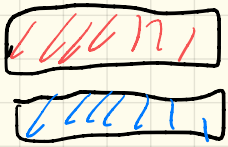
Clustering Validation & Assessment

10/2

3 types of tasks:
- evaluation: assessing the quality of the clustering
- stability: understand sensitivity of of clustering
- tendency: suitability for clustering (generally hard) ← not going to cover much

Types of measures:
- external: use criterion not in the dataset
  e.g. labels proved by an expert
- internal: use criterion from the data
  e.g. "compactness" of cluster
- relative: compare algos
  e.g. 1 param set vs another

External measures
Assume correctness is already known
- labeld data
- synthetic data



Let $\mathbb{D} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ $\vec{x}_i \in \mathbb{R}^d$

$\Rightarrow$ points in $\mathbb{D}$ are partitioned into k clusters

$y_i \in \{1, \ldots, k\}$ is the ground truth (label) for point $\vec{x}_i$

$\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ are ground truth partitioning

e.g. $T_j = \{\vec{x}_i \in \mathbb{D} \mid y_i = j\}$ each T_j is a partition

$m_j = |T_j|$

And $\mathcal{C} = \{C_1, C_2, \ldots, C_r\}$ be a clustering produced by some algo

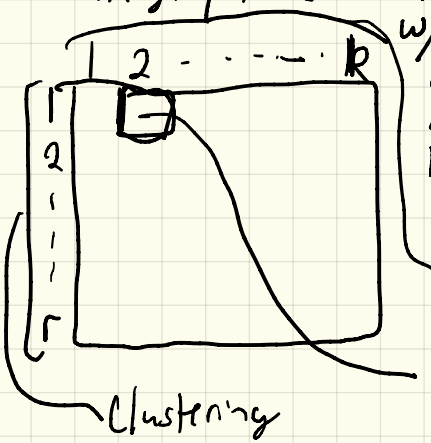
C_i is a cluster

$\hat{y}_i \in \{1, \ldots, r\}$ be the label (from clustering) for $\vec{x}_i$

$n_i = |C_i|$

Contingency table $r \times k$ matrix w/ i,j entry containing # of points in C_i and T_j

$N(i,j) = n_{ij} = |C_i \cap T_j|$



$N(1,2) = |C_1 \cap T_2|$

How long to compute contingency table?

$\Theta(kr)$ for init to 0s

$\Theta(n)$ for incrementing

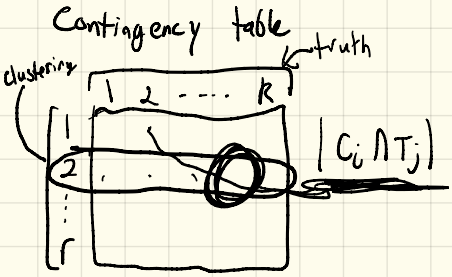
Matching based measures

Purity - measure aggreement w/ ground truth
value between $[0,1]$ w/ 1 being all points are in 1 partition

$$\text{purity}_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}$$

$$\text{purity} = \sum_{i=1}^{r} \frac{n_i}{n} \text{purity}_i$$

$$= \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k} \{n_{ij}\}$$



$[0,1]$ w/ 1 $\Rightarrow$ each cluster contains points from only 1 partition

Internal measures

No labels so use intrinsic props:
- similarity
- compactness
- separation

Given a pairwise dist matrix

$$W = \{\|\vec{x}_i - \vec{x}_j\|\}_{i,j=1}^{n}$$

Note 1: $\|\vec{x}_i - \vec{x}_j\| = \|\vec{x}_j - \vec{x}_i\|$ } upper Δ w/ diag
$\|\vec{x}_i - \vec{x}_i\| = 0$

Note 2: W can also be thought of as a complete undirected graph w/ edge weights as dist between pts

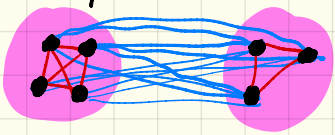
Consider as a k-way cut

Let $V = \bigcup_i C_i$

Given subsets $S, R \subset V$

define $W(S,R) = \sum_{\vec{x}_i \in S} \sum_{\vec{x}_j \in R} w_{ij}$

E.g. pick S & R to be clusters



- edges between clusters
  intercluster edges
  N_{out} # of intercluster edges

- edges w/in cluster
  intracluster edges
  N_{in} # of intracluster edges

Intracluster weight

$$W_{in} = \frac{1}{2} \sum_{l=1}^{k} W(C_l, C_l)$$

Intercluster weight

Let $\bar{S} = V \setminus S$

$$W_{out} = \frac{1}{2} \sum_{i=1}^{k} W(C_i, \bar{C}_i)$$