



Codificación de variables categóricas

.....

Dr. Gaddiel Desirena López

Creación de variables binarias

Variables categóricas ordinales

Conteos o frecuencias de categorías

Codificación en base a la media

Feature hashing

One-Hot

- ▶ Cada bit representa una categoría posible.
- ▶ Una variable categórica con k categorías posibles se codifica como un vector de características de longitud k .
- ▶ Se crea una dependencia lineal. Esto permite múltiples modelos válidos para el mismo problema.
- ▶ Los datos faltantes se pueden codificar como el vector de todos ceros.

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.

Codificación de efectos

- ▶ la categoría de referencia ahora está representada por el vector cuyos elementos son todos -1 .
- ▶ Resulta costoso tanto para el almacenamiento como para el cálculo.

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.
 - ▶ Se limita a $N - 1$ grados del polinomio.
 - ▶ Se pueden usar simultáneamente en el mismo modelo.
 - ▶ Se requiere que las representaciones numéricas sean ortogonales.
- ▶ 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.

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.

Codificación en base a la media

- ▶ Usar el promedio o mediana de la respuesta de la categoría.
- ▶ Usar el efecto principal en la respuesta de codificación de efectos.
 - ▶ 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.

Feature hashing

- ▶ Es una forma de combinar categorías de un conjunto de valores a otro más pequeño.
- ▶ Las características son identificadas con un valor llamado *keys*.
- ▶ A cada key se le aplica una función hash.

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

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

- ▶ La función hash es unidireccional.
- ▶ 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.