CS247: Advanced Data Mining (Winter 2024) January 30, 2024

Assignment 3 (Part 1)

Due Date: **February 7, 2023**

Instructions

- Submit your answer on Gradescope as a PDF file. Both typed and scanned handwritten answers are acceptable.
- Submit your solutions to Part 1 and Part 2 through GradeScope in BruinLearn separately.
- Late submissions are allowed up to 24 hours post-deadline with a penalty factor of $\mathbf{1}(t \leq 24)e^{-(\ln(2)/12)t}$.
- Ensure all sources are cited appropriately; plagiarism will be reported.

Problems

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:

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: