Problem 1. Derive the E-step and M-step update equations of EM algorithm for estimating the Gaussian mixture model $p(X;\theta) = \sum_{k=1}^{K} \pi_k N(x;\mu_k,\sigma_k^2)$ where π_k is the mixture weight with $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, and μ_k, σ_k^2 are the mean and variance of the Gaussian distribution corresponding to cluster k.

Here we derive the EM algorithm fully, in some places the notion is changed to follow the books notation.

1 Define Gaussian Mixture Model in terms of z

We start with the definition of the Gaussian mixture model
$p(X;\theta) = \sum_{k=1}^{K} \pi_k N(x;\mu_k,\sigma_k^2)$
We will introduce a discrete latent variable z such that we may write a joint distribution as
$p(x,z) = p(x)p(x|z)$
The z variable is a binary K dimensional and has has 1-of-K representation subject to
$z_k \in \{0,1\}$ with $\sum_k z_k = 1$
For a given z_k we have the marginal distribution in terms of the mixing coefficient as
$p(z_k = 1) = \pi_k$
For a proper probabilistic distribution we constrain it to
$0 \leqslant \pi_k \leqslant 1$ with $\sum_k \pi_k = 1$
We can now see that for a set of z we may write the total probability distribution as:
$p(z) = p(z_k = 1)p(z_k = 2)...p(z_k = K) = \prod_{k=1}^{K} \pi_k^{z_k}$
Now for the conditional probability we can simply set it to a Guaussian for a given z
$p(x|z_k = 1) = N(x|\mu_k,\sigma_k^2)$
Since we can take a product of probabilities for a set of z we can write as
$p(x|z) = \prod_{k=1}^{K} N(x|\mu_k,\sigma_k^2)^{z_k}$
Now we can see that the joint distribution we want is a sum over all z
$p(x) = \sum_z p(z)p(x|z) = \sum_z \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} N(x|\mu_k,\sigma_k^2)^{z_k}$
because our z only pics out each state at $z_k = 1$ we can reduce this to
$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k,\sigma_k^2)$
We see now that our original Guassian mixture model is reproduced from the joint distribution of the discrete latent variable z
$p(x) = \sum_z p(z)p(x|z) = \sum_{k=1}^{K} \pi_k N(x|\mu_k,\sigma_k^2) = p(X;\theta)$

2 Defining z_{ik}

Now we can prove the following for z_{ik} :
Using Bayes theorem we know that

$$p(z|x) = \frac{p(z)p(x|z)}{p(x)}$$

and we showed that

$$p(x) = \sum_z p(z)p(x|z)$$

pluging in the values for $p(z)$,$p(x|z)$, and $p(x)$from above and we have that

$$z_{ik} = P(z_i = k|X, \mu, \sigma, \pi) = \frac{p(z)p(x|z)}{p(x)} = \frac{\pi_k p(x_i; \mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k p(x_i; \mu_k, \sigma_k^2)}$$

For a Gaussian mixture defined as above $N(x_n|\mu_k, \sigma_k^2)$ we have:

$$= \frac{\pi_k N(x_n|\mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k N(x_n|\mu_k, \sigma_k^2)}$$

3 Maximizing π_k and μ_k

Now we shall follow the same procedure to find the maximum rate of change of both the variables π_k and μ_k, namely we will take the derivatives and set them equal to 0 subject to constrains which will be set by the Lagrange multiplier.

First we will use our above discussion and apply it to a whole data set $X = \{x1...x_N\}$

Since we have

$$p(x) = \sum_{k=1}^K \pi_k N(x|\mu_k, \sigma_k^2)$$

then for a set X we can write

$$p(X|\pi, \mu, \sigma) = \prod_{n=1}^N p(x_n) = \prod_{n=1}^N \sum_{k=1}^K \pi_k N(x_n|\mu_k, \sigma_k^2)$$

Because we wish to maximize this function we are better of writing a likelihood function in terms of the Log which turns the product into a sum component:

$$\ln p(X|\pi, \mu, \sigma) = \ln \prod_{n=1}^N \sum_{k=1}^K \pi_k N(x|\mu_k, \sigma_k^2) = \sum_{n=1}^N \ln \sum_{k=1}^K \pi_k N(x_n|\mu_k, \sigma_k^2)$$

3.1 μ_k

Now we take the derivative of the above equation for $\ln p(X|\pi, \mu, \sigma)$ with respect to μ_k and set to 0:

$$\partial_\mu \ln p(X|\pi, \mu, \sigma) = \partial_\mu \sum_{n=1}^N \ln \sum_{k=1}^K \pi_k N(x_n|\mu_k, \sigma_k^2) =$$
$$\sum_{n=1}^N \frac{\pi_k}{\sum_{k=1}^K \pi_k N(x_n|\mu_k, \sigma_k^2)} \partial_\mu \{N(x_n|\mu_k, \sigma_k^2)\}$$

Where we have used the distributive property of the partial derivative ∂_μ on the sum and Log. Next we can show that for a Gaussian mixture:

$$\partial_\mu \{N(x_n|\mu_k, \sigma_k^2)\} = \partial_\mu \frac{1}{2\pi^{D/2}} \frac{1}{\sigma^{1/2}} \exp\left[-\frac{1}{2}(x_n - \mu_k)^2 \sigma_k^2\right] = -(x_n - \mu_k)\sigma_k^2 N(x_n|\mu_k, \sigma_k^2)$$

So we have then that:

$$\partial_\mu \ln p(X|\pi, \mu, \sigma) = \sum_{n=1}^N \frac{\pi_k}{\sum_{k=1}^K \pi_k N(x_n|\mu_k, \sigma_k^2)} \partial_\mu \{N(x_n|\mu_k, \sigma_k^2)\} = 0$$
$$0 = -\sum_{n=1}^N \frac{\pi_k}{\sum_{k=1}^K \pi_k N(x_n|\mu_k, \sigma_k^2)} (x_n - \mu_k)\sigma_k^2 N(x_n|\mu_k, \sigma_k^2)$$

Or introducing z_{ik} from above we can write this as:

$$0 = -\sum_{n=1}^N \frac{\pi_k N(x_n|\mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k N(x_n|\mu_k, \sigma_k^2)} (x_n - \mu_k)\sigma_k^2 = -\sum_{n=1}^N z_{nk}(x_n - \mu_k)\sigma_k^2$$

we can solve for μ_k

$$0 = -\sum_{n=1}^N z_{nk}(x_n - \mu_k)\sigma_k^2$$

We multiple by $1/\sigma_k^2$, noting that it is now dependent only inside $z_n k$ function

$$0 = -\sum_{n=1}^N z_{nk}(x_n - \mu_k)$$
$$0 = -\sum_{n=1}^N z_{nk} x_n + \sum_{n=1}^N z_{nk}\mu_k$$

$$\sum_{n=1}^{N} z_{nk}x_n = \sum_{n=1}^{N} z_{nk}\mu_k$$

Because μ_k has no dependence on n we can factor it out of the $\sum$

$$\mu_k = \frac{1}{\sum_{n=1}^{N} z_{nk}} \sum_{n=1}^{N} z_{nk}x_n = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk}x_n$$

Where we have defined:

$$N_k = \sum_{n=1}^{N} z_{nk}$$

3.2 π_k

As above we take the derivative and set to 0, but subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$ which we will introduce as a Lagrange multiplier:

$$\ln p(X|\pi,\mu,\sigma) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

Taking the dirivative with respect to π_k and setting to 0:

$$\partial_{\pi_k}\left[\ln p(X|\pi,\mu,\sigma) + \lambda(\sum_{k=1}^{K} \pi_k - 1)\right]$$

We have then:

$$\partial_{\pi_k} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\sigma_k^2) + \partial_{\pi_k}\lambda(\sum_{k=1}^{K} \pi_k - 1) = \sum_{n=1}^{N} \frac{N(x_n|\mu_k,\sigma_k^2)}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\sigma_k^2)} + \lambda = 0$$

We will rewrite this in terms of N_k by noting that we only need π_k in the numerator

$$\pi_k * (\sum_{n=1}^{N} \frac{N(x_n|\mu_k,\sigma_k^2)}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\sigma_k^2)} + \lambda) = 0 * \pi_k$$

$$\sum_{n=1}^{N} z_{nk} + \lambda\pi_k = 0$$

Take a sum over all k where N_k defined above are the number of data in cluster k we have:

$$\sum_k N_k + \lambda \sum_k \pi_k = N + \lambda = 0$$

Thus we have that

$$\lambda = -N$$

Plug this back in and solve for π_k we have:

$$\sum_{n=1}^{N} z_{nk} + (-N)\pi_k = N_k - N\pi_k = 0$$

We have that the max is at:

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

4 EM Algorithm

We have everything needed to define the steps for the EM Algorithm which is as follows (NOTE: we are not asked to derive or update σ_k, so we exclude that discussion) :

1. Initialize the log likelihood function with initial parameters μ_k^i, σ_k^i, and π_k^i
$\ln p(X|\pi,\mu,\sigma) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\sigma_k^2)$

2. E-Step Evaluate z_{nk}^i
$z_{nk} = \dfrac{\pi_k N(x_n|\mu_k,\sigma_k^2)}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\sigma_k^2)}$

3. M-step update the new values with current z_{nk}^i using the maximized parameters

$\mu_k^{i+1} = \frac{1}{\sum_{n=1}^{N} z_{nk}} \sum_{n=1}^{N} z_{nk} x_n = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} x_n$
$\pi_k^{i+1} = \frac{N_k}{N}$

4. Re evaluate the log likelihood function with the new parameters.
$\ln p(X|\pi, \mu, \sigma)^{i+1}$

5. Repeat steps 2-4 until convergence