ITESO, Universidad Jesuita de Guadalajara

Departamento de Matemáticas y Física

Ingeniería de características

Codificación de variables categóricas

Dr. Gaddiel Desirena López

Primavera 2024

Para poder representar una variable categórica computacionalmente, es necesario que cada elemento diferente de ésta (llámese a este elemento categoría) sea numérico. Sin embargo, para conservar lo más posible la esencia de la variable, es decir, que no tenga sentido hablar de un mapeo entre una categoría (o variable categórica) y otra; establecemos dos puntos a considerar en la codificación:

- Las categorías se deben representar numéricamente.
- Si no existe una correlación entre categorías, ésta no se debe generar.

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 todas las variables.

Si solo nos importa codificar que, cada categoría es diferente a la otra, una solución es crear variables Booleanas que indiquen la pertenencia a dicha categoría.

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.
- Debido a la dependencia lineal, permite múltiples modelos válidos para el mismo problema.
- Los datos faltantes se pueden codificar como el vector de ceros para todas las variables creadas.

Por ejemplo, en la Tabla 1 se muestra la variable categórica "Ciudad", donde solo nos importa codificar si las categorías son diferentes entre sí; además una variable numérica con la cual, a través de un modelo de regresión lineal, se quiere predecir el precio de renta basado en la ciudad, por lo que la variable categórica debe codificarse para que ésta sea una entrada numérica en el modelo, justo como la Tabla 2.

Tabla 1: Muestra de variables categóricas.

	Ciudad	Precio de renta
0	Guadalajara	4000
1	Zapopan	3000
2	Zapopan	6000
3	Guadajalara	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.

De forma análoga a la codificación One–Hot, la codificación se muestra en la Tabla 4.

Tabla 3: Muestra de variables categóricas.

	Ciudad	Precio de renta
0	Guadalajara	4000
1	Zapopan	3000
2	Zapopan	6000
3	Guadajalara	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 referencia, 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.

Así mismo, para la regresión lineal y comparar las tres codificaciones de Creación de variables binarias, la codificación propia se muestra en la Tabla 6.

Tabla 5: Muestra de variables categóricas.

	Ciudad	Precio de renta
0	Guadalajara	4000
1	Zapopan	3000
2	Zapopan	6000
3	Guadajalara	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 $x_1, x_2 \ldots, x_n$, 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 $\{a_1, a_2, \ldots, a_n\}$ 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 $\{a_1, a_2, \ldots, a_n\}$ 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, esto es el promedio de las proyecciones de los puntos donde el hiperplano toca el eje de cada categoría en 1, en otra palabras, el promedio de los promedios de las rentas de cada categoría; mientras que los coeficientes son las diferencias entre el promedio de las rentas en cada ciudad y el promedio global.
- Para la codificación ficticia, la intersección es el promedio de las rentas de la categoría 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 [4].

    Si tenemos mas de una variable categorica no correlacionada, lo ideal es que esta independencia se mantenga, es decir, que no sea posible encontrar una solucion para a_0 y a_1, al problema

    $$a_0 + a_1 X_1 = X_2 + \varepsilon$$

    siendo X_1 y X_2 las variables a codificar. Esto es, para cada categoria de $X_1 = \{x_{11}, x_{12}, \ldots, x_{1n}\}$ y $X_2 = \{x_{21}, x_{22}, \ldots, x_{2n}\}$, se genera un sistema de ecuaciones de la siguiente forma

    $$\begin{aligned} a_0 + a_1 x_{11} &= x_{21} + \varepsilon_1 \\ a_0 + a_1 x_{12} &= x_{22} + \varepsilon_2 \\ &\vdots \\ a_0 + a_1 x_{1n} &= x_{2n} + \varepsilon_n \end{aligned}$$

    resolviendo para a_0 y a_1 de forma que se minimice $\varepsilon = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T$, se ve

    $$n a_0 + a_1(x_{11} + x_{12} + \cdots + x_{1n}) = (x_{21} + x_{22} + \cdots + x_{2n}) + \sum_{i=1}^{n} \varepsilon_i,$$

    por esta razon, como los coeficientes estan multiplicados por cero, no es posible encontrar una solucion.
  - 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.

2.1. Creacion de coeficientes para el contraste polinomico

La codificacion debe satisfacer un polinomio del tipo $a_0 + a_1 x_i + a_2 x_i^2 + \ldots$, donde x_i es la categoria i a codificar; de manera que al sustituirlas, la suma de los valores codificados sea cero.

Para cumplir este objetivo, se impone la condicion de que las categorias se distribuyan de forma equitativa para los valores positivos como negativos. Esto es, si la cardinalidad de X es par, la mitad de las categorias son positivas y la otra mitad negativas; si es impar, una categoria se le asigna el valor cero, al resto se distribuye de igual forma que el proceso anterior. Entonces los valores numericos para la codificacion se pueden asignar de la siguiente forma

$$X = \begin{cases} -(n-1) : 2 : n-1, & n = par \\ -\dfrac{n-1}{2} : 1 : \dfrac{n-1}{2}, & n = impar, \end{cases}$$

la primera ecuacion parte de $-(n-1)$, da saltos de 2 en 2 y finaliza en $n-1$; la segunda inicia en $-\frac{n-1}{2}$, da saltos de 1 en 1 y termina en $\frac{n-1}{2}$.

Sustituyendo estos valores numericos en el polinomio deseado para todas las categorias, se impone la restriccion de que la suma sea cero; para polinomios impares $a_0 = 0$, para polinomios pares, los coeficientes impares se simplifican. Por ejemplo:

Para el polinomio de primer grado

$$\begin{array}{c|c} n = 3 & n = 4 \\ \begin{array}{ccc} X = \{-1 & 0 & 1\} \\ a_0 + a_1(-1) & a_0 + a_1(0) & a_0 + a_1(1) \\ a_0 - a_1 & a_0 & a_0 + a_1 \end{array} & \begin{array}{cccc} \text{X=\{-3} & \text{-1} & 1 & 3\} \\ a_0 + a_1(-3) & a_0 + a_1(-1) & a_0 + a_1(1) & a_0 + a_1(3) \\ a_0 - 3a_1 & a_0 - a_1 & a_0 + a_1 & a_0 + 3a_1 \end{array} \\ a_0 - \cancel{a_1} + a_0 + a_0 + \cancel{a_1} = 0 & a_0 - \cancel{3a_1} + a_0 - \cancel{a_1} + a_0 + \cancel{a_1} + a_0 + \cancel{3a_1} = 0 \\ a_0 = 0,\ a_1 \neq 0 & a_0 = 0,\ a_1 \neq 0 \end{array}$$

Para el polinomio de segundo grado

$$n = 3$$

$$\begin{array}{ccc} X = \{-1 & 0 & 1\} \\ a_0 + a_1(-1) + a_2(-1)^2 & a_0 + a_1(0) + a_2(0)^2 & a_0 + a_1(1) + a_2(1)^2 \\ a_0 - a_1 + a_2 & a_0 & a_0 + a_1 + a_2 \end{array}$$

$$a_0 - \cancel{a_1} + a_2 + a_0 + a_0 + \cancel{a_1} + a_2 = 0$$

$$n a_0 + a_2 \sum_{i=X} i^2 = 0$$

$$n = 4$$

$$\begin{array}{cccc} \text{X=\{-3} & \text{-1} & 1 & 3\} \\ a_0 + a_1(-3) + a_2(-3)^2 & a_0 + a_1(-1) + a_2(-1)^2 & a_0 + a_1(1) + a_2(1)^2 & a_0 + a_1(3) + a_2(-3)^2 \\ a_0 - 3a_1 + 9a_2 & a_0 - a_1 + a_2 & a_0 + a_1 + a_2 & a_0 + 3a_1 + 9a_2 \end{array}$$

$$a_0 - \cancel{3a_1} + 9a_2 + a_0 - \cancel{a_1} + a_2 + a_0 + \cancel{a_1} + a_2 + a_0 + \cancel{3a_1} + 9a_2 = 0$$

$$n a_0 + a_2 \sum_{i=X} i^2 = 0$$

En en todos los casos, el coeficiente mas significativo debe ser diferente de cero.

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	Guadajalara	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 $[1, m]$.

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

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: Diccionario de codificación para las variables categóricas.

Ciudad	ID
A	228
B	214
C	112
D	221
E	110
F	14
G	110
H	101
I	28
J	11

Tabla 16: Hashes usando la función MurmurHash3 de 32 bits [6].

ID	Mapeo	
	$m = 4$	$m = 7$
228	0	96
214	99	-11
112	249	0
221	121	228
110	0	121
14	99	35
110	249	0
101	121	96
28	0	-11
11	99	0

5.1. Criterio de selección de categorías

El principio de Pareto establece que, para muchos resultados, el 80 % de las consecuencias provienen del 20 % de las causas. Este principio ayuda a identificar el orden de importancia, por ejemplo: Puede referirse al impacto financiero de un problema o el número relativo de ocurrencia de éste [5].

1. Determinar cómo se juzgará la importancia relativa. Es decir, si se deberá basar sobre una variable numérica dependiente o sobre la frecuencia de ocurrencia.
2. Ordenar las categorías.
3. Calcular la frecuencia acumulativa de las categorías en el orden seleccionado.
4. Si el 20 % de las categorías están presentes en el 80 % de esta acumulación, entonces existe un excedente de categorías.

Referencias

[1] Kuhn M. Johnson K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models. Chapman and Hall/CRC Press, pp. 93–120.

[2] Galli, S. (2022). Python feature engineering cookbook: over 70 recipes for creating, engineering, and transforming features to build machine learning models. Packt Publishing Ltd., pp. 92–139.

[3] Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc.", pp. 77–97.

[4] `https://en.wikipedia.org/wiki/Contrast_(statistics)`

[5] `http://red.unal.edu.co/cursos/ciencias/2001065/html/un1/cont_120_20.html`

[6] `https://github.com/aappleby/smhasher`