Problem 1: Probabilistic Latent Semantic Analysis (10 points)

You are provided with a toy dataset consisting of two documents and a vocabulary of four words: $\{1 : A, 2 : B, 3 : C, 4 : D\}$. The documents are represented in a bag-of-words model as follows:

- Document d_1: (4,3,2,1) — indicating 4 occurrences of A, 3 of B, 2 of C, and 1 of D.
- Document d_2: (2,2,3,1) — indicating 2 occurrences of each A and B, 3 of C, and 1 of D.

Let θ_{ij} be the probability of topic j in document i (e.g., $P(z_1 = 1 \mid d_1) = \theta_{11}$). Let β_{zw} be the probability of word w given topic z. Initialize the parameters as follows:

- $\theta_{11}^{(0)} = 0.3$, $\theta_{21}^{(0)} = 0.4$.
- $\beta_1^{(0)} = (1, 0, 0, 0)$, $\beta_2^{(0)} = (0, 0.4, 0.3, 0.3)$.

1. **(5 points)** E-Step Calculation: Compute $P(z = 1 \mid w, d_1)$ for all words in d_1 using the initialized values.
2. **(5 points)** M-Step Calculation: Given the additional information for document d_2:
   - $P(z = 1 \mid A, d_2) = 1$
   - $P(z = 1 \mid B, d_2) = 0$
   - $P(z = 1 \mid C, d_2) = 0$
   - $P(z = 1 \mid D, d_2) = 0$

   Use your results from the E-step to compute the new values of β_{11}, β_{12}, θ_{11}, and θ_{12}.

Answer 1

To compute $P(z = 1|w, d_1)$ for all words in d_1, need to use the E-step of the Expectation-Maximization (EM) algorithm in Probabilistic Latent Semantic Analysis (PLSA). In the E-step, we calculate the posterior probability of the latent variable z given the observed variables w and d_1.

Given the initialized values:

$$\theta_{11}^{(0)} = 0.3$$

$$\theta_{21}^{(0)} = 0.4$$

$$\beta_1^{(0)} = (1,0,0,0)$$

$$\beta_2^{(0)} = (0,0.4,0.3,0.3)$$

Let's calculate $P(z = 1|w, d_1)$ for each word in d_1:

For word A:

$P(z = 1|w = A, d_1)$ = (P(w = A | z = 1) * P(z = 1 | d_1) / P(w = A | d_1)

P(w = A | z = 1) = $\beta_{11}^{(0)}$ = 1

P(z = 1 | d_1) = $\theta_{11}^{(0)}$ = 0.3

To calculate P(w = A | d_1), we use the law of total probability:

$P(w = A \mid d_1) = P(w = A, z = 1 \mid d_1) + P(w = A, z = 2 \mid d_1)$
$= P(w = A \mid z = 1) * P(z = 1 \mid d_1) + P(w = A \mid z = 2) * P(z = 2 \mid d_1)$
$= \beta_{11}^{(0)} * \theta_{11}^{(0)} + \beta_{12}^{(0)} ** \theta_{21}^{(0)}$
Substituting the values, we get:
$P(w = A \mid d) = 1 * 0.3 + 0 * 0.4 = 0.3$
Now we can calculate $P(z = 1 \mid w = A, d_1)$:
$P(z = 1 \mid w = A, d_1) = (1 * 0.3) / 0.3 = 1$

For word B:

$P(z = 1 \mid w = B, d_1) = (P(w = B \mid z = 1) * P(z = 1 \mid d_1)) / P(w = B \mid d_1)$

$P(w = B \mid z = 1) = \beta_{21}^{(0)} = 0$

$P(z = 1 \mid d_1) = \theta_{11}^{(0)} = 0.3$

$P(w = B \mid d_1) = \beta_{21}^{(0)} * \theta_{11}^{(0)} + \beta_{22}^{(0)} * \theta_{21}^{(0)} = 0 * 0.3 + 0.4 * 0.4 = 0.16$

$P(z = 1 \mid w = B, d_1) = (0 * 0.3) / 0.16 = 0$

For word C:

$P(z = 1 \mid w = C, d_1) = (P(w = C \mid z = 1) * P(z = 1 \mid d_1)) / P(w = C \mid d_1)$

$P(w = C \mid z = 1) = \beta_{31}^{(0)} = 0$

$P(z = 1 \mid d_1) = \theta_{11}^{(0)} = 0.3$

$P(w = C \mid d_1) = \beta_{31}^{(0)} * \theta_{11}^{(0)} + \beta_{32}^{(0)} * \theta_{21}^{(0)} = 0 * 0.3 + 0.3 * 0.4 = 0.12$

$P(z = 1 \mid w = C, d_1) = (0 * 0.3) / 0.12 = 0$

For word D:

$P(z = 1 \mid w = D, d_1) = (P(w = D \mid z = 1) * P(z = 1 \mid d_1)) / P(w = D \mid d_1)$

$P(w = D \mid z = 1) = \beta_{41}^{(0)} = 0$

P(z = 1 | d_1) $= \theta_{11}^{(0)} = 0.3$ Apologies for the abrupt interruption. To continue the calculation:

$P(w = D \mid d_1) = \beta_{41}^{(0)} * \theta_{11}^{(0)} + \beta_{42}^{(0)} * \theta_{21}^{(0)} = 0 * 0.3 + 0.3 * 0.4 = 0.12$

$P(z = 1 \mid w = D, d_1) = (0 * 0.3) / 0.12 = 0$
Therefore, the values of $P(z = 1 \mid w, d_1)$ for all words in d_1 are as follows:
$P(z = 1 \mid w = A, d_1) = 1$
$P(z = 1 \mid w = B, d_1) = 0$
$P(z = 1 \mid w = C, d_1) = 0$
$P(z = 1 \mid w = D, d_1) = 0$

Answer 2

Update the values of β_{11}, β_{12}, θ_{11}, and θ_{12} as follows:

Update β_{11}:
β_{11} = (sum over d1 of $P(z = 1 \mid w, d_2)$ * count of word w in d_2) divided by (sum over d1 of $P(z = 1 \mid w, d_2)$ * total word count in d_2)
Using the values from the E-step:
$P(z = 1 \mid w = A, d_2) = 1$
$P(z = 1 \mid w = B, d_2) = 0$
$P(z = 1 \mid w = C, d_2) = 0$
$P(z = 1 \mid w = D, d_2) = 0$

Count of A in d1 = 4

Total word count in d_2 = 4 + 3 + 2 + 1 = 10

$\beta_{11} = (1 * 4) / (1 * 10) = 0.4$

Update β_{12}:
Since β_{12} represents the probability of word w given topic z = 2, we need to calculate $P(z = 2 \mid w, d_2)$ for all words in d1:
$P(z = 2 \mid w = A, d_2) = 1 - P(z = 1 \mid w = A, d_2) = 1 - 1 = 0$
$P(z = 2 \mid w = B, d_2) = 1 - P(z = 1 \mid w = B, d_2) = 1 - 0 = 1$
$P(z = 2 \mid w = C, d_2) = 1 - P(z = 1 \mid w = C, d_2) = 1 - 0 = 1$
$P(z = 2 \mid w = D, d_2) = 1 - P(z = 1 \mid w = D, d_2) = 1 - 0 = 1$

Using these values, we can update β_{12} using a similar formula as for β_{11}:

Count of B in d_2 = 3
Count of C in d_2 = 2
Count of D in d_2 = 1

$\beta_{12} = ((1 * 3) + (1 * 2) + (1 * 1)) / (1 * 10) = 0.6$

Update θ_{11}:
θ_{11} = (sum over d1 of $P(z = 1 \mid w, d_2)$ * count of topic z in d_2) divided by (sum over d_2 of count of topic z in d_2)
Count of z = 1 in d_2 = 4

$\theta_{11} = (1 * 4) / (4) = 1$

Update θ_{12}:
Since θ_{12} represents the probability of topic z = 2 in document d_2, calculate it as:
$\theta_{12} = 1 - \theta_{11} = 1 - 1 = 0$

Therefore, the new values of the parameters are:

$\beta_{11} = 0.4$
$\beta_{12} = 0.6$
$\theta_{11} = 1$
$\theta_{12} = 0$

Problem 2: Multinomial Mixture Models (25 points)

One effective approach for understanding and categorizing these documents is by using a multinomial mixture model. This model assumes that each document is generated by a mixture of topics (clusters), where each topic is characterized by a distinct multinomial distribution over words.

Consider a dataset of N documents, where each document i is represented as a bag-of-words vector x_i. Assume there are K clusters (topics) in the dataset and each document's cluster label z_i is sampled from a Categorical distribution: $z_i \sim \text{Categorical}(\pi)$, where π is a probability vector with $P(z = k) = \pi_k$. Further, each cluster z is a multinomial distribution with parameters β_k and the word distribution x_i belonging to cluster z_i is given by $x_i \mid z_i \sim \text{Multinomial}(\beta_k)$.

Your task is to derive the Expectation-Maximization (EM) algorithm for soft document clustering under a multinomial mixture model.

1. **(10 points)** In the E-step, please compute the posterior probabilities of the cluster assignments given the current parameter estimates. Please derive the formula to compute the posterior probability $P(z_i = k \mid x_i; \beta, \pi)$ for each document i and cluster k.

2. **(15 points)** In the M-step, you will re-estimate the parameters β_k and π based on the new posterior probabilities obtained from the E step. Please derive the update rules for the parameters β_k for each cluster k and the mixing proportions π.

Answer 1

In the E-step of the Expectation-Maximization (EM) algorithm for soft document clustering under a multinomial mixture model, we compute the posterior probabilities of the cluster assignments given the current parameter estimates. To derive the formula for computing the posterior probability $P(z_i = k|x_i; \beta, \pi)$ for each document i and cluster k, make use of Bayes' theorem.

Bayes' theorem states:

P(A | B) = (P(B | A) * P(A)) / P(B),

where P(A | B) is the posterior probability of event A given event B, P(B | A) is the likelihood of event B given event A, P(A) is the prior probability of event A, and P(B) is the probability of event B.

In our case, we want to compute the posterior probability $P(z_i = k|x_i; \beta, \pi)$, which represents the probability that document i belongs to cluster k given its feature vector x_i and the current parameter estimates β and π.

Using Bayes' theorem, we can write:

$P(z_i = k|x_i; \beta, \pi) = (P(x_i|z_i = k; \beta, \pi) * P(z_i = k; \beta, \pi))/ P(x_i; \beta, \pi)$
Where $P(x_i|z_i = k; \beta, \pi)$ is the likelihood of observing feature vector x_i given that document i belongs to cluster k, $P(z_i = k; \beta, \pi)$ is the prior probability of document i belonging to cluster k, and $P(x_i; \beta, \pi)$ is the probability of observing feature vector x_i.

The likelihood $P(x_i|z_i = k; \beta, \pi)$ can be obtained from the multinomial distribution:
$P(x_i|z_i = k; \beta, \pi)$=Multinomial($x_i$;$\beta_k$)

where MultiMultinomial(x_i ; β_k)esents the probability mass function of the multinomial distribution with parameters β_k, $\beta_k$ch is the word distribution of cluster k.
The prior probability$P(z_i = k; \beta, \pi)$can be computed as:

$$P(z_i = k; \beta, \pi) = \pi_k$$

where π_k is the mixing proportion or weight associated with cluster k.

The probability of observing feature vector x_i can be written as:

$$P(x_i k; \beta, \pi) = \sum_k (P(x_i|\ z_i = k; \beta, \pi) * P(z_i = k; \beta, \pi))$$

which represents the sum of the likelihoods weighted by the prior probabilities over all clusters.
Putting it all together, the formula for the posterior probability $P(z_i = k; \beta, \pi)$ is

$$P(z_i = k|x_i; \beta, \pi) = (\mathrm{Multinomial}(x_i; \beta_k) * \pi_k)/\sum_k \mathrm{Multinomial}(x_i; \beta_k) * \pi_k$$

Answer 2
In the M-step of the Expectation-Maximization (EM) algorithm for soft document clustering under a multinomial mixture model, we re-estimate the parameters β_k and π based on the new posterior probabilities obtained from the E-step. The update rules for the parameters are as follows:

Updating the word distribution parameters β_k:
The word distribution parameters β_k represent the probabilities of each word in the vocabulary for cluster k. To update these parameters, we can use the weighted maximum likelihood estimator, where the weights are the posterior probabilities from the E-step.

The update rule for β_k is given by:

$$\beta_k = (\sum_i w_{ik} * x_i)/\sum_i \sum_j w_{ij} x_i$$

Where w_{ik} represents the posterior probability of document i belonging to cluster k obtained from the E-step, and x_i is the feature vector of document i. The summation is over all documents i and all words j in the vocabulary.

Essentially, we compute the weighted sum of the feature vectors for all documents assigned to cluster k, and then normalize it by the sum of the weighted feature vectors over all documents and words.

Updating the mixing proportions π:
The mixing proportions π represent the probabilities of each cluster in the mixture. To update these proportions, we can compute the average of the posterior probabilities for each cluster.

The update rule for π_k is given by:

$$\pi_k = (\sum_i w_{ik})/N$$

where w_{ik} represents the posterior probability of document i belonging to cluster k obtained from the E-step, and N is the total number of documents.

compute the sum of the posterior probabilities for all documents assigned to cluster k and normalize it by the total number of documents.

After updating the parameters β_k and π in the M-step, we repeat the E-step and M-step iteratively until convergence, where the convergence criteria can be based on changes in the log-likelihood or the parameters.

These update rules for the parameters β_k and π ensure that the model parameters are iteratively refined based on the updated assignments of documents to clusters, leading to a better fit of the multinomial mixture model to the data.