DATA SCIENCE

Federico Baiocco
baioccofede@gmail.com
3512075440

Clase 14 - Agenda

VARIABLES CATEGÓRICAS

- One hot encoding
- Label / Ordinal encoder
- Discretización

¿ Dudas de la clase pasada ?

¿ Todos pudieron terminar ?

Variables categóricas

La clase anterior vimos como completar nulos..
Otra parte en el proceso de feature engineering consiste transformar las variables categóricas en algo que las computadoras puedan entender.

Variables categóricas

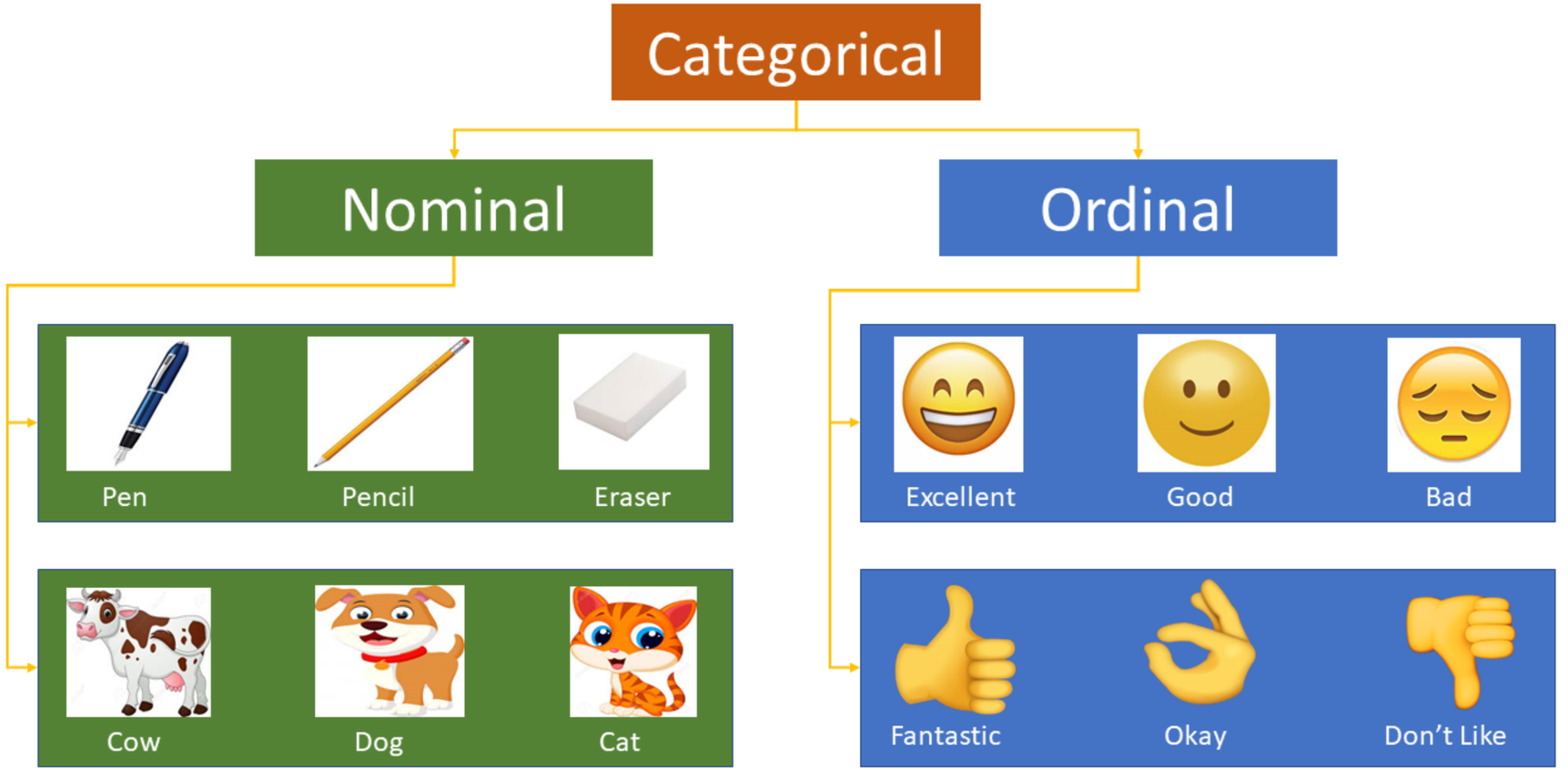
Sabemos que existen distintos tipos de variables.

Cuando tenemos variables que son numéricas, podemos meterlas en un modelo de machine learning sin problema. Sin embargo, casi siempre vamos a tener alguna variable categórica como "género", "nombre", etc.

A estas variables hay que tratarlas para poder usarlas.

Hay distintas maneras de encarar el problema dependiendo de los datos.

Variables categóricas



¿ Cómo las tratamos ?

Scikit learn y pandas nos brindan varios métodos para tratar estas variables.

Algunos de ellos son:

- One hot encoding
- Label encoding

One hot encoding

En este método, cada categoría se convierte en una columna del dataframe. Los valores de estas nuevas columnas pasan a ser 0s o 1s según corresponda.

Por ejemplo: Si tenemos una columna "género" que puede tomar los valores "F", "M", "N/A". Podemos crear a partir de esta columna otras 3 columnas: "F", "M", "N/A" que contengan 0s y 1s y eliminar la columna original.

Problemas con este método:
- Si la variable tiene alta cardinalidad, se generan muchisimas columnas

¿ Cómo implementarlo ?
- Pandas: get_dummies()
- Scikit learn: preprocessing.OneHotEncoder()

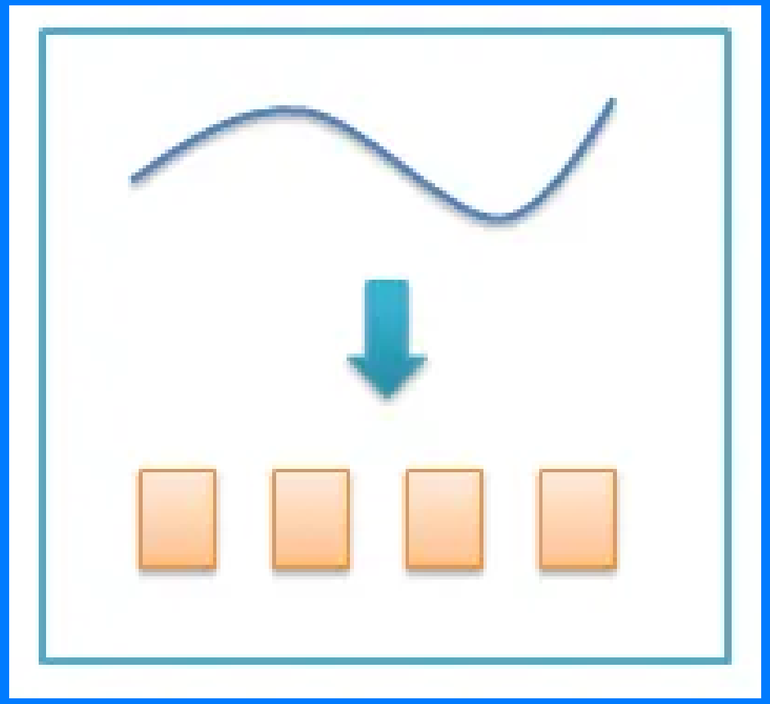
Label encoder

En este método a cada categoría se le asigna un valor numérico de 1 a N (donde N = número de categorías).

La ventaja de este método es que representamos todo en una sola columna y no creamos muchas como en el caso de one hot encoder.

La principal desventaja y por lo que solemos utilizar one hot encoder para variables categóricas no-ordinales es que si no hay una relación u orden entre las categorías, al asignar valores numéricos crecientes, los algoritmos de ML pueden considerar que hay algún orden o relación.

Por ejemplo, si tenemos una variable categórica "color" que puede valer: azul, rojo, verde, amarillo y naranja, estas variables no representan un orden y al utilizar un label encoder se puede entender que hay algún orden.



Discretización

También conocido como "binning". Transformamos variables continuas en valores discretos.

Discretización

Como dijimos, discretizar significa pasar variables continuas a discretas. Para esto, primero decidimos en cuantos bins o rangos queremos dividir nuestra variable, luego, a cada valor continuo lo asignamos a un bin.
De esta forma, en lugar de tener por ejemplo la edad de una persona, podemos tener una variable que nos diga si su edad esta entre [0 y 10], [11 y 30], [31 y 50] o [51 o más] años.

¿ Para qué haríamos esto ?

Muchas veces nos puede servir para algún análisis, pero se suele utilizar muchas veces para entrenar modelos de ML.

Muchas veces, al transformar una variable continua en otra que tiene one hot encoding aplicado, obtenemos un dataset mas "expresivo" que es mejor para entrenar un modelo.

Abrimos notebook variables categóricas

Preguntas?

Para la próxima clase:

- Terminar de tratar las variables categóricas del dataset