CS 7641 Machine Learning: Assignment 2

October 11, 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 π_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 μ_k, covariance matrix Σ_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), π_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

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

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

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

1 (c).
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 μ_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,
$0 = -\Sigma_{n=1}^{N} \gamma(z_{nk}) \Sigma_k (x_n - \mu_k)$
$\mu_k \Sigma_{n=1}^{N} \gamma(z_{nk}) = \Sigma_{n=1}^{N} \gamma(z_{nk}) x_n$

$$\mu_{\mathbf{k}} = \frac{\mathbf{1}}{\mathbf{N_k}} \mathbf{\Sigma_{n=1}^{N} \gamma(z_{nk}) x_n} \text{ ,where } N_k = \Sigma_{n=1}^{N} \gamma(z_{nk})$$

Derivation likelihood function w.r.t Σ_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 ${\Sigma_k}^{-1}$,

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

Differentiating loglikelihood function w.r.t π_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 π_k:
$$0 = \lambda + \Sigma_{n=1}^{N} \frac{N(x_n|\mu_k,\Sigma_k)}{\Sigma_{j=1}^{K} \pi_j N(x|\mu_j,\Sigma_j)} \quad (1.3.2)$$

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

$\boldsymbol{\pi_k = \frac{N_k}{N}}$

1(d).
For a Gaussian Mixture Model where all components have covariance ϵI,
$p(x|\mu_k,\Sigma_k) = \frac{1}{\sqrt{2\pi\epsilon}}exp(\frac{-||x-\mu_k||^2}{2\epsilon})$
The posterior probablities are given by,
$\gamma(z_{nk}) = \frac{\pi_k exp(-||x_n-\mu_k||^2/2\epsilon)}{\Sigma_{j=1}\pi_j exp(-||x_n-\mu_j||^2/2\epsilon)}$
Considering the limit, $\epsilon \mapsto 0$ and assuming that none of the π_k are 0, in the denominator the terms $||x_n-\mu_j||^2$ with smallest values tend to zero most slowly. So all the posterior probabilites, $\gamma(z_{nj})$ for x_n will go to zero, except for kth component where $\gamma(z_{nk})$ is equal to 1. This is transition from soft assignment of EM to hard assignment where $\gamma(z_{nk}) = r_{nk}$.
So each data point is associated to nearest center like in K-means.
Substituting the above equation in the mean of GMM,
$\mu_k = \frac{1}{N_k}\Sigma_{n=1}^{N} r_{nk}x_n$
The mean assignment for this model is similar to k-means.
The log Likelihood function for this GMM tends to,
$E_z(lnp(X,Z|\mu,\Sigma,\pi)) = E_z(lnp(Z|X,\mu,\Sigma,\pi)p(X))$

as $\epsilon \mapsto 0$ $E(lnp(X,Z|\mu,\Sigma,\pi)) \mapsto -\frac{1}{2}\Sigma_{n=1}^{N}\Sigma_{k=1}^{K} r_{nk}||x_n-\mu_k||^2 + const$
which is similar to cost function in K-means which is $J = \Sigma_n\Sigma_k r_{nk}||x_n-\mu_k||^2$

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 Δ_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, $\boldsymbol{\mathbf{lnp(X|i)} = \Sigma_{n=1}^{N}\mathbf{ln(h_{n(i)})}}$

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,
$\mathbf{h_j = \frac{n_j}{N\Delta_j}}$

2(c).

- False. Non parametric estimation will have parameters that grow with increase in training data and model is not entirely dependent on paramters instead is dependent in parameter and 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
$I(X;Y) = \Sigma_x \epsilon X, y \epsilon Y p(x,y) log \frac{p(x,y)}{p(x)p(y)}$
$\Sigma_{x,y} p(x,y) log(\frac{p(x,y)}{p(x)}) - p(y) logp(x,y)$
$\Sigma_x p(x)(\Sigma_y p(y|x) logp(y|x)) - \Sigma_y logp(y) \Sigma_x p(x,y))$
$= -\Sigma_x p(x) H(Y|X = x) - \Sigma_y p(y) logp(y)$
$= H(Y) - H(Y|X)$
$H(Y|X) = \Sigma_{x,y} p(y|x) logp(y|x)$
$= \Sigma_{x,y} \frac{p(x,y)}{p(x)} log \frac{p(x,y)}{p(x)}$
$= H(X) - H(X,Y)$
Subtituting the above equation, we get
$I(X;Y) = H(X) + H(Y) - H(X,Y)$

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

3(c).
$Z = X + Y$
Entropy, $H(Z) = -\Sigma_{z \epsilon Z} z logz = -\Sigma_{z \epsilon X+Y} p(z) logp(z)$
$= -\Sigma_{z \epsilon X} p(z) logp(z) - \Sigma_{z \epsilon Y} p(z) logp(z) - \Sigma_{z \in xy} p(z) logp(z)$
$= -\Sigma_{z \epsilon X} p(z) logp(z) - \Sigma_{z \epsilon Y} p(z) logp(z) + \Sigma_{z \in X \& Y} p(z) logp(z)$
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.
The maximum accuracy observed is 84.75. And Average accuracy being 78.

Reference
Pattern Recognition and Machine Learning by Bishop
https://en.wikipedia.org/wiki/Informationtheory
https://en.wikipedia.org/wiki/Joint probability distribution
http://mathworld.wolfram.com/NonparametricEstimation.html
Kernel density estimation of reliability with applications to extreme value distribution Branko Miladinovic 2008