18/12/2023

Unit 06 – Data cleaning

Los modelos tiene 2 problemas a la hora de aceptar datos:

* No acceptan datos
* No hacen datos categoricos (strings)

Datos nulos

Hay 3 principals opciones para lidiar con datos nulos

* Eliminar la columna entera. Interesante cuando hay muchos nulos en esa columna
* Eliminar la fila entera. Es una solución sencilla, pero si llega a producción un dato con nulo, el modelo no será capaz de funcionar para ese caso. Si se toma esta decisión, es muy importante indicarlo en el modelo
* Inferir el valor: con la media, mediana, mínimo…

Datos categóricos

La mayoría de modelos no aceptan datos categóricos o no es lo ideal. Para transformar datos categóricos en numéricos

LabelEncoder: Asignar un numero a cada categoría. No tiene orden.

OrdinalEncoder: se le puede agregar un orden predefinido cuando se quiere que el modelo entienda ese orden

OneHotEncoder: divide las categorías de una columna en una columna para cada categoría

Preparar datos

El columna transformer permite definir los transformadores que se quiere aplicar a las columas

08/01/2024

Pasos a seguir

* Obtener los datos
* Echarles un vistado (con la funciones head, info o describe). Hay que tener en cuenta que la mayoría de modelo no admiten datos categóricos, hay que pasarlo a numérico)
* Tratar los datos: los modelos usualmente no admiten nulos y tampoco admiten datos categóricos. No hay que fiarse de los posibles nulos que muestre el DF, hay que mirar realmente los valores que tiene (el nulo puede ser un espacio o un ‘?’)

Datos nulos

* Eliminar la columna que tiene los datos nulos (usualemnte se hace cuando una columna tiene muchos datos)
* Eliminar la fila con el dato nulo (peligroso, porque el modelo siempre va a tener en cuenta esos datos nulos cuando le vengan en producción)
* Darle valor a los datos nulos: se puede poner la media, la mediana, crear un modelo de ML que prediga el valor de ese dato.

Datos categóricos

La mayoría de modelos solo aceptan números, no aceptan datos categóricos (texto). Los datos de texto que no son categóricos (el nombre por ejemplo, usualmente no aportan nada).

Hay que transformar los datos categóricos en numéricos, utilizando:

* LabelEncoder: asigna un valor a cada categoría, sin tener en cuenta el orden. Usualmente se utiliza para el target. No tiene en cuenta el order
* OrdinalEncoder: mantiene un orden
* OneHotEncoder: crea tantas columnas como posibles valores tenga la columna categórica. La desventaja de este algoritmo es si existen muchos posibles valores, se crearan muchas nuevas features

Hay dos funciones cuando se hace un codificadro. Fit para entrenar el codificador y el transform para utilizar ese transformador ya entrenado. Si se encuentra un valor nuevo que no estaba contenplada en el tests, hay un hyperparametro para indicar si lanzar un error o asignar un valor por defecto

Prepararar daots

Una vez el modelo esta en producción, hay que palicar las mismas tranformaciones a los datos que nos llegan. Una forma de hacer estas transformaciones es utilizar el columna transformer