5-1-2025

Captura y Preparación de Datos.

PEC4

Juan Luis Acebal Rico
GRADO DE CIENCIA DE DATOS APLICADA

Indice

Parte teórica

Q1.

¿Qué diferencia hay entre selección o extracción de características y reducción de la dimensionalidad?

Selección de características: se elige un subconjunto de las variables originales, sin modificar su significado (de ellas, aunque puede que del dataset ya que tiene menos columnas), por ejemplo si tenemos una variable que es booleana que representa el sexo pero luego tenemos una variable que representa sexo pero tambien LGTBI, podriamos quedarnos con la segunda (o la primera, depende el contexto) ya que tiene informacion repetida.

Extracción de características: se generan nuevas variables (transformaciones o combinaciones de las originales), frecuentemente con el objetivo de reducir la dimensionalidad, por ejemplo: patrimonio (que seria la columna activos - pasivos).

Reducción de la dimensionalidad: bajar el número de variables (dimensiones) mediante transformaciones que capturan la mayor parte de la información original (PCA es la mas conocida aunque tambien hemos visto SVD, LDA,…).

Q2.

En los métodos de selección de características, ¿En qué se diferencian los métodos de envoltura de los métodos embebidos?

Métodos de envoltura (wrapper): Entrenan un modelo buscando muchsa combinaciones de caracteristicas.

Métodos embebidos (embedded): La selección de características sucede dentro del proceso de entrenamiento del modelo.

¿Y los filtros de los métodos embebidos?

Filtros: aplican criterios estadísticos (correlación, chi-cuadrado, etc) sin depender de un modelo concreto.

Embebidos: la importancia de las características se determina dentro del propio entrenamiento del modelo.

Q3.

¿Por qué se recomienda ser más conservadores con los filtros en el proceso de extracción de características?

Los filtros descartan variables basándose en métricas más o menos simples, que además, por ejemplo, tengan poca correlación con una variable objetivo, pero luego puede darse el caso que en combinación con otras, SI que tengan relación con la variable objetivo.

¿Cómo funcionan los filtros?

Se calcula una puntuación para cada característica que luego se puede seleccionar por un ranking o por un umbral

¿Sirven los mismos filtros para un problema de regresión y de clasificación?

Si, pero se usan distintos puntuajes, por ejemplo, Pearson, y chi cuadrado para regresión y clasificación respectivamente.

Q4.

¿Cuáles son los métodos más conocidos para la codificación de las variables categóricas? Descríbelos brevemente y pon un ejemplo de cada uno.

One-Hot Encoding (*)

Crea una columna binaria para cada categoría

Ejemplo: “Color” con categorías {Rojo, Azul} se convierte en Color_Rojo, Color_Azul. El problema es que puede crear muchas columnas si hay muchas categorias., ya que si hay 10 colores para un coche, tendriamos 9 de ellos como false y 1 de ellos como true.

Label Encoding

Asigna un número entero a cada categoría.

Ejemplo: “Color” = {Rojo=0, Azul=1, Verde=2}.

Ordinal Encoding

Parecido a Label Encoding, pero para categorías con orden natural (p.ej., Tallas S < M < L).

Ejemplo: Talla = {S=1, M=2, L=3}. Tambien como la parte practica, education de 1 al 16 segun el nivel de estudios….

Target (Mean) Encoding

Sustituye cada categoría por el promedio de la variable objetivo para esa categoria

Ejemplo: si Color=Rojo suele asociarse a un valor medio de 0.8 en la variable objetivo, se reemplaza por 0.8. Si educacion doctorado es en promedio un income de >50mil, que seria un 1, podriamos decir que education_doctorado seria cercano a 1. Siendo 1 en income >50mil y el numero en education_doctorado seria el promedio de las personas que tienen doctorado en el valor de la variable objetivo.

Bibliografia

(*) https://codificandobits.com/blog/como-codificar-datos-categoricos/

https://www.hackersrealm.net/post/target-or-mean-encoding-python