① Objective of k-means:

Given $d_1, d_2, \cdots, d_N$ documents, we want to find a partition of $C = \{ C_1, C_2, \cdots, C_k \}$, such that

$$\arg\min_{C} \sum_{j=1}^{k} \sum_{i \in C_j} \| x_i - \mu_j \|^2$$

② Purity for clustering:

N: total # of documents

k: total # of clusters

n_i: # of documents in cluster i

m_{ij}: # of instances in cluster i that belong to golden class j.

$$\text{purity} = \frac{1}{N} \sum_{i=1}^{k} \left\{ \max_{j} m_{ij} \right\}.$$

Alternative definition:

clusters $C_1 \quad C_2 \quad \cdots \quad C_k$

golden classes $g_1 \quad g_2 \quad \cdots \quad g_t$

$$\text{Purity} = \frac{1}{N} \sum_{i=1}^{k} \left\{ \max_{j=1}^{t} | C_i \cap g_j | \right\}.$$

③ Mixture model:

$p(x)$: prob. model for text data

$p(x)$ is mixture model $\rightarrow$ means a mixture of uni-model distribution

Gaussian mixture model: k- Gaussian components

$$p(x_n) = \sum_{k} p(x_n | z_n = k) \cdot p(z_n = k)$$
$$= \sum_{k} N(x_n | \mu_k, \Sigma_k) \cdot \pi_k$$

This probability model describes how each data point x_n can be generated:

- Step. 1: flip k-sided die, with prob. π_k for kth side to select cluster c.
- Step 2: generate the values of the data point from $N(\mu_c, \Sigma_c)$.

para: $\theta = \{\mu_k, \Sigma_k, \pi_k, k=1,2,\cdots k\}$.

$\theta_i = \{\mu_i, \Sigma_i, \pi_i\}$.

x_i: observed sample data

z_i: $\{z_i^1, \cdots, z_i^k\}$. unobserved cluster labels.

$z_i^j \in \{0,1\}$.

④. Submodular function:

$\mathcal{U} = \{s_1, s_2, \cdots s_n\}$.

$F: 2^n \to \mathbb{R}$.

Let $V = [n]$, $A \subseteq B \subseteq V \setminus v$.

$F(A \cup v) - F(A) \geq F(B \cup v) - F(B)$ For $\forall A, B$.

diminishing return.