

CS 307

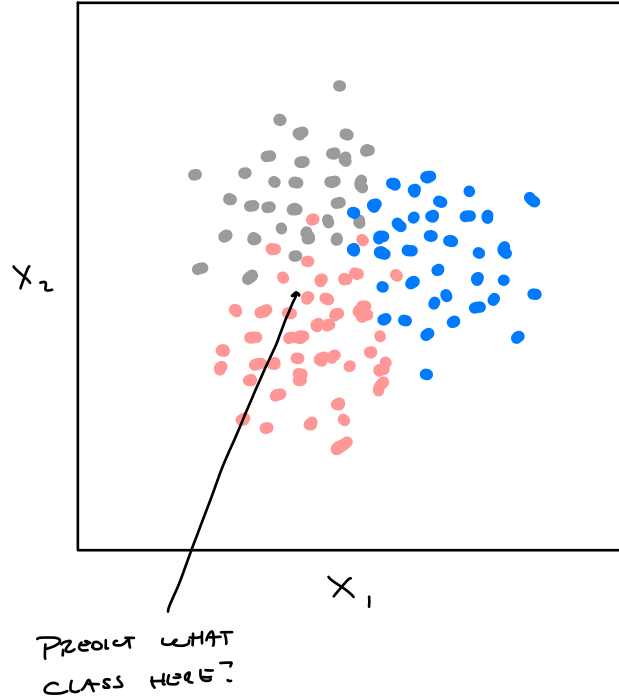
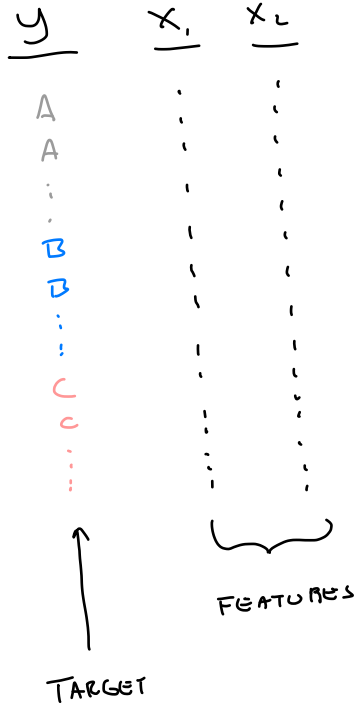
FALL 2023

DALPIAZ

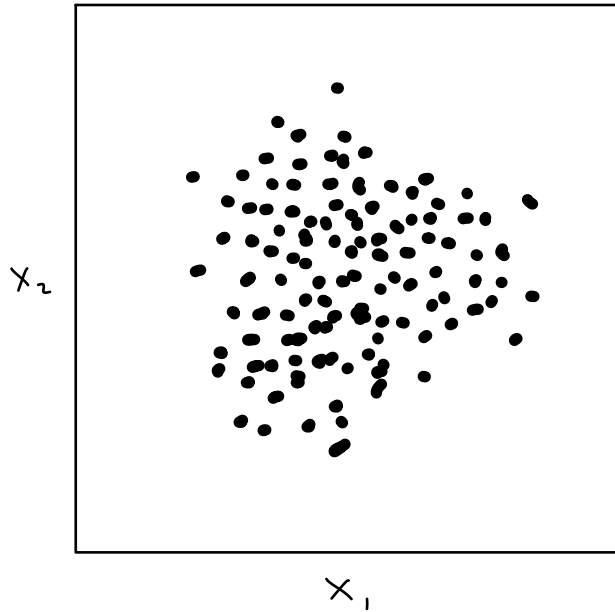
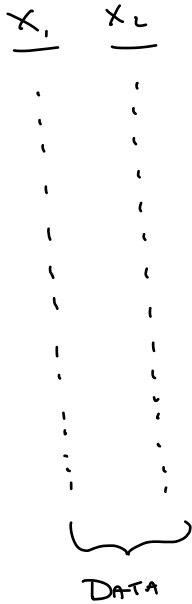
WEEK 11

UNSUPERVISED LEARNING

SUPERVISED LEARNING



UNSUPERVISED LEARNING



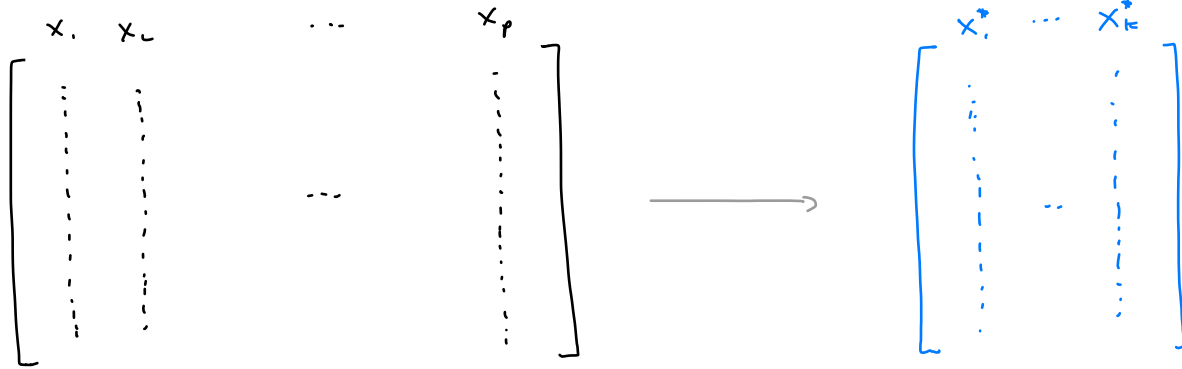
???

UNSUPERVISED LEARNING

- DIMENSION REDUCTION PCA
- CLUSTERING K-MEANS, HIERARCHICAL CLUSTERING
- DENSITY ESTIMATION KDE, MIXTURES
- OUTLIER DETECTION ONE-CLASS SVM, ISOLATION FOREST

DIMENSION REDUCTION

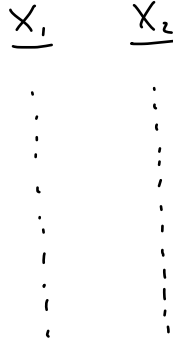
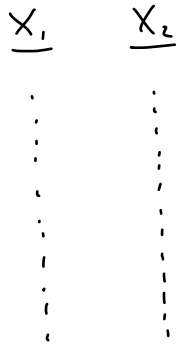
$$1 \leq k \leq p$$



OFTEN USED AS PREPROCESSING FOR SUPERVISED LEARNING

EXAMPLE: PCA

CLUSTERING



CLUSTER ASSIGNMENT

A
B
B
C
B
A
A
C

NOT A TARGET

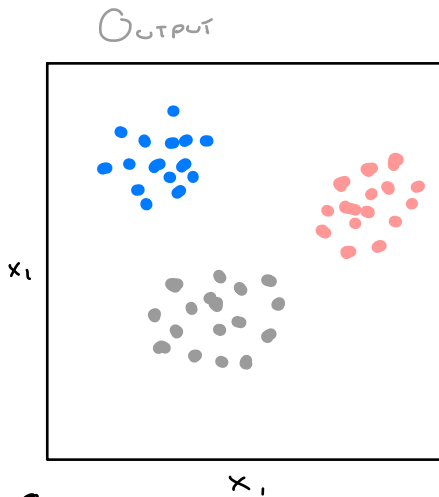
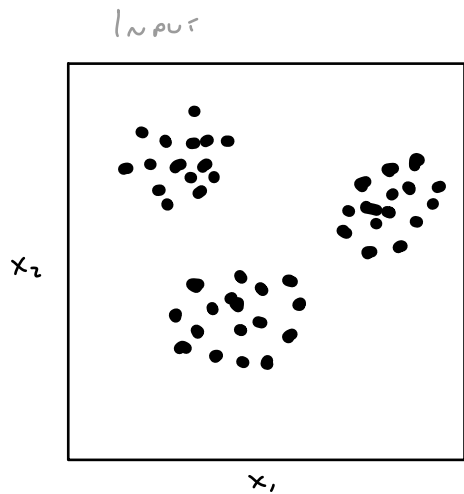
HARD TO VALIDATE

- NO STANDARD TO COMPARE TO IN PRACTICE
- ORDER / LABELS MEANINGLESS

EXAMPLES

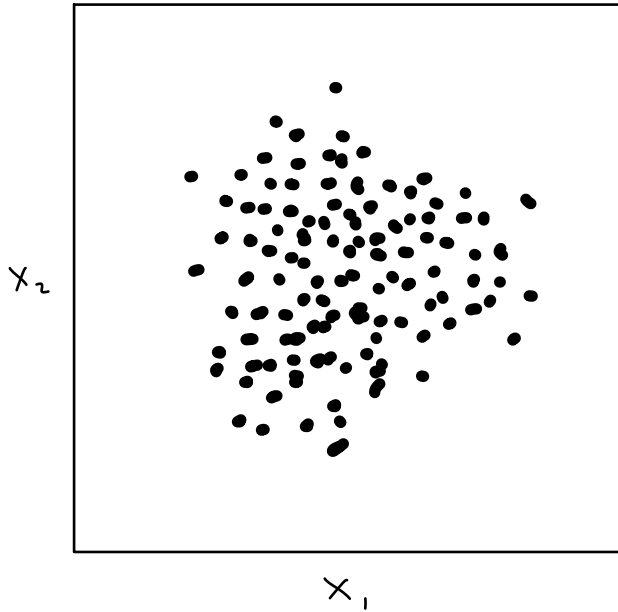
K-MEANS
HIERARCHICAL

CLUSTERING



↑
COLORS HAVE NO MEANING!
COULD BE SHUFFLED!

CLUSTERING



???

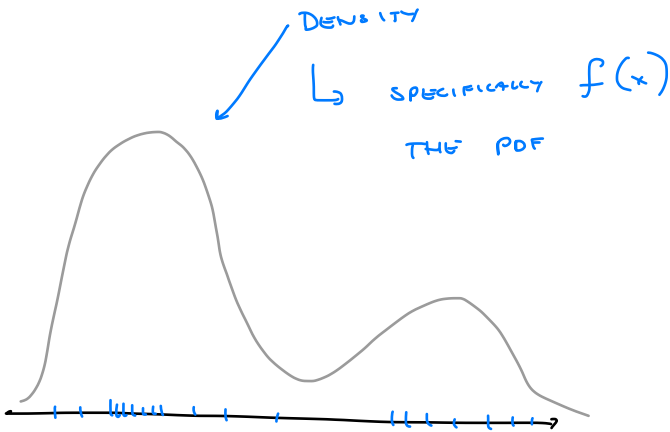
• How many clusters?

OFTEN NEEDS TO BE SET

DENSITY ESTIMATION

X

⋮



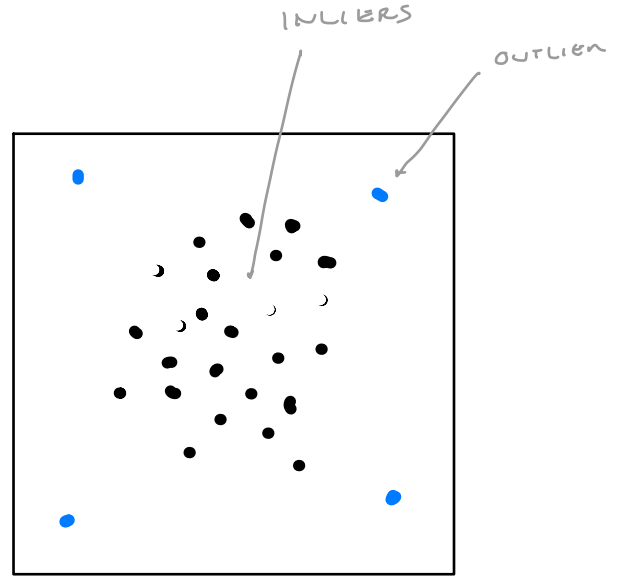
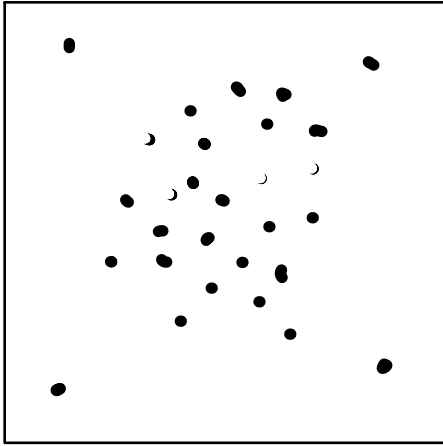
RUG PLOT

EXAMPLES

KDE

MIXTURE MODELS

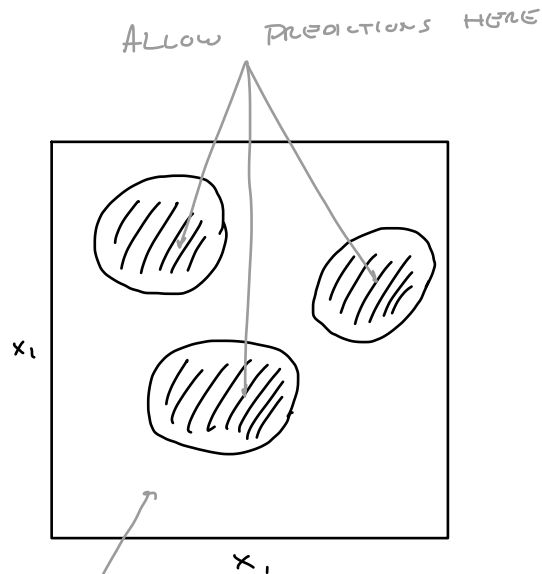
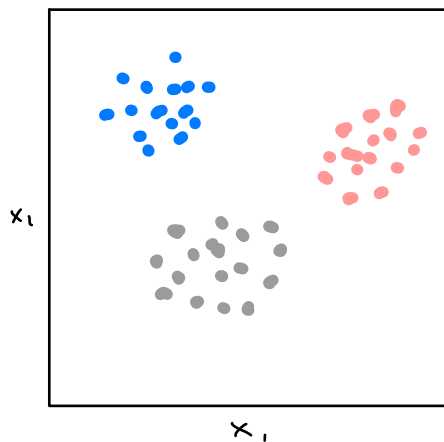
OUTLIER DETECTION



OUTLIER DETECTION

- WHEN SHOULD OUTLIERS BE REMOVED FROM TRAINING DATA? ONLY VERY CAREFULLY!
- ELIMINATE "OUTLIERS" AT TEST TIME?
 - ↳ NOVELTY DETECTOR
 - ↳ USEFUL WITHIN SUPERVISED PIPELINES

NOVELTY DETECTION



REFUSE TO PREDICT HERE

