Joel Cheverie
1002924393

CSC 2515: Assignment 2

1. The Gaussian mixture model is:

$$p(t) = \sum_{k=1}^{K} \pi_k N(X \mid \mu_k, \Sigma_k)$$

Let $\gamma(z_k) = p(z_k = 1 \mid X) = \dfrac{p(z_k=1)\,p(X \mid z_k=1)}{\sum_{j=1}^{K} p(z_j=1)\,p(X \mid z_j=1)}$ (by Bayes Rule)

$$= \frac{\pi_k N(X \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(X \mid \mu_j, \Sigma_j)}$$

So then, given X, we are trying to maximize the likelihood of parameters μ_k and Σ_k, where $\Sigma_k = \Sigma \;\forall k$.

So we can rewrite $p(X \mid \mu_k, \pi_k, \Sigma_k)$ to a product of sums of marginals, to find the likelihood:

$$p(X \mid \pi_k, \mu_k, \Sigma_k) = \prod_{n=1}^{N} \left[\sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma) \right.$$

Taking logs, we can consider the log likelihood, as this will be easier to work with.

$$l(X) = \ln(p(X \mid \pi_k, \mu_k, \Sigma)) = \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k N(X \mid \mu_k, \Sigma)\right)$$

Similarly, note that

$$N(X \mid \mu_k, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(X^{(n)} - \mu_k)^T \Sigma^{-1} (X^{(n)} - \mu_k))}$$

First we can take the derivative of $l(X)$ w.r.t. μ_k to maximize μ_k.

$$\frac{\partial \ell(\theta)}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k N(X|\mu_k, \Sigma)\right) =$$

$$\frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{|\Sigma|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}((X^{(n)}-\mu_k)^T \Sigma^{-1}(X^{(n)}-\mu_k))}\right)$$

$$= \frac{\partial}{\partial \mu_k}\left[\sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k e^{-\frac{1}{2}((X^{(n)}-\mu_k)^T \Sigma^{-1}(X^{(n)}-\mu_k))}\right) - K\sum_{n=1}^{N} \ln\left(\sqrt{2\pi^D}|\Sigma|\right)\right]$$

(by the property of logs that $\log(\frac{A}{B}) = \log(A) - \log(B)$ and as the denominator does not depend on the index k, it can be treated as a constant and separated out of the sum).

Now the second term above does not depend on μ_k, so it can be ignored, as the derivative w.r.t. μ_k there is 0. Looking at the first term,

$$= \sum_{n=1}^{N} \frac{1}{\sum_{j=1}^{K} \pi_j e^{-\frac{1}{2}((X^{(n)}-\mu_j)^T \Sigma^{-1}(X^{(n)}-\mu_j))}} \pi_k e^{-\frac{1}{2}((X^{(n)}-\mu_k)^T \Sigma^{-1}(X^{(n)}-\mu_k))} \cdot \left(-\Sigma^{-1}(X^{(n)}-\mu_k)\right)$$

Setting this to 0, we can find μ_k.

$$0 = \sum_{n=1}^{N} \frac{\pi_k e^{-\frac{1}{2}((X^{(n)}-\mu_k)^T \Sigma^{-1}(X^{(n)}-\mu_k))}}{\sum_{j=1}^{K} \pi_j e^{-\frac{1}{2}((X^{(n)}-\mu_j)^T \Sigma^{-1}(X^{(n)}-\mu_j))}} \cdot (X^{(n)}-\mu_k)$$ ($-\Sigma^{-1}$ is just a scaling factor and can be removed).

$$0 = \sum_{n=1}^{N} \frac{\pi_k \left(\frac{1}{\sqrt{2\pi^D}|\Sigma|}\right) e^{-\frac{1}{2}((X^{(n)}-\mu_k)^T \Sigma^{-1}(X^{(n)}-\mu_k))}}{\sum_{j=1}^{K} \pi_j \left(\frac{1}{\sqrt{2\pi^D}|\Sigma|}\right) e^{-\frac{1}{2}((X^{(n)}-\mu_j)^T \Sigma^{-1}(X^{(n)}-\mu_j))}} \cdot (X^{(n)}-\mu_k))$$ (as scaling top and bottom by $\frac{1}{\sqrt{2\pi^D}|\Sigma|}$ does not affect the derivative being zero)

$$0 = \sum_{n=1}^{N} \gamma(z_{nk})(X^{(n)}-\mu_k)$$ (by definition of $\gamma(z_k)$)

$$\Leftrightarrow \sum_{n=1}^{N} \gamma(z_{nk})(X^{(n)}) = \sum_{n=1}^{N} \gamma(z_{nk})\mu_k \quad \Leftrightarrow \quad \mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) X^{(n)}}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

Looking next at Σ,

$$\frac{\partial \ell(\theta)}{\partial \Sigma} = \frac{\partial}{\partial \Sigma} \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k N(X^{(n)}|\mu_k, \Sigma)\right)$$

$$= \sum_{n=1}^{N} \frac{\pi_k}{\sum_{k=1} \pi_k N(x^{(n)}|\mu_k, \Sigma)} \cdot \frac{\partial}{\partial \Sigma} N(x^{(n)}|\mu_k, \Sigma) \qquad (*)$$

Now $\frac{\partial}{\partial \Sigma} N(x^{(n)}|\mu_k, \Sigma) =$

$$\frac{\partial}{\partial \Sigma} \frac{1}{\sqrt{2\pi^D |\Sigma|}} e^{-\frac{1}{2}((x^{(n)}-\mu_k)^T \Sigma^{-1} (x^{(n)}-\mu_k))}$$

$$= \left(\frac{\partial}{\partial \Sigma} \frac{1}{\sqrt{2\pi^D |\Sigma|}}\right) e^{-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1} (x^{(n)}-\mu_k))} + \left(\frac{1}{\sqrt{2\pi^D |\Sigma|}}\right)\left(\frac{\partial}{\partial \Sigma} e^{-\frac{1}{2}((x^{(n)}-\mu_k)^T \Sigma^{-1} (x^{(n)}-\mu_k))}\right)$$

(By Product Rule)

$$= \frac{1}{\sqrt{2\pi^D}} \left(-\frac{1}{2}\right)(|\Sigma|)^{-\frac{1}{2}} (\Sigma^{-1})^T e^{-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1} (x^{(n)}-\mu_k))} + \left(\frac{1}{\sqrt{2\pi^D |\Sigma|}}\right) e^{-\frac{1}{2}((x^{(n)}-\mu_k)^T \Sigma^{-1} (x^{(n)}-\mu_k))} \cdot (\Delta),$$

where $(\Delta) = \frac{\partial}{\partial \Sigma}\left(-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1} (x^{(n)}-\mu_k)\right) = \frac{1}{2}\left(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1}\right)^T$

So putting it together

$$= \frac{-\frac{1}{2}(\Sigma^{-1})^T}{\sqrt{2\pi^D |\Sigma|}} e^{-\frac{1}{2}((x^{(n)}-\mu_k)^T \Sigma^{-1} (x^{(n)}-\mu_k))} + N(x^{(n)}|\mu_k, \Sigma)\left(\frac{1}{2}(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1})^T\right.$$

$$= -\frac{1}{2} N(x^{(n)}|\mu_k, \Sigma)\left[(\Sigma^{-1})^T - (\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1})^T\right]$$

Putting this result back into $(*)$, we have:

$$\frac{\partial \ell(\theta)}{\partial \Sigma} = \sum_{n=1}^{N} \frac{\pi_k N(x^{(n)}|\mu_k, \Sigma)}{\sum_{k=1} \pi_k N(x^{(n)}|\mu_k, \Sigma)} \left[-\frac{1}{2}((\Sigma^{-1})^T - (\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1})^T\right]$$

To maximize this, set it to 0.

$$0 = -\frac{1}{2}\sum_{n=1}^{N} \gamma(z_{nk})\left[(\Sigma^{-1})^T - (\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1})^T\right]$$

$$= -\frac{1}{2}(\Sigma^{-1})^T \sum_{n=1}^{N} \gamma(z_{nk})\left[1 - (\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T)^T\right.$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) - \sum_{n=1}^{N} \gamma(z_{nk})(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T)^T$$

(as the scaling factor $-\frac{1}{2}(\Sigma^{-1})^T$ can just be removed)

But as $\Sigma_k = \Sigma \ \forall k$, we know that Σ is symmetric.

Hence, $\sum_{n=1}^{N} \gamma(z_{nk}) = \sum_{n=1}^{N} \gamma(z_{nk}) (x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T \Sigma^{-1}$

So, $\Sigma = \dfrac{\sum_{n=1}^{N} \gamma(z_{nk}) (x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$

Finally, to obtain π_k, we can also use MLE, with a few extra tricks.

$$\frac{\partial \ln(p(x|\pi,\mu,\Sigma))}{\partial \pi_k}, \quad \text{with constraint } \sum_k \pi_k = 1$$

As suggested in the textbook, we can use a Lagrange multiplier.

$$\frac{\partial \ln(p(x|\pi,\mu,\Sigma))}{\partial \pi_k} = 0 \quad \text{(as we want to find the MLE)}$$

$$\iff 0 = \frac{\partial}{\partial \pi_k}\left[\sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma_k)\right) + \lambda\left(1-\sum_k \pi_k\right)\right] \quad (\lambda \text{ is the Lagrange multiplier})$$

$$0 = \sum_{n=1}^{N} \frac{\partial}{\partial \pi_k} \ln\left(\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma_k)\right) + \lambda \frac{\partial}{\partial \pi_k}\left(1-\sum_k \pi_k\right)$$

$$0 = \sum_{n=1}^{N} \frac{1}{\left(\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma_k)\right)} \cdot \left(N(x^{(n)}|\mu_k,\Sigma_k)\right) + \lambda(-1)$$

$$0 = \sum_{n=1}^{N} \frac{N(x^{(n)}|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma_k)} - \lambda$$

$$0 = \pi_k \cdot \sum_{n=1}^{N} \frac{N(x^{(n)}|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma_k)} - \lambda \cdot \pi_k \quad \text{(as multiplying by a constant doesn't change it being 0).}$$

$$0 = \sum_{n=1}^{N} \gamma(z_{nk}) - \pi_k \lambda \quad (*)$$

So then, one can sum over k, to yield

$$0 = \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk}) - \lambda \sum_{k=1}^{K} \pi_k$$

$$\iff 0 = N - \lambda \quad \left(\text{as } \sum_{k=1}^{K} \pi_k = 1\right) \iff N = \lambda$$

Hence by (*), we get that $\pi_k = \frac{1}{N}\sum_{n=1}^{N} \gamma(z_k^{(n)})$

So putting it all together, the equations which maximize the likelihood function for all of the parameters are:

$$\boxed{\mu_K = \frac{\sum_{n=1}^{N} \gamma(z_{nK})\, x^{(n)}}{\sum_{n=1}^{N} \gamma(z_{nK})}, \quad \Sigma_K = \frac{\sum_{n=1}^{N} \gamma(z_{nK})\, (x^{(n)} - \mu_K)(x^{(n)} - \mu_K)^T}{\sum_{n=1}^{N} \gamma(z_{nK})} \qquad \pi_K = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nK})}$$