# CS 7641 Machine Learning: Assignment 2

October 9, 2016

## 1

Mixture of K Gaussian Mixture Models are given by

$p(x) = \Sigma_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)(1.1.1)$
where $\pi_k$ represents the probability of data point belong to $k^{th}$ component and $N(x|\mu_k, \Sigma_k)$ represents normal distribution of x with mean $\mu_k$, covariance matrix $\Sigma_k$

Let's introduce latent variables $z^{(1)}, z^{(2)}, ..., z^{(K)}$ where
$z^{(1)} = [1\ 0\ 0...\ 0]$
$z^{(2)} = [0\ 1\ 0...\ 0]$
.
.
$z^{(K)} = [0\ 0\ 0...\ 1]$
If a data point $x_n$ belong to $2^{nd}$ component, then $z^n = z^{(2)} = [0\ 1\ 0...\ 0]$
where $p(z) = \Pi_{k=1}^{K} {\pi_k}^{z_k}$ for $k$ in [1,K] (1.1.2)
and $p(x|z) = \Pi_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$ (1.1.3)

### 1 (a).

P(d belongs to $k^{th}$ component), $\pi_k$, can be re-written as $\Pi {\pi_k}^{z_k}$ ( as $z_k$ take values either 0 or 1 and will be equal to 0 if the data point doesn't belong to $k$)
$\pi_k = \Pi_{k=1}^{K} {\pi_k}^{z_k}$ (1.1.4)
Similarly $N(x|\mu_k, \Sigma_k)$ can be re-written as
$N(x|\mu_k, \Sigma_k) = \Pi_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$ (1.1.5)
Substituting Equations 1.1.4 and 1.1.5 in Equation 1.1.1,
$p(x) = \Sigma_{k=1}^{K} (\Pi \pi_k z^k)(\Pi N(x|\mu_k, \Sigma_k)^{z_k})$ (1.1.6)
Substituting Equations 1.1.2 and 1.1.3 in Equation 1.1.6,

$\mathbf{p(x) = \Sigma p(z) p(x|z)}$

### 1 (b).

Using Bayes rule,
$p({z_k}^n|x_n) = \frac{p(x_n|{z_k}^n)p({z_k}^n)}{p(x_n)}$

where, $p(x_n) = \Sigma_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$ from Equation 1.1.1

From Eq 1.1.2, probability of ${z_k}^n$ can be derived as,
$p({z_k}^n) = \Pi_{k=1}^{K} {\pi_k}^{{z_k}^n} = \pi_k$
and $p(x_n|{z_k}^n) = \Pi_{k=1}^{K} N(x_n|\mu_k, \Sigma_k)^{{z_k}^n} = N(x_n|\mu_k, \Sigma_k)$ as $z_k$ will be zero when data point does not belong to $k^{th}$ component

$$p({z_k}^n|x_n) = \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\Sigma_{l=1}^{K} \pi_l N(x_n|\mu_l,\Sigma_l)} \quad (1.2.1)$$

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

$$\mathbf{p({z_k}^n|x_n) = \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\Sigma_{l=1}^{K} \pi_l N(x_n|\mu_l,\Sigma_l)} = \frac{\pi_k \frac{1}{\sqrt{|2\pi\Sigma_k|}} e^{-\frac{(x_n-\mu_k){\Sigma_k}^{-1}(x_n-\mu_k)^T}{2}}}{\Sigma_{l=1}^{K} \pi_l \frac{1}{\sqrt{|2\pi\Sigma_l|}} e^{-\frac{(x_n-\mu_l){\Sigma_k}^{-1}(x_n-\mu_l)^T}{2}}}}$$

## 1 (c).

Expectation Maximization for Gaussian Mixture Models:
$p(x, z) = \Pi {\pi_k}^{z_k} N(x|\mu_k, \Sigma_k)^{z_k}$
$p(x, z) = \Pi {\pi_k}^{z_k} \sqrt{\frac{|\Lambda|}{2\pi}}^{z_k} e^{-\frac{z_k(x-\mu_k)^T \Lambda_k (x-\mu_k)}{2}}$ where $\Lambda_k$ is the inverse of covariance matrix $\Sigma_k$
$p(x, z) = \frac{1}{\sqrt{2\pi^{z_k}}} e^{\Sigma z_k log\pi_k + \frac{z_k}{2} log|\Lambda_k| - \frac{z_k}{2}(X^T \Lambda_k X - 2{\mu_k}^T \Lambda_k X + {\mu_k}^T \Lambda_k \mu_k)}$
By reducing the above equation, we get
$p(X, z) = \frac{1}{\sqrt{2\pi^{z_k}}} e^{(\Sigma z_k \beta_k) - \frac{1}{2}\Sigma(z_k X X^T \Lambda_k) + \Sigma(z_k X^T \Lambda_k \mu_k)}$
where $\beta_k$ is dependent only on the fixed parameters and is equal to $(log\pi_k + 0.5 * log|\Lambda_k| - 0.5{\mu_k}^T \Lambda_k \mu_k)$
Probability Density Function $p(X_1, X_2, ..X_n, z_1, z_2, .., z_n) = \Pi_{i=1}^{n} p(X_i, z_i)$
$= \frac{1}{\sqrt{{2\pi^{z_k}}^n}} e^{\Sigma_{k=1}^{K}(\Sigma_{i=1}^{n} z_{ik})\beta_k - \frac{1}{2}\Sigma_{k=1}^{K}(\Sigma_{i=1}^{n} z_{ik} X_i {X_i}^T)\Lambda_k + \Sigma_{k=1}^{K}(\Sigma_{i=1}^{n} z_{ik} {X_i}^T)\Lambda_k \mu_k}$
Log of probability density function:
$ln(p) = \Sigma_{k=1}^{K}(\Sigma_{i=1}^{n} z_{ik})\beta_k - \frac{1}{2}\Sigma_{k=1}^{K}(\Sigma_{i=1}^{n} z_{ik} X_i {X_i}^T)\Lambda_k + \Sigma_{k=1}^{K}(\Sigma_{i=1}^{n} z_{ik} {X_i}^T)\Lambda_k \mu_k$
In the above equation sufficient statistics functions are $\Sigma z_{ik}, \Sigma z_{ik} X_i, \Sigma z_{ik} X_i {X_i}^T$
According to Expectation Maximization $E_{\theta_{t-1}}((S_j(X, Z)|X = x) = E_{\theta_t}(X, Z)$ where $S_j$ are sufficient statistics
Now let's apply the above equation to the sufficient statistics functions of EM of GMMs
(i) To $\Sigma z_{ik}$:
$E_{\theta_0}(\Sigma_{i=1}^{N} z_{ik}|X = x_i) = E_\theta(\Sigma_{i=1}^{N} z_{ik})$
$E_\theta(\Sigma_{i=1}^{N} z_{ik}) = \Sigma_{i=1}^{N}(E_\theta z_{ik}) = \Sigma_{i=1}^{N} \pi_k = N\pi_k$ (7)
$E_{\theta_0}(\Sigma_{i=1}^{N} z_{ik}|X = x_i) = \Sigma_{i=1}^{N} E_{\theta_0}(z_{ik}|X = x_i)$
As derived from 1(b)
$p(z_k^n|x_n) = \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\Sigma \pi_l N(x_n|\mu_l,\Sigma_l)}$, let this be equal to $\gamma(z_{nk})$
So $\Sigma_{i=1}^{N} E_{\theta_0}(z_{ik}|X = x_i) = \Sigma_{i=1}^{N} \gamma(z_{ik}) = N_k$ (8)
Equation equation 7 and 8, we get:
$N\pi_k = N_k$, this implies $\pi_k = \frac{N_k}{N}$
(ii) To $\Sigma z_{ik} X_i$:

$E_{\theta_0}(\Sigma_{i=1}^N z_{ik} X_i | X = x_i) = E_\theta(\Sigma_{i=1}^N z_{ik} X_i)$ (9)
$E_\theta(\Sigma_{i=1}^N z_{ik} X_i) = \Sigma_{i=1}^N (E_\theta(z_{ik} X_i))$
$E_\theta(z_{ik} X_i) = E_\theta(E_\theta(z_{ik} X_i | z_{ik}))$
$E_\theta(z_{ik} X_i | z_{ik}) = E_\theta(X_i | z_{ik} = 1) = \pi_k \mu_k$
$E_\theta(E_\theta(z_{ik} X_i | z_{ik})) = \pi_k \mu_k$
This implies $E_\theta(\Sigma_{i=1}^N z_{ik} X_i) = \Sigma_{i=1}^N (E_\theta(z_{ik} X_i)) = N \pi_k \mu_k = N_k \mu_k$ (10)
$E_{\theta_0}(\Sigma_{i=1}^N z_{ik} X_i | X = x_i) = \Sigma_{i=1}^N E_{\theta_0}(z_{ik} X_i | X = x_i) = \Sigma_{i=1}^N \gamma_{ik} X_i$ (11)
Substituting eq. 10 and 11 in 9,
$N_k \mu_k = \Sigma_{i=1}^N \gamma(z_{ik}) X_i$
This implies $\mu_k = \frac{\Sigma_{i=1}^N \gamma_{ik} X_i}{N_k}$
(iii) To $\Sigma z_{ik} X_i {X_i}^T$:
$E_{\theta_0}(\Sigma_{i=1}^N z_{ik} X_i X_i^T | X_i = x_i) = E_\theta(\Sigma_{i=1}^N z_{ik} X_i X_i^T)$ (12)
$E_\theta(\Sigma_{i=1}^N z_{ik} X_i X_i^T) = \Sigma_{i=1}^N E_\theta(E_\theta(z_{ik} X_i X_i^T | Z_{ik}))$
$E_\theta(z_{ik} X_i X_i^T | Z_{ik}) = E_\theta(X_i X_i^T) = \Sigma_k + \mu_k \mu_k^T$
as covariance$(X_i), \Sigma_i = E(X_i X_i^T) - E(X_i) E(X_i^T)$
$\Sigma_{i=1}^N E_\theta(E_\theta(z_{ik} X_i X_i^T | Z_{ik})) = \Sigma_{i=1}^N E_\theta(\Sigma_k + \mu_k \mu_k^T)$
$= \Sigma_{i=1}^N \pi_k(\Sigma_i + \mu_i \mu_i^T) = N \pi_k (\Sigma_i + \mu_i \mu_i^T) = N_k(\Sigma_i + \mu_i \mu_i^T)$ (13)
$E_{\theta_0}(\Sigma_{i=1}^N z_{ik} X_i X_i^T | X_i = x_i) = \Sigma_{i=1}^N \gamma_{ik} X_i X_i^T$ (14)
Substituting eq. 13 and 14 in 12,
$N_K(\Sigma_i + \mu_i \mu_i^T) = \Sigma_{i=1}^N \gamma_{ik} X_i X_i^T$
Therefore $\Sigma_i = \frac{1}{N_k}(\Sigma_{i=1}^N \gamma_{ik} X_i X_i^T - \mu_i \mu_i^T)$

**1 (d).**
Data log likelihood in EM is given by,
$lnp(X|\pi, \mu, \Sigma) = \Sigma_{n=1}^N ln(\Sigma_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k))$
Derivating above likelihood function w.r.t $\mu_k$ and equating it to 0
$0 = -2\Sigma_{n=1}^N \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\Sigma_{j=1}^K \pi_j N(x|\mu_j, \Sigma_j)} \Sigma_k (x_n - \mu_k) \quad 1.3.1$
From equation 1.2.1, $p({z_k}^n | x_n) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\Sigma_{l=1}^K \pi_l N(x_n|\mu_l, \Sigma_l)} = \gamma(z_{nk})$
Substituting the above equation in Eq 1.3.1,
$0 = -\Sigma_{n=1}^N \gamma(z_{nk}) \Sigma_k (x_n - \mu_k)$
$\mu_k \Sigma_{n=1}^N \gamma(z_{nk}) = \Sigma_{n=1}^N \gamma(z_{nk}) x_n$

$\boldsymbol{\mu_k = \frac{1}{N_k} \Sigma_{n=1}^N \gamma(z_{nk}) x_n}$ ,where $N_k = \Sigma_{n=1}^N \gamma(z_{nk})$

Derivation likelihood function w.r.t $\Sigma_k$ and equating it to 0
Using matrix differentiation priciples, $\frac{d|A|}{dA} = |A|(A^{-1})^T \; and \; \frac{dA^{-1}}{dA} = -(A^{-1})(A^{-1})$
$0 = -\Sigma_{n=1}^N \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\Sigma_{j=1}^K \pi_j N(x_n|\mu_j, \Sigma_j)} (1 + {\Sigma_k}^{-1} (x_n - \mu_k)(x_n - \mu_k)^T)$
Implies,
$N_k = {\Sigma_k}^{-1} \Sigma_{n=1}^N \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$

Multiplying on both sides by ${\Sigma_k}^{-1}$ ,

$\boldsymbol{\Sigma_k = \frac{1}{N_k} \Sigma_{n=1}^N \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}$

Differentiating loglikelihood function w.r.t $\pi_k$ and equating it to 0
Using Lagrange multiplier for maximization by taking constraint $\Sigma_{k=1}^{K}\pi_k = 1$
Loglikelihood function $= \Sigma_{n=1}^{N} ln(\Sigma_{k=1}^{K}\pi_k N(x_n|\mu_k, \Sigma_k)) + \lambda(\Sigma_{k=1}^{K}\pi_k - 1)$
Differentiating w.r.t $\pi_k$:
$0 = \lambda + \Sigma_{n=1}^{N} \frac{N(x_n|\mu_k,\Sigma_k)}{\Sigma_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)}$ (1.3.2)
Multiplying $\Sigma_{k=1}^{K}\pi_k$ on both sides of equation
$-\lambda\Sigma_{k=1}^{K}\pi_k = \Sigma_{n=1}^{N} \frac{\Sigma_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)}{\Sigma_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)}$
This implies $\lambda = -N$ as $\Sigma_{k=1}^{K}\pi_k = 1$ (1.3.3)
Substituting Eq(1.3.3) in 1.3.2 we get

$\pi_\mathbf{k} = \frac{\mathbf{N_k}}{\mathbf{N}}$

# 2

### 2(a).

In a histogram like density model in which space x is divided into fixed regions where density $p(x)$ takes constant value $h_i$ over $ith$ region, volume of region $i$ $\Delta_i$ and $n_i$ of N observations fall in region $i$, the probability that a data point $x_n$ belongs to $jth$ region is given by,
$p(X = x_j) = h_j$
Probability Density function $= p(x_{1i})p(x_{2i})...p(x_{Ni}) = \Pi_{n=1}^{N} h_{n(i)}$, for data points fall in $i_{th}$ region
Log likelihood function is given by, $\mathbf{lnp(X|i) = \Sigma_{n=1}^{N} ln(h_{n(i)}})$

### 2(b).

The histogram-like density model optimization is subject to density constraint given by $\Sigma_{i=1}^{N} h_i\Delta_i = 1$
Using Lagrange multiplier to maximize Log Likelihood function:
$lnp(X|i) = \Sigma_{n=1}^{N} ln(h_{n(i)}) + \lambda(\Sigma h_i\Delta_i - 1)$
Derivating the above equation w.r.t $h_j$ and equating it to 0:
$0 = \Sigma_{n=1}^{N} \frac{1}{h_j} + \lambda\Delta_j = \frac{n_j}{h_j} + \lambda\Delta_j$
This implies, $h_j = \frac{-n_j}{\lambda\Delta_j}$ (2.2.1)
$\Sigma_{j=1}^{N} n_j = N$ ,this implies $-\Sigma_{j=1}^{N} h_j\lambda\Delta_j = N$

This implies, $\lambda\Sigma_{j=1}^{N} h_j\Delta_j = \lambda = -N$ (2.2.2)

Substituting 2.2.2 in 2.2.1,
$\mathbf{h_j} = \frac{\mathbf{n_j}}{\mathbf{N\Delta_j}}$

### 2(c).

- True. Non parametric estimation usually doesn't have any parameters and the estimation is purely done on the basis of data.

- True. The Epanechnikov kernel is the optimal kernel function for all data as it produces smooth curves when compared to uniform or triangular (sharp) kernels, has better smoothing than Gaussian kernels and results in least Mean Square Error.
- False. Histogram is not an efficient way to estimate density for high-dimensional data as histograms can't capture subtle differences in data over various dimensions and leads to statistical error for each bin.
- True. Parametric density estimation assumes shape of probability function.Given a parametric model or shape of probability, parametric density estimation fits data in model.

## 3

### 3(a).

Joint Entropy, $H(X,Y) = -\Sigma_{x\epsilon X}\Sigma_{y\epsilon Y}p(x,y)logp(x,y) = -E[logp(x,y)]$ (3.1.1)

By Bayes' rule $p(x,y) = p(x|y)p(y)$ (3.1.2)

Substituting 3.1.2 in 3.1.1,

$H(X,Y) = -E[logp(x,y)] = -E[log(p(x|y)p(y))]$

$= -E[logp(x|y) + logp(y)] = -E[logp(x|y)] - E[logp(y)]$ (3.1.3)

By Bayes rule$p(x) = p(x|y)p(y) + p(x|\tilde{y})p(\tilde{y})$ and from this equation it can be inferred that $p(x) >= p(x|y)$

Substituting the above inequality in Equation 3.1.3

$H(X,Y) = -E[logp(x|y)] - E[logp(y)] <= -E[logp(x)] - E[logp(y)]$ (3.1.4)

where $-E[logp(x)]$ represents $H(X)$.

Replacing $-E[logp(x)]$ with $H(X)$ and $-E[logp(y)]$ with $H(Y)$ in Eq 3.1.4

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

### 3(b).

Mutual Information, $I(X;Y) = -E_{X,Y}[SI(x,y)]$ (3.2.1)

where SI(x,y) is point-wise mutual information

$I(X;Y) = \Sigma_x \epsilon X, y\epsilon Y p(x,y)log\frac{p(x,y)}{p(x)p(y)}$

$= \Sigma_{x\epsilon X,y\epsilon Y}p(x,y)(logp(x,y) - logp(x) - logp(y))$

$= -\Sigma_{x\epsilon X}p(x,y)logp(x) - \Sigma_{y\epsilon Y}p(x,y)logp(y) + \Sigma_{x\epsilon X,y\epsilon Y}p(x,y)logp(x,y)$

$= -\Sigma_{x\epsilon X}p(x)logp(x) - \Sigma_{y\epsilon Y}p(y)logp(y) + \Sigma_{x\epsilon X,y\epsilon Y}p(x,y)logp(x,y)$

this implies $I(X;Y) = H(X) + H(Y) - H(X,Y)$

### 3(c).

$Z = X + Y$

Entropy, $H(Z) = -\Sigma_{z\epsilon Z}zlogz = -\Sigma_{z\epsilon X+Y}p(z)logp(z)$

$= -\Sigma_{z\epsilon X}p(z)logp(z) - \Sigma_{z\epsilon Y}p(z)logp(z) - \Sigma_{z\in xy}p(z)logp(z)$

$= -\Sigma_{z\epsilon X}p(z)logp(z) - \Sigma_{z\epsilon Y}p(z)logp(z) + \Sigma_{z\in X\&Y}p(z)logp(z)$

For the above statement should be equivalent to $H(X)+H(Y) = -\Sigma_{x\epsilon X}p(x)logp(x)-$

$\Sigma_{y\epsilon Y}p(y)logp(y)$

$-\Sigma_{z\epsilon X}p(z)logp(z)-\Sigma_{z\epsilon Y}p(z)logp(z)+\Sigma_{z\in x\&y}p(z)logp(z) = -\Sigma_{x\epsilon X}p(x)logp(x)- \Sigma_{y\epsilon Y}p(y)logp(y)$

This implies for $z\epsilon X\&Y, p(z)$ should be equal to 0

So the necessary condition to be met is, either X or Y or both X and Y are events with zero occurences and are equal to zero; or X and Y must be independent and mutually exclusive.

# 4 References

https://en.wikipedia.org/wiki/Information$_{t}heory$
$https://en.wikipedia.org/wiki/Joint_{p}robability_{d}istribution$
$http://mathworld.wolfram.com/NonparametricEstimation.html$
$KerneldensityestimationofreliabilitywithapplicationstoextremevaluedistributionBrankoMiladinovic2$
$http://scicomp.stackexchange.com/questions/7572/is-there-a-fast-way-to-compute-histograms-for-high-dimensional-large-datasets$