Lecture 26 announcements

- HW 13 is due Tuesday, 11/27/2018, 2:00 PM
- Project report deadline is Tuesday, 12/4, 2:00 PM (new date)

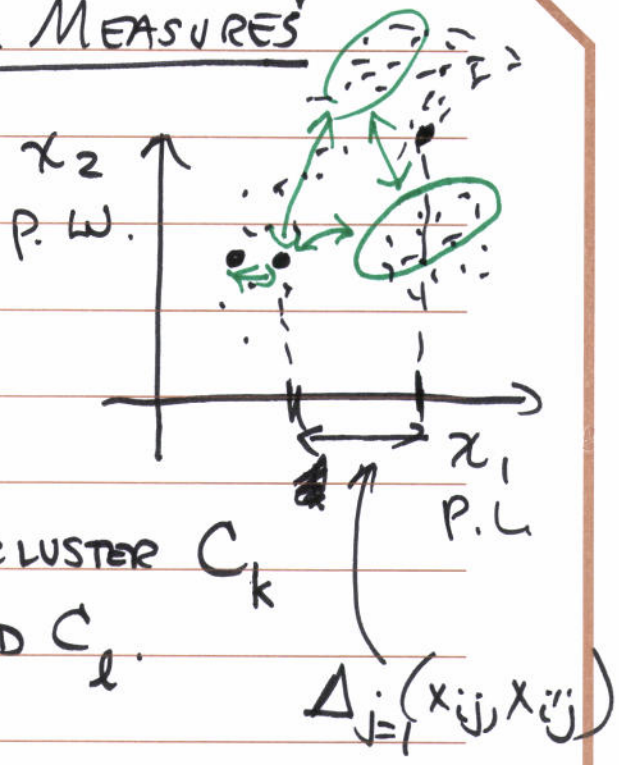
Lecture 26 outline

- Unsupervised learning (part 2)
  - Similarity / dissimilarity measures
  - Hierarchical clustering
    - Agglomerative

SIMILARITY / DISSIMILARITY MEASURES

NEED MEASURES FOR SIMILARITY/DISSIM. BETWEEN:

- 2 POINTS $\underline{x}_i$ AND $\underline{x}_j$
- A POINT $\underline{x}_i$ AND A CLUSTER C_k
- TWO CLUSTERS C_k AND C_l.



ALSO: MEASURE FOR QUALITY OF A CLUSTERING

- DIFFICULT PROBLEM (MORE LATER).

LET $\Delta(\underline{x}_i, \underline{x}_{i'}) = \Delta_{ii'}$ ($= d_{ii'}$) DENOTE A DISSIMILARITY FUNCTION

CAN USE A DISTANCE FUNCTION, E.G.:

$$\Delta(\underline{x}_i, \underline{x}_{i'}) = d_{\bullet}(\underline{x}_i, \underline{x}_{i'}) \bullet$$

$$= \sum_{j=1}^{D} \Delta_j(x_{ij}, x_{i'j})$$

$\Delta_j(x_{ij}, x_{i'j})$ CAN BE:

$(x_{ij} - x_{i'j})^2$ (EUCL. DIST.^{2})

$|x_{ij} - x_{i'j}|$ (ℓ_1 NORM DIST. OR CITY-BLOCK DIST. OR MANHATTAN DIST.)

FOR NOMINAL FEATURES (SYMBOLS, CATEGORICAL, OR LABELS):

$\Delta(x_{ij}, x_{i'j})$ = # OF FEATURES THAT ARE DIFFERENT

$$= \sum_{j=1}^{p} \mathbb{I}(x_{ij} \neq x_{i'j})$$

= HAMMING DISTANCE.

CAN LET $\Delta(x_i, x_{i'}) =$ THIS OR OTHER $d_{ii'}$, s.t. $d_{ii'} \geq 0 \ \forall i, i'$, AND $d_{ii} = 0 \ \forall i$.

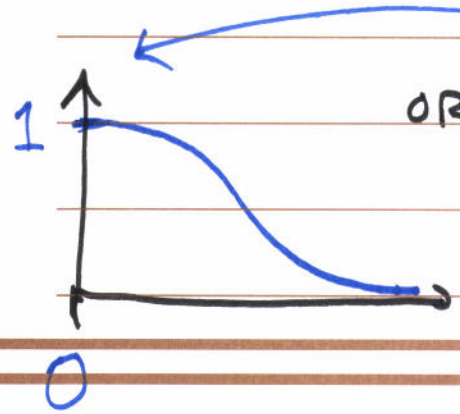
OR $f(d_{ii'})$, $f =$ ANY MONOTONICALLY INCREASING FUNCTION.

LET $s(\underline{x}_i, \underline{x}_{i'}) = s_{ii'}$ DENOTE A SIMILARITY FCN.

CAN CHOOSE $s_{ii'} = g[d_{ii'}]$, g = ANY MONOTONICALLY DECREASING FUNCTION.

E.G.: $s(\underline{x}_i, \underline{x}_{i'}) = \left(\max_{l,l'} d_{ll'}\right) - d_{ii'}$

OR $= \exp\left\{\frac{-d_{ii'}^2}{\sigma^2}\right\}$



OTHER SIMILARITY MEASURES:

FOR BINARY FEATURES

(e.g., CONTAINS AN ATTRIBUTE OR NOT):

$$s(\underline{x}_i, \underline{x}_{i'}) = \frac{\underline{x}_i^T \underline{x}_{i'}}{\underline{x}_i^T \underline{x}_i + \underline{x}_{i'}^T \underline{x}_{i'} - \underline{x}_i^T \underline{x}_{i'}}$$

= % OF ATTRIBUTES THAT ARE SHARED.

FOR SIGNALS WITH (SPATIAL OR TEMPORAL) STRUCTURE;

CORRELATION COEFFICIENT, E.G. (PEARSON):

LET j BE INDEX OVER SPATIAL LOCATION OR TIME

$$r_{ii'} = \frac{\sum_{j=1}^{D} (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\left[\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{ij} - \bar{x}_{i'})^2\right]^{1/2}}$$

IN WHICH:

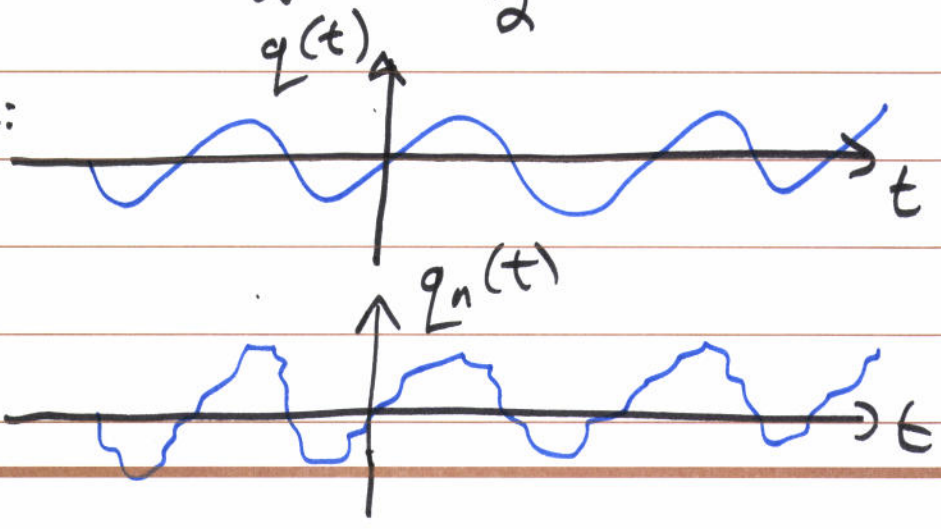
$$\bar{x}_i = \frac{1}{D}\sum_{j=1}^{D} x_{ij}.$$

NOTE: $-1 \le r_{ii'} \le 1$ ALWAYS.

CAN LET $S_{ii'} = r_{ii'}$. FOR SIMILARITY.

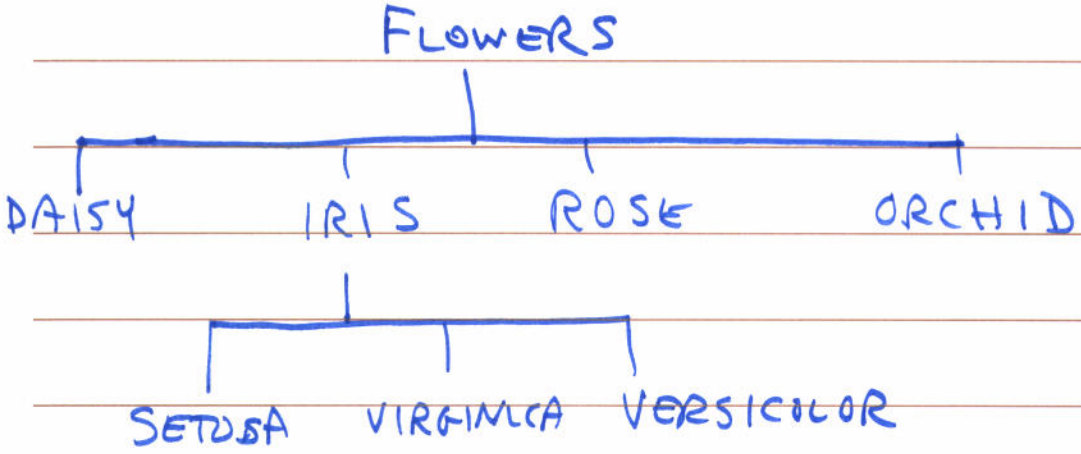
OR: $d_{ii'} = \frac{1 - r_{ii'}}{2}$

E.G.:

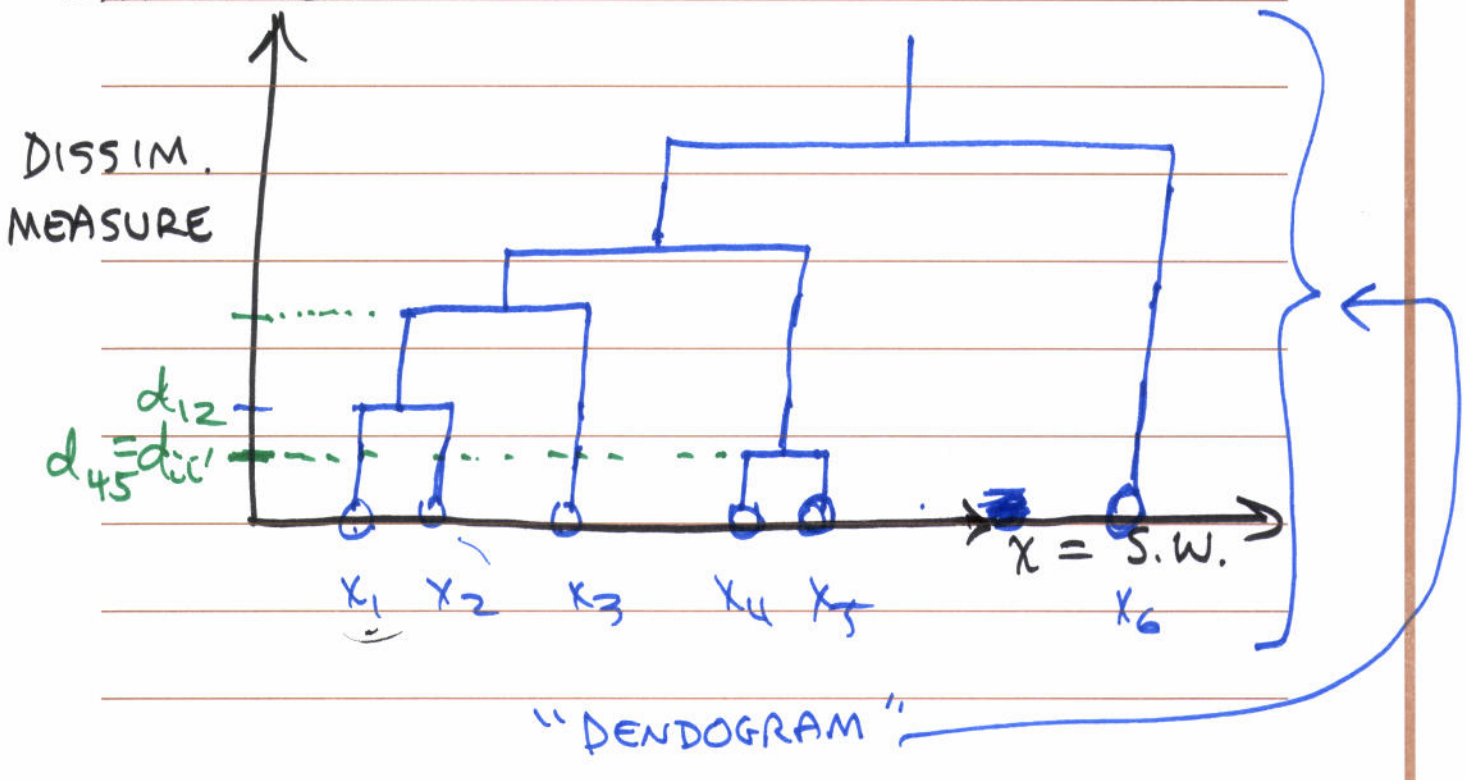


HIERARCHICAL CLUSTERING [MURPHY 25.5]

SOMETIMES DATA IS HIERARCHICAL IN NATURE!



EX. IN 1D:



TWO BASIC APPROACHES:

- AGGLOMERATIVE (BOTTOM UP).
- DIVISIVE (TOP DOWN)

→ UNDERLYING ASSUMPTION: IF 2 DATA POINTS ARE IN THE SAME CLUSTER AT ONE LEVEL, THEN THEY ARE IN THE SAME CLUSTER AT ALL HIGHER LEVELS.

AGGLOMERATIVE HIERARCHICAL CLUSTERING PROCEDURE

LET δ_{jk} = DISTANCE OR DISSIMILARITY BETWEEN CLUSTERS C_j AND C_k.

$\hat{K}$ = CURRENT # OF CLUSTERS.

1. CHOOSE HALTING CONDITION (H.C.)
2. INITIALIZE $\hat{K} = N$, CLUSTER $C_i = \{x_i\}$, $i = 1, 2, \cdots, N$; ITERATION $m = 1$.
3. REPEAT UNTIL H.C. IS MET:
   4. FIND NEAREST (= LEAST DISSIMILAR) PAIR OF CLUSTERS:

$$j', k' = \underset{j,k}{\operatorname{argmin}}\ \delta_{jk}, \text{ AND } \delta' = \min_{j,k} \delta_{jk}$$

(RESOLVE TIES RANDOMLY)

   5. OPTIONALLY OUTPUT $m, \hat{K}, \delta', j', k'$
   6. IF H.C. IS BASED ON δ', TEST FOR IT (HALT IF TRUE) ($\rightarrow$ e.g., $\delta' \geq \delta_{halt}$)
   7. MERGE CLUSTERS $C_{j'}$ AND $C_{k'}$ TO FORM A NEW CLUSTER C_l [APPLY MERGE RULE]

8. UPDATE $\hat{K} \leftarrow \hat{K} - 1$, ITERATION $m \leftarrow m+1$

9. IF H.C. IS BASED ON $\hat{K}$, TEST FOR IT (HALT IF TRUE).

10. OUTPUT FINAL CLUSTERS C_ℓ, $\ell = 1, \cdots, \hat{K}_{FINAL}$; $\hat{K}_{FINAL}$; DISSIMILARITY $\delta(m)$.

IF $\hat{K}_{FINAL} = 1$, THE RESULTING HIERARCHY IS A DENDOGRAM.

USEFUL DISTANCE OR DISSIMILARITY MEASURES: (BETWEEN CLUSTERS C_k, C_ℓ):

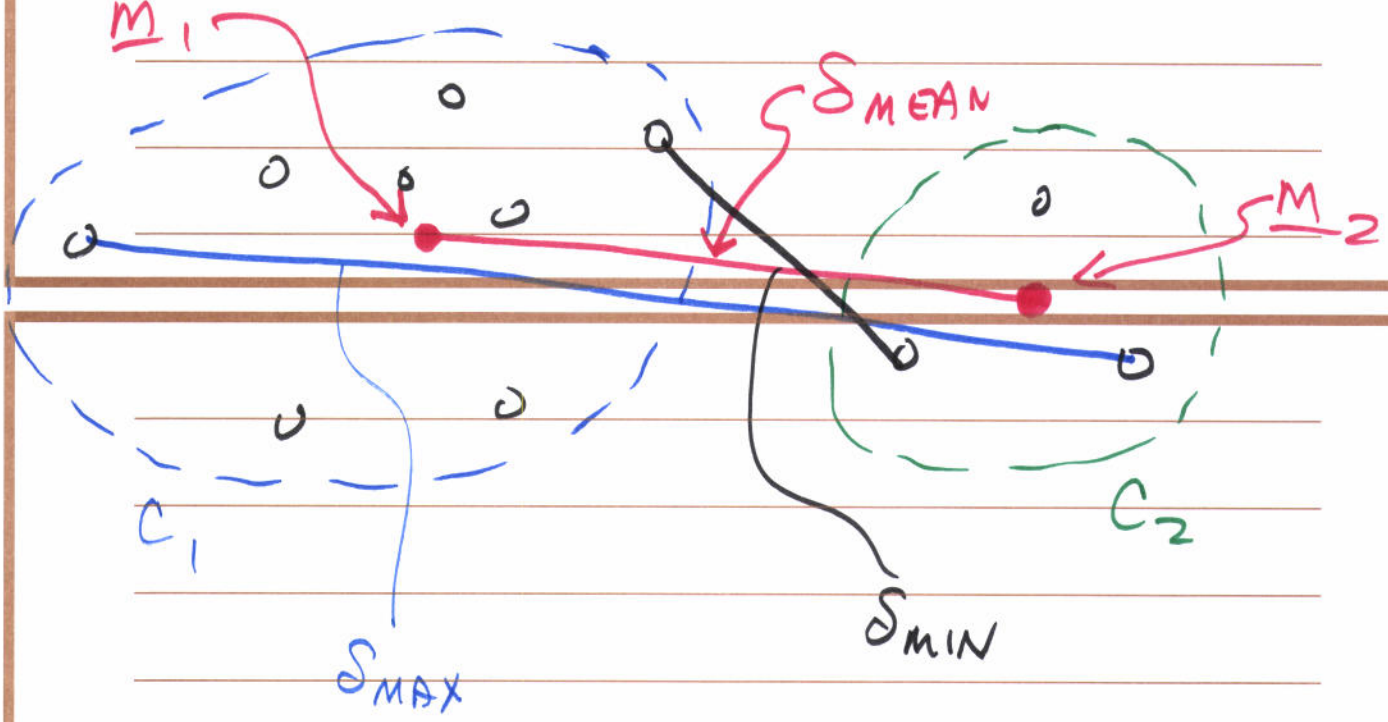
$$\delta_{MEAN}(C_k, C_\ell) \triangleq \| \underline{m}_k - \underline{m}_\ell \|_2$$

$$\delta_{AVG}(C_k, C_\ell) \triangleq \frac{1}{N_k N_\ell} \sum_{\underline{x} \in C_k} \sum_{\underline{x}' \in C_\ell} \| \underline{x} - \underline{x}' \|_2$$

(MEAN OF DISTANCES BETWEEN ALL PAIRS OF PTS, ONE FROM EACH CLUSTER).

$$\delta_{MIN}(C_k, C_l) = \min_{\substack{\underline{x} \in C_k \\ \underline{x}' \in C_l}} \| \underline{x} - \underline{x}' \|_2$$

$$\delta_{MAX}(C_k, C_l) = \max_{\substack{\underline{x} \in C_k \\ \underline{x}' \in C_l}} \| \underline{x} - \underline{x}' \|_2$$



ξ $\delta_{AVG.}$ = MEAN OF ALL $d_{\underline{x}\underline{x}'}$

$\underline{x} \in C_1$

$\underline{x}' \in C_2$.