CIS 678 Machine Learning

Features encoding!

Outline

- Label Encoding
- One-Hot Encoding

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)
10	1.3	45
30	1.7	67
65	1.5	57
…	…	…

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)
10	1.3	45
30	1.7	67
65	1.5	57
…	…	…

Numeric (R) values at different scales

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)	Sex
10	1.3	45	F
30	1.7	67	M
65	1.5	57	F
…	…	…	…

Categorical Feature

Data/Feature Encoding

Given a Data Table

Age (Yr)	**Height (M)**	**Weight (Kg)**	**Sex**
10	1.3	45	F
30	1.7	67	M
65	1.5	57	F
…	…	…	…

Categorical Feature

{F, M}

{F=0, M=1}

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)	Sex	Color preference
10	1.3	45	F	Red
30	1.7	67	M	Blue
65	1.5	57	F	Green
…	…	…	…	…

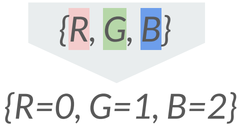
Categorical Feature

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)	Sex	Color preference
10	1.3	45	F	Red
30	1.7	67	M	Blue
65	1.5	57	F	Green
…	…	…	…	…

Categorical Feature



{R, G, B}

{R=0, G=1, B=2}

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)	Sex	Color preference
10	1.3	45	F	Red
30	1.7	67	M	Blue
65	1.5	57	F	Green
…	…	…	…	…

{F, M}

{F=0, M=1}

{R, G, B}

{R=0, G=1, B=2}

Label encoder

Categorical Feature

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)	Sex	Color preference
10	1.3	45	F	Red
30	1.7	67	M	Blue
65	1.5	57	F	Green
…	…	…	…	…

{R, G, B}

{R=0, G=1, B=2}

Label encoder

Do you see any issue here?

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)	Sex	Color preference
10	1.3	45	F	Red
30	1.7	67	M	Blue
65	1.5	57	F	Green
…	…	…	…	…

- *We are saying, Green is closer to Red than Blue,*

{R, G, B}

{R=0, G=1, B=2}

Do you see any issue here?

Label encoder

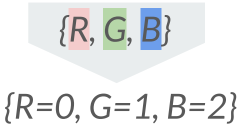
Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)	Sex	Color preference
10	1.3	45	F	Red
30	1.7	67	M	Blue
65	1.5	57	F	Green
…	…	…	…	…

Do you see any issue here?

- *We are saying, Green is closer to Red than Blue,*
- *And enforcing Bias in the Vector Space*



{R, G, B}

{R=0, G=1, B=2}

Label encoder

Data/Feature Encoding

Given a Data Table

Age (Yr)	Height (M)	Weight (Kg)	Sex	Color preference
10	1.3	45	F	Red
30	1.7	67	M	Blue
65	1.5	57	F	Green
…	…	…	…	…

- *We are saying, Green is closer to Red than Blue,*
- *And enforcing Bias in the Vector Space*
- *Which doesn't seem ok.*

{R, G, B}

{R=0, G=1, B=2}

Do you see any issue here?

Label encoder

Data/Feature Encoding

One Hot Encoding

- *How many colors do we have?*

Data/Feature Encoding

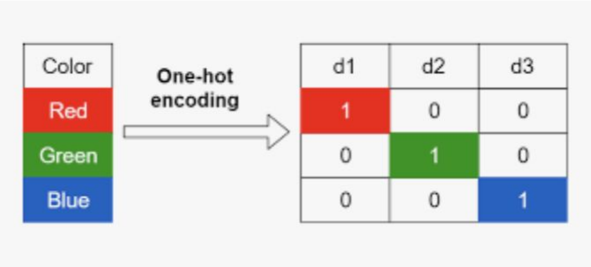
One Hot Encoding

- *How many colors do we have?*
- *3* *{R, G, B}*

Data/Feature Encoding

One Hot Encoding

- *How many colors do we have?*
- *3 {R, G, B}*
- *So, we use 3 Bits to define each color*



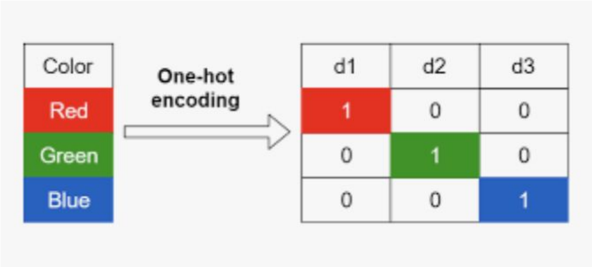
Color
Red
Green
Blue

d1	d2	d3
1	0	0
0	1	0
0	0	1

Data/Feature Encoding

One Hot Encoding

- *How many colors do we have?*
- *3 {R, G, B}*
- *So, we use 3 Bits to define each color*
- *How does it solves the Vector Space Distance Problem?*



Color
Red
Green
Blue

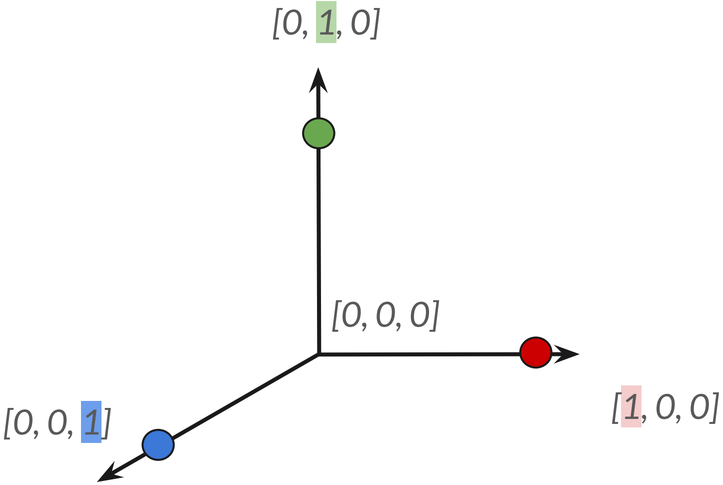
One-hot encoding

d1	d2	d3
1	0	0
0	1	0
0	0	1

Data/Feature Encoding

One Hot Encoding

- *How many colors do we have?*
- *3* {R, G, B}
- *So, we use 3 Bits to define each color*
- *How does it solves the Vector Space Distance Problem?*
- *We have debiased the definition. Isn’t it Cool!*



[0, 1, 0]

[0, 0, 0]

[1, 0, 0]

[0, 0, 1]

One hot encoding (cont.)

Classification task:

- Binary example {Cat vs Dog}
- Set size is 2
  - Cat (0, 1)
  - Dog (1, 0)
  - Or vice versa
- Same rule applies every categorical data

Notebook Presentation!

Regression task with categorical variables.

QA