia lineal entre las variables.