16 OCT

REVIEW: UNSUPERVISED LEARNING

K-MEANS
INPUT: SET OF PTS
"SEED" CENTROIDS
NUMBER K

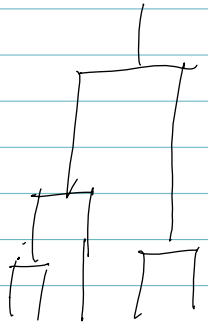
- ASSIGNS PTS TO NEAREST CENT.
- RECOMPUTE CENTROID
- LOOP UNTIL CONVG
  (NO REASSIGN, MIN SSE, ETC.)

ADV:
FAST
LOT OF EVID. OF PRACT USE

DISADV:
SENS TO OUTLIERS
SENS TO INIT SEEDS
LOCAL OPTIMAL
USER MUST SEL K

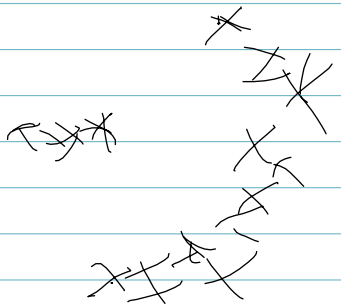
HIERARCH CLUSTERING

CLUSTER CONTAINE INSIDE EACH OTHER



(-) GENERALL n^2 RUNTIME

(+) HANDLES CLUSTERS THAT ARE CONCAVE



DATA STANDARDIZATION

NUMERIC ATTRIBUTES

RANGE (min, max)
Z-SCORE (σ, μ)

INTERVAL vs. RATIO

BINARY vs. NOMINAL

COMM FINDING

COMPONENT-BASED

MODULARITY MAXIMIZATION (M)

GREEDY

LOUVAIN — FASTER, MAXIMIZES M BETTER

LABEL PROPAGATN

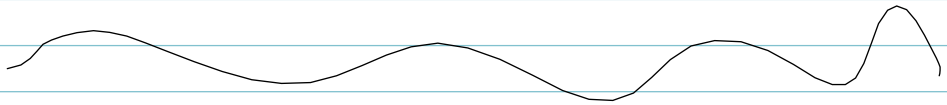
- VERY FAST
- LABEL PROP LIKE IN DIFFS.
- DEGENER. SOLN.

$$P(D_1 \mid I) = 0.8$$

$$P(D_2 \mid I) = 0.5$$

$$P(\neg D_1 \wedge \neg D_2 \mid I) = 0.1$$

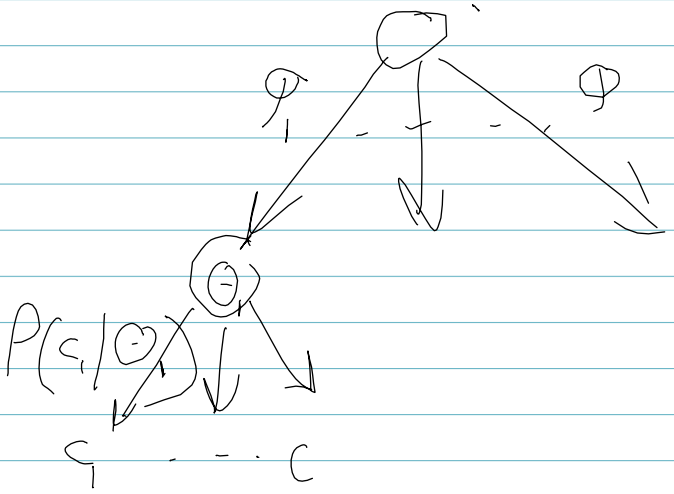
$$(0.2 \cdot 0.5)$$



$$P(D_2 \mid \neg D_1 \wedge I) = 0$$

$$p_j = Pr(c_j \mid \theta)$$

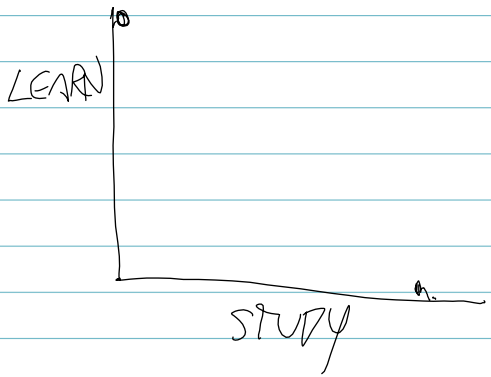
$$Pr(d_i \mid \theta) = \sum_{j=1}^{|C|} P(c \mid \theta) P(d \mid c, \theta)$$

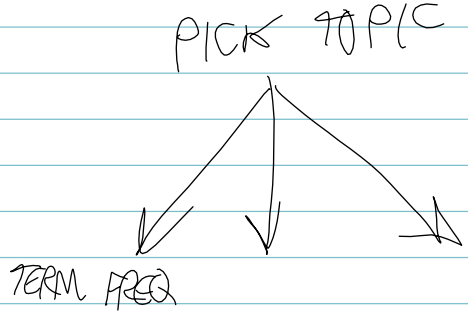


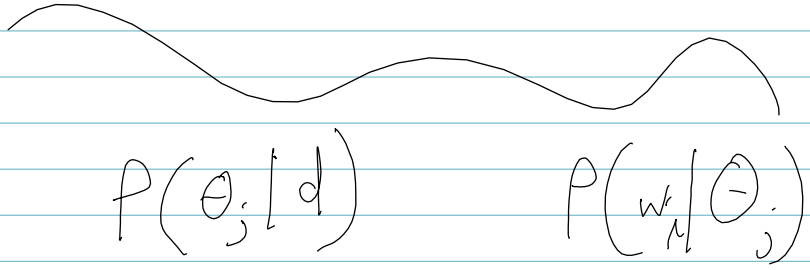
$$P(w \mid c, \theta)$$

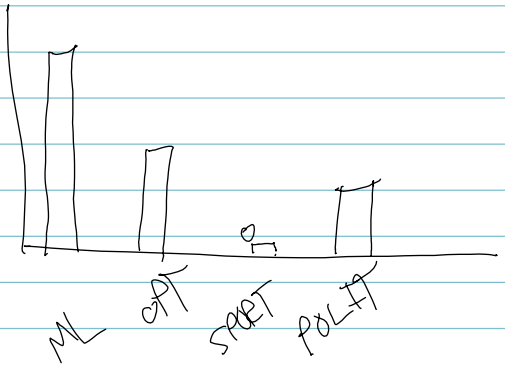
$P(d_i | c_j)$

$P(d_i | c_j, \theta_i) =$


LEARN
STUDY


PICK TOPIC
TERM FREQ


P(θ_j | d)
P(w_i | θ_j)


ML
OPT
SMOOT
POS-FT