Encoding

30 May 2023 20:39

How to Convert categorical data into numerical?

* Label Encoder
* One hot Encoding / Get dummies

Label Encoder

Label Encoder is one of the preprocessing method in machine learning which is used to convert categorical data into numerical data.

Gender	Gender
Male	1
Female	0
Male	1
Female	0

Note

* Label Encoder assigns numbers from 0 to ..... in alphabetical order

Height	Height
Short	1
Medium	0
Tall	2

Color	Color
Red	2
green	1
Blue	0

Limitation

Gender	Gender
Male	1
Female	0

* Label Encoder leads to priority issue.
* A label with high value will be considered as to have priority than the label with less priority..

1 > 0

Marks	Review
Fail — 0	bad — 0
Pass — 1	Good — 1

For these data Label Encoding.

One Hot Encoding / Get Dummies.

It is one of the preprocessing method used to convert categorical data into numerical data.

* It is used for nominal data [Data having order]
* It creates number of columns equal to the number of categories.

Gender | Gender_male | Gender_female

1 → Male

* It creates number of columns equal to

Gender	Gender_male	Gender_female
Male	1	0
Female	0	1
Male	1	0
Female	0	1

[1, 0] → Male
[0 -1] → Female

Color	Color_red	Color_green	Color_blue
Red	1	0	0
green	0	1	0
blue	0	0	1
Red	1	0	0
green	0	1	0

[1 0 0] - red
[0 1 0] - green
[0 0 1] → blue

Limitation

* No of columns increases and results in overfitting.
Overfitting → Training score is good and testing score is less.

Breaks 10 - 10:10 Pm

How to overcome?

* Use one hot Encoding when categories are less than 3.
* Drop one of the dummy columns which you have created.

Color	Color_red X	Color_green	Color_blue	Color_Orange
red	1 X	0	0	0
green	0	1	0	0
blue	0	0	1	0
orange	0	0	0	1
red	1	0	0	0
green	0	1	0	0
blue	0	0	1	0

[0 0 0] - red
[1 0 0] - green
[0 1 0] - blue
[0 0 1] - Orange