UNIVERSITAS Miguel Hernández

Apendice 1. Variables Dummies

José L. Sainz-Pardo Auñón

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones.

Variables Dummies

- **Definición**:
  - Las variables dummies son variables binarias (0 o 1) que se utilizan para representar categorías en modelos estadísticos y de machine learning.
  - Permiten incluir variables categóricas en modelos que requieren datos numéricos.

Tipos de Variables Dummies

- **Dummies Binarias**: Representan dos categorías (por ejemplo, "Sí" o "No").
- **Dummies de Múltiples Clases**: Representan más de dos categorías. Por ejemplo, para la variable "Color" con valores "Rojo", "Verde" y "Azul", se crearían tres variables:
  - D_{Rojo}
  - D_{Verde}
  - D_{Azul}

Ejemplo de Codificación con Redundancia

- Supongamos la variable categórica "Color" con tres valores: Rojo, Verde, Azul.
- Podemos crear una variable dummy para cada categoría:

Color	D_{Rojo}	D_{Verde}	D_{Azul}
Rojo	1	0	0
Verde	0	1	0
Azul	0	0	1

- Cada color se representa con una combinación única de las tres variables dummies.

Ejemplo de Codificación sin Redundancia

- En algunos modelos, una de las variables es redundante. Podemos codificar usando solo dos variables dummies:

Color	D_{Rojo}	D_{Verde}
Rojo	1	0
Verde	0	1
Azul	0	0

- En este caso, el valor "Azul" es implícito cuando ambas variables dummies son 0.
- Esta codificación elimina la redundancia, ya que una tercera variable sería linealmente dependiente de las otras dos.

Eliminación de Redundancia en Variables Dummies

Es imprescindible eliminar la redundancia en:

- **Regresión lineal** y **logística**: Evita la **multicolinealidad**, que causa inestabilidad en los coeficientes y dificultades en la interpretación.
- **Análisis Discriminante Lineal (LDA)**: También sensible a la colinealidad.

No es necesario eliminar la redundancia en:

- **Árboles de decisión**, **Random Forests**, **SVM**, **k-NN**, **Redes Neuronales**.
- Estos algoritmos no se ven afectados por la colinealidad y seleccionan las variables más relevantes automáticamente.

Dado que:

- eliminar o no (cuando no sea imprescindible) las variables redundantes finalmente produce resultados similares;
- es más eficiente eliminar la redundancia en términos de tiempo de computación y memoria;

resulta siempre **recomendable eliminar las variables redudantes**.

Formas de Crear Variables Dummies

Método Manual: Crear las columnas manualmente y asignar 0 o 1 según corresponda. Supongamos que tenemos una columna 'Color' y queremos crear una variable dummy manualmente para 'Rojo':

```
df['Rojo_dummy'] = df['Color'].apply(lambda x: 1 if x
    == 'Rojo' else 0)
```

Usando librerías:

- Sin eliminar redundancia:

```
1 dummies = pd.get_dummies(df['Color'])
```

- Eliminando redundancia:

```
1 dummies_drop_first = pd.get_dummies(df['Color'],
    drop_first=True)
```

Cuándo Usar Variables Dummies

- **Modelos de Regresión**: Cuando se incluyen variables categóricas en modelos de regresión lineal o logística.
- **Árboles de Decisión**: Para representar categorizaciones en modelos de árboles de decisión y ensemble.
- **Redes Neuronales**: Para convertir categorías en un formato numérico que pueda ser procesado por la red.
- **Machine Learning**: En algoritmos como kNN, SVM, y otros que requieren entradas numéricas.