CS 7641 CSE/ISYE 6740 Homework 2

Deadline: 10/10 Mon, 11:55 pm

- Submit your answers as an electronic copy on T-square.
- No unapproved extension of deadline is allowed. Late submission will lead to 0 credit.
- Typing with Latex is highly recommended. Typing with MS Word is also okay. If you handwrite, try to be clear as much as possible. No credit may be given to unreadable handwriting.
- Explicitly mention your collaborators if any.
- Recommended reading: PRML[1] Section 1.5, 1.6, 2.5, 9.2, 9.3

1 EM for Mixture of Gaussians

Mixture of K Gaussians is represented as

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{1}$$

where π_k represents the probability that a data point belongs to the kth component. As it is probability, it satisfies $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$. In this problem, we are going to represent this in a slightly different manner with explicit latent variables. Specifically, we introduce 1-of-K coding representation for latent variables $z^{(k)} \in \mathbb{R}^K$ for $k = 1, ..., K$. Each $z^{(k)}$ is a binary vector of size K, with 1 only in kth element and 0 in all others. That is,

$$\begin{aligned} z^{(1)} &= [1; 0; ...; 0] \\ z^{(2)} &= [0; 1; ...; 0] \\ &\vdots \\ z^{(K)} &= [0; 0; ...; 1]. \end{aligned}$$

For example, if the second component generated data point x^n, its latent variable z^n is given by $[0; 1; ...; 0] = z^{(2)}$. With this representation, we can express $p(z)$ as

$$p(z) = \prod_{k=1}^{K} {\pi_k}^{z_k},$$

where z_k indicates kth element of vector z. Also, $p(x|z)$ can be represented similarly as

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}.$$

[1] Christopher M. Bishop, Pattern Recognition and Machine Learning, 2006, Springer.

By the sum rule of probability, (1) can be represented by

$$p(x) = \sum_{z \in Z} p(z)p(x|z). \tag{2}$$

where $Z = \{z^{(1)}, z^{(2)}, ..., z^{(K)}\}$.

(a) Show that (2) is equivalent to (1). [5 pts]

Answer:
(a) For $z = z^{(i)}$,

$$\begin{aligned}
p(z^{(i)}) &= \pi_1^{z_1} \cdots \pi_K^{z_K} = \pi_1^0 \cdots \pi_i^1 \cdots \pi_K^0 \\
&= \pi_i \\
p(x|z^{(i)}) &= \mathcal{N}(x|\mu_1, \Sigma_1)^{z_1} \cdots \mathcal{N}(x|\mu_K, \Sigma_K)^{z_K} = \mathcal{N}(x|\mu_1, \Sigma_1)^0 \cdots \mathcal{N}(x|\mu_i, \Sigma_i)^1 \cdots \mathcal{N}(x|\mu_K, \Sigma_K)^0 \\
&= \mathcal{N}(x|\mu_i, \Sigma_i)
\end{aligned}$$

Then to calculate the $p(x)$:

$$\begin{aligned}
p(x) &= \sum_{z \in Z} p(z)p(x|z) \\
&= p(z^{(1)})p(x|z^{(1)}) + p(z^{(2)})p(x|z^{(2)}) + \cdots + p(z^{(K)})p(x|z^{(K)}) \\
&= \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \Sigma_2) + \cdots + \pi_K \mathcal{N}(x|\mu_K, \Sigma_K) \\
&= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)
\end{aligned}$$

Therefore equation (2) is equivalent to (1).

(b) In reality, we do not know which component each data point is from. Thus, we estimate the responsibility (expectation of z_k^n) in the E-step of EM. Since z_k^n is either 1 or 0, its expectation is the probability for the point x_n to belong to the component z_k. In other words, we estimate $p(z_k^n|x_n)$. Derive the formula for this estimation by using Bayes rule. Note that, in the E-step, we assume all other parameters, i.e. π_k, μ_k, and Σ_k, are fixed, and we want to express $p(z_k^n|x_n)$ as a function of these fixed parameters. [10 pts]

Answer:
(b) According to the Bayes rule:

$$P(z|x) = \frac{P(x|z)P(z)}{P(x)} = \frac{P(x,z)}{\sum_{z'} P(x,z')}$$

Then we can obtain:

$$P(z_k^n|x_n) = \frac{P(x_n|z_k^n)P(z_k^n)}{P(x_n)} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_i \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)}$$

(c) In the M-Step, we re-estimate parameters π_k, μ_k, and Σ_k by maximizing the log-likelihood. Given N i.i.d (Independent Identically Distributed) data samples, derive the update formula for each parameter. Note that in order to obtain an update rule for the M-step, we fix the responsibilities, i.e. $p(z_k^n|x_n)$, which we have already calculated in the E-step. [15 pts]

Hint: Use Lagrange multiplier for π_k to apply constraints on it.

Answer:

(c) From the previous E-step, we can obtain the objective function:

$$= E_{q(z^1,z^2\cdots z^n|x^i)}[\log \prod_{i=1}^{n} p(x^i, z^i|\theta)]$$

$$= E_{q(z^1,z^2\cdots z^n|x^i)}[\log \prod_{i=1}^{n} \pi_{z^i} N(x^i|\mu_{z^i}, \Sigma_{z^i})]$$

$$= E_{q(z^1,z^2\cdots z^n|x^i)}[\sum_{i=1}^{n} \log \pi_{z^i} + \log \frac{1}{(2\pi)^{d/2}|\Sigma_{z^i}|^{1/2}} \exp[-\frac{1}{2}(x^i - \mu_{z^i})^T \Sigma_{z^i}^{-1}(x^i - \mu_{z^i})]]$$

$$= \sum_{i=1}^{n} E_{q(z^i|x^i)}[\log \pi_{z^i} - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{z^i}| - \frac{1}{2}(x^i - \mu_{z^i})^T \Sigma_{z^i}^{-1}(x^i - \mu_{z^i})]$$

Let $\tau_k^i = p(z_k^i = k|x^i)$ and simplify the formula:

$$f(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_k^i[\log \pi_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x^i - \mu_k)^T \Sigma_k^{-1}(x^i - \mu_k) - \frac{d}{2}\log 2\pi]$$

Form Lagrangian constrains:

$$L = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_k^i[\log \pi_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x^i - \mu_k)^T \Sigma_k^{-1}(x^i - \mu_k) - \frac{d}{2}\log 2\pi] + \lambda(1 - \sum_{k=1}^{K}\pi_k)$$

(1)Take partial derivative with respect to π_k and let it equal to 0:

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{n} \frac{\tau_k^i}{\pi_k} - \lambda = 0 \Rightarrow \pi_k = \frac{\sum_{i=1}^{n}\tau_k^i}{\lambda}$$

Since $\sum_{k=1}^{K}\pi_k^i = 1$, then we can plugin to π_k:

$$\sum_{k=1}^{K}\frac{1}{\lambda}\sum_{i=1}^{n}\tau_k^i = 1 \Rightarrow \frac{1}{\lambda}n = 1 \Rightarrow \lambda = n \Rightarrow \pi_k = \frac{\sum_{i=1}^{n}\tau_k^i}{n}$$

(2)Take partial derivative with respect to μ_k and let it equal to 0:

$$\frac{\partial L}{\partial \mu_k} = \sum_{i=1}^{n}\tau_k^i \Sigma_k^{-1}(x^i - \mu_k) = 0 \Rightarrow \mu_k = \frac{\sum_{i=1}^{n}\tau_k^i x^i}{\sum_{i=1}^{n}\tau_k^i}$$

(3)To partial derive the equation with respect to Σ_k, we need to recall some results from matrix algebra. Since $x^i - \mu_k)^T\Sigma_k^{-1}(x^i - \mu_k)$ is scalar, the trace of a scalar equals that scalar. Also, $tr(AB) = tr(BA)$. Thus $tr((x^i - \mu_k)^T\Sigma_k^{-1}(x^i - \mu_k)) = tr(\Sigma_k^{-1}(x^i - \mu_k)^T(x^i - \mu_k))$. Therefore

$$\frac{\partial}{\partial \Sigma_k^{-1}}\log|\Sigma_k^{-1}| = \Sigma_k^T \quad \texttt{Using C.28 from PRML appendix}$$

$$\frac{\partial}{\partial \Sigma_k^{-1}} tr(\Sigma_k^{-1}(x^i - \mu_k)^T(x^i - \mu_k)) = (x^i - \mu_k)(x^i - \mu_k)^T \quad \texttt{Using C.24 from PRML appendix}$$

$$\frac{\partial L}{\partial \Sigma_k} = \sum_{i=1}^{n}\tau_k^i\Sigma_k - \sum_{i=1}^{n}\tau_k^i(x^i - \mu_k)(x^i - \mu_k)^T = 0 \Rightarrow \Sigma_k = \frac{\sum_{i=1}^{n}\tau_k^i(x^i - \mu_k)(x^i - \mu_k)^T}{\sum_{i=1}^{n}\tau_k^i}$$

(d) EM and K-Means [10 pts]

K-means can be viewed as a particular limit of EM for Gaussian mixture. Considering a mixture model in which all components have covariance ϵI, show that in the limit $\epsilon \to 0$, maximizing the expected complete data log-likelihood for this model is equivalent to minimizing objective function in K-means:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \|x_n - \mu_k\|^2,$$

where $\gamma_{nk} = 1$ if x_n belongs to the k-th cluster and $\gamma_{nk} = 0$ otherwise.
(d)Answer: Since all components have covariance $\Sigma_k = \epsilon I$, so that the mixture of K Gaussians is:

$$N(X|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}(\epsilon)^{1/2}} \exp\{-\frac{1}{2\epsilon}(X-\mu)^T(X-\mu)\}$$

The posterior probability is:

$$\begin{aligned} \tau_k^n &= \frac{\pi_k N(x^n|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} N(x^n|\mu_{k'}, \Sigma_{k'})} \\ &= \frac{\pi_k \exp\{-(x^n-\mu_k)^T(x^n-\mu_k)/2\epsilon\}}{\sum_{k'=1}^{K} \pi_{k'} \exp\{-(x^n-\mu_{k'})^T(x^n-\mu_{k'})/2\epsilon\}} \end{aligned}$$

If the limit $\epsilon \to 0$, then for a particular data point x^i which can minimize $||x^i - \mu_k||$(closest to 0), it can let:

$$\begin{aligned} \tau_k^i &= \frac{\pi_k \exp\{-(x^i-\mu_k)^T(x^i-\mu_k)/2\epsilon\}}{\sum_{k'=1}^{K} \pi_{k'} \exp\{-(x^i-\mu_{k'})^T(x^i-\mu_{k'})/2\epsilon\}} \\ &= \frac{\pi_k}{\sum_{k'=1}^{K} \pi_{k'}(0+\cdots+1+\cdots+0)} = 1 \end{aligned}$$

The expected complete-data log likelihood becomes:

$$\begin{aligned} f(\theta) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_k^i [\log \pi_k - \frac{1}{2}\log\epsilon - \frac{1}{2\epsilon}(x^i-\mu_k)^T(x^i-\mu_k) - \frac{d}{2}\log 2\pi] \\ &= -\frac{1}{2\epsilon} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_k^i (x^i-\mu_k)^T(x^i-\mu_k) + \text{OTHER TERMS} \end{aligned}$$

Therefore, to maximize the log likelihood with respect to μ_k or τ_k^i is equal to minimize $(x^i-\mu_k)^T(x^i-\mu_k)$ which can represent in another way, $||x^i - \mu_k||^2$. In the limit $\epsilon \to 0$, maximizing the expected complete data log-likelihood for this model is equivalent to minimizing objective function in K-means.

2 Density Estimation

Consider a histogram-like density model in which the space x is divided into fixed regions for which density $p(x)$ takes constant value h_i over ith region, and that the volume of region i in denoted as Δ_i. Suppose we have a set of N observations of x such that n_i of these observations fall in regions i.

(a) What is the log-likelihood function? [8 pts]

(a)Answer: log-likelihood function, supposed there are totally M regions:

$$\log \prod_{n=1}^{N} p(x^n) = \log \prod_{i=1}^{M} {h_i}^{n_i} = \sum_{i=1}^{M} n_i \log h_i$$

(b) Derive an expression for the maximum likelihood estimator for h_i. [10 pts]

Hint: This is a constrained optimization problem. Remember that $p(x)$ must integrate to unity. Since $p(x)$ has constant value h_i over region i, which has volume Δ_i. The normalization constraint is $\sum_i h_i \Delta_i = 1$. Use Lagrange multiplier by adding $\lambda (\sum_i h_i \Delta_i - 1)$ to your objective function.
(b)Answer:

$$L = \sum_{i=1}^{M} n_i \log h_i + \lambda \left(\sum_i h_i \Delta_i - 1 \right)$$

Partial derive the likelihood function with respect to h_i, and let it equal to 0:

$$\frac{\partial L}{\partial h_i} = \frac{n_i}{h_i} + \lambda \Delta_i = 0 \Rightarrow h_i = -\frac{n_i}{\lambda \Delta_i}$$

Partial derive the likelihood function with respect to λ, and let it equal to 0:

$$L = \sum_{i=1}^{M} n_i(-\log n_i - \ log\lambda - \log \Delta_i) + \lambda(-\sum_{i=1}^{M} \frac{n_i}{\lambda} - 1)$$

$$\frac{\partial L}{\partial \lambda} = -\frac{\sum_{i=1}^{M} n_i}{\lambda} - 1 = -\frac{N}{\lambda} - 1 = 0 \Rightarrow \lambda = -N$$

Therefore:

$$h_i = \frac{n_i}{N \Delta_i}$$

(c) Mark T if it is always true, and F otherwise. Briefly explain why. [12 pts]

- Non-parametric density estimation usually does not have parameters.
  Answer: F. "Non-parametric" does not mean there are no parameters. "Non-parametric" means it can not be described by a fixed number of parameters.

- The Epanechnikov kernel is the optimal kernel function for all data.
  Answer: F. The Epanechnikov kernel is optimal in a mean square error sense. So it depends on the data.

- Histogram is an efficient way to estimate density for high-dimensional data.
  Answer: F. According to the computation consideration, if total number of bins n^d is larger than number of sample m, then most of the bins are empty. That would be inefficient.

- Parametric density estimation assumes the shape of probability density.
  Answer: T. Parametric density estimation assumes the shape of probability by choosing reasonable parameters.

3 Information Theory

In the lecture you became familiar with the concept of entropy for one random variable and mutual information. For a pair of discrete random variables X and Y with the joint distribution $p(x, y)$, the *joint entropy* $H(X, Y)$ is defined as

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) \tag{3}$$

which can also be expressed as

$$H(X,Y) = -\mathbb{E}[\log p(X,Y)] \tag{4}$$

Let X and Y take on values $x_1, x_2, ..., x_r$ and $y_1, y_2, ..., y_s$ respectively. Let Z also be a discrete random variable and $Z = X + Y$.

(a) Prove that $H(X, Y) \leq H(X) + H(Y)$ [4 pts]

(a)Answer:

$$\begin{aligned}
H(X,Y) &= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log[p(x)p(y|x)] \\
&= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \\
&= -\sum_{x \in X} \log p(x) \sum_{y \in Y} p(x,y) - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \\
&= -\sum_{x \in X} \log p(x) p(x) - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

By definition of mutual information:

$$\begin{aligned}
I(X,Y) &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(y|x)}{p(y)} \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) - \sum_{y \in Y} p(y) \log p(y) \\
&= -H(Y|X) + H(Y)
\end{aligned}$$

Then combine the two equations:

$H(X,Y) = H(X) + H(Y|X) = H(X) + H(Y) - I(X,Y) \Rightarrow$ **To verify:** $I(X,Y) = H(X) + H(Y) - H(X,Y) \geq 0$

Relative entropy:

$$D(p(x)||q(x)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Jessen's inequality:

$$E[f(x)] \geq f(E[x])$$

Since

$$\begin{aligned}I(X,Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\\ &= D(p(x,y)||p(x)p(y))\end{aligned}$$

Proof of non-negativity of relative entropy: Let $p(x)$ and $q(x)$ be two arbitrary distribution, then:

$$\begin{aligned}D(p(x)||q(x)) = \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} &= -\sum_{x\in X} p(x)\log\frac{q(x)}{p(x)} = -E[\log\frac{q(x)}{p(x)}]\\ \geq -\log(E[\frac{q(x)}{p(x)}]) &= -\log(\sum_{x\in X} p(x)\frac{q(x)}{p(x)}) = -\log(\sum_{x\in X} q(x)) = 0\end{aligned}$$

Therefore, $I(X,Y) = D(p(x,y)||p(x)p(y)) \geq 0 \Rightarrow H(X)+H(Y) \geq H(X,Y)$
(b) Show that $I(X;Y) = H(X)+H(Y)-H(X,Y)$. [2 pts]
(b)Answer: According to the solution of (a), we can obtain:

$$\begin{aligned}H(X,Y) &= H(X)+H(Y|X)\\ I(X,Y) &= H(Y)-H(Y|X)\\ &\Rightarrow I(X,Y) = H(Y)+(H(X)-H(X,Y)) = H(X)+H(Y)-H(X,Y)\end{aligned}$$

(c) Under what conditions does $H(Z) = H(X)+H(Y)$. [4 pts]
(c)Answer:
Since $Z = X+Y$, then $p(Z=z|X=x) = p(Y=z-x|X=x)$

$$\begin{aligned}H(Z|X) &= -\sum_{x\in X} p(X=x)\sum_{z\in Z} p(Z=z|X=x)\log p(Z=z|X=x)\\ &= -\sum_{x\in X} p(X=x)\sum_{z\in Z} p(Y=z-x|X=x)\log p(Y=z-x|X=x)\\ &= H(Y|X)\end{aligned}$$

By symmetry we can obtain $H(Z|Y) = H(X|Y)$. Since Z is a combination of X and Y, such that:

$$H(Z) = H(Z|X)+H(Z|Y) = H(Y|X)+H(X|Y)$$

Therefore, if X and Y is independent, then, $H(Y|X) = H(Y)$ and $H(X|Y) = H(X)$

$$H(Z) = H(Y|X)+H(X|Y) = H(X)+H(Y)$$

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.

Answer:

- For $\mu_j c$, I initialized it by MATLAB's built-in function `rand()` and normalize it.
- For π_c, I evenly initialized it considering the constraint. So for 4 cluster the $\pi_c = [0.25, 0.25, 0.25, 0.25]$.
- The stopping criteria for my implementation depends on number of iteration.(iteration = 200)
- I run the algorithm for run = 20 times, comparing to the provided true clusters, the mean accuracy is 79.7125%, the max accuracy is 88.25% and the min accuracy is 66.5%. For different initial assignment, the results are sensitive and different.

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.