Document Clustering

I. Why directly conducting Maximum Likelihood method for latent variable models is problematic and how the Expectation Maximization (EM) algorithm solves that problem.

Solution:

Expectation Maximization (EM) is a maximum likelihood algorithm for probabilistic models with latent variables. The goal of the EM algorithm is to find maximum likelihood solution for models having latent variables.

The original log-likelihood function for latent variable models if given by:

$$\ln p(X \mid \theta) = \ln \sum_{Z} p(X, Z \mid \theta) = \ln \sum_{Z} p(Z \mid \theta)\, p(X \mid Z, \theta)$$

For the above log-likelihood function, it is not feasible to directly maximize the function since the latent variable Z for the corresponding dataset X is not known, i.e., we cannot use the complete-data log likelihood.

However, to maximize the likelihood of incomplete data, our state of knowledge of the values of the latent variables in Z is given only by the posterior distribution $p(Z \mid X, \theta)$. Since we cannot use the complete-data log likelihood, we consider instead its *expected* value under the posterior distribution of the latent variable, which corresponds to the E- step of the EM algorithm.

Thus, the EM algorithm is as follows:

- In the *E-step*, we use the current parameter values θ^{old} to find the posterior distribution of the latent variables given by $p(Z \mid X, \theta^{old})$. We then use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value θ. This expectation, denoted by $\mathrm{Q}(\theta, \theta^{old})$, is given by:

$$Q(\theta, \theta^{old}) := \sum_{Z} p(Z \mid X, \theta^{old}) \ln p(X \mid Z, \theta)$$

- In the *M step*, we determine the revised parameter estimate θ^{new} by maximising the Q function:

$$\theta^{new} \leftarrow \arg max_{\theta}\, \mathrm{Q}(\theta, \theta^{old})$$

Thus, the EM algorithm aids in solving latent variables problem by maximizing the expected value of the complete-data log likelihood.

Original maximization expression:

$$\arg max_{\theta} : \ln \sum_{Z} p(X, Z \mid \theta)$$

New maximization expression:

$$\arg max_{\theta} : \sum_{Z} p(Z \mid X, \theta^{old}) \ln p(X \mid Z, \theta)$$

II. Derive **Expectation** and **Maximization** steps of the hard-EM algorithm for Document Clustering. In Include all the model parameters that should be learnt and the exact expression that should be used to update these parameters during the learning process (ie., E step, M step and assignments).

Solution:

For complete data, the probability of generating a pair of document and its cluster (k, d) is

$$p(k,d) = p(k)p(d|k)$$
$$= \varphi_k \prod_{w \epsilon A} {\mu_{k,w}}^{c(w,d)}$$

$\varphi_k \leftarrow$ cluster proportion
$\mu_{k,w} \leftarrow$ word proportion corresponding to cluster k
$c(w,d) \leftarrow$ occurrences of word w in doc d

Since the document clusters (k) are not given,

For parameter θ, **data likelihood** $\mathrm{p}(d_n|\,\theta) = \sum_{k=1}^{K} \varphi_k \prod_{w\epsilon A} {\mu_{k,w}}^{c(w,d)}$

∴ **Probability of observed documents:**

$\mathrm{p}(d_1, d_2, \ldots., d_N|\,\theta) = \prod_{n=1}^{N} \mathrm{p}(d_n|\,\theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} (\varphi_k \prod_{w\epsilon A} {\mu_{k,w}}^{c(w,d)})$

Posterior ($\gamma(\,z_{nk})$):

$p(\,z_{nk} = 1\,|d_n, \theta) = \frac{joint\ probability}{data\ likelihood} = \frac{p(z_{nk}=1, d_n\,|\,\theta)}{\mathrm{p}(d_n|\,\theta)}$

$\gamma(\,z_{nk}) = \frac{\varphi_k \prod_{w\epsilon A} {\mu_{k,w}}^{c(w,d)}}{\sum_{k=1}^{K} \varphi_k \prod_{w\epsilon A} {\mu_{k,w}}^{c(w,d)}}$

Maximum log-likelihood:

$$\ln \mathrm{p}(d_1, d_2, \ldots., d_N|\,\theta) = \sum_{n=1}^{N} \ln\,(\mathrm{p}(d_n|\,\theta))$$
$$= \sum_{n=1}^{N} \ln\, \sum_{k=1}^{K} p(z_{n,k} = 1, d_n)$$
$$= \sum_{n=1}^{N} \ln\, \sum_{k=1}^{K} (\varphi_k \prod_{w\epsilon A} {\mu_{k,w}}^{c(w,d)})$$

Since the above expression contains latent variables, we use EM algorithm.

$$\mathbf{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) := \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{nk} = 1\,|\,d_n,\ \theta^{old})\ \ln p(z_{nk} = 1,\ d_n\,|\,\theta)$$
$$:= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})\,(\ln \varphi_k + \sum_{w\epsilon A} c(w, d_n) \ln \mu_{k,w})$$

Where,
$\theta \coloneqq (\varphi,\ \mu_1,\ \mu_2, \ldots, \mu_k)$
$\gamma(z_{nk}) \coloneqq p(z_{nk} = 1\,|\,d_n,\ \theta^{old})$

To maximise the Q function, form the Lagrangian to enforce the constraints, and set the derivatives to zero which leads to the following solution for the model parameters:

- The mixing components: $\varphi_k = \frac{N_k}{n}$, where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

- The word proportion parameters for each cluster: $\mu_{k,w} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\ c(w,d_n)}{\sum_{w' \in A}\sum_{n=1}^{N}\gamma(z_{nk})\ c(w',d_n)}$

Hard EM for Document Clustering

- Choose an initial setting for the parameters $\theta^{old} = (\varphi^{old},\ \mu_1{}^{old},\ \mu_2{}^{old}, \dots, \mu_k{}^{old})$
- While the convergence is not met:
  - **E-Step:** Set $\forall$n, $\forall$k : $\boldsymbol{\gamma(z_{nk})}$ based on θ^{old}

    $\arg max_z \boldsymbol{\gamma(z_{nk})} = \arg max_z\ p(z_{nk} = 1 \mid d_n, \theta^{old})$

    $Z^* = \begin{cases} \mathbf{1}, & \forall \arg max_z \gamma(z_{nk}) \\ \mathbf{0}, & rest\ of\ \gamma(z_{nk}) \end{cases}$

  - **M-Step:** $\theta^{new} \leftarrow \arg max_\theta\ \mathrm{Q}(\theta, \theta^{old})$

    The revised parameter estimates after maximising $\mathrm{Q}(\theta, \theta^{old})$ are:

    $\varphi_k{}^{new} = \frac{N_k}{n}$

    $\mu_{k,w}{}^{new} = \frac{\sum_{n=1}^{N} Z^* \times c(w,d_n)}{\sum_{w' \in A}\sum_{n=1}^{N} Z^* \times c(w',d_n)}$

  - $\theta^{old} \leftarrow \theta^{new}$