Ans We already know the cluster for first N inputs (labelled) while last M are unlabelled.

So, we can club the technique used in generative classification and latent variable model (using EM). We will assume z_n for $n > N$, as our latent variables. EM will help us in simulating this condition by making guesses about values of z_n. We will first compute the expectation of each z_n (for $n > N$), and then using these will estimate θ via MLE.

(b.) From the notes, we have.

$$\hat{\theta} = \arg\max_{\theta} E_{p(Z|X,\theta)}[\log p(X, Z|\theta)].$$

$$= \text{''} \sum_{n=1}^{N+M} E_{p(z_n|x_n,\theta)}[\log p(x_n, z_n|\theta)]$$

$$= \arg\max\Big(\sum_{n=1}^{N} \log[\mathcal{N}(x_n|\mu_{y_n}, \Sigma_{y_n})] \Big] = \text{say } (a).$$ labelled

$$\sum_{n=N+1}^{N+M} E_{p(z_n|x_n,\theta)} \log p(x_n, z_n|\theta)\Big). + \cancel{\sum_{n=N+1}^{N+M} E_{p(\;)}[\log}$$

$$= (a) + \underbrace{\sum_{n=N+1}^{N+M} \sum_{k=1}^{K} E[z_{nk}][\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k)]}_{\text{using lecture.}}$$

$$\therefore CLL = (a) + (b)$$

gaussian distribution

$$(a) = \sum_{n=1}^{N} \log(\mathcal{N}(x_n|\mu_{y_n}, \Sigma_{y_n})).$$

$$(b) = \sum_{n=N+1}^{N+M} \sum_{k=1}^{K} \log \pi_k + \log \mathcal{N}(x_n|\mu_{k}, \Sigma_{k}) \cdot E[z_{nk}]$$

defined on next page.

Page No.
Date: / /20

where $E[z_{nk}] = \dfrac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x_n | \mu_i, \Sigma_i)}$,

3.) EM algo is given by.

Initialisation
↓
use labelled exaples ie $n = 1$ to N.
μ_k :- mean of labelled example.
Σ_k :- covariance " " "
π_k :- fraction of labelled example in kth cluster ie $\frac{N_k}{N}$.

$\theta_0 = \{\pi_k, \mu_k, \Sigma_k\}$

E-step 2. Compute expectation of each z_n for $n \geq N$, through above formula

M-step Given responsibilite, reestimate θ via M.LE on CLL defined in part (b)