4 Programming: Text Clustering

In this problem, we will explore the use of EM algorithm for text clustering. Text clustering is a technique for unsupervised document organization, information retrieval. We want to find how to group a set of different text documents based on their topics. First we will analyze a model to represent the data.

Bag of Words

The simplest model for text documents is to understand them as a collection of words. To keep the model simple, we keep the collection unordered, disregarding grammar and word order. What we do is counting how often each word appears in each document and store the word counts into a matrix, where each row of the matrix represents one document. Each column of matrix represent a specific word from the document dictionary. Suppose we represent the set of n_d documents using a matrix of word counts like this:

$$D_{1:n_d} = \begin{pmatrix} 2 & 6 & ... & 4 \\ 2 & 4 & ... & 0 \\ \vdots & & \ddots & \end{pmatrix} = T$$

This means that word W_1 occurs twice in document D_1 . Word W_{n_w} occurs 4 times in document D_1 and not at all in document D_2.

Multinomial Distribution

The simplest distribution representing a text document is multinomial distribution(Bishop Chapter 2.2). The probability of a document D_i is:

$$p(D_i) = \prod_{j=1}^{n_w} \mu_j^{T_{ij}}$$

Here, μ_j denotes the probability of a particular word in the text being equal to w_j, T_{ij} is the count of the word in document. So the probability of document D_1 would be $p(D_1) = \mu_1^2 \cdot \mu_2^6 \cdot ... \cdot \mu_{n_w}^4 \cdot$

Mixture of Multinomial Distributions

In order to do text clustering, we want to use a mixture of multinomial distributions, so that each topic has a particular multinomial distribution associated with it, and each document is a mixture of different topics. We define $p(c) = \pi_c$ as the mixture coefficient of a document containing topic c, and each topic is modeled by a multinomial distribution $p(D_i|c)$ with parameters μ_{jc}, then we can write each document as a mixture over topics as

$$p(D_i) = \sum_{c=1}^{n_c} p(D_i|c)p(c) = \sum_{c=1}^{n_c} \pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}$$

EM for Mixture of Multinomials

In order to cluster a set of documents, we need to fit this mixture model to data. In this problem, the EM algorithm can be used for fitting mixture models. This will be a simple topic model for documents. Each topic is a multinomial distribution over words (a mixture component). EM algorithm for such a topic model, which consists of iterating the following steps:

1. Expectation

   Compute the expectation of document D_i belonging to cluster c:

$$\gamma_{ic} = \frac{\pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}}{\sum_{c=1}^{n_c} \pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}}$$

2. Maximization

   Update the mixture parameters, i.e. the probability of a word being W_j in cluster (topic) c, as well as prior probability of each cluster.

$$\mu_{jc} = \frac{\sum_{i=1}^{n_d} \gamma_{ic} T_{ij}}{\sum_{i=1}^{n_d} \sum_{l=1}^{m_w} \gamma_{ic} T_{il}}$$

$$\pi_c = \frac{1}{n_d} \sum_{i=1}^{n_d} \gamma_{ic}$$

Task [20 pts]

Implement the algorithm and run on the toy dataset `data.mat`. You can find detailed description about the data in the `homework2.m` file. Observe the results and compare them with the provided true clusters each document belongs to. Report the evaluation (e.g. accuracy) of your implementation.

Hint: We already did the word counting for you, so the data file only contains a count matrix like the one shown above. For the toy dataset, set the number of clusters $n_c = 4$. You will need to initialize the parameters. Try several different random initial values for the probability of a word being W_j in topic c, μ_{jc}. Make sure you normalized it. Make sure that you should not use the true cluster information during your learning phase.

Extra Credit: Realistic Topic Models [20pts]

The above model assumes all the words in a document belongs to some topic at the same time. However, in real world datasets, it is more likely that some words in the documents belong to one topic while other words belong to some other topics. For example, in a news report, some words may talk about "Ebola" and "health", while others may mention "administration" and "congress". In order to model this phenomenon, we should model each word as a mixture of possible topics.

Specifically, consider the log-likelihood of the joint distribution of document and words

$$\mathcal{L} = \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} T_{dw} \log P(d, w), \tag{5}$$

where T_{dw} is the counts of word w in the document d. This count matrix is provided as input.

The joint distribution of a specific document and a specific word is modeled as a mixture

$$P(d, w) = \sum_{z \in \mathcal{Z}} P(z) P(w|z) P(d|z), \tag{6}$$

where $P(z)$ is the mixture proportion, $P(w|z)$ is the distribution over the vocabulary for the z-th topic, and $P(d|z)$ is the probability of the document for the z-th topic. And these are the parameters for the model.

The E-step calculates the posterior distribution of the latent variable conditioned on all other variables

$$P(z|d, w) = \frac{P(z)P(w|z)P(d|z)}{\sum_{z'} P(z')P(w|z')P(d|z')}. \tag{7}$$

In the M-step, we maximizes the expected complete log-likelihood with respect to the parameters, and get the following update rules

$$P(w|z) = \frac{\sum_d T_{dw} P(z|d, w)}{\sum_{w'} \sum_d T_{dw'} P(z|d, w')} \tag{8}$$

$$P(d|z) = \frac{\sum_w T_{dw} P(z|d, w)}{\sum_{d'} \sum_w T_{d'w} P(z|d', w)} \tag{9}$$

$$P(z) = \frac{\sum_d \sum_w T_{dw} P(z|d, w)}{\sum_{z'} \sum_{d'} \sum_{w'} T_{d'w'} P(z'|d', w')}. \tag{10}$$

Task

Implement EM for maximum likelihood estimation and cluster the text data provided in the `nips.mat` file you downloaded. You can print out the top key words for the topics/clusters by using the `show_topics.m` utility. It takes two parameters: 1) your learned conditional distribution matrix, i.e., $P(w|z)$ and 2) a cell array of words that corresponds to the vocabulary. You can find the cell array `wl` in the `nips.mat` file. Try different values of k and see which values produce sensible topics. In assessing your code, we will use another dataset and observe the produces topics.