

## Codificación de variables categóricas

Dr. Carlos Augusto Arellano Muro

Primavera 2022

- Las categorías se pueden representar numéricamente.
- Los valores de una variable categórica no se pueden ordenar entre sí.

Es tentador simplemente asignar un número entero, digamos de 1 a  $k$ , a cada una de las  $k$  categorías posibles, pero los valores resultantes podrían ordenarse entre sí, lo que no debería ser permisible para las categorías.

## 1. Creación de variables binarias

### 1.1. One-Hot

- Cada bit representa una categoría posible. Siendo '0' si no pertenece o '1' si lo sí lo hace.
- Si la variable no puede pertenecer a varias categorías a la vez, solo un bit del grupo puede ser 1. Esto se denomina codificación one-hot.
- Una variable categórica con  $k$  categorías posibles se codifica como un vector de características de longitud  $k$ .
- Se crea una dependencia lineal. Debido a que una de las nuevas variables debe ser 1, el algoritmo usa una variable más de la estrictamente necesaria, es decir, **la suma de todos los bits debe ser igual a uno**, lo que genera una dependencia lineal entre características.
- Permite múltiples modelos válidos para el mismo problema.
- Los datos faltantes se pueden codificar como el vector de todos ceros.

Tabla 1: Muestra de variables categóricas.

	Ciudad	Precio de renta
0	Guadalajara	4000
1	Zapopan	3000
2	Zapopan	6000
3	Guadalajara	3500
4	Ciudad de México	9000
5	Monterrey	6000
6	Ciudad de México	18000
7	Zapopan	12000

Tabla 2: Datos codificados usando One-Hot.

	Zapopan	Guadalajara	Ciudad de México	Monterrey
0	0	1	0	0
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	1	0	0	0

## 1.2. Codificación ficticia

- Da lugar a modelos únicos e interpretables.
- Una característica se coloca debajo del bus y se representa mediante el vector de todos los ceros (Categoría de referencia).
- No puede manejar fácilmente los datos faltantes.

Tabla 3: Muestra de variables categóricas.

	Ciudad	Precio de renta
0	Guadalajara	4000
1	Zapopan	3000
2	Zapopan	6000
3	Guadalajara	3500
4	Ciudad de México	9000
5	Monterrey	6000
6	Ciudad de México	18000
7	Zapopan	12000

Tabla 4: Datos codificados usando variable ficticia.

	Zapopan	Ciudad de México	Monterrey
0	0	0	0
1	1	0	0
2	1	0	0
3	0	0	0
4	0	1	0
5	0	0	1
6	0	1	0
7	1	0	0

La categoría de Guadalajara, en los índices 0 y 3 se codifica como  $\{0,0,0\}$ .

## 1.3. Codificación de efectos

- la categoría de referencia ahora está representada por el vector cuyos elementos son todos  $-1$ .
- Da lugar a modelos únicos e interpretables.
- El vector de todos los  $-1$  es un vector denso, que resulta costoso tanto para el almacenamiento como para el cálculo.

Tabla 5: Muestra de variables categóricas.

	Ciudad	Precio de renta
0	Guadalajara	4000
1	Zapopan	3000
2	Zapopan	6000
3	Guadalajara	3500
4	Ciudad de México	9000
5	Monterrey	6000
6	Ciudad de México	18000
7	Zapopan	12000

Tabla 6: Codificación de efectos.

	Zapopan	Ciudad de México	Monterrey
0	-1	-1	-1
1	1	0	0
2	1	0	0
3	-1	-1	-1
4	0	1	0
5	0	0	1
6	0	1	0
7	1	0	0

## 1.4. Regresión lineal

Para ver las diferencias entre cada codificación, se hace una regresión lineal para cada punto:

- A cada categoría se le asigna una variable  $x_1, x_2, \dots, x_n$ , siguiendo el orden ‘Zapopan’, ‘Guadalajara’, ‘Ciudad de México’, etc., según sea el caso.
- la respuesta de la regresión

$$y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + i$$

corresponde a la variable numérica ‘Precio de renta’.

- Se añade también el término  $i$  (intersección) para evitar una salida en cero cuando todas las variables son cero.

- Se desea encontrar los coeficientes  $\{a_1, a_2, \dots, a_n\}$  que, al introducir, por ejemplo  $\{1, 0, 0, 0\}$  en la ecuación que represente a la codificación One-Hot, arroje como resultado el estimado de renta para ‘Zapopan’.

Los coeficientes  $\{a_1, a_2, \dots, a_n\}$  y el valor de intersección  $i$  se observan en la Tabla 7 Como se puede ver,

Tabla 7: Comparación entre codificadores.

	One-Hot	C. Ficticia	C. De efectos
$a$	$[-562.5, -3812.5, 5937.5, -1562.5]$	$[3250, 9750, 2250]$	$[-562.5, 5937.5, -1562.5]$
$i$	7562.5	3750	7562.5

los resultados son fácilmente interpretables. Para el caso de One-Hot, la intersección es el promedio global de las rentas, mientras que los coeficientes son las diferencias entre el promedio global y el promedio de las rentas en cada ciudad. Para la codificación ficticia, la intersección es el promedio de las rentas de la variable de referencia ‘Guadalajara’, de igual forma, los coeficientes son la diferencia entre los promedios de cada categoría y la intersección. Por último, en la codificación de efectos, el termino de intersección es el promedio de todos los promedios individuales y los coeficientes individuales indican cuánto difieren las medias de las categorías individuales de la intersección (Esto se llama **efecto principal** de la categoría, de ahí el nombre “codificación de efecto”). Ninguna característica única representa la categoría de referencia, por lo que el efecto de la categoría de referencia debe calcularse por separado como la suma negativa de los coeficientes de todas las demás categorías.

## 2. Variables categóricas ordinales

- Pueden codificarse asignando un número a cada categoría mas se pierde una posible relación relativa.
- Contraste polinómico: relación polinomial entre categorías.
  - Sus coeficientes deben sumar cero.
    - Permite la comparación de diferentes tratamientos [1].
  - Se limita a  $N - 1$  grados del polinomio (siendo  $N$  el número de categorías diferentes).
  - Se pueden usar simultáneamente en el mismo modelo.
    - Se requiere que las representaciones numéricas sean ortogonales.
- Algunos coeficientes de contraste polinómicos son los presentados en la Tabla 8.

Tabla 8: Algunos coeficientes de contraste polinómico.

Polinomio	3 categorías	4 categorías	5 categorías
Lineal	$[-1, 0, 1]$	$[-3, -1, 1, 3]$	$[-2, -1, 0, 1, 2]$
Cuadrático	$[1, -2, 1]$	$[1, -1, -1, 1]$	$[2, -1, -2, -1, 2]$
Cúbico	–	$[-1, 3, -3, 1]$	$[-1, 2, 0, -2, 1]$

- Algunos coeficientes de polinomios ortonormales son los presentados en la Tabla 9.

Tabla 9: Algunos coeficientes de polinomios ortonormales.

Polinomio	
Lineal	$[-0.71, 0.0, 0.71]$
Cuadrático	$[0.41, -0.82, 0.41]$

- Tratar las categorías como desordenadas.
  - Si el patrón subyacente es polinómico, puede no adaptarse el predictor a estos datos.
- Calificar las características.
  - Por ejemplo, puntuar del 1 al 10 una falla.

### 3. Conteos o frecuencias de categorías

- Requieren independencia entre atributos.
- Representan la probabilidad de que ocurra el suceso y no el suceso en sí.
  - Se necesitan datos de sucesos anteriores.
- Se representa con el número de veces de aparición.
  - Comúnmente se hacen *hashes* independientes para los conteos menos comunes.

#### 3.1. Bolsa-de-X

- Es deseable que el resultado sea simple y fácilmente interpretable.
- Se puede codificar cada palabra, una frase común, una acción, etc.

Bolsa-de-Palabras (Bag-of-Words BoW)

- Una lista de recuentos de palabras. Una palabra mencionada muchas veces en un texto puede ser un indicador de la importancia de ésta.
- No tiene una secuencia de elementos, simplemente recuerda cuántas veces aparece cada palabra en el texto.

Tabla 10: Vector generado por codificación BoW.

"es un cachorro y es extremadamente lindo"	es	2
	son	0
	cachorro	1
	gato	0
	y	1
	extremadamente	1
	lindo	1
	...	

- No representa ningún concepto de jerarquía de palabras. Por ejemplo, el concepto de "animal" incluye "perro", "gato", "cuervo", etc., pero en una representación de bolsa-de-palabras son todos elementos iguales.
- Cada palabra representa un eje en el espacio de características
- Siguiendo este hilo, cada vector de palabras, como el de la Tabla 3.1, puede representar un eje y el recuento de este vector sería una Bolsa-de-Documentos.

Bolsa-de-n-Gramas

- Un n-Grama. Es una secuencia de n-tokens. Una palabra es esencialmente un 1-grama o unigrama.

Tabla 11: Vector generado por tokenización de 2-Gramas.

"es un cachorro y es extremadamente lindo"	es un	1
	un cachorro	1
	cachorro y	1
	y es	1
	es extremadamente	1
	extremadamente lindo	1

- Es más probable que resulten elementos únicos entre más se aumenta el valor de 'n', lo que se traduce en mayor costo computacional.

## 4. Codificación en base a la media

- Usar el promedio o mediana de la respuesta de la categoría.

Tabla 12: Muestra de variables categóricas.

	Ciudad	Precio de renta
0	Guadalajara	4000
1	Zapopan	3000
2	Zapopan	6000
3	Guadalajara	3500
4	Ciudad de México	9000
5	Monterrey	6000
6	Ciudad de México	18000
7	Zapopan	12000

Tabla 13: Codificación usando el promedio.

Característica original	Codificación
Zapopan	7000
Guadalajara	3750
Ciudad de México	13000
Monterrey	6000

- Usar el efecto principal en la respuesta de codificación de efectos.

Tabla 14: Codificación usando el efecto de la codificación de efectos.

Característica original	Codificación
Zapopan	-562.5
Guadalajara	7562.5
Ciudad de México	5937.5
Monterrey	-1562.5

- Codificar la categorías con el algoritmo de Codificación de efectos.
- Realizar una regresión lineal (o logística).
- Usar los coeficientes correspondientes de cada categoría como codificación.

## 5. Feature hashing

Es una forma de combinar categorías de un conjunto de valores a otro más pequeño.

Las características son identificadas con un valor (ID) que se obtiene de un modelo, ya sea caracterización de color, de textura o de forma, como también de codificación como ASCII, UTF-8 o Unicode. A estos valores de identificación se les conoce como *keys*. La **función hash** se construye para cualquier objeto que se pueda representar computacionalmente.

Función hash: Es una función determinista que asigna un entero potencialmente ilimitado a un rango de enteros finito  $[1, m]$ .

$$h : \mathbb{N} \rightarrow \Omega, \quad \Omega = [1, m] \subset \mathbb{N}$$

Función hash uniforme: Asegura que se mapee aproximadamente el mismo número de números en cada uno de los  $m$  contenedores.

- Es unidireccional. Una vez que se ejecuta la función hash, no es posible obtener las categorías de regreso.
- Es posible que se asignen varios números a la misma salida (Colisión).
- El Feature hashing comprime el vector de características original en un vector  $m$ -dimensional aplicando una función hash al ID de la característica *key*.

Tabla 15: Muestra de variables categóricas.

Ciudad	ID
Zapopan	582753783
Guadalajara	1166288024
Ciudad de México	-157684639
Monterrey	-1267876914

Tabla 16: Hashes usando 16 y 256 columnas.

Característica original	Codificación	
	$m = 16$	$m = 256$
Zapopan	8	248
Guadalajara	9	153
Ciudad de México	2	98
Monterrey	15	207

## Referencias

- [1] [https://en.wikipedia.org/wiki/Contrast\\_\(statistics\)](https://en.wikipedia.org/wiki/Contrast_(statistics))