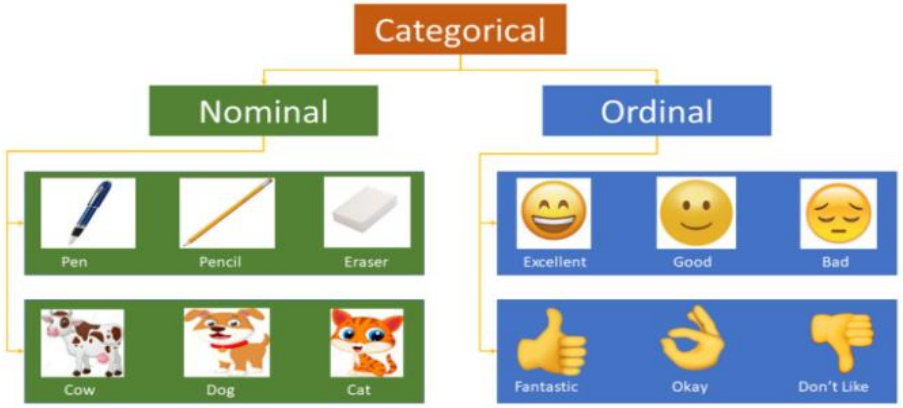
Basic Categorical Encoding

A Summary Note

By Chiranjit Pathak

Categorical Encoding :

Machine learning algorithms are unable to handle the categorical variables unless they are converted in to numerical forms.



- **Binary** : Yes /No; True/False
- **Nominal**: no numerical importance
- **Ordinal**: associated with numeric order

Binary : $f(Cat1, Cat2) \rightarrow f(0,1)$

Ordinal: $f(Cat\ 1, Cat\ 2 \ldots., Cat\ n.) \rightarrow f(0,1, \ldots..n)$

Nominal: $f(Cat1, Cat2 \ldots. Cat\ n.) \rightarrow f([\![01]\!]) \rightarrow \begin{matrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & .. & 0 \\ .. & .. & .. & .. \\ 0 & 0 & .. & 1 \end{matrix}$

Categorical Methods Summary :

Method	Tools	Pros	Cons
One Hot Encoding	get_dummies (Pandas) OneHotEncoder(Sklearn)	Easy to use	May slow down the learning for more numbers of features.
Label Encoding	factorize (Pandas) LabelEncoder(Sklearn)	Less memory utilization	Algorithm will consider the categories are in some order
Ordinal Encoding (User defined)	Dictionary{ } and .map()	Easy and flexible technique	Manual intervention Not suitable for more number of features.
Helmert Encoder	category_encoders.HelmertEncoder	This can be useful in certain situations where categorical variable are ordered from lowest to highest, or from smallest to largest.	Categories without order may get problems sometime.
Binary Encoder	category_encoders.BinaryEncoder	Less number of coded features i.e. 7 columns will take up to 128 numbers features.	

Categorical Methods Summary :

Method	Tools	Pros	Cons
Frequency Encoder	groupby() and .map()	Easy to use	May stuck if some features are having same frequencies.
Target Encoder: Mean	groupby(), . mean() and .map()	it does not affect the volume of the data and helps in faster learning	Randomness gets missed out during this grouping so become notorious for over-fitting
Target Encoder: Smooth	groupby(), . mean() and .map()	it does not affect the volume of the data and helps in faster learning	Randomness gets missed out during this grouping so become notorious for over-fitting
Target Encoder: Weight of Evidence	groupby(), .mean() and .log()	WoE transformation orders the categories on a “logistic” scale which is natural for Logistic Regression. It can be compared across categories and variables as it includes standardization too.	Loss of information (variation) due to binning to a few categories. Correlation between independent variables are not being considered.
Target Encoder: Probability Ratio	groupby() and .mean()	Similar to WoE and easy	If Probability of failure value is zero, this may become irritating.

Thanks for reading!

Lets collaborate and happy to receive any feedback/suggestion/comment at.....

pathak.chiranjit@gmail.com