CS 7641 CSE/ISYE 6740 Homework 2

Rakesh Surapaneni

10/16 Fri, 6 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]**

Solution:
(2) states that

$$P(x) = \sum_{z \epsilon Z} P(z)P(x|z)$$

(by substituting values of $P(z)$ and $P(x|z)$)

$$P(x) = \sum_{z \epsilon Z} [(\prod_{k=1}^{K} {\pi_k}^{z_k})(\prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k})]$$

$$= \sum_{z \epsilon Z} (\prod_{k=1}^{K} [{\pi_k}^{z_k} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}])$$

We know that $Z = \{z^{(1)}, z^{(2)}, ..., z^{(K)}\}$. Hence we get

$$P(x) = \sum_{i=1}^{K} (\prod_{k=1}^{K} [{\pi_k}^{z_k^{(i)}} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k^{(i)}}])$$

$z_k^{(i)} = 1$ if and only if k = i. Other wise it is 0. Hence we know that

$$\prod_{k=1}^{K} [{\pi_k}^{z_k^{(i)}} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k^{(i)}}] = \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$$

Therefore combining above two equations we get

$$P(x) = \sum_{i=1}^{K} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i) = \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Therefore

$$P(x) = \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Which is nothing but (1)

Hence (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]

Solution:

We know that from (1)

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

$$P(x_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{3}$$

Also we know that

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

Therfore

$$P(z_k^n = 1) = \pi_k$$

Or simply

$$P(z_k^n) = \pi_k \tag{4}$$

Also given that

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

This implies that

$$P(x_n|z_k^n) = \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{5}$$

Since $z_k = 1$.

By Bayes rule

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

Combining with the equations (3), (4) and (5) we get.

$$\boxed{P(z_k^n|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}}$$

(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.
Solution:

Likelihood function on the Given Domain,

$$P(X|\pi,\mu,\Sigma) = \prod_{n=1}^{N}\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)$$

$$= \prod_{n=1}^{N}\sum_{k=1}^{K}\pi_k\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}exp(\frac{-1}{2}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))$$

Log likelihood function,

$$L(X;\pi,\mu,\Sigma) = log(P(X|\pi,\mu,\Sigma))$$

$$L = \sum_{n=1}^{N}log(\sum_{k=1}^{K}(\pi_k\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}exp(\frac{-1}{2}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))))$$

calculation of π_k:
We know that $\sum_{k=1}^{K}\pi_k = 1$. Hence We can use Lagrange multiplier for π_k to apply constraints on it.

$$L = \sum_{n=1}^{N}log(\sum_{k=1}^{K}(\pi_k\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}exp(\frac{-1}{2}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k)))) + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N}\frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{i=1}^{K}\pi_i\mathcal{N}(x_n|\mu_i,\Sigma_i)} + \lambda$$

$$\frac{\partial L}{\partial \pi_k} = 0 =>$$

$$\sum_{n=1}^{N}\frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{i=1}^{K}\pi_i\mathcal{N}(x_n|\mu_i,\Sigma_i)} + \lambda = 0$$

Let T_k denote the LHS part. Now

$$T_k = 0$$

$$0 = \sum_{k=1}^{K}\pi_k T_k = \sum_{k=1}^{K}\pi_k\sum_{n=1}^{N}\frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{i=1}^{K}\pi_i\mathcal{N}(x_n|\mu_i,\Sigma_i)} + \lambda = \sum_{n=1}^{N}\frac{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{i=1}^{K}\pi_i\mathcal{N}(x_n|\mu_i,\Sigma_i)} + \sum_{k=1}^{K}(\pi_k\lambda)$$

$$= \sum_{n=1}^{N}1 + (\sum_{k=1}^{K}\pi_k)\lambda = N + \lambda$$

$$\lambda = -N$$

From part b, we know that

$$P(z_k^n|x_n) = \frac{\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)}$$

Using this equation in $T_k = 0$ fetches us

$$\sum_{n=1}^{N} \frac{1}{\pi_k} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)} + \lambda = 0$$

$$\sum_{n=1}^{N} \frac{P(z_k^n|x_n)}{\pi_k} - N = 0$$

$$\boxed{\pi_k = \sum_{n=1}^{N} \frac{P(z_k^n|x_n)}{N}}$$

calculation of μ_k:

$$\frac{\partial L}{\partial \mu_k} = 0 =>$$

$$\sum_{n=1}^{N} \left(\frac{\pi_k}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)} \frac{\partial \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\partial \mu_k}\right) = 0$$

$$\sum_{n=1}^{N} \left(\frac{\pi_k}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)} \left(\frac{-1}{2}(x_n - \mu_k)^T(\Sigma_k^{-1} + (\Sigma_k^{-1})^T)\right)\mathcal{N}(x_n|\mu_k, \Sigma_k)\right) = 0$$

$$\sum_{n=1}^{N} P(z_k^n|x_n) \frac{-1}{2}(x_n - \mu_k)^T(\Sigma_k^{-1} + (\Sigma_k^{-1})^T)) = 0$$

Ignoring constants and multiplying by $(\Sigma_k^{-1} + (\Sigma_k^{-1})^T)^{-1}$, we get

$$\sum_{n=1}^{N} P(z_k^n|x_n)(x_n - \mu_k)^T = 0$$

Transpose both sides

$$\sum_{n=1}^{N} P(z_k^n|x_n)(x_n - \mu_k) = 0$$

$$\boxed{\mu_k = \frac{\sum_{n=1}^{N} P(z_k^n|x_n)(x_n)}{\sum_{n=1}^{N} P(z_k^n|x_n)}}$$

calculation of Σ_k:

$$\frac{\partial L}{\partial \Sigma_k} = 0$$

$$\Sigma_{n=1}^{N} \frac{\pi_k}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)} \frac{\partial \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\partial \Sigma_k} = 0$$

From jacobi's formulae, we know that

$$\frac{\partial det(A)}{\partial A_{ij}} = det(A)(A^{-1})_{ij}$$

By matrix calculus

$$\frac{\partial det(\Sigma_k)}{\partial \Sigma_k} = det(\Sigma_k)(\Sigma_k^{-1})$$

$$\frac{\partial (x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k)}{\partial \Sigma_k} = -\Sigma_K^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_K^{-1}$$

$$\frac{\partial \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\partial \Sigma_k} = \frac{1}{\sqrt{(2\pi)^K}}(\frac{1}{\sqrt{|\Sigma_k|}}\frac{-1}{2}(-\Sigma_K^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_K^{-1})(e^{\frac{-1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)})+e^{\frac{-1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)}*\frac{-1}{2}|\Sigma$$

$$= \frac{1}{\sqrt{(2\pi)^K}}(\frac{1}{\sqrt{|\Sigma_k|}}\frac{-1}{2}(-\Sigma_K^{-1}(x_n-\mu_k)(x_n-\mu_k)^T\Sigma_K^{-1})(e^{\frac{-1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)})+e^{\frac{-1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)}*\frac{-1}{2}|\Sigma_k|^{(-1/2)}(\Sigma_k^{-1}))$$

$$= -1/2 * \mathcal{N}(x_n|\mu_k, \Sigma_k)(-\Sigma_K^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_K^{-1} + \Sigma_k^{-1})$$

$$\frac{\partial L}{\partial \Sigma_k} = 0$$

$$= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)}(\Sigma_k^{-1} - \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}) * -1/2$$

Ignoring constant terms such as -1/2 and multiplying by Σ_k twice front and back, we get

$$\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)}(\Sigma_k - (x_n - \mu_k)(x_n - \mu_k)^T) = 0$$

$$\boxed{\Sigma_k = \frac{\sum_{n=1}^{N} P(z_n^k|x_n)(x_n - \mu_n)(x_n - \mu_n)^T}{\sum_{n=1}^{N} P(z_n^k|x_n)}}$$

There fore M-step involves following computations:

$$\boxed{\pi_k = \sum_{n=1}^{N} \frac{P(z_k^n|x_n)}{N}}$$

$$\boxed{\mu_k = \frac{\sum_{n=1}^{N} P(z_k^n|x_n)(x_n)}{\sum_{n=1}^{N} P(z_k^n|x_n)}}$$

$$\boxed{\Sigma_k = \frac{\sum_{n=1}^{N} P(z_n^k|x_n)(x_n - \mu_n)(x_n - \mu_n)^T}{\sum_{n=1}^{N} P(z_n^k|x_n)}}$$

(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.
Solution:

$$\Sigma_k = \epsilon I$$

$$|\Sigma_k| = |\epsilon I| = \epsilon^K$$

$$\mathcal{N}(x_n|\mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi|\Sigma_k|}} e^{(\frac{-1}{2}(x_n-\mu_k)^T \Sigma^{-1}(x_n-\mu_k))}$$

$$P(z_k^n|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n|\mu_i, \Sigma_i)}$$

Since $\Sigma_i = \epsilon I$

$$P(z_k^n|x_n) = \frac{\pi_k e^{\frac{-1}{2}\frac{||x_n-\mu_k||^2}{\epsilon}}}{\sum_{i=1}^{K} \pi_i e^{\frac{-1}{2}\frac{||x_n-\mu_i||^2}{\epsilon}}}$$

if x_n is nearest to some μ_k i.e,kth cluster center, then, the distance to μ_k is less than μ_j for all j which is not k. Let us say the factor be 1+α

$$(1+\alpha)||x_n - \mu_k|| < ||x_n - \mu_j||$$

$$P(z_k^n|x_n) = \frac{\pi_k e^{\frac{-1}{2}\frac{||x_n-\mu_k||^2}{\epsilon}}}{\sum_{i=1}^{K} \pi_i e^{\frac{-1}{2}\frac{||x_n-\mu_i||^2}{\epsilon}}} \geq \frac{\pi_k e^{\frac{-1}{2}\frac{||x_n-\mu_k||^2}{\epsilon}}}{\sum_{i\neq k} \pi_i e^{\frac{-1}{2}\frac{||x_n-\mu_k||^2(1+\alpha)^2}{\epsilon}} + \pi_k e^{\frac{-1}{2}\frac{||x_n-\mu_k||^2}{\epsilon}}}$$

This term goes to 1 as ϵ goes to 0 since

$$\frac{\pi_j e^{\frac{-1}{2}\frac{||x_n-\mu_j||^2}{\epsilon}}}{\pi_k e^{\frac{-1}{2}\frac{||x_n-\mu_k||^2}{\epsilon}}} \approx 0$$

when ϵ goes to 0.
Therefore for nearest mean cluster 'k' as ϵ goes to 0,$P(z_k^n|x_n) \approx 1$ and the for non nearest clusters 'j' this prior is 0.

Loglikelihood function,

$$L = \sum_{n=1}^{N} log(\sum_{k=1}^{K}(\pi_k \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} exp(\frac{-1}{2}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))))$$

$$= \sum_{n=1}^{N} [log(\frac{1}{\sqrt{(2\pi\epsilon)^k}}) + log(\sum_{k=1}^{K}(\pi_k exp(\frac{-1}{2}\frac{||x_n-\mu_k||^2}{\epsilon})))]$$

$$= constant + \sum_{n=1}^{N} log(\sum_{k=1}^{K}(\pi_k exp(\frac{-1}{2}\frac{||x_n - \mu_k||^2}{\epsilon})))$$

By similar logic as above, only one term survives as ϵ goes to 0.rest becomes insignificant. Hence

$$L \approx constant_1 - constant_2 * \sum_{n=1}^{N} \gamma_{nk}||x_n - \mu_k||^2$$

Where γ_{nk} is 1 for nearest cluster, 0 otherwise.
Hence as ϵ goes to 0, maximising L becomes minimizing $\sum_{n=1}^{N} \gamma_{nk}||x_n - \mu_k||^2$ which is K-means objective function.

(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 $p(x|k)$.
Solution:
Note: I am currently solving this problem assuming continuous distribution. The solution for discrete distribution is same, except the symbol of integration need to be replaced with summation symbol.
The idea and logic is same and to avoid repetition, I will not repeat solution for discrete distribution. Let Mean of the distribution be m

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

We know that mean of k'th cluster is μ_k.
Hence mean,

$$\boxed{m = \sum_{k=1}^{K} \pi_k \mu_k}$$

which is nothing but a weighted Mean.
co-variance: Let 'S' denote the co-variance of the whole distribution.

$$S_{ij} = E[(x_i - m_i)(x_j - m_j)],$$

Where E represents Expected value.

$$S_{ij} = \int_x p(x)(x_i - m_i)(x_j - m_j)dx = \sum_{k=1}^{K} \pi_k \int_x P(x|k)((x_i - \mu_{k_i}) + (\mu_{k_i} - m_i))((x_j - \mu_{k_j}) + (\mu_{k_j} - m_j))dx$$

$$= \sum_{k=1}^{K} \pi_k * [\int_x P(x|k)(x_i - \mu_{k_i})(x_j - \mu_{k_j})dx + (\mu_{k_i} - m_i) \int_x P(x|k)(x_j - \mu_{k_j} + \mu_{k_j} - m_j)dx + (\mu_{k_j} - m_j) \int_x P(x|k)(x_i - \mu_{k_i}$$

Note that $\int_x P(x|k)(x_i)dx = \mu_{k_i}$, Hence

$$\int_x P(x|k)(x_i - \mu_{k_i} dx = \int_x P(x|k)(x_j - \mu_{k_j} dx = 0$$

$$S_{ij} = \sum_{k=1}^{K} \pi_k [(\Sigma_k)_{ij} + (\mu_{k_j} - m_j) * (\mu_{k_i} - m_i)]$$

$$= \sum_{k=1}^{K} \pi_k (\Sigma_k)_{ij} + \sum_{k=1}^{K} \pi_k \mu_{k_i} \mu_{k_j} + m_i m_j \sum_{k=1}^{K} \pi_k - m_i \sum_{k=1}^{K} \pi_k \mu_{k_j} - m_j \sum_{k=1}^{K} \pi_k \mu_{k_i}$$

$$= \sum_{k=1}^{K} \pi_k (\Sigma_k)_{ij} + \sum_{k=1}^{K} \pi_k \mu_{k_i} \mu_{k_j} + m_i m_j * 1 - m_i \sum_{k=1}^{K} m_j - m_j \sum_{k=1}^{K} m_i$$

$$= \sum_{k=1}^{K} \pi_k (\Sigma_k)_{ij} + \sum_{k=1}^{K} \pi_k \mu_{k_i} \mu_{k_j} + m_i m_j - 2 m_i m_j$$

$$\boxed{S_{ij} = \sum_{k=1}^{K} \pi_k (\Sigma_k)_{ij} + \sum_{k=1}^{K} \pi_k \mu_{k_i} \mu_{k_j} - m_i m_j}$$

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]

Solution:

Likelihood function,

$$P(X|h) = \prod_{n=1}^{N} P(x_n|h)$$

Let r_n denote region in which x_n lies in. Log-likelihood function,

$$L = ln(P(X|h)) = \sum_{n=1}^{N} ln(P(x_n|h)) = \sum_{n=1}^{N} ln(h_{r_n}) = \sum_i n_i ln(h_i)$$

Hence log likelihood,

$$\boxed{L = \sum_i n_i ln(h_i)}$$

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

Solution:

We know that

$$\sum_i h_i \Delta_i = 1$$

Hence using Largange multiplier to result in part A, we get

$$L = \sum_i n_i ln(h_i) + \lambda(\sum_i h_i \Delta_i - 1)$$

At the point where L is maximized through h_i,

$$\frac{\partial L}{\partial h_i} = 0$$

$$=> \frac{n_i}{h_i} + \lambda \Delta_i = 0$$

$$=> \lambda = \frac{-n_i}{h_i \Delta_i} \tag{6}$$

We know $\sum_i h_i \Delta_i = 1$, using this in above equation we get,

$$=> \sum_i h_i \Delta_i = \frac{\sum_i -n_i}{\lambda} = 1$$

$$=> \lambda = -N$$

Putting this in equation 6, we get

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

(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.
  Solution: F
  In Non-parametric density estimation, parameters may exist though they are not fixed and increase with data set. But they may have parameters.

- The Epanechnikov kernel is the optimal kernel function for all data.
  Solution: T
  The Epanechnikov kernel is a kernel which is optimal in terms of Mean square error. It is the optimal kernel function if Mean sq error is used to measure the accuracy.

- Histogram is an efficient way to estimate density for high-dimensional data.
  Solution: F
  Histogram divides the space into small reqions and calculate expected probabilities in that region. In high dimensional space, the complexity and number of subregions increase exponentially and hence it is not efficient.

- Parametric density estimation assumes the shape of probability density.
  Solution: T
  Parametric density estimation compresses data into finite parameters by assuming shape of data. Using the shape of data and parameters, probability distribution can be recovered. Hence it assumes shape of probability density.

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]
Solution:

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

$$= -\sum_{x \epsilon X} \sum_{y \epsilon Y} P(x)P(y|x) log(P(x) * P(y|x))$$

$$= -\sum_{x \epsilon X} P(x) \sum_{y \epsilon Y} P(y|x)(log(P(x)) + log(P(y|x)))$$

$$= -\sum_{x \epsilon X} P(x) log(P(x)) \sum_{y \epsilon Y} P(y|x) - \sum_{x \epsilon X} P(x) \sum_{y \epsilon Y} P(y|x) log(P(y|x))$$

We know that $\sum_{y \epsilon Y} P(y|x) = 1$, Hence

$$H(X,Y) = H(x) + \sum_{x \epsilon X} P(x)H(y|x)$$

We know that $H(y|x) \leq H(y)$ since we are observing a variable x, Hence

$$H(X,Y) = H(x) + \sum_{x \epsilon X} P(x)H(y|x) \leq H(x) + \sum_{x \epsilon X} P(x)H(y) = H(x) + H(y)$$

Hence,

$$\boxed{H(X,Y) \leq H(x) + H(y)}$$

(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]
Solution:
In part a, we have seen that

$$H(X,Y) = -\sum_{x \epsilon X} P(x)log(P(x)) \sum_{y \epsilon Y} P(y|x) - \sum_{x \epsilon X} P(x) \sum_{y \epsilon Y} P(y|x)log(P(y|x)) = H(x) - \sum_{x \epsilon X} P(x) \sum_{y \epsilon Y} P(y|x)log(P(y|x))$$

Since X,Y are independent, P(y—x) = P(y), Hence

$$H(X,Y) = -\sum_{x \epsilon X} P(x)log(P(x)) \sum_{y \epsilon Y} P(y|x) - \sum_{x \epsilon X} P(x) \sum_{y \epsilon Y} P(y|x)log(P(y|x))$$

$$= H(x) - \sum_{x \epsilon X} P(x) \sum_{y \epsilon Y} P(y)log(P(y))$$

$$= H(x) - \sum_{x \epsilon X} P(x)H(Y) = H(X) - H(Y) \sum_{x \epsilon X} P(x) = H(X) - H(Y)$$

Hence

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

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

$$I(X;Y) = H(Y) - H(Y|X)$$

$$H(Y|X) = -\sum_x \sum_y P(x,y)log(P(y|x))$$

$$= -\sum_x \sum_y P(x,y)log(\frac{P(x,y)}{P(x)})$$

$$= -\sum_x \sum_y P(x,y)log(P(y|x)) + \sum_x \sum_y P(x,y)log(P(x))$$

$$= H(X,Y) + \sum_x log(P(x)) \sum_y P(x,y)$$

$$= H(X,Y) + \sum_x log(P(x))P(x)$$

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

Hence

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - (H(X,Y) - H(X))$$

Therefore

$$\boxed{I(X;Y) = H(X) + H(Y) - H(X,Y)}$$

(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]
Solution:
We know that $P(z = z_i|x = x_i) = P(y = (z_i - x_i)|x = x_i)$ since y = z-x

$$H(Z|X) = -\sum_{x,z} P(x,z)log(P(z|x)) = -\sum_{x,z} P(x)P(z|x)log(P(z|x)) = -\sum_{x,y} P(x)P(y|x)log(P(y|x)) = H(Y|X)$$

Hence

$$\boxed{H(Z|X) = H(Y|X)}$$

Now we know that $H(Z) \geq H(Z|X)$ since entropy reduces once an observation is made.
Hence given X,Y are independent

$$H(Z) \geq H(Z|X) = H(Y|X) = -\sum_{x,y} P(x)P(y|x)log(p(y|x))$$

We know $P(y|x) = P(y)$ since X,Y are independent.

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

Hence

$$\boxed{H(Z) \geq H(Y)}$$

By applying similar logic we get

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

Hence

$$\boxed{H(Z) \geq H(Y)}$$

Hence addition of two independent variables add uncertainty.

(e) Under what conditions does $H(Z) = H(X) + H(Y)$. [4 pts]
Solution: We know Z = X+Y.
Now X,Y has higher information than X+Y and hence entropy is lost, it is easy to see that

$$H(Z) = H(fn(X,Y)) \leq H(X,Y)$$

$$H(Z) = H(fn(X,Y)) \leq H(X,Y) \leq H(X) + H(Y)$$

from part b, we know that H(X,Y) = H(X) + H(Y), when X,Y are independent. now H(Z) = H(X)+ H(Y) when they are **independent**. Also the first equality occurs when we are able to fully recover X,Y. For ex: X is real numbers and Y a complex numbers. Hence H(Z) = H(X) + H(Y) if

- X,Y can be recovered from Z, X = f(Z) and Y = g(Z) and
- X,Y are independent random variables.

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]

Solution: Let $P(y = 1|x) = a$, then $P(y = -1|x) = 1 - a$, Let f(x) be the classification function. Given any point x, if we classify it as 1(f(x) = 1), the expected loss,

$$E_1 = E(L(f(x), y)) = L(1, -1) * P(y = -1|x) = p * (1 - a)$$

Loss if we classify point as -1 (f(x) = -1)

$$E_{-1} = E(L(f(x), y)) = L(-1, 1) * P(y = 1|x) = q * a$$

Hence to minimize loss, we device f such that it minimizes loss

$$E_1 < E_{-1} => p * (1 - a) < q * a => p < (p + q) * a => a > \frac{p}{p + q}$$

Hence we classify 1 if $P(y = 1|x) > \frac{p}{p+q}$, -1 otherwise.

$$\boxed{f(x) = \begin{cases} 1, & \text{if } P(y = 1|x) > \frac{p}{p+q} \\ -1, & \text{otherwise} \end{cases}}$$

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]

Solution:

$$P(y = 1|x) = \frac{\pi_1 \mathcal{N}(x|\mu_1, \Sigma_1)}{\pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_{-1} \mathcal{N}(x|\mu_{-1}, \Sigma_{-1})}$$

In the previous classifier, we classify as 1 if $p * P(y = -1|x) < q * P(y = 1|x)$, (another way to representation). Let Σ_1,μ_1,π_1 and Σ_{-1},μ_{-1},π_{-1} denote the Gaussian parameters for conditional distribution . this is same as checking the ratio

$$r = \frac{q(P(y = 1|x)}{p(P(y = -1|x))}$$

If r < 1, then we return 1, we classify -1 otherwise. Since r is always positive, (P(y=1—x) is non 0 since it is Gaussian)

$$r < 1 <=> h(x) = log(r) < 0$$

$$h(x) = log(r(x)) = log(q/p) - \frac{1}{2} log(\frac{|\Sigma_{-1}|}{|\Sigma_1|}) - \frac{1}{2}(x - \mu_1)\Sigma_1^{-1}(x - \mu_1)^T - \frac{1}{2}(x - \mu_{-1})\Sigma_{-1}^{-1}(x - \mu_{-1})^T$$

$$\boxed{h(x) = log(q/p) - \frac{1}{2} log(\frac{|\Sigma_{-1}|}{|\Sigma_1|}) - \frac{1}{2}(x - \mu_1)\Sigma_1^{-1}(x - \mu_1)^T - \frac{1}{2}(x - \mu_{-1})\Sigma_{-1}^{-1}(x - \mu_{-1})^T}$$

The geometric shape of the boundary is h(x) = 0, which is second degree in terms of x, hence it is a quadratic boundary.

(b) Repeat (a) but assume the two Gaussians have identical covariance matrices. What is the geometric shape of the decision boundary? [10 pts]
Solution: If Σ_1 is identical to Σ_{-1}, then the above equation simplifies to

$$h(x) = log(q/p) - \frac{1}{2}log(\frac{|\Sigma_{-1}|}{|\Sigma_1|}) - \frac{1}{2}(x-\mu_1)\Sigma_1^{-1}(x-\mu_1)^T - \frac{1}{2}(x-\mu_{-1})\Sigma_{-1}^{-1}(x-\mu_{-1})^T$$
$$= log(q/p) - 0 - \frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)^T - \frac{1}{2}(x-\mu_{-1})\Sigma^{-1}(x-\mu_{-1})^T$$
$$= log(q/p)+\frac{1}{2}(x^T\Sigma^{-1}x-\mu_{-1}^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu_{-1}+\mu_{-1}^T\Sigma^{-1}\mu_{-1})-\frac{1}{2}(x^T\Sigma^{-1}x-\mu_1^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu_1+\mu_1^T\Sigma^{-1}\mu_1)$$

canceling terms and grouping x and x^T terms together, we get

$$\boxed{h(x) = (log(q/p) + 1/2 * \mu_{-1}^T\Sigma^{-1}\mu_{-1} - 1/2 * \mu_1^T\Sigma^{-1}\mu_1) + \frac{1}{2}[(\mu_1-\mu_{-1})^T\Sigma^{-1}x + x^T\Sigma^{-1}(\mu_1-\mu_{-1})]}$$

The boundary of of format $c_1 + c_2x + x^Tc_3$ which is a **linear decision boundary in x**.

(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] **Solution:** If the covariance matrix is Identity matrix, then we have

$$h(x) = (log(q/p) + 1/2 * \mu_{-1}^TI\mu_{-1} - 1/2 * \mu_1^TI\mu_1) + \frac{1}{2}[(\mu_1-\mu_{-1})^TIx + x^TI(\mu_1-\mu_{-1})]$$

Note that $\mu_{-1}^T\mu_{-1} = \mu_1^T\mu_1$, Hence

$$\boxed{h(x) = log(q/p) + (\mu_1-\mu_{-1})^Tx}$$

This is a **perpendicular bisector** hyper plane which perpendicularly bisects the line connecting μ_1 and μ_{-1}.

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.

report:

Following is the report of evaluation of the algorithm. Iterations is number of iterations EM algorithm permitted to run.

I have experimented with various cluster sizes and following are the results.

k = 4

Iterations	Accuracy
10	85.75
50	82.75
100	80.5
400	74.5
1600	61.75

k = 5

Iterations	Accuracy
10	74
50	44.25
100	61.5
400	71.5
1600	74.75

k = 7

Iterations	Accuracy
10	76
50	43
100	44
400	40

k = 10

Iterations	Accuracy
10	32.75
50	26.25
100	45
400	38.75
1600	41

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

$$\mathcal{L} = \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} T_{dw} \log P(d, w), \tag{9}$$

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

$$P(d, w) = \sum_{z \in \mathcal{Z}} P(z) P(w|z) P(d|z), \tag{10}$$

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

$$P(z|d, w) = \frac{P(z) P(w|z) P(d|z)}{\sum_{z'} P(z') P(w|z') P(d|z')}. \tag{11}$$

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

$$P(w|z) = \frac{\sum_d T_{dw} P(z|d, w)}{\sum_{w'} \sum_d T_{dw'} P(z|d, w')} \tag{12}$$

$$P(d|z) = \frac{\sum_w T_{dw} P(z|d, w)}{\sum_{d'} \sum_w T_{d'w} P(z|d', w)} \tag{13}$$

$$P(z) = \frac{\sum_d \sum_w T_{dw} P(z|d, w)}{\sum_{z'} \sum_{d'} \sum_{w'} T_{d'w'} P(z'|d', w')}. \tag{14}$$

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.

Solution:

Top few words printed are

W 1: space,field,approach,optimal,algorithms,noise,

W 2: units,local,order,learning,recognition,high,

W 3: point,classification,cell,structure,neuron,functions,

W 4: layer,rate,inputs,features,functions,gaussian,