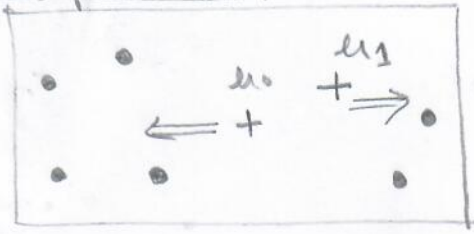
Semi-supervised GMM

(a)



Qualitatively, μ_0 will move leftwards and μ_1 will move rightwards

Justification:

- In the E-step, the four points on the left will have a higher probability of being assigned to cluster 0, compared to cluster 1. Reverse for cluster 1 with the two points on the right.
- In the M-step, the points with higher likelihood of cluster assignment will influence the mean more.

(b) GMM expresses the likelihood of the datapoint $p(x)$ as

$$p(x) = \sum_{i=0}^{1} p(c=i)\, p(x \mid c=i)$$

<u>E-step</u>:

Let p_{ij} denote the likelihood $p(c=i \mid x_j)$, the probability that x_j was generated by component $i \in \{0,1\}$

$$p_{ij} = p(c=i \mid x_j) \propto \underbrace{p(x_j \mid c=i)}_{N(x^j \mid \mu_i, \Sigma_i)}\ \underbrace{p(c=i)}_{\pi_i} \quad i \in \{0,1\}$$

$n_{ij} = \sum_j p_{ij}$ (Effective no. of datapoints that have been assigned to cluster i

<u>M-step</u>:

Compute the new mean, covariance & the component weights as

- Re-estimated mean: $\mu_i \leftarrow \dfrac{\sum p_{ij} x_j}{n_i} \quad i \in \{0,1\}$
- Re-estimated variance: $\Sigma_i \leftarrow \dfrac{\sum_j p_{ij} (x_j - \mu_i)(x_j - \mu_i)^T}{n_i}$
- Re-estimate weight $w_i \leftarrow \left(\dfrac{n_i}{N}\right)$

(c) Modified E-step:

- For the m-points for which the labels are known

$$p_{ij} = \begin{cases} 1 & \text{if } j = y^j \\ 0 & \text{otherwise} \end{cases}, \quad 0 \le j \le m-1$$

- For the remaining $(n-m)$ points for which the label is unknown

$$p_{ij} = p(c=i \mid x_j) \propto p(x_j \mid c=i)\, p(c=i)$$

$$, m \le j \le n-1$$

Justification:

- When the labels are given, there is no uncertainty over the label of the point. Hence, no expectation is to be taken.
- When the labels are not provided, then the likelihood of the data point belonging to a cluster is computed (same as in the standard EM).

Modified M-step

- There will be no change in the equations, except that the updates will take all the n-points in the estimation.

mean, $\mu_i \leftarrow \dfrac{\sum_j p_{ij} x_j}{n_i}$ $\quad i \in \{0,1\}$ $\quad \underbrace{0 \le j \le n-1}_{\text{all points}}$

$$\Sigma_i \leftarrow \frac{\sum_j p_{ij} (x_j - \mu_i)(x_j - \mu_i)^T}{n_i}$$

$$\pi_i \leftarrow \left(\frac{n_i}{N} \right)$$