Px → Numerical

eg:

cid	income	cs	#of loan	Status
1	14	750	1	Y
2	15	650	2	Y
3	16	350	0	N

X (income, cs, #of loan) ; Y (Status)

(X, Y) → MODEL → f() → good ← testing

Numerical

BUT

cid	income	cs	# of loan	city	Status
1	14	750	1	Tier-I	Yes
2	15	650	2	Tier-I	Yes
3	16	350	0	Tier-III	Yes

X

KNN / Los → sklearn

(X, Y) → f() fail

convert

Numerical Rep.

f() ← M()

eg. $\overbrace{\text{weight} \quad \text{hair-color} \quad \text{country}}^{X}$ --- $\vec{Y}$

64 Black US 180

$(X, Y) \rightarrow \boxed{} \rightarrow M(\ )$ ← X (testing)

$\hat{Y}$

categorical → Numbers

Numerical

let – hair-color → { Bl, Br, Red, Green, Grey } etc

'1' '2' '3' '4' '5'

hair-color

$D_n(X, Y)$

at the end

weight hue-colour — — —

~~Bk~~ 1
~~Green~~ 4
~~Bk~~ 1
~~Green~~ 4
~~Grey~~ 5

X

Numbers

categorical

problem

$D_n(X,Y) \to D_n'(X,Y)$

ML()

categorical to Numerica

1 < 4 →

Black grey X

Dn → 'Rating'

good
v good
good
Bad
Bad
good

↑

categorical

transfor →

Rating
avg
2
1
0
0
1

→ ML

↑
OK →
↑

Bad → 0
good → 1
v-goo - 2

ONE HOT ENCODING

simple → hair-colour

5 unique → {Bl, Br, Red, Green, Grey}

5 → 6D

2D → 6D ↑

$f'()$

$M(X, Y)$

Dn

weight	hair-colour	Bl	Br	Red	Green	Grey
54	Black	1	0	0	0	0
56	Black	1	0	0	0	0
51	Green	0	0	0	1	0

4

Country

IND

Asia

200

G1
G2
G3
G4

4-D

ONE HOT encoding

200

category

sklearn

pd.getdum

200-D

IND	
1	0 0 0 0 0 0 0 0

W ↑ M ↓

good → col → 20 unique → Domain

categorical

Dimension

↳ Missing value imputation

$350 + 410 + 780 + 360 / \rightarrow 354$

$D_n \rightarrow$

income	CS	Status
3500	350	yes
17	410	yes
~~Na~~ Na	780	yes
14	~~Na~~ 354	yes
9	360	No

fail

(X, Y

treat

↳ mean →
↳ median
↳ mode

i) take all non-empty valu

ii) compute mean | median | mode and put in place of missing valu

$\frac{35 + 17 + 14 + 9}{4} \approx 21$

Bias-Variance Tradeoff

$\boxed{K=n}$ $\boxed{K=1}$

underfitting

D_n 1000

now we can understand mathematically

Bias-variance Tradeoff

$D_x \to \boxed{?} \to F() - ?$

ML/DL

generalisation error $= Bias^2 +$ irreducible error $+$ variance

{future unseen data}

Log Regres → assumption

high bias

$LR()$

$D_x \to$ XYZ dis

LR

high variance

tr

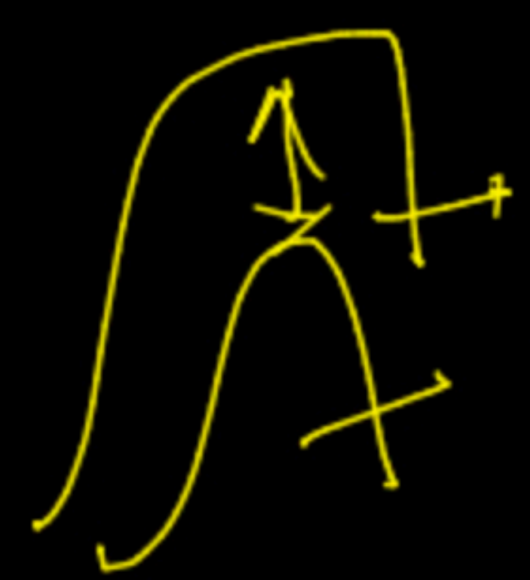
train

Notedown

90%

test

71%



90%

tr