Stat3001 Assignment 3

Dominic Scocchera

April 2023

Q1

i)

Bayesian

ii)

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:

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

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)

Given that $Q(\Psi;\Psi^{(k)})$ has the same form as the complete-data log likelihood with the unobserved z_{ij} replaced by $\tau_i(w_j;\Psi^{(k)})$, the updated iterate $\Psi^{(k+1)}$ is given by replacing the unobserved z_{ij} with the $\tau_i(w_j;\Psi^{(k)})$. we also note that the ML estimate is the one derived in tutorial sheet 5, Q3 (ii).

$$\begin{aligned}\pi_1^{(k+1)} &= \frac{\sum_{j=1}^{N}\omega_1(w_j;\Psi^{(k)})}{N}\\ \mu_i^{(k+1)} &= \frac{\sum_{j=1}^{N}\omega_i(w_j;\Psi^{(k)})w_j}{\sum_{j=1}^{N}\omega_i(w_j;\Psi^{(k)})} \quad (i=1,2)\\ {\sigma_i^{(k+1)}}^2 &= \frac{\sum_{j=1}^{N}\omega_i(w_j;\Psi^{(k)})(w_j-\bar{w}_i)^2}{\sum_{j=1}^{N}\omega_i(w_j;\Psi^{(k)})} \quad (i=1,2)\end{aligned}$$

Where:

$$\omega_i(w_j;\Psi^{(k)}) = \begin{cases}\tau_i(w_j;\Psi^{(k)}) & \text{if } z_{ij} \text{ was not observed}\\ z_{ij} & \text{else}\end{cases}$$

iv)

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, $\pi_2 = 1-\pi_1$, $\mu_i \;\; (i \in \{1,2\})$ 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 estimates for the categorical distribution:

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

Q2

i)

ii)