Clustering 성능 기준

1) 정확한 답(cluster의 갯수 및 소속)을 알고있는 경우.

X\Y	Y_1	Y_2	----	Y_s	Sums
X_1	n_{11}	n_{12}	----	n_{1s}	a_1
X_2	n_{21}	n_{22}	----	n_{2s}	a_2
⋮	⋮	⋮	⋱	⋮	⋮
X_r	n_{r1}	n_{r2}	----	n_{rs}	a_r
Sums	b_1	b_2	----	b_s	

- Adjusted Rand Index

1) n_{ij} : X clustering 에서는 cluster X_i 였는데, Y clustering 에서는 같은 cluster Y_j 인 Data의 수.

2) a_i : X clustering 에서는 같은 cluster X_i 였는데, Y clustering 에서도 같은 cluster의 Data 쌍의 수

3) b_j : X clustering 에서는 다른 cluster 였는데, Y cluster 에서도 다른 cluster Y_j 인 데이터 쌍의 수.

$$\underset{\text{ARI}}{\text{(Adjusted Index)}} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\left(^{a_i}\right.\right.}{\underbrace{\tfrac{1}{2}\left[\sum_i\binom{a_i}{2} + \sum\binom{b_i}{2}\right]}_{\text{Max Index}} - \underbrace{\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}}_{\text{Expected Index}}}$$

- Adjusted Mutual Information.

$X = \{X_1, X_2, \cdots, X_r\}$

$Y = \{V_1, V_2, \cdots, V_s\}$

$P(i) = \frac{|X_i|}{N}$

$P(i,j) = \frac{|X_i \cap Y_j|}{N}$

$$MI(X,Y) = \sum_{i=1}^{r}\sum_{j=1}^{s} P(i,j)\log\frac{P(i,j)}{P(i)P(j)}$$