Scaling, Encoding, Missing Value

Sunday, December 25, 2022 10:04 PM

Scaling

Data preprocessing

Make changes / Manipulate the data before giving it to the model

- Numeric
- Categorical

Scaling

Numerical

KM	M
5	5000
6	6000
7	7000

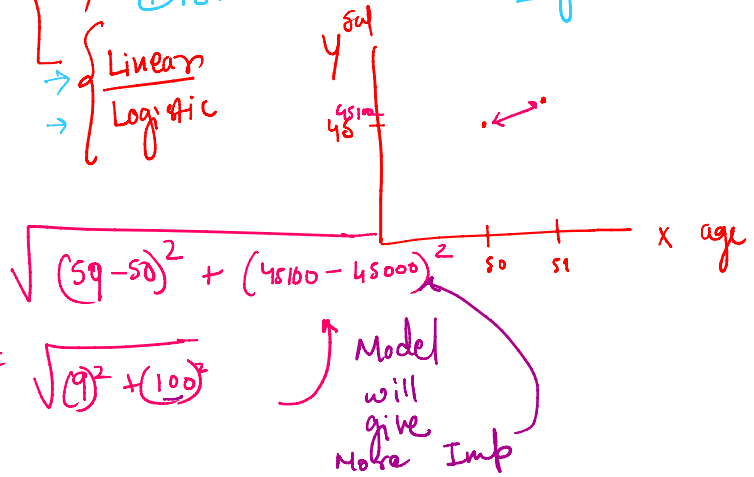
age	Salary
20	50
T	
60	59
	45000
	45100

20 T → 2 L

$\frac{5000}{5000} = 5$

Distance based algo

- Linear
- Logistic



$\sqrt{(59-50)^2 + (45100-45000)^2}$

$= \sqrt{(9)^2 + (100)^2}$

Model will give More Imp

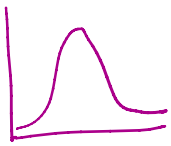
age	Salary in Thousand
50	45
59	45.1

$$\sqrt{(59-50)^2 + (45.1-45)^2}$$

$$\sqrt{(9)^2 + (0.1)^2}$$

Scaling : will Bring all the values in Same range

Image → Min Max (Normalization)

PCA → Standarization ←  If your data follow ND

M	KM
5000	5
7000	7
8000	8

$$\frac{x_i - a_{min}}{a_{max} - a_{min}}$$

$$\frac{8000 - 5000}{8000 - 5000}$$

M	KM
0	0
0.6	0.66
1	1

0 — 1

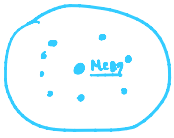
Mean, Variance, Std dev

$$Z = \frac{x - u}{\sigma}$$ (u: Mean, σ: Std dev)

D = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

$$\text{Mean} \rightarrow \frac{1+2+3+4+5+6+7+8+9+10}{10}$$

$$u = 5.5$$



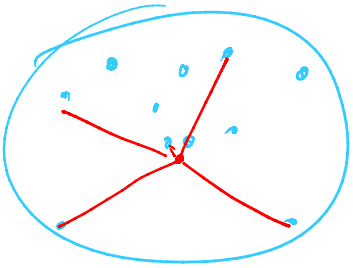
$$\text{Variance} = \frac{\Sigma (x-u)^2}{n}$$

how much your data is Spread

$$\sigma^2 = \frac{(1-5.5)^2 + (2-5.5)^2 \cdots (10-5.5)^2}{10}$$

$$\sigma^2 = 9.6$$

$$\text{Std dev} = \sqrt{\sigma^2} \quad \sqrt{9.6}$$



0 = Mean
1 = Std

$$\begin{bmatrix} -2 \\ 1 \\ 0 \\ 3.5 \\ -3.7 \end{bmatrix}$$

$$Z_1 = \frac{1 - 5.5}{\sqrt{9.6}}$$ Scale

0 = Mean
1 = Std

0
3.5
-3.7

$z_1 = \frac{1}{\sqrt{9.6}}$

$z_2 = \frac{2-5.5}{\sqrt{9.6}}$

$z_{10} = \frac{10-5.5}{\sqrt{9.6}}$

Scale
Mean = 0
Std = 1

Encoding

D_C D_T D_S D_SI
1 0 0 0

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Categorical

Computer cannot understand anything other than number

Encode the categorical Data

Dummy / one hot encoding

Label encoding

Gender | Male_Gender | Female_Gender
Male | 1 | 0
Female | 0 | 1
M | 1 | 0
F | 0 | 1
M | 1 | 0
M | 1 | 0

• ordinal

Grade
1 A 2
2 B 1
3 A 1
4 C 0

2 > 1 > 0

Male - 0
Female - 1

1 > 0

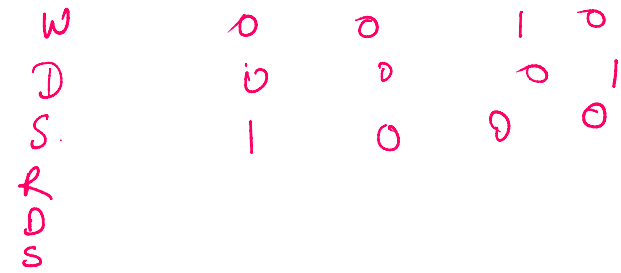
F > M

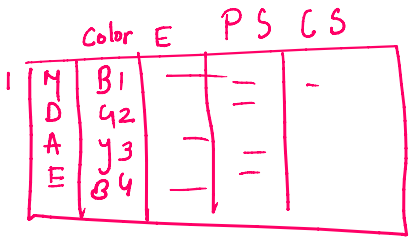
F = M

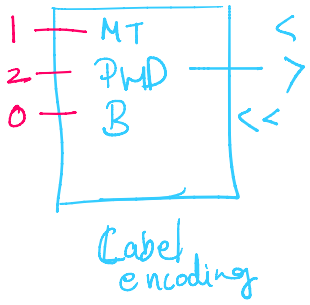
Nominal

Black = Green

Weather | W_S | W_R | W_W | W_D
S | 1 | 0 | 0 | 0
R | 0 | 1 | 0 | 0







- Outlier
- Confusion Matrix
- Missing Value
- Feature selection

Coding

↳ Dataset

Reg → Classific

- Miss
- Std
- out
- Enc

Linear DEC

Miss Std out Enc

Logis Random

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis