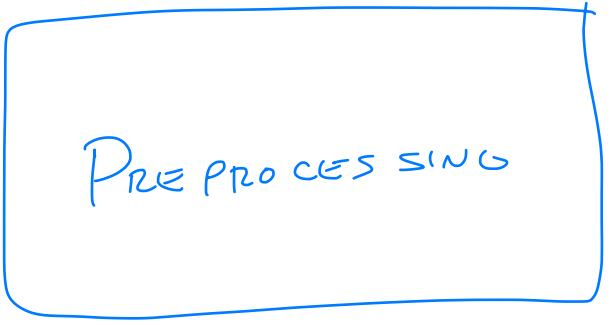
CS 307

Spring 2024

Dalpiaz



IMPUTATION → HOW TO DEAL WITH MISSING DATA

TRAINING DATA

X_1	X_2
5.3	DOG
4.2	DOG
1.1	NAN
NAN	CAT
6.3	DOG

.transform

REPLACE WITH DOG

MISSING

MISSING →

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

TEST DATA

X_1	X_2
NAN	NAN

REPLACE WITH VALUES LEARNED VIA TRAIN

.transform

FOR TRAINING DATA, LEARN HOW TO REPLACE MISSING

NUMERIC: MEDIAN, MEAN, ETC

CATEGORICAL: MODE / MOST FREQUENT

.fit

Scaling → Putting numeric features on the same scale

X_1	X_2
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: $\frac{x_i - \min(x)}{\max(x)}$

Like imputation:

Learn with train

Apply to train and test

ENCODING → HOW TO MAKE CATEGORIES INTO NUMBERS