

CS 307

SPRING 2024

DALPIAZ



PRE PROCESSING

IMPUTATION → HOW TO DEAL WITH MISSING DATA

TRAINING DATA

<u>X₁</u>	<u>X₂</u>
5.3	DOG
4.2	DOG
1.1	NAN ← MISSING
NAN ← MISSING	CAT
6.3	DOG

REPLACE WITH DOG

REPLACE WITH MEDIAN (5.3, 4.2, 1.1, 6.3)

`.transform`

TEST DATA

<u>X₁</u>	<u>X₂</u>
NAN	NAN

REPLACE WITH VALUES LEARNED VIA TRAIN

`.transform`

FOR TRAINING DATA, LEARN HOW TO REPLACE MISSING

NUMERIC: MEDIAN, MEAN, STD

CATEGORICAL: MODE / MOST FREQUENT

`.fit`

SCALING

→ PUTTING NUMERIC FEATURES ON THE SAME SCALE

<u>X_1</u>	<u>X_2</u>
5.3	1000
4.2	2000
1.1	3000
4.6	5000
6.3	1500

↑
WOULD "DOMINATE" A
DISTANCE CALCULATION

STANDARDIZE :
$$\frac{X_i - \text{MEAN}(X)}{\text{STD}(X)}$$

MIN-MAX :
$$\frac{X_i - \text{MIN}(X)}{\text{MAX}(X)}$$

LIKE IMITATION :

LEARN WITH TRAIN

APPLY TO TRAIN AND TEST

ENCODING → How to make CATEGORIES INTO NUMBERS

ONE-HOT ENCODING

<u>X₁</u>	<u>X₂</u>	<u>X_{cat}</u>	<u>X_{dog}</u>	<u>X_{black}</u>	<u>X_{brown}</u>	<u>X_{tan}</u>
DOG	BROWN	0	1	0	1	0
CAT	BLACK	1	0	1	0	0
DOG	TAN	0	1	0	1	0
DOG	BROWN	0	1	0	1	0
CAT	BLACK	1	0	1	0	0

REFERENCE LEVEL

REFERENCE LEVEL

OMIT FOR "DUMMY" ENCODING