XCS299i Problem Set #5

JUAN RICARDO PEDRAZA ESCOBAR

Machine Learning

Stanford Center for Professional Development

September, 2021

1.b

In the original representation, each pixel requires $3 \times 8 = 24$ bits to be fully characterized. In the compressed representation, each pixel can be mapped to one of 16 clusters and to represent one of 16 colors requires $log_2 16 = 4$ bits per pixel. Therefore, images are compressed by factor of 24/4 = 6.

2.a

Definition,

$$\ell\left(\theta^{(t+1)}\right) = \alpha\ell_{sup}\left(\theta^{(t+1)}\right) + \ell_{unsup}\left(\theta^{(t+1)}\right)$$

Jensen's inequality,

$$\geq \alpha\ell_{sup}\left(\theta^{(t+1)}\right) + \sum_{i=1}^{n}\sum_{z^{(i)}} Q_i^{(t)}\, z^{(i)} \log\frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)} z^{(i)}}$$

M step,

$$\geq \alpha\ell_{sup}\left(\theta^{(t+1)}\right) + \sum_{i=1}^{n}\sum_{z^{(i)}} Q_i^{(t)}\, z^{(i)} \log\frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)} z^{(i)}}$$

$$= \alpha\ell_{sup}\left(\theta^{(t)}\right) + \ell_{unsup}\left(\theta^{(t)}\right)$$

$$\ell\left(\theta^{(t+1)}\right) = \ell\left(\theta^{(t)}\right)$$

2.b

From the lecture notes:

$$w_j^{(i)} = Q_i\left(z^{(i)} = j\right) = \frac{p\left(x^{(i)} | z^{(i)} = j\,;\, \mu, \Sigma\right) p(z^{(i)} = j)}{\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l; \mu, \Sigma)\, p(z^{(i)} = j)}$$

$$= \frac{\frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \phi_j}{\sum_{l=1}^{k} \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_l\right)^T \Sigma_l^{-1}(x^{(i)} - \mu_l)\right) \phi_l}$$

$$w_j^{(i)} = Q_i\left(z^{(i)} = j\right) = \frac{\left|\Sigma_j\right|^{-1/2} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \phi_j}{\sum_{l=1}^{k} \left|\Sigma_l\right|^{-1/2} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_l\right)^T \Sigma_l^{-1}(x^{(i)} - \mu_l)\right) \phi_l}$$

2.c

List the parameters which need to be re-estimated in the M-step:

In order to simplify derivation, it is useful to denote

$$w_j^{(i)} = Q_i^{(t)}(z^{(i)} = j)$$

And

$$\tilde{w}_j^{(i)} = \begin{cases} \alpha & \tilde{z}^{(i)} = j \\ 0 & \text{otherwise.} \end{cases}$$

We further denote $S = \Sigma - 1$, and note that because of chain rule of calculus, $\nabla S^{\ell} = 0 \Rightarrow \nabla \Sigma^{\ell} = 0$. So, we choose to rewrite the M-step in terms of S and maximize it w.r.t S, and re-express the resulting solution back in terms of Σ. Based on this, the M-step becomes:

$$\phi^{(t+1)}, \mu^{(t+1)}, S^{(t+1)} = \arg\max_{\phi,\mu,S} \sum_{i=1}^{n} \sum_{j=1}^{k} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, S)}{Q_i^{(t)}(z^{(i)})} + \alpha \sum_{i=1}^{\tilde{n}} \log p(\tilde{x^{(i)}}, \tilde{z^{(i)}}; \phi, \mu, S)$$

$$= \arg\max_{\phi,\mu,S} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{|S_j|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left(\left(-\frac{1}{2} \left(x^{(i)} - \mu_j\right)^T S_j \left(x^{(i)} - \mu_j\right)\right) \phi_j\right.}{w_j^{(i)}}$$

$$+ \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{k} \tilde{w}_j^{(i)} \log \frac{\frac{|S_j|^{1/2}}{(2\pi)^{d/2}} \exp\left(\left(-\frac{1}{2} \left(\tilde{x}^{(i)} - \mu_j\right)^T S_j \left(\tilde{x}^{(i)} - \mu_j\right)\right) \phi_j\right.}{\tilde{w}_j^{(i)}}$$

Now, calculate the update steps by maximizing the expression within the argmax for each parameter (We will do the first for you). ϕ_j : We construct the Lagrangian including the constraint that $\sum_{j=1}^{k} \phi_j = 1$, and absorbing all irrelevant terms into constant C:

$$\mathcal{L}(\phi, \beta) = C + \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{k} \tilde{w}_j^{(i)} \log \phi_j + \beta \left(\sum_{j=1}^{k} \phi_j - 1 \right)$$

$$\nabla_{\phi_j}\mathcal{L}(\phi,\beta) = \sum_{i=1}^{n} w_j^{(i)}\frac{1}{\phi_j} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\frac{1}{\phi_j} + \beta = 0$$
$$\Rightarrow \phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{-\beta}$$

$$\nabla_{\beta}\mathcal{L}(\phi,\beta) = \sum_{j=1}^{k}\phi_j - 1 = 0$$
$$\Rightarrow \sum_{j=1}^{k}\frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{-\beta} = 1$$
$$\Rightarrow -\beta = \sum_{j=1}^{k}\left(\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\right)$$

$$\Rightarrow \phi_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{\sum_{j=1}^{k}\left(\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\right)}$$
$$= \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{n + \alpha\tilde{n}}$$

μ_j : Next, derive the update for μ_j . Do this by maximizing the expression with the argmax above with respect to μ_j .

First, we calculate the gradient with respect to μ_j and next we set the gradient to zero and solve for μ_j

$$0 = -\nabla_{\mu_j}\left(C + \sum_{i=1}^{n} w_j^{(i)}\left(x^{(i)} - \mu_j\right)^T \mathrm{S}_j\left(x^{(i)} - \mu_j\right) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\left(\tilde{x}^{(i)} - \mu_j\right)^T \mathrm{S}_j\,(\tilde{x}^{(i)} - \mu_j)\right)$$

$$0 = -2\left(\sum_{i=1}^{n} w_j^{(i)}\left(-S\,x^{(i)} + S\,\mu_j\right) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\,(-S\tilde{x}^{(i)} + S\mu_j)\right)$$

$$0 = -2S\left(\sum_{i=1}^{n} w_j^{(i)}\,x^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\tilde{x}^{(i)}\right) - 2S\left(\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\right)\mu_j$$

solve for μ_j,

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_j^{(i)}\,x^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\tilde{x}^{(i)}}{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}$$

Σ_j: Finally, derive the update for Σ_j via S_j. Again, Do this by maximizing the expression with the argmax above with respect to S_j.

$$0 = \nabla_{S_j}\left(C + \sum_{i=1}^{n} w_j^{(i)}\left(\log|\mathrm{S}_j| - \left(x^{(i)} - \mu_j\right)^T \mathrm{S}_j\left(x^{(i)} - \mu_j\right)\right) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\log|\mathrm{S}_j| - \left(\tilde{x}^{(i)} - \mu_j\right)^T \mathrm{S}_j\,(\tilde{x}^{(i)} - \mu_j)\right)$$

$$0 = \left(\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\,S_j^{-1}\right) - \left(\sum_{i=1}^{n} w_j^{(i)}\left(x^{(i)} - \mu_j\right)^T \mathrm{S}_j\left(x^{(i)} - \mu_j\right) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\left(\tilde{x}^{(i)} - \mu_j\right)^T \mathrm{S}_j\,(\tilde{x}^{(i)} - \mu_j)\right)$$

$$S_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_j^{(i)}\left(x^{(i)} - \mu_j\right)^T \mathrm{S}_j\left(x^{(i)} - \mu_j\right) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\left(\tilde{x}^{(i)} - \mu_j\right)^T \mathrm{S}_j\,(\tilde{x}^{(i)} - \mu_j)}{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}$$

This results in the final set of update expressions:

$$\phi_j := \frac{\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = j\}}{n + \alpha \tilde{n}}$$

$$\mu_j := \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\left\{\tilde{z}^{(i)} = j\right\} x^{(i)}}{\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)} = j\}}$$

$$\Sigma_j := \frac{\sum_{i=1}^{n} w_j^{(i)} \left(x^{(i)} - \mu_j\right)^T \mathrm{S}_j \left(x^{(i)} - \mu_j\right) + \alpha \sum_{i=1}^{\tilde{n}} 1\left\{\tilde{z}^{(i)} = j\right\} \left(x^{(i)} - \mu_j\right)^T \mathrm{S}_j \left(x^{(i)} - \mu_j\right)}{\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\left\{\tilde{z}^{(i)} = j\right\}}$$

2.f

1. The unsupervised GMM took more iterations (150 iterations) to converge than the semi-supervised GMM took 25 iterations.

2. Semi-supervised GMM was more stable than the unsupervised version, clusters were almost exactly in all three experiments conducted. However, the clustering result for unsupervised GMM varied among three experiments.

3. Unsupervised EM suffers from a few types of errors. For example, it fails discovering there is just a single relatively high-variance Gaussian distribution in the mixture and in the other hand the Semi-supervised EM almost exactly uncovers the underlying distribution. Therefore, Semi-supervised finds a higher quality assignment overall.