Stat3001 Assignment 3

Dominic Scocchera

April 2023

Q1

i)

The complete-data log likelihood is given by:

$$\log L_c(\Psi) = \sum_{i=1}^{2}\sum_{j=1}^{N} z_{ij}\{\log \pi_i + \log \phi(w_j; \mu_i, \sigma_i^2)\}$$

We observe the data:

$$\mathbf{y} = (w_1, ..., w_{n+m}, \mathbf{z}_1^T, ..., \mathbf{z}_n^T)^T$$

And we are missing the component-indicator variables:

$$\mathbf{Z} = (\mathbf{z}_{n+1}^T, ..., \mathbf{z}_{n+m}^T)^T$$

ii)

We see that the complete-data log likelihood is linear in the missing component-indicator variables. The conditional expectation of $\log L_c(\Psi)$ and hence the Q-function is obtained simply by replacing each missing z_{ij} by its conditional expectation given the observed data y. Thus we have that:

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) = &\sum_{j=1}^{n_1} z_{1j}\{\log \pi_1^{(k)} + \log \phi(w_j; \mu_1^{(k)}, {\sigma_1^{(k)}}^2)\} \\ &+ \sum_{j=n_1+1}^{n} z_{2j}\{\log \pi_2^{(k)} + \log \phi(w_j; \mu_2^{(k)}, {\sigma_2^{(k)}}^2)\} \\ &+ \sum_{i=1}^{2}\sum_{j=n+1}^{n+m} \tau_i(w_j; \Psi^{(k)})\{\log \pi_i^{(k)} + \log \phi(w_j; \mu_i^{(k)}, {\sigma_i^{(k)}}^2)\} \end{aligned}$$

Where:

$$\begin{aligned}\tau_i(w_j;\Psi^{(k)}) &= \mathbb{E}_{\Psi^{(k)}}\{z_{ij}|\mathbf{y}\} \\ &= \mathbb{P}_{\Psi^{(k)}}\{z_{ij}=1|\mathbf{y}\} \\ &= \frac{\pi_i^{(k)}\phi(w_j;\mu_i^{(k)},{\sigma_i^{(k)}}^2)}{\sum_{h=1}^{2}\pi_h^{(k)}\phi(w_j;\mu_h^{(k)},{\sigma_h^{(k)}}^2)}\end{aligned}$$

is the posterior probability that $z_{ij}=1$ given the observed value w_j.

iii)

In order to find the updates we take the partial derivatives of the $Q-$function wrt the parameters. First we note that $\log\phi(\Psi)(w_j;\mu_i,\sigma_i^2) = -\frac{1}{2}\log(2\pi\sigma_i^2) - \frac{(w_j-\mu_i)^2}{2\sigma_i^2}$. Now let $\mathcal{L}(\psi) = Q(\Psi;\Psi^{(k)}) + \lambda(\sum_{i=1}^{2}\pi_i - 1)$. Now:

$$\begin{aligned}\frac{\partial\mathcal{L}(\psi)}{\partial\pi_1} &= \sum_{j=1}^{n_1}\frac{z_{1j}}{\pi_1} + \sum_{j=n+1}^{n+m}\frac{\tau_1(w_j;\Psi)}{\pi_1} + \lambda = 0 \\ \iff \hat{\pi_1} &= \frac{\sum_{j=1}^{n_1}z_{1j} + \sum_{j=n+1}^{n+m}\tau_1(w_j;\Psi)}{-\lambda}\end{aligned}$$

The estimator for π_2 is similar and we also have:

$$\begin{aligned}\sum_{i=1}^{2}\pi_i = 1 &= \frac{\sum_{j=1}^{n_1}z_{1j} + \sum_{n+1}^{n+m}\tau_1(w_j;\Psi)}{-\lambda} + \frac{\sum_{j=n_1+1}^{n}z_{2j} + \sum_{n+1}^{n+m}\tau_2(w_j;\Psi)}{-\lambda} \\ \iff -\lambda &= \sum_{j=1}^{n_1}z_{1j} + \sum_{n+1}^{n+m}\tau_1(w_j;\Psi) + \sum_{j=n_1+1}^{n}z_{2j} + \sum_{n+1}^{n+m}\tau_2(w_j;\Psi) \\ &= n_1 + (n-n_1) + (N-n) \\ &= N\end{aligned}$$

So we get:

$$\begin{aligned}\pi_1^{(k+1)} &= \frac{\sum_{j=1}^{n_1}z_{1j} + \sum_{j=n}^{n+m}\tau_1(w_j;\Psi^{(k)})}{N} \\ &= \frac{n_1 + \sum_{j=n}^{n+m}\tau_1(w_j;\Psi^{(k)})}{N}\end{aligned}$$

Now we take the derivative wrt μ_1:

$$\frac{\partial Q(\Psi;\Psi^{(k)})}{\partial \mu_1} = \sum_{j=1}^{n_1} \frac{z_{1j}(w_j-\mu_1)}{\sigma_1^2} + \sum_{j=n+1}^{n+m} \frac{\tau_1(w_j;\Psi)(w_j-\mu_1)}{\sigma_1^2} = 0$$

$$\iff \mu_1 \left(\sum_{j=1}^{n_1} z_{1j} + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi) \right) = \sum_{j=1}^{n_1} z_{1j} w_j + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi) w_j$$

$$\iff \mu_1^{(k+1)} = \frac{\sum_{j=1}^{n_1} z_{1j} w_j + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi^{(k)}) w_j}{\sum_{j=1}^{n_1} z_{1j} + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi^{(k)})}$$

$$= \frac{\sum_{j=1}^{n_1} w_j + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi^{(k)}) w_j}{n_1 + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi^{(k)})}$$

Similarly:

$$\mu_2^{(k+1)} = \frac{\sum_{j=n_1+1}^{n} w_j + \sum_{j=n+1}^{n+m} \tau_2(w_j;\Psi^{(k)}) w_j}{n_2 + \sum_{j=n+1}^{n+m} \tau_2(w_j;\Psi^{(k)})}$$

Now we calculate the derivative wrt to the variance:

$$\frac{\partial Q(\Psi;\Psi^{(k)})}{\partial \sigma_1^2} = \sum_{j=1}^{n_1} -\frac{z_{1j}}{2\sigma_1^2} + \frac{z_{1j}(w_j-\mu_1)^2}{2(\sigma_1^2)^2} + \sum_{j=n+1}^{n+m} -\frac{\tau_1(w_j;\Psi)}{2\sigma_1^2} + \frac{\tau_1(w_j;\Psi)(w_j-\mu_1)^2}{2(\sigma_1^2)^2} = 0$$

$$\iff \sum_{j=1}^{n_1} z_{1j} + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi) = \sum_{j=1}^{n_1} \frac{z_{1j}(w_j-\mu_1)^2}{\sigma_1^2} + \sum_{j=n+1}^{n+m} \frac{\tau_1(w_j;\Psi)(w_j-\mu_1)^2}{\sigma_1^2}$$

$$\iff \sigma_1^{2^{(k+1)}} = \frac{\sum_{j=1}^{n_1} z_{1j}(w_j-\mu_1^{(k)})^2 + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi^{(k)})(w_j-\mu_1^{(k)})^2}{\sum_{j=1}^{n_1} z_{1j} + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi^{(k)})}$$

$$= \frac{\sum_{j=1}^{n_1} (w_j-\mu_1^{(k)})^2 + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi^{(k)})(w_j-\mu_1^{(k)})^2}{n_1 + \sum_{j=n+1}^{n+m} \tau_1(w_j;\Psi^{(k)})}$$

Similarly:

$$\sigma_2^{2^{(k+1)}} = \frac{\sum_{j=n_1+1}^{n} (w_j-\mu_2^{(k)})^2 + \sum_{j=n+1}^{n+m} \tau_2(w_j;\Psi^{(k)})(w_j-\mu_2^{(k)})^2}{n_2 + \sum_{j=n+1}^{n+m} \tau_2(w_j;\Psi^{(k)})}$$

iv)

From lecture notes 7 we know that if we assume:

(1) Ω is a subset in d-dimensional Euclidean space $\mathbb{R}^d$

(2) $\Omega_{\Psi_o} = \{\Psi \in \Omega : L(\Psi) \geq L(\Psi_o)\}$ is compact for any $L(\Psi_o) > -\infty$

(3) $L(\Psi)$ is continuous and differentiable in the interior of Ω

that we will get convergence of the EM algorithm. In our case the second regularity condition does not hold as each data point w_j gives rise to a singularity in $L(\Psi)$ on the edge of the parameter space Ω. More specifically, if $\mu_1(\mu_2)$ is set equal to w_j , then $L(\Psi)$ tends to infinity as $\sigma_1^2(\sigma_2^2)$ tends to zero. Thus in this example, if Ψ_o is any point in Ω with μ_1 or μ_2 set equal to w_j, then clearly Ω_{Ψ_o} is not compact. This space can be made compact by imposing the constraint $\sigma_i^2 > \epsilon (i = 1, 2)$. So the maximum we want, $\Psi^{(k+1)}$ will be a solution of $\frac{\partial Q(\Psi;\Psi^{(k)})}{\partial \Psi} = \mathbf{0}$ and hence we can assume that the Q-function is globally maximised. We also note that this will not hold for EM sequences started sufficiently close to the boundary of the modified parameter space. We fix this issue in V) by setting obvious intial values to the parameters.

v)

The obvious initialisation is to first consider the labelled data and set π_1 to be the proportion of that data labelled as being in class 1, $\pi_2 = 1 - \pi_1$, μ_i $(i \in \{1, 2\})$ the mean of the data from class i and σ_i^2 the variation of the data from class i. The mean and variation can be calculated through the standard ML estimates for the normal distribution:

$$\mu_1 = \frac{\sum_{j=1}^{n_1} w_j}{n_1}$$

$$\mu_2 = \frac{\sum_{j=n_1+1}^{n} w_j}{n_2}$$

$$\sigma_1^2 = \frac{1}{n_1} \sum_{j=1}^{n_1} (w_j - \mu_1)^2$$

$$\sigma_2^2 = \frac{1}{n_2} \sum_{j=n_1+1}^{n} (w_j - \mu_2)^2$$

And π_1 can be calculated through the ML estimate for the categorical distribution:

$$\pi_1 = \frac{n_1}{n}$$

We also see that these are exactly the ML extimates for known data from a mixture of Gaussians with known classes as derived in tutorial sheet 5 Q3 (ii).

Q2

i)

Assuming the classes are independent, the first n_1 data points come from the first normal distribution, the second n_2 data points from the second normal distribution and the remaining m data points from the mixture, the complete-data log likelihood is then given by:

$$L_c(\Psi) = \sum_{j=1}^{n_1} \phi(w_j; \mu_1, \sigma_1^2) + \sum_{j=n_1}^{n} \phi(w_j; \mu_2, \sigma_2^2) + \sum_{i=1}^{2} \sum_{j=n+1}^{n+m} z_{ij} \log \phi(w_j; \mu_i, \sigma_i^2)$$

As the complete-data log likelihood is linear in the missing components we get:

$$\begin{aligned} Q(\Psi; \Psi^{(k)}) = &\sum_{j=1}^{n_1} \log \phi(w_j; \mu_1^{(k)}, \sigma_2^{2^{(k)}}) \\ &+ \sum_{j=n_1+1}^{n} \log \phi(w_j; \mu_2^{(k)}, \sigma_2^{2^{(k)}}) \\ &+ \sum_{i=1}^{2} \sum_{j=n+1}^{n+m} \tau_i(w_j; \Psi^{(k)}) \log \phi(w_j; \mu_i^{(k)}, \sigma_i^{2^{(k)}}) \end{aligned}$$

Where:

$$\begin{aligned} \tau_i(w_j; \Psi^{(k)}) &= \mathbb{E}_{\Psi^{(k)}}\{z_{ij}|\mathbf{y}\} \\ &= \mathbb{P}_{\Psi^{(k)}}\{z_{ij} = 1|\mathbf{y}\} \\ &= \frac{\pi_i \phi(w_j; \mu_i^{(k)}, \sigma_i^{(k)^2})}{\sum_{h=1}^{2} \pi_h \phi(w_j; \mu_h^{(k)}, \sigma_h^{(k)^2})} \end{aligned}$$

ii)

To find the updates we take the derivatives wrt the parameters, however we notice we will get the same updates as in Q1 for the parameters μ_1, μ_2, σ_1^2 and σ_2^2. This is because the only difference in the Q-function is the $\log \pi_i$ terms which disappear when taking the derivative wrt anything that isn't π_i anyways. Now we will take a similar approach as in Q1 to derive the update for π_1.

Consider $\mathcal{L}(\Psi) = Q(\Psi; \Psi^{(k)}) + \lambda(\sum_{i=1}^{2} \pi_i - 1)$. Now:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\Psi)}{\partial \pi_i} &= \sum_{j=n+1}^{n+m} \frac{\tau_i(w_j; \Psi)}{\pi_i} + \lambda = 0 \\
\iff \pi_i &= \frac{\sum_{j=n+1}^{n+m} \tau_i(w_j; \Psi)}{-\lambda}
\end{aligned}$$

We also now have:

$$\begin{aligned}
\sum_{i=1}^{2} \pi_i = 1 &= \frac{\sum_{i=1}^{2} \sum_{j=n+1}^{n+m} \tau_i(w_j; \Psi)}{-\lambda} \\
\iff -\lambda &= \sum_{i=1}^{2} \sum_{j=n+1}^{n+m} \tau_i(w_j; \Psi) \\
&= m
\end{aligned}$$

So the update for π_1 is:

$$\pi_1^{(k+1)} = \frac{\sum_{j=n+1}^{n+m} \tau_i(w_j; \Psi^{(k)})}{m}$$