$\mu$ = mean
$\sigma$ = S.D.

[→ * LabelEncoder Class = 1-D input (classification predictive modelling)

10/07/20

## ENCODING CATEGORICAL DATA

2 common approaches:

I) (natural ordering) Ordinal coding: conversion of categorical data into integer codes ranging from 0 to (no. of categories − 1)

- easily reversible
- for categorical variables, it imposes an ordinal encoding w/o meaningful relationship. ∴ in such cases ⇒ ONE HOT ENCODING is used.

The ordinal encoding transform is available in the scikit-learn Python machine learning library via the OrdinalEncoder class. (for matrix) - row - column

By default, it will assign integers to labels in the order that is observed in the data. If a specific order is desired, it can be specified via the "categories" argument as a list with the rank order of all expected labels.

Eg → converting colors categories "red", "green", & "blue" into integers. First, the categories are sorted, then numbers are applied.

For strings, this means the labels are sorted alphabetically: blue = 0, green = 1, red = 2

// Ordinal encoding implicitely assumes an order across categories.

II) One-hot encoding - "Each bit represents a possible category. If the variable cannot belong to multiple categories at once, then only 1 bit in the group can be "on". "

Eg - Color variable example - 3 categories (blue, green, red) ∴ 3 variables are needed. A "1" value is placed in the ~~binary~~ binary variable for the color & "0" values for other colors.

Python - scikit-learn ML library
OneHotEncoder class

III) Dummy variable Encoding

OneHotEncoding creates 1 binary variable for each category. - this representation includes redundancy.

Eg [1, 0, 0] : blue
[0, 1, 0] : green

then we don't need another binary variable to represent - "red", instead we could use 0 values for both "blue" & "green" along.. eg [0, 0]

This is called dummy variable encoding, & always represents 'C' categories with 'C-1' binary variables.

- (For linear regression models)

## IMAGE DATA

An image is comprised of small tiles called "pixels".
Color of each pixel is represented with a set of values:

### Grayscale Images

- Each pixel is represented by a single number.
- Range : 0 to 255

Eg 0 ⇒ black
255 ⇒ bright white

- depth = 1 (only 1 channel)

### RGB Images

- Each Pixel is represented by a vector of 3 numbers.
- Range : 0 to 255.

Eg Purple - a mix of red & blue with no green.
(128, 0, 128).

- depth = 3 (R G B)
3 channels

### Encoding an image

3 things required to reproduce an image :

- Horizontal position of each pixel.
- Vertical position of each pixel.
- Color of each pixel.

Size of a vector required for a given image ⇒ Height * Width * Depth

### Other Processing Steps (Preprocessing steps).

- Uniform aspect ratio - (by making sure that all input images are square in shape).
- Normalized - subtract mean pixel value in a channel from each pixel value in that channel.

Others - rotation, cropping, resizing, denoising & centering the image.