Lab Meeting 9/16/13

Nature Methods UCLA
digitally scanned light sheet
(original sheet done w/ elliptical lens
→ susceptible to higher order aberrations)

Rand derivation notes

Contingency table for clusterings

$X = \{\{x_1, x_2\}, \{x_3, x_4, x_5\}, \ldots\}$ (X_1, X_2, X_i)

$Y = \{\{x_1\}, \{x_2, x_3\}, \{x_4, x_5, x_6, x_7\}, \ldots\}$ (Y_1, Y_2, Y_j)

X/Y	Y_1	Y_2	...	Y_j
X_1	n_{11}	n_{12}	...	n_{1j}
X_2	n_{21}	n_{22}		
⋮	⋮	⋮	⋱	
X_i	n_{i1}			n_{ij}

n_{ij} = # of elements in common btw cluster X_i and Y_j

Important identity: $\binom{N}{2} = \frac{N(N-1)}{2}$

See Fowlkes + Mallows:

~~$T_K =$~~

See Meilă:

N_{11} # point PAIRS in same cluster in X + Y
N_{01} " " " in same cluster in ~~X~~ Y, diff clusters in X
N_{10} " " " " " " in X, " " " ~~Y~~
N_{00} " " " in diff clusters in X and Y

~~$RI(X,Y) =$~~ $N_{11} + N_{01} + N_{10} + N_{00} = \binom{N}{2}$

where N is # of points (total # possible pairs)

$$\text{Rand}(X,Y) = \frac{\#\text{ pairs together in } X/Y \text{ or apart in } X/Y}{\#\text{ possible pairs}}$$

$$= \frac{N_{11} + N_{00}}{\binom{N}{2}}$$

See Fowlkes + Mallows:

$$T_K = 2N_{11} = 2\sum_i \sum_j \binom{n_{ij}}{2} \rightarrow \#\text{ of pairs together in } X \text{ and } Y = \tfrac{1}{2}T_K$$

$$= 2\sum\sum \frac{n_{ij}(n_{ij}-1)}{2}$$

$$= \sum\sum n_{ij}^2 - \sum\sum n_{ij} = N \ (\#\text{ pts})$$

$$P_K = 2\sum_i \binom{\sum_j n_{ij}}{2} \rightarrow \#\text{ of pairs in } Y = \tfrac{1}{2}P_K$$

$$= \sum_i \left(\sum_j n_{ij}\right)^2 - N$$

$$Q_K = 2\sum_j \left(\sum_i n_{ij}\right)^2 - N \quad \#\text{ of pairs in } X = \tfrac{1}{2}Q_K$$

$$\#\text{pairs in } Y + \#\text{pairs in } X = 2N_{11} + N_{01} + N_{10}$$

2 b/c Common pairs counted twice in sum

$$\Rightarrow \tfrac{1}{2}P_K + \tfrac{1}{2}Q_K = 2N_{11} + N_{01} + N_{10} \qquad 2N_{11} = T_K$$

$$\textcircled{1} \quad N_{01} + N_{10} = \tfrac{1}{2}[P_K + Q_K] - T_K$$

$$N_{11} + N_{01} + N_{10} + N_{00} = \binom{N}{2}$$

$$N_{11} + N_{00} = \binom{N}{2} - [N_{01} + N_{10}] \quad (2)$$

$$(1) + (2) \Rightarrow N_{11} + N_{00} = \binom{N}{2} - \underbrace{\left[\tfrac{1}{2}(P_K + Q_K) - T_K\right]}_{N\text{'s cancel out}}$$

$$\rightarrow \frac{N_{11} + N_{00}}{\binom{N}{2}} = \text{Rand}(X, Y)$$

$$= \frac{\binom{N}{2} - \frac{1}{2}\left[\sum_i \left(\sum_j n_{ij}\right)^2 + \sum_j \left(\sum_i n_{ij}\right)^2\right] + \sum\sum n_{ij}}{\binom{N}{2}}$$

Ranges $\emptyset \rightarrow 1$, Adjusted Rand Index (normalized after subtract expected value)

$$= \frac{\text{Rand}(X, Y) - E[R]}{1 - E[R]}$$ Meilă

$\Rightarrow$ subtract expected value for randomized clusters and then normalize by range

So new range is still $\emptyset \rightarrow 1$ although some really ~~bad~~ "different" clusterings can produce negative vals