CS 7641 CSE/ISYE 6740 Homework 2

Parminder Bhatia

Deadline: 10/28 Tue, 1:30pm (before starting the class)

- 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]**

Now, by substituting $p(z)$ and $p(x|z)$ in $p(x)$, we get:

$$\begin{aligned} p(x) &= \sum_{z \in Z} p(z)p(x|z) \\ &= \sum_{z \in Z} \prod_{k=1}^{K} {\pi_k}^{z_k} \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \\ &= \sum_{z \in Z} \prod_{k=1}^{K} (\pi_k \mathcal{N}(x|\mu_k, \Sigma_k))^{z_k} \end{aligned}$$

For each $z \in Z$, only one of its entry is equal to 1, and rest of elements in the vector are equal to 0. Thus, for all the vectors , there exist a particular $k \in 1 \dots K$ such that $z_k = 1$ and is 0 for the other values of k. So for a particular $p(z)$ or$p(x/z)$, only contribution would come from that element in the vector which is 1, for all the others, number raise to power zero would fetch 1.

Therefore, from this explanation, we get:

$$\begin{aligned} p(x) &= \sum_{z \in Z} (\pi_1 \mathcal{N}(x|\mu_1, \Sigma_1))^{z_1} \dots (\pi_K \mathcal{N}(x|\mu_K, \Sigma_K))^{z_K} \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \end{aligned}$$

Thus, (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]

Applying the Bayes Rule, we get:

$$p(z_k^n|x_n) = \frac{p(x_n|z_k^n)p(z_k^n)}{p(x_n)}$$

Note that z_k^n is a point and not a vector , so can't be used with above formulas directly. In the 1-of-K coding representation, z_k^n will thus be represented by a vector where the k^{th} element in the vector will be equal to 1 and the rest will be zero, i.e., [0,0,. . . ,1,. . . ,0,0] which is same as $z^{(k)}$

Therefore, $p(x_n|z_k^n)$, $p(z_k^n)$ and $p(x_n)$ are given by:

$$
\begin{aligned}
p(x_n|z_k^n) &= \prod_{k'=1}^{K} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{(z_k^n)_{k'}} \\
p(z_k^n) &= \prod_{k'=1}^{K} \pi_k{}^{(z_k^n)_{k'}}
\end{aligned}
$$

Note that $(z_k^n)_{k'}$ is basically same as $z^{(k)}$ After substituting these equations into the equation for $p(z_k^n|x_n)$

$$
\begin{aligned}
p(z_k^n|x_n) &= \frac{p(x_n|z_k^n)p(z_k^n)}{p(x_n)} \\
&= \frac{\prod_{k'=1}^{K} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{(z_k^n)_{k'}} \prod_{k'=1}^{K} \pi_k{}^{(z_k^n)_{k'}}}{\sum_{k'=1}^{K} \pi_{k'}\mathcal{N}(x_n|\mu_{k'}, \Sigma_{k'})} \\
&= \frac{\prod_{k'=1}^{K} \left(\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)^{(z_k^n)_{k'}}}{\sum_{k'=1}^{K} \pi_{k'}\mathcal{N}(x_n|\mu_{k'}, \Sigma_{k'})}
\end{aligned}
$$

Note that $(z_k^n)_{k'}$ is basically same as $z^{(k)}$ Thus, in 1-of-K coding representation, only one of its entry is equal to 1, and rest of them are equal to 0. Thus, there exist a particular $k' \in 1 \ldots K$ such that $(z_k^n)_{k'} = 1$ and is 0 for the other values of k.

Therefore, we get:

$$
\begin{aligned}
p(z_k^n|x_n) &= \frac{\prod_{k'=1}^{K} \left(\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)^{z^{(k')}}}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \\
&= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}
\end{aligned}
$$

(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:

$$
\begin{aligned}
p(z_k^i = k|x^i) &= \frac{p(z^i = k, x^i)}{\sum_{k=1..K} p(z^i = k, x^i)} \\
&= \frac{\pi_k \mathcal{N}(x^i|\mu_k, \Sigma_k)}{\sum_{k=1..K'} \pi_k \mathcal{N}(x^i|\mu_k, \Sigma_k)}
\end{aligned}
$$

$$
\begin{aligned}
f(\theta) &= E_{q(z^1,z^2,..,z^m)}[\log \prod_{i=1}^{m} p(x^i, z^i|\theta)] \\
&= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}[\log p(x^i, z^i|\theta)] \\
&= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}[\log \pi_{z^i} \mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i})]
\end{aligned}
$$

Expanding the log of Gaussian, we get:

$$log\mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i})$$

$$\begin{aligned}
f(\theta) &= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}[log\pi_{z^i} - (x^i - \mu_{z^i})^T \Sigma_{z^i}(x^i - \mu_{z^i}) + log\Sigma_k + c] \\
&= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}[log\pi_{z^i} - (x^i - \mu_{z^i})^T \Sigma_{z^i}(x^i - \mu_{z^i}) + log\Sigma_{z^i} + c] \\
&= \sum_{i=1}^{m} E_{p(z^i=k|x^i)}[log\pi_k - (x^i - \mu_k)^T \Sigma_k(x^i - \mu_k) + log\Sigma_k + c] \\
f(\theta) &= \sum_{i=1}^{m} \sum_{k=1}^{K} E_{p(z^i=k|x^i)}[\log \pi_k - (x^i - \mu_k)^T \Sigma_k(x^i - \mu_k) + \log \Sigma_k + c]
\end{aligned}$$

Hence, to find π_k, we apply the Lagrange multiplier, and use the information that $\sum_{i=1}^{K} \pi_k = 1$
Forming the Lagrangian, we obtain:

$$L = \sum_{i=1}^{m} \sum_{k=1}^{K} E_{p(z^i=k|x^i)}[\log \pi_k + terms] + \lambda(1 - \sum_{i=1}^{K} \pi_k)$$

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} \tau_k^i[\log \pi_k + terms] + \lambda(1 - \sum_{i=1}^{K} \pi_k) \tag{3}$$

Taking partial derivative and equating to 0, we get:

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{m} \frac{p(z^i = k|x^i)}{\pi_k} - \lambda = 0$$

$$\pi_k = \frac{1}{\lambda} \sum_{i=1}^{m} p(z^i = k|x^i)$$

$$\lambda = m$$

For the other two parts, we need to expand the middle term
Expansion of squared Mahalanobis distance.

$$\begin{aligned}
&(x - \mu_k)^t \Sigma_k^{-1}(x - \mu_k) \\
&\qquad = x^t \Sigma_k^{-1} x - x^t \Sigma^{-1} u_k - u_k^t \Sigma^{-1} x + u_k^t \Sigma^{-1} u_k \\
&\qquad = x^t \Sigma^{-1} x - 2(\Sigma^{-1} u_k)^t x + u_k^t \Sigma^{-1} u_k
\end{aligned}$$

here the term $x^t \Sigma^{-1} x$ is common in all and thus gets removed, thus we get the following equation (note we ignore λ as it was just dependent on π_k)

$$f(\theta) = \sum_{i=1}^{m} \sum_{k=1}^{K} \tau_k^i[\log \pi_k - x^t \Sigma_k^{-1} x - 2(\Sigma_k^{-1} u_k)^t x + u_k^t \Sigma_k^{-1} u_k + \log \Sigma_k + c] \tag{4}$$

Now taking derivative wrt μ_k , we get

$$\begin{aligned}\frac{\partial L}{\partial \mu_k} =& \sum_{i=1}^{m} \tau_k^i(-2\Sigma_k^{-1}x + 2\Sigma_k^{-1}\mu_k) = 0\\ \sum_{i=1}^{m} \tau_k^i 2x =& \sum_{i=1}^{m} \tau_k^i(2\mu_k)\end{aligned}$$

Here we removed Σ_k^{-1} as it was independent of the number of points. Similarly, we in the second summation , we can right μ_k outside the summation

$$\begin{aligned}\frac{\partial L}{\partial \mu_k} =& \sum_{i=1}^{m} \tau_k^i(-2\Sigma_k^{-1}x + 2\Sigma_k^{-1}\mu_k) = 0\\ \sum_{i=1}^{m} \tau_k^i x =& \mu_k \sum_{i=1}^{m} \tau_k^i\\ \mu_k =& \frac{\sum_{i=1}^{m} \tau_k^i x}{\sum_{i=1}^{m} \tau_k^i}\end{aligned}$$

Similarly as in the last equation, we may right the last equation as

$$f(\theta) = \sum_{i=1}^{m}\sum_{k=1}^{K} \tau_k^i[\log \pi_k - x^t\Sigma_k^{-1}x - 2(\Sigma_k^{-1}u_k)^t x + u_k^t\Sigma_k^{-1}u_k - \log \Sigma_k^{-1} + c] \tag{5}$$

Taking derivative wrt to Σ_k^{-1} , we get

$$\begin{aligned}\frac{\partial L}{\partial \Sigma_k^{-1}} =& \sum_{i=1}^{m} \tau_k^i(x^t x - 2(\mu_k)^t x + \mu_k^t\mu_k - \frac{1}{\Sigma_k^{-1}}) = 0\\ \sum_{i=1}^{m} \tau_k^i((x-\mu_k)^t(x-\mu_k)) =& \Sigma_k \sum_{i=1}^{m} \tau_k^i\\ \Sigma_k =& \frac{\sum_{i=1}^{m} \tau_k^i((x-\mu_k)^t(x-\mu_k))}{\sum_{i=1}^{m} \tau_k^i}\end{aligned}$$

(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.

In E step we calculate $\gamma(k)$, which is basically p(k/x). From the above expressions we got :

$$p(x) = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}.$$

$$p(k/x) = \frac{\pi_k \exp\left(-\frac{(x-\mu_k)^t\Sigma^{-1}(x-\mu_k)}{2}\right)}{\sum_{j=1}^{K} \pi_j \exp\left(-\frac{(x-\mu_j)^t\Sigma^{-1}(x-\mu_j)}{2}\right)}.$$

Inverse of Covariance Matrix

$$\Sigma_k^{-1} = (1/\epsilon) * I$$

, So its only effect is to scale vector product by $1/\epsilon$ Thus,

$$\begin{aligned} p(k/x) =& \frac{\pi_k \exp.\left(-\frac{(x-\mu_k)^t(x-\mu_k)}{2\epsilon}\right)}{\sum_{j=1}^K \pi_j \exp\left(-\frac{(x-\mu_j)^t(x-\mu_j)}{2\epsilon}\right)} \\ =& \frac{\pi_k \exp\left(-\frac{||(x-\mu_k)||^2}{2\epsilon}\right)}{\sum_{j=1}^K \pi_j \exp\left(-\frac{||(x-\mu_j)||^2}{2\epsilon}\right)}. \end{aligned}$$

Now , in order to get the limiting value when $\epsilon = 0$, let us divide numerator and denominator by $\pi_k \exp\left(-\frac{(x-\mu_k)^t(x-\mu_k)}{2\epsilon}\right)$
Thus we get the equation of the form :

$$\begin{aligned} p(k/x) =& \frac{1}{\frac{\sum_{j=1}^K \pi_j \exp\left(-\frac{||(x-\mu_j)||^2}{2\epsilon}\right)}{\pi_k \exp.\left(-\frac{||(x-\mu_k)||^2}{2\epsilon}\right)}} \\ =& \frac{1}{1+\sum_{j=1,j\neq k}^K (\pi_j/\pi_k)\exp\left(-\frac{||(x-\mu_j)||^2 - ||(x-\mu_k)||^2}{2\epsilon}\right)} \end{aligned}$$

Now for the above equation, for the k with the maximum posterior probability $||(x-\mu_k)||^2$ is will smallest as it will be closest to the point x , thus when at limit $\epsilon = 0$, then all the vaues in summation will be zero as $||(x-\mu_j)||^2 - ||(x-\mu_k)||^2$ will be greater than zero for all j's, thus negative of exponential of that number would tend to 0.
However , for all the other values of k , there will be at least one value in summation such that $||(x-\mu_j)||^2 - ||(x-\mu_k)||^2$ will be less than zero, thus its negotiation will be less than zero and so at limit would blow to infinity(denominator). Thus p(k/x) for all the others will be zero. This is exactly what happens in K means as $\gamma(k)$ would be 1 for only one cluster and zero for the rest. Thus it exactly corresponds to the equation given above.

(e) General setting [10 pts]

Consider a mixture of distribution of the form

$$P(x) = \sum_{k=1}^K \pi_k p(x|k)$$

where the elements of x could be discrete or continuous or a combination of these. Express the mean and covariance of the mixture distribution using the mean μ_k and covariance Σ_k of each component distribution $p(x|k)$.

Now, the expectation of the mixture distribution can be given by:

$$
\begin{aligned}
E[x] &= \sum_{x} xP(x) \\
&= \sum_{x} x \sum_{k=1}^{K} \pi_k p(x|k) \\
&= \sum_{k=1}^{K} \pi_k \sum_{x} xp(x|k) \\
&= \sum_{k=1}^{K} \pi_k E[x|k] \\
E[x] &= \sum_{k=1}^{K} \pi_k \mu_k
\end{aligned}
$$

Note that above, for Expection of x given k is basically the mean for the distribution k cluster. For finding the covariance, the following general relation can be used:

$$\Sigma = E[xx^T] - E[x]E[x]^T$$

Therefore, we get:

$$
\begin{aligned}
\Sigma &= E[xx^T] - E[x]E[x]^T \\
\Sigma &= \sum_{k=1}^{K} \pi_k E_k[xx^T] - E[x]E[x]^T
\end{aligned}
$$

Therefore, we get:

$$
\begin{aligned}
\Sigma &= \sum_{k=1}^{K} \pi_k E_k[xx^T] - E[x]E[x]^T \\
\Sigma &= \sum_{k=1}^{K} \pi_k (\Sigma_k + \mu_k \mu_k^T) - E[x]E[x]^T
\end{aligned}
$$

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]

Probability is basically mass which can also be represented by density and volume. Now, over a region i where volume of the region is given by Δ_i:

$$P(x) = h_i \Delta_i \qquad \textit{over } i^{th} \textit{ region}$$

And we are given a set of N observations of x such that n_i of these observations fall in regions i. Therefore, the likelihood is given by:

$$P(x|data) = \prod_i (h_i \Delta_i)^{n_i}$$

Now, the log-likelihood is given by:

$$\log P(x|data) = \log \prod_i (h_i \Delta_i)^{n_i} = \sum_i n_i \log(h_i \Delta_i)$$

(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 $\lambda \left(\sum_i h_i \Delta_i - 1\right)$ to your objective function.

As this is a constrained optimization problem, we need to use Lagrange multiplier. So, the following equation is:

$$\log P(x|data) = \sum_i n_i \log(h_i \Delta_i) + \lambda \left(\sum_i h_i \Delta_i - 1 \right)$$

Now differentiating the log-likelihood with respect to h_i, we get:

$$\begin{aligned}
\frac{\partial \log P(x|data)}{\partial h_i} &= \frac{\partial}{\partial h_i} \left\{ \sum_i n_i \log(h_i \Delta_i) + \lambda \left(\sum_i h_i \Delta_i - 1 \right) \right\} \\
&= n_i \frac{1}{h_i \Delta_i} \Delta_i + \lambda \Delta_i \\
&= \frac{n_i}{h_i} + \lambda \Delta_i
\end{aligned}$$

Equating this to 0, we get:

$$h_i = -\frac{n_i}{\lambda \Delta_i}$$

Substituting this value back in the constraint $\sum_i h_i \Delta_i = 1$, we get:

$$\begin{aligned}
\sum_i \left(-\frac{n_i}{\lambda \Delta_i} \right) \Delta_i = 1 \\
\lambda = -\sum_i n_i \\
\lambda = -N
\end{aligned} \tag{6}$$

Therefore, we get h_i:

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

(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.

  FALSE. Because Non-parametric density estimation has parameters dependent on the number of data points. The difference between parametric model and non-parametric model is that the former has a fixed number of parameters, while the latter grows the number of parameters with the amount of training data. As an Example, Histograms make no assumptions about the probability distributions of the variables being assessed.

- The Epanechnikov kernel is the optimal kernel function for all data.

  TRUE. Epanechnikov kernel is the optimal kernel since it minimizes the AMISE (Asymptotic Mean Integrated Squared Error) and hence is the optimal kernel function for all data. Kernel efficiency is measured in comparison to the Epanechnikov Kernel. This Kernel function minimizes the risk as well, hence it can be considered the optimal kernel.

- Histogram is an efficient way to estimate density for high-dimensional data.

  FALSE. Histograms are not an efficient since in case there is high-dimensional data, there are too many bins and it gets in the problem of sparse bins where weight of bins is sufficient in only a few number of bins. For instance, if bin size is $\frac{1}{n}$ and there are m i.i.d samples of data and the data is split into bins $= n^d$ if $n^d > m$ then most bins are empty, hence it is not a good utilization of memory or space as well.
  Another reason is that statistically histograms are not better than other models like KDE. The integrated risk is much more in case of histogram than other models like the KDE. Integrated risk converges at the rate of $O(m^{-\frac{2}{3}})$ in case of histograms where as for the KDE it converges at the rate of $O(m^{-\frac{4}{5}})$. This impact is even more significant when the data is in high dimensions.

- Parametric density estimation assumes the shape of probability density.

  TRUE. In case of the parametric approach, we assume that the density function takes on a particular form, and where its corresponding parameters are estimated by data. When the parametric model consists of absolutely continuous distributions, it is often specified in terms of corresponding probability density functions.

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

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) \tag{7}$$

which can also be expressed as

$$H(X,Y) = -\mathbb{E}[\log p(X,Y)] \tag{8}$$

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]

We can write the joint probability of x and y as:

$$p(x,y) = p(y|x)p(x)$$

Substituting this in $H(x,y)$, we get:

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

Then, we get:

$$\begin{aligned} H(Y) - H(Y|X) &= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y) + \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \\ &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(y|x)}{p(y)} \\ &\geq 0 \end{aligned}$$

Therefore, we get:

$$\begin{aligned} H(Y) - H(Y|X) &\geq 0 \\ H(Y|X) &\leq H(Y) \end{aligned}$$

Thus, using this inequality, we get that:

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

(b) If X and Y are independent, i.e. $P(X,Y) = P(X)P(Y)$, then $H(X,Y) = H(X) + H(Y)$ [4 pts]

Answer:

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

Since X and Y are independent:

$$P(X,Y) = P(X)P(Y)$$

, Thus we get

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x)p(y)\log(p(x)p(y)) \tag{9}$$

Using log(p(x)p(y)) as log(p(x)) + log(p(y), we get:

$$\begin{aligned}
&= -\sum_{x\in X}\sum_{y\in Y} p(x)p(y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x)p(y)\log p(y) \\
&= -\sum_{y\in Y} p(y)\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X} p(x)\sum_{y\in Y} p(y)\log p(y)
\end{aligned}$$

Summation of x and y in their space is one.Thus, p(y) and p(x) in the respective terms above clearly sum up to 1.

$$= -1 * \sum_{x\in X} p(x)\log p(x) - 1 * \sum_{y\in Y} p(y)\log p(y)$$

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

Since,

$$-\sum_{x\in X} p(x)\log p(x) = H(X)$$

$$-\sum_{y\in Y} p(y)\log p(y) = H(Y) \tag{10}$$

Thus , we get

$$= H(X) + H(Y) \tag{11}$$

Hence, proved that

$$H(X,Y) = H(X) + H(Y)$$

(c) Show that $I(X;Y) = H(X) + H(Y) - H(X,Y)$. [4 pts]

Answer: Mutual information of two discrete random variables X and Y is defined as :

$$\begin{aligned}
I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log(p(x,y)) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log(p(x)) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log(p(y)) \\
&= -H(X,Y) - \sum_{x\in X}\log(p(x))\sum_{y\in Y} p(x,y) - \sum_{y\in Y}\log(p(y))\sum_{x\in X} p(x,y)
\end{aligned} \tag{12}$$

In the above equation we took $\log(p(y))$ and $\log(p(x))$ out of summation as their are constant wrt to other summation.
Note that second terms are marginalization equations x and y respectively,

$$\sum_{y\in Y} p(x,y) = p(x)$$

, Thus we can say that:

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

and

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

Hence, the earlier equation can be written as:

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

(d) Show that $H(Z|X)=H(Y|X)$. Argue that when X,Y are independent, then $H(X)\leq H(Z)$ and $H(Y)\leq H(Z)$. Therefore, the addition of *independent* random variables add uncertainty. [4 pts]

Answer: Condition entropy over Random variables can be expressed as summation of individual conditional entropies in space. Thus,

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

Hence, proved that:

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

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

Since X and Y are independent, we can

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

First summation across the domain would be 1. Hence, we get:

$$=-\sum_{y\in Y}p(y)\log p(y)=H(Y)$$

Hence, we have proved that:

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

Since Z is a function of X and Y , so knowing X or conditioning would reduce entropy, thus we get:

$$H(Z) \geq H(Z|X) = H(Y|X) = H(Y)$$

$$H(Z) \geq H(Z)$$

Rewriting, we get:

$$Hence, H(Y) \leq H(Z)$$

Similarly, we can also prove that

$$H(X) \leq H(Z)$$

Hence, proved.

(e) Under what conditions does $H(Z) = H(X) + H(Y)$. [4 pts]

Answer:
H(Z) can be represented as H(X+Y). We have already proved that If X and Y are independent, i.e. $P(X,Y) = P(X)P(Y)$, then $H(X,Y) = H(X) + H(Y)$
Also, we can see here that $H(Z) <= H(X,Y)$, because $H(Z)$ is a function for X,Y and for a function there may be multiple ways to get a particular value. Lets say Z=5, then we can get 5 from $(1,4)$ or (2,3). If we want the equality to hold, that is

$$H(X+Y) = H(X,Y) \tag{13}$$

if the addition is invertible, that is, if there are unique X and Y for any $X+Y$; this would happen, for example, if $\mathcal{X} = \{1,2\}$ and $\mathcal{Y} = \{1,3\}$ since the possible values of the sum are 2, 3, 4 and 5 and each corresponds to a different choice of X and Y; however, if $\mathcal{X} = \{1,2\}$ and $\mathcal{Y} = \{1,2\}$ then $X = 1$, $Y = 2$ and $X = 2$, $Y = 1$ both give $X + Y = 3$. Now

$$H(X,Y) = H(X) + H(Y|X) \tag{14}$$

Thus, the conditions for independence are that function for $X + Y$ should be bijective and X and Y should be independent for the condition to hold true.

4 Bayes Classifier

4.1 Bayes Classifier With General Loss Function

In class, we talked about the popular 0-1 loss function in which $L(a,b) = 1$ for $a \neq b$ and 0 otherwise, which means all wrong predictions cause equal loss. Yet, in many other cases including cancer detection, the asymmetric loss is often preferred (misdiagnosing cancer as no-cancer is much worse). In this problem, we assume to have such an asymmetric loss function where $L(a,a) = L(b,b) = 0$ and $L(a,b) = p, L(b,a) = q, p \neq q$. Write down the the Bayes classifier $f : X \to Y$ for binary classification $Y \in \{-1,+1\}$. Simplify the classification rule as much as you can. [20 pts]

The optimal solution is the one which minimizes the loss function. However, the loss function depends on the true class, which is unknown. For a given input vector x, our uncertainty in the true class is expressed through the joint probability distribution p(x, Ck) and so we seek instead to minimize the average loss, where the average is computed with respect to this distribution, which is given by

$$E[L] = \sum_k \sum_j \int_{R_j} L_{kj} p(x, C_k)$$

Each x can be assigned independently to one of the decision regions R_j . Our goal is to choose the regions R_j in order to minimize the expected loss , which implies that for each x we should minimize

$$\sum_k L_{kj} p(x, C_k).$$

Thus , we can do the following

$$Min \sum_k L_{kj} p(x, C_k)$$
$$=> Min \sum_k L_{kj} p(C_k|x) * p(x) => Min \sum_k L_{kj} p(C_k|x)$$

Note – We can ignore p(x) as summation is across k thus it is a same constant in all terms
Let us denote this term by risk so for a particular class k

$$R(k_1) = \sum_k L_{kk_1} p(C_k|x)$$

Thus Bayesian Rule for the Binary class will be

$$R(k_1) < R(k_2)$$

where

$$R(k_1) = L_{k_1k_1} p(C_{k_1}|x) + L_{k_2k_1} p(C_{k_2}|x)$$
$$R(k_2) = L_{k_2k_2} p(C_{k_2}|x) + L_{k_1k_2} p(C_{k_1}|x)$$

We are given that L(a,a) and L(b,b) are zero. Thus, our decision rule becomes

$$Min(L(b,a)p(a|x), L(a,b)p(b|x))$$
$$Min(p * p(a|x)), q * p(b|x)$$

We can convert this into class condition using

$$p(a|x) = p(x|a)p(a)/p(x)$$

Here again we may ignore p(x).Thus our equation becomes and this is out decision boundary for the given case.

$$Min((p * p(x|a) * p(a)), (q * p(x|b) * p(b)))$$

4.2 Gaussian Class Conditional distribution

(a) Suppose the class conditional distribution is a Gaussian. Based on the general loss function in problem 4.1, write the Bayes classifier as $f(X) = \text{sign}(h(X))$ and simplify h as much as possible. What is the geometric shape of the decision boundary? [10 pts]

(b) Repeat (a) but assume the two Gaussians have identical covariance matrices. What is the geometric shape of the decision boundary? [10 pts]

(c) Repeat (a) but assume now that the two Gaussians have covariance matrix which is equal to the identity matrix. What is the geometric shape of the decision boundary? [10 pts]

Answer For all the 3 parts , we will first define a log for the earlier decision boundary for better understanding of the problem

$$g_k(x) = \ln(loss_k) + \ln(p(x|k) + \ln(p(a))$$

,

Here we have Normal Distribution for a class

$$p(x) = \frac{1}{(2\pi)^{d/2}\Sigma^{1/2}} \exp\left(-\frac{(x-\mu)^t\Sigma^{-1}(x-\mu)}{2}\right).$$

Thus ,

$$g_k(x) = \ln(loss_k) + \frac{(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)}{2} - d/2\ln(2\pi) - 1/2\ln\Sigma_k + \ln(p(k))$$

This is our basic equation which will be used across all the three parts .

Case a : Arbitary Now for the first case, we may ignore $d/2\ln(2\pi)$ as it is same across all the classes.

Expansion of squared Mahalanobis distance.

$$\begin{aligned}
&(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k) \\
&\qquad = x^t\Sigma_k^{-1}x - x^t\Sigma_k^{-1}u_k - u_k^t\Sigma_k^{-1}x + u_k^t\Sigma_k^{-1}u_k \\
&\qquad = x^t\Sigma_k^{-1}x - 2(\Sigma_k^{-1}u_k)^t x + u_k^t\Sigma_k^{-1}u_k
\end{aligned}$$

the last step comes from symmetry of the covariance matrix and thus its inverse:

Thus putting this value in equation we get a quadratic equation:

$$x^t(\Sigma_k^{-1}/2)x - ((\Sigma_k^{-1}u_k)^t)x + u_k^t\Sigma_k^{-1}u_k + \ln(loss_k) - 1/2\ln\Sigma_k + \ln(p(k))$$

The above is a quadratic equation in terms of the variable x and Decision Boundary which basically corresponds to

$$g_1(x) = g_2(x)$$

can thus are hyperquadratics, and they can assume any of the general forms: hyperquadrics: can be hyperplanes,

1. hyperplane pairs,
2. hyperspheres,
3. hyperellipsoids,
4. hyperparabaloids,
5. hyperhyperparabaloids .

Case b : Identical Covariance

$$g_k(x) = \ln(loss_k) + \frac{(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)}{2} - d/2\ln(2\pi) - 1/2\ln\Sigma_k + \ln(p(k))$$

For identical covariance $\Sigma_k = \Sigma$ for all k's.
Thus we can remove terms $-d/2\ln(2\pi) - 1/2\ln\Sigma_k$, so our equation now is :

$$g_k(x) = \ln(loss_k) + \frac{(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)}{2} + \ln(p(k))$$

Expansion Expansion of squared Mahalanobis distance.

$$\begin{aligned}&(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)\\ &\qquad\qquad = x^t\Sigma_k^{-1}x - x^t\Sigma^{-1}u_k - u_k^t\Sigma^{-1}x + u_k^t\Sigma^{-1}u_k\\ &\qquad\qquad = x^t\Sigma^{-1}x - 2(\Sigma^{-1}u_k)^tx + u_k^t\Sigma^{-1}u_k\end{aligned}$$

here the term

$$x^t\Sigma^{-1}x$$

is common in all and thus gets removed, thus we get the following equation

$$g_i(x) = ((\Sigma^{-1}u_k)^t)x + u_k^t\Sigma^{-1}u_k + \ln(loss_k) - 1/2\ln\Sigma + \ln(p(k))$$

Decision Boundary which basically corresponds to

$$g_1(x) - g_2(x) = 0$$

, So we get the following equation:

$$((\Sigma^{-1}u_1 - \Sigma^{-1}u_0)^t)x = \frac{-(u_1\Sigma^{-1}u_1 - u_0\Sigma^{-1}u_0)}{2} + \ln\frac{p*p(a)}{q*p(b)}$$

By Multiplying

$$\frac{(\mu_1-\mu_0)^t\Sigma^{-1}(\mu_1-\mu_0)}{(\mu_1-\mu_0)^t\Sigma^{-1}(\mu_1-\mu_0)}$$

to $\frac{p*p(a)}{q*p(b)}$ and manipulating the common variables in the equation, we get the following equation

$$((\Sigma^{-1}u_1 - \Sigma^{-1}u_0)^t)x = (\Sigma^{-1}u_1 - \Sigma^{-1}u_0)^t)(\frac{\mu_1+\mu_0}{2} + (\mu_1-\mu_0) * \frac{\ln\frac{p*p(a)}{q*p(b)}}{(\mu_1-\mu_0)^t\Sigma^{-1}(\mu_1-\mu_0)})$$

Thus can be written as norm form of equation

$$w^t(x - x_0) = 0$$

where x_0 is

$$(\frac{\mu_1+\mu_0}{2} + (\mu_1-\mu_0) * \frac{\ln\frac{p*p(a)}{q*p(b)}}{(\mu_1-\mu_0)^t\Sigma^{-1}(\mu_1-\mu_0)})$$

Because the equations are linear, the resulting decision boundaries are hyperplanes.
Thus , we will have a straight line as the boundary between the two. A line through the point x_0 defines this decision boundary between 1 and 0. If the product of loss(i,j) prior probabilities (example p*p(a)) are equal

then x0 is halfway between the means. If the products are not equal, the optimal boundary hyperplane is shifted away from the more likely mean .The decision boundary is in the direction orthogonal defined by

$$(\Sigma^{-1}u_1 - \Sigma^{-1}u_0)^t$$

Thus, the boundary line will be tilted depending on how the 2 features covary and their respective variances.

Case c : Identidy Covariance $\sigma^2 I$

$$g_k(x) = \ln(loss_k) + \frac{(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)}{2} - d/2\ln(2\pi) - 1/2\ln\Sigma_k + \ln(p(k))$$

We can remove terms $-d/2\ln(2\pi) - 1/2\ln\Sigma_k$, so our equation now is :

$$g_k(x) = \ln(loss_k) + \frac{(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)}{2} + \ln(p(k))$$

Inverse of Covariance Matrix

$$\Sigma_k^{-1} = 1/\sigma^2 * I$$

, So its only effect is to scale vector product by $1/\sigma^2$ Thus,

$$\begin{aligned} g_k(x) &= \ln(loss_k) + \frac{(x-\mu_k)^t(x-\mu_k)}{2\sigma^2} + \ln(p(k)) \\ &= \frac{x^tx - 2\mu_k^t x + \mu_k^t\mu_k}{2\sigma^2} + \ln(loss_k) + \ln(p(k)) \end{aligned} \tag{15}$$

Here also we may ignore x^tx

$$g_k(x) = \frac{\mu_k^t x}{2\sigma^2} + \frac{\mu_k^t\mu_k}{2\sigma^2} + \ln(loss_k) + \ln(p(k)) \tag{16}$$

Which is again a linear equation in x, So taking the decision boundary and writing in norm of a line form

$$w^t(x - x_0) = 0$$

, we get

$$g_k(x) = \frac{\mu_k^t x}{2\sigma^2} + \frac{\mu_k^t\mu_k}{2\sigma^2} + \ln(loss_k) + \ln(p(k)) \tag{17}$$

$$((u_1 - u_0)^t)x = (u_1 - u_0)^t)(\frac{\mu_1 + \mu_0}{2} + (\mu_1 - \mu_0) * \frac{\sigma^2 \ln\frac{p*p(a)}{q*p(b)}}{(\mu_1 - \mu_0)^t(\mu_1 - \mu_0)})$$

Geometrically, this defines a hyperplane through the point x_0 that is orthogonal to the vector w. But since w= $\mu_1 - \mu_0$ then the hyperplane which seperates the regions is orthogonal to the line that links their means. If $p * p(a) = q * p(b)$, the second term on the right vanishes, and thus the point x_0 is halfway between the means (equally divide the distance between the 2 means, with a decision region on either side), and the hyperplane is the perpendicular bisector of the line between the means .

5 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:

$$\gamma = \frac{\pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}}{\sum_{c=1}^{n_d} \pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}}$$

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

Here, I am running the instance for 100 iterations(should be reduced larger datasets).Files used are homework2.m(running file) and mycluster.m(code for EM). I initialize the variables using repmat on random normal gaussian distribution across column/row , so that they are greater than 0 and sum to 1 in a row/column depending upon the requirement. I am getting from 80 to 90 percent accuracy for the above toy data set, which fluctuates depending upon the initializations for the prior probabilities. Code is vectored, however commented code is based on loops for the understanding of the code.

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

$$\mathcal{L} = \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} T_{dw} \log P(d, w), \tag{18}$$

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

$$P(d, w) = \sum_{z \in \mathcal{Z}} P(z) P(w|z) P(d|z), \tag{19}$$

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

$$P(z|d,w) = \frac{P(z)P(w|z)P(d|z)}{\sum_{z'} P(z')P(w|z')P(d|z')}. \tag{20}$$

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

$$P(w|z) = \frac{\sum_d T_{dw} P(z|d,w)}{\sum_{w'} \sum_d T_{dw'} P(z|d,w')} \tag{21}$$

$$P(d|z) = \frac{\sum_w T_{dw} P(z|d,w)}{\sum_{d'} \sum_w T_{d'w} P(z|d',w)} \tag{22}$$

$$P(z) = \frac{\sum_d \sum_w T_{dw} P(z|d,w)}{\sum_{z'} \sum_{d'} \sum_{w'} T_{d'w'} P(z'|d',w')}. \tag{23}$$

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.

Answer

Files for running this part are extrahomework2.m (running file) and topicmycluster.m (contains code for generating the topics). Here I have tried getting clusters/topics corresponding to different values of K. Like in the data set provided, too less clusters example 2 gave confusing topics as it kind of merged the topics. I was getting reasonable topics when I was using 4 to 6 clusters . For clusters greater than 10, it seemed that similar topics are getting distributed in different clusters.Code is vectored, however commented code is based on loops for the understanding of the code.Currently, I am running for 25 iterations.