CSCI567 2013 Homework Assignment 5

Student Name	Arpit Bansal
Last 4 digits of USC ID	0979
I have collaborated with	Jia Li Udit Agrawal Yin Ray Rick Huang

1 Clustering

1.1 Q1

We are given,

$$D = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||x_n - \boldsymbol{\mu}_k||_2^2$$

So, minimizing it w.r.t. $\boldsymbol{\mu}_k$,

$$\begin{aligned}
\frac{\partial D}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} -2r_{nk}(x_n - \boldsymbol{\mu}_k) &= 0 \\
\sum_{n=1}^{N} r_{nk}(x_n - \boldsymbol{\mu}_k) &= 0 \\
\mu_k = \frac{\sum_{n=1}^{N} r_{nk}x_n}{\sum_{n=1}^{N} r_{nk}}
\end{aligned}$$

Here, we can see that $\boldsymbol{\mu}_k$ is the mean of all data points assigned to cluster k.

1.2 Q2

Suppose there are n points $x_1, x_2 ... x_n \ \epsilon \ R^D$ associated with the k^{th} cluster. So loss function for this cluster for a prototype point μ_k will be,

$$D_k = \sum_{n=1}^{N} |x_n - \mu_k|$$

Now, as specified in the problem, the median for a dataset is defined as elementwise median of all data points assigned to the k^{th} cluster. So for any j^{th} element of datapoints, the loss function for the k^{th} cluster will be

$$D_{kj} = \sum_{n=1}^{N} |x_{nj} - \mu_{kj}| = \sum_{x_{nj} < \mu_{kj}} (\mu_{kj} - x_{nj}) + \sum_{x_{nj} > \mu_{kj}} (x_{nj} - \mu_{kj})$$

By taking the derivative w.r.t. to μ_{kj} excluding the points where the above defined loss function is not continous i.e. $x_{nj} = \mu_{kj}$, we get

$$\frac{\partial D_{kj}}{\partial \mu_{kj}} = \sum_{n=1}^{N} I(x_{nj} < \mu_{kj}) - \sum_{n=1}^{N} I(x_{nj} > \mu_{kj})$$

Now, lets take m_j as the median corresponding to j^{th} element of n data points. The above equaion just specify how many data points are in the left and right of the chosen μ_{kj}. So it arises two cases.

Case 1: n is odd
It again arises two cases as below,

$$when\ \mu_{kj} < m_j, \frac{\partial D_{kj}}{\partial \mu_{kj}} < 0$$

$$when\ \mu_{kj} > m_j, \frac{\partial D_{kj}}{\partial \mu_{kj}} > 0$$

So the minimum value for loss function will only be at $\mu_{kj} = m_j$.

Case 2: n is even
It again arises the same two cases as earlier.
In this case, we get the minimum when $m_{j1} < \mu_{k,j} < m_{j2}$ i.e. the average of m_{j1} and m_{j2}. Here m_{j1} and m_{j2} are the left and right value of median m_j.

New iterative algorithm: First select k random start points as the center, then
Step 1: for all data points, find its closest center, and assign itself to that center.
Step 2: calculate the element wise median for each cluster as the new center.
Repeat step 1 and 2, until it converges.

1.3 Q3

We are given $\Sigma_k = \sigma^2 I$. So we can define $p(z = k|x)$ as,

$$p(z = k|x) = \frac{\omega_k p(x|z)}{\sum_{k'} \omega'_k p(x|z)} = \frac{\omega_k e^{-\frac{1}{2\sigma^2}|x-\mu_k|_2^2}}{\sum_{k'} \omega_{k'} e^{-\frac{1}{2\sigma^2}|x-\mu_{k'}|_2^2}} = \frac{1}{\sum_{k'} \frac{\omega_{k'}}{\omega_k} e^{-\frac{1}{2\sigma^2}(|x-\mu_{k'}|_2^2 - |x-\mu_k|_2^2)}}$$

Now, as $\sigma \to 0$, it arises several cases to look for. so if
$(|x - \mu_{k'}|_2^2 - |x - \mu_k|_2^2) > 0$, then $\frac{\omega_{k'}}{\omega_k} e^{-\frac{1}{2\sigma^2}(|x-\mu_{k'}|_2^2 - |x-\mu_k|_2^2)} = \frac{\omega_{k'}}{\omega_k} e^{-\infty} = 0$
or if $(|x - \mu_{k'}|_2^2 - |x - \mu_k|_2^2) < 0$, then $\frac{\omega_{k'}}{\omega_k} e^{-\frac{1}{2\sigma^2}(|x-\mu_{k'}|_2^2 - |x-\mu_k|_2^2)} = \frac{\omega_{k'}}{\omega_k} e^{\infty} = \infty$
or if $(|x - \mu_{k'}|_2^2 - |x - \mu_k|_2^2) = 0$, then $\frac{\omega_{k'}}{\omega_k} e^{-\frac{1}{2\sigma^2}(|x-\mu_{k'}|_2^2 - |x-\mu_k|_2^2)} = \frac{\omega_{k'}}{\omega_k} e^{0} = \frac{\omega_{k'}}{\omega_k}$
So here if μ_k is not closest to x, then there will always be some point $|x - \mu_{k'}|_2^2 < |x - \mu_k|_2^2$. So this condition will satisfy the second equation which makes the $p(z = k|x) = 0$. Also, when μ_k is the closest to x, then in denominator, only the $k' = k$e ratio remains. All other terms satisfy the first equation which contributes equal to zero. In other words, we will have $p(z = k|x) = \frac{1}{\frac{\omega_k}{\omega_k} + \sum_{k' \neq k} 0} = 1$.
This proves the given problem.

2 Gaussian mixture models and EM Algorithm

2.1 Q4

$$\begin{aligned}
Q(\theta, \theta^{old}) &= \sum_{n,k} \gamma_{nk} \log p(x_n|z_n = k, \theta) p(z_n = k|\theta) \\
&= \sum_{n,k} \gamma_{nk} \log w_k [(2\pi)^{\frac{-D}{2}} |\Sigma|^{\frac{-1}{2}} e^{-\frac{1}{2}(x_n-\mu_k)^T \Sigma_k^{-1} (x_n-\mu_k)}] \\
&= \sum_{n,k} \gamma_{nk} \log w_k + \sum_{n,k} -\frac{D}{2} \gamma_{nk} \log 2\pi + \sum_{n,k} -\frac{1}{2} \gamma_{nk} log|\Sigma_k| + \sum_{n,k} -\frac{1}{2} \gamma_{nk} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)
\end{aligned}$$

solve for w_k One constraint on w_k is $\sum_k w_k = 1$. The above function is a concave function. So we first negate function Q to make it convex, then use Lagrange optimization by adding the above specified constraint as $\lambda(\sum_k w_k - 1)$ where λ is a Lagrange mutiplier. So now by taking the derivative w.r.t. w_k and λ, we get

$$\sum_n \frac{\gamma_{nk}}{w_k} - \lambda = 0 \Rightarrow w_k = \sum_n \frac{\gamma_{nk}}{\lambda}$$

and

$$\sum_n w_k = 1$$

By solving above two equations , we get

$$\lambda = \sum_{n,k} \gamma_{nk} = \sum_k N_k$$

$$w_k = \frac{N_k}{\sum_k N_k}$$

Solve for μ_k

$$\begin{aligned}
\frac{\partial Q(\theta, \theta^{old})}{\partial \mu_k} = -\frac{1}{2} \sum_n 2\gamma_{nk} \Sigma_k^{-1} (x_n - \mu_k) &= 0 \\
\Sigma_k^{-1} \sum_n \gamma_{nk} (x_n - \mu_k) &= 0 \\
\mu_k &= \frac{\sum_n \gamma_{nk} x_n}{N_k}
\end{aligned}$$

Solve for Σ_k Here we substitute the covariance matrix Σ with the precision matrix $\Lambda = \Sigma^{-1}$

$$
\begin{aligned}
\frac{\partial Q(\theta, \theta^{old})}{\Lambda_k} = \sum_n \frac{1}{2}\gamma n, k \frac{\partial log|\Lambda|}{\partial \lambda} + \sum_n -\frac{1}{2}\gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T &= 0 \\
\sum_n \gamma_{nk}(\Lambda^{-1})^T - \sum_n \gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T &= 0 \\
\Sigma_k &= \frac{\sum_n \gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_n \gamma_{nk}} \\
&= \frac{1}{N_k} \sum_n \gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T
\end{aligned}
$$

2.2 Q5

This question is same as the previous question except the covariance matrix for all the Gaussian Mixture components is same here. So the only difference in this problem will be about how we update covariance matrix Σ now.

$$
Q(\theta, \theta^{old}) = \sum_{n,k} \gamma_{nk} \log w_k + \sum_{n,k} -\frac{D}{2}\gamma_{nk} \log 2\pi + \sum_{n,k} -\frac{1}{2}\gamma_{nk} log|\Sigma| + \sum_{n,k} -\frac{1}{2}\gamma_{nk}(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k)
$$

By repeating the same process as in the previous problem for deriving w_k, μ_k and Σ, we get

$$
\begin{aligned}
w_k &= \frac{N_k}{\sum_k N_k} \\
\mu_k &= \frac{\sum_n \gamma_{nk} x_n}{N_k} \\
\Sigma &= \frac{1}{\sum_k N_k} \sum_{n,k} \gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T
\end{aligned}
$$