Categorical encoding

Algoritmos entendem números

Categorical encoding é o processo de transformar categorias em números

Duas Formas:

Label encoding

One-hot encoding

Label encoding

Cada categoria recebe um número, normalmente em ordem alfabética

Color	Size	Price
Red	Small	10
Green	Medium	20
Blue	Large	30
Red	Large	25
Green	Small	15

Color	Size	Price
Red	Small	10
Green	Medium	20
Blue	Large	30
Red	Large	25
Green	Small	15

Category	Encoded Value
Color: Red	0
Color: Green	1
Color: Blue	2

Category	Encoded Value
Size: Small	0
Size: Medium	1
Size: Large	2

Color_Encoded	Size_Encoded	Price
0	0	10
1	1	20
2	2	30
0	2	25
1	0	15

Label encoding

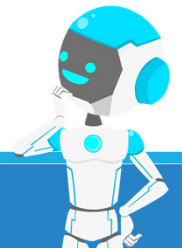
Problema: o algoritmo pode correlacionar os dados como uma ordem de grandeza!

Color_Encoded	Size_Encoded	Price
0	0	10
1	1	20
2	2	30
0	2	25
1	0	15

One-hot encoding

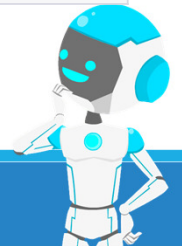
Cada categoria é transformada em outro atributo: dummy variable

Um valor binário informa a ocorrência



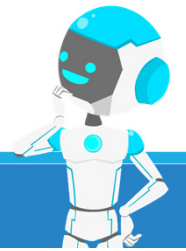
Color	Size	Price
Red	Small	10
Green	Medium	20
Blue	Large	30
Red	Large	25
Green	Small	15

Color_Red	Color_Green	Color_Blue	Size_Small	Size_Medium	Size_Large	Price
1	0	0	1	0	0	10
0	1	0	0	1	0	20
0	0	1	0	0	1	30
1	0	0	0	0	1	25
0	1	0	1	0	0	15



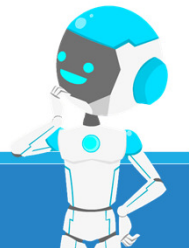
One-hot encoding

- Muitas colunas podem gerar um espaço de características de alta dimensão, que pode causar super ajuste e ter um custo computacional muito alto
- Maldição da Dimensionaldiade: Dados esparços, muitas colunas com valor zero, tornando difícil encontrar valores nos dados
- Dummy Variable Trap: valores de colunas binárias podem ser previstos a partir dos valores de outras colunas



Qual valor?

Color_Red	Color_Green	Color_Blue
1	?	?
0	?	0
?	?	1
?	0	0
0	?	0



Dummy Variable Trap

- O valor dos atributos se torna altamente previsível
- Resultado, correlação entre as variáveis independentes: multicolinearidade
- Solução: Excluir um dos atributos ou combinar colunas binárias

Color_Red	Color_Green	Color_Blue	Size_Small	Size_Medium	Size_Large	Price
1	0	0	1	0	0	10
0	1	0	0	1	0	20
0	0	1	0	0	1	30
1	0	0	0	0	1	25
0	1	0	1	0	0	15

Qual usar?

Label encoding	One-hot encoding
Há ordem (progr. Junior, Pleno, Sênior)	Não há ordem
Grande Número de categorias, não da pra usar One-hot encoding	Número de categorias é pequeno