Stanford CS229 ps3 sol

Shuifan Wang

August 2025

1 A Simple Neural Network

(a) The activation function σ satisfies that $\sigma' = \sigma(1-\sigma)$, therefore we have

$$\begin{aligned}\frac{\partial l}{\partial w_{1,2}^{[1]}} =& \frac{1}{m}\frac{\partial \sum_{i=1}^{m}(o^{(i)}-y^{(i)})^2}{\partial w_{1,2}^{[1]}} \\ =& \frac{\partial l}{\partial o^{(i)}}\frac{\partial o^{(i)}}{\partial (W^{[2]}h^{(i)}+W_0^{[2]})}\frac{\partial (W^{[2]}h^{(i)}+W_0^{[2]})}{\partial h_2^{(i)}}\frac{\partial h_2^{(i)}}{\partial (W^{[1]}x^{(i)}+W_0^{[1]})}\frac{\partial (W^{[1]}x^{(i)}+W_0^{[1]})}{\partial w_{1,2}^{[1]}} \\ =& \frac{2}{m}\sum_{i=1}^{m}(o^{(i)}-y^{(i)})o^{(i)}(1-o^{(i)})w_2^{[2]}h_2^{(i)}(1-h_2^{(i)})x_1^{(i)}.\end{aligned}$$

In the last equation, we use $o = \sigma(W^{[2]}h + W_0^{[2]})$ and $h = \sigma(W^{[1]}x + W_0^{[1]})$. So the gradient descent update should be

$$w_{1,2}^{[1]} := w_{1,2}^{[1]} - \alpha\frac{2}{m}\sum_{i=1}^{m}(o^{(i)}-y^{(i)})o^{(i)}(1-o^{(i)})w_2^{[2]}(W^{[1]}x^{(i)}+W_0^{[1]})(1-(W^{[1]}x^{(i)}+W_0^{[1]}))x_1^{(i)}.$$

(b) Yes we have. Let the three lines of classifier lie on the three sides of the dataset triangle. Then the sample will be classified into the triangle if it is greater than zero (at least one is greater than zero) and out of the triangle if it is less than zero, w.r.t these three linear classifiers:

$$\begin{aligned}w_{1,1}^{[1]}x_1 + w_{2,1}^{[1]}x_2 + w_{0,1}^{[1]} = 0 \\ w_{1,2}^{[1]}x_1 + w_{2,2}^{[1]}x_2 + w_{0,2}^{[1]} = 0 \\ w_{1,3}^{[1]}x_1 + w_{2,3}^{[1]}x_2 + w_{0,3}^{[1]} = 0.\end{aligned}$$

Then we can solve from the figure that (the solution is not unique) $w_{1,1}^{[1]} = 1, w_{2,1}^{[1]} = 1, w_{0,1}^{[1]} = -4; w_{1,2}^{[1]} = -1, w_{2,2}^{[1]} = 0, w_{0,2}^{[1]} = 0.5; w_{1,3}^{[1]} = 0, w_{2,3}^{[1]} = -1, w_{0,3}^{[1]} = 0.5$. And $w_0^{[2]} = -0.5, w_1^{[2]} = 1, w_2^{[2]} = 1, w_3^{[2]} = 1$.

(c) No, it's not possible. If it is possible, then weight $w^{[2]}$ must satisfies that $o = \sigma(W^{[2]}h + W_0^{[2]}) = \sigma(W^{[2]}W^{[1]}x + W^{[2]}W_0^{[1]} + W_0^{[2]})$. It is just one linear classifier and cannot classify non-linear shape triangle.

2 KL divergence and Maximum Likelihood

(a) It is obvious that $f(x) = -\log x$ is a convex function. Therefore from Jensen's inequality we have

$$D_{KL}(P||Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} = \mathbb{E}(-\log(\frac{Q(x)}{P(x)})) \geq -\log(\mathbb{E}(\frac{Q(x)}{P(x)})) = -\log\sum_{x\in\mathcal{X}} Q(x) = 0.$$

The equation holds if and only if

$$\frac{P(x)}{Q(x)} = \text{Constant}, \forall x \in \mathcal{X} \Leftrightarrow P = Q,$$

since

$$\sum_{x\in\mathcal{X}} P(x) = 1 = \sum_{x\in\mathcal{X}} Q(x).$$

(b)

$$\begin{aligned}
D_{KL}(P(Y|X)||Q(Y|X)) =& \sum_y P(y)(\sum_x P(x|y)\log\frac{P(x|y)}{Q(x|y)}) \\
=& \sum_y P(y)(\sum_x \frac{P(x,y)}{P(y)}\log\frac{P(x,y)/P(y)}{Q(x,y)/Q(y)}) \\
=& \sum_y P(y)(\sum_x \frac{P(x,y)}{P(y)}(\log\frac{P(x,y)}{Q(x,y)} - \log\frac{P(y)}{Q(y)})) \\
=& \sum_{x,y} P(x,y)\log\frac{P(x,y)}{Q(x,y)} - \sum_y P(y)\log\frac{P(y)}{Q(y)}(\sum_x \frac{P(x,y)}{P(y)}) \\
=& D_{KL}(P(X,Y)||Q(X,Y)) - D_{KL}(P(X)||Q(X)).
\end{aligned}$$

The second last equation is due to the fact that $\log(P(y)/Q(y))$ is irrelative to x. Thus we prove the chain rule for KL divergence.
(c)

$$\begin{aligned}
\arg\min_\theta D_{KL}(\hat{P}||P_\theta) =& \arg\min_\theta \sum_x \hat{P}(x)\log\frac{\hat{P}(x)}{P_\theta(x)} \\
=& -\arg\min_\theta \sum_x \hat{P}(x)\log P_\theta(x) \\
=& \arg\max_\theta \frac{1}{m}\sum_x (\sum_{i=1}^m 1\{x^{(i)} = x\}\log P_\theta(x)) \\
=& \arg\max_\theta \sum_{i=1}^m \sum_x (1\{x^{(i)} = x\}\log P_\theta(x)) \\
=& \arg\max_\theta \sum_{i=1}^m \log P_\theta(x^{(i)}).
\end{aligned}$$

3 KL Divergence, Fisher Information, and the Natural Gradient

(a)

$$\begin{aligned}
\mathbb{E}_{y\sim p(y;\theta)}[\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}] =& \mathbb{E}_{y\sim p(y;\theta)}[\frac{\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)}] \\
=& \int_{-\infty}^{\infty}[\frac{\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)}]p(y;\theta)dy \\
=& \int_{-\infty}^{\infty}\nabla_{\theta'}p(y;\theta')|_{\theta'=\theta}dy \\
=& \nabla_{\theta'}(\int_{-\infty}^{\infty}p(y;\theta')dy)|_{\theta'=\theta} \\
=& 0.
\end{aligned}$$

(b)

$$\begin{aligned}
\mathcal{I}(\theta) =& \mathrm{Cov}_{y\sim p(y;\theta)}[\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}] \\
=& \mathbb{E}_{y\sim p(y;\theta)}[\nabla_{\theta'}\log p(y;\theta')\nabla_{\theta'}\log p(y;\theta')^T|_{\theta'=\theta}] \\
& - \mathbb{E}_{y\sim p(y;\theta)}[\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}]\mathbb{E}_{y\sim p(y;\theta)}[\nabla_{\theta'}\log p(y;\theta')^T|_{\theta'=\theta}] \\
=& \mathbb{E}_{y\sim p(y;\theta)}[\nabla_{\theta'}\log p(y;\theta')\nabla_{\theta'}\log p(y;\theta')^T|_{\theta'=\theta}].
\end{aligned}$$

(c) Similar to (a), we have

$$\begin{aligned}
& \mathbb{E}_{y\sim p(y;\theta)}[-\nabla^2_{\theta'}\log p(y;\theta')|_{\theta'=\theta}]_{ij} \\
=& -\mathbb{E}_{y\sim p(y;\theta)}[\frac{\partial^2\log p(y;\theta')}{\partial\theta'_i\partial\theta'_j}|_{\theta'=\theta}] \\
=& \mathbb{E}_{y\sim p(y;\theta)}[(\frac{1}{p^2(y;\theta')}\frac{\partial p(y;\theta')}{\partial\theta'_i}\frac{\partial p(y;\theta')}{\partial\theta'_j} - \frac{1}{p(y;\theta')}\frac{\partial^2 p(y;\theta')}{\partial\theta'_i\partial\theta'_j})|_{\theta'=\theta}] \\
=& \mathbb{E}_{y\sim p(y;\theta)}[\frac{1}{p^2(y;\theta')}\frac{\partial p(y;\theta')}{\partial\theta'_i}\frac{\partial p(y;\theta')}{\partial\theta'_j}|_{\theta'=\theta}] - \int_{-\infty}^{\infty}\frac{1}{p(y;\theta)}\frac{\partial^2 p(y;\theta)}{\partial\theta_i\partial\theta_j}p(y;\theta)dy \\
=& \mathbb{E}_{y\sim p(y;\theta)}[\frac{1}{p^2(y;\theta')}\frac{\partial p(y;\theta')}{\partial\theta'_i}\frac{\partial p(y;\theta')}{\partial\theta'_j}|_{\theta'=\theta}] - \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int_{-\infty}^{\infty}p(y;\theta)dy \\
=& \mathbb{E}_{y\sim p(y;\theta)}[\frac{1}{p^2(y;\theta')}\frac{\partial p(y;\theta')}{\partial\theta'_i}\frac{\partial p(y;\theta')}{\partial\theta'_j}|_{\theta'=\theta}].
\end{aligned}$$

At the same time, we have

$$\begin{aligned}
\mathcal{I}(\theta) =& \mathbb{E}_{y\sim p(y;\theta)}[\nabla_{\theta'}\log p(y;\theta')\nabla_{\theta'}\log p(y;\theta')^T|_{\theta'=\theta}]_{ij} \\
=& \mathbb{E}_{y\sim p(y;\theta)}[\frac{1}{p^2(y;\theta')}\frac{\partial p(y;\theta')}{\partial\theta'_i}\frac{\partial p(y;\theta')}{\partial\theta'_j}|_{\theta'=\theta}].
\end{aligned}$$

Therefore $\mathcal{I}(\theta) = \mathbb{E}_{y\sim p(y;\theta)}[-\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}]$.
(d) With the Taylor Series expansion,

$$\begin{aligned}
D_{KL}(p_\theta || p_{\theta+d}) &= \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\theta)] - \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\theta+d)] \\
&\approx -\mathbb{E}_{y\sim p(y;\theta)}(d^T[\nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta}] + \frac{1}{2} d^T[\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}]d) \\
&= -d^T \mathbb{E}_{y\sim p(y;\theta)}[\nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta}] - \frac{1}{2} d^T \mathbb{E}_{y\sim p(y;\theta)}[\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}]d \\
&= \frac{1}{2} d^T \mathcal{I}(\theta) d.
\end{aligned}$$

(e) With the Taylor approximations substituted in, we solves

$$d^* = \arg\max_d (\log p(y;\theta) + d^T[\nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta}]) \quad \text{subject to} \quad \frac{1}{2} d^T \mathcal{I}(\theta) d = c.$$

Then we construct the Lagrangian

$$\mathcal{L}(d,\lambda) = d^T[\nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta}] - \lambda(d^T \mathcal{I}(\theta) d - 2c),$$

and solves that

$$\nabla_d \mathcal{L}(d,\lambda) = [\nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta}] - 2\lambda \mathcal{I}(\theta) d = 0 \Leftrightarrow \tilde{d} = \mathcal{I}(\theta)^{-1} \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{2\lambda p(y;\theta)}.$$

Moreover, note that $\mathcal{I}(\theta)$ is symmetric,

$$\begin{aligned}
\nabla_\lambda \mathcal{L}(d,\lambda) &= 2c - d^T \mathcal{I}(\theta) d \\
&= 2c - \frac{[\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}]^T}{2\lambda p(y;\theta)} \mathcal{I}(\theta)^{-1} \mathcal{I}(\theta) \mathcal{I}(\theta)^{-1} \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{2\lambda p(y;\theta)} \\
&= 0,
\end{aligned}$$

so

$$\lambda = \frac{1}{2p(y;\theta)} \sqrt{\frac{[\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}]^T \mathcal{I}(\theta)^{-1} \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{2c}}.$$

Plug this into d, we have

$$d^* = \sqrt{\frac{2c}{[\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}]^T \mathcal{I}(\theta)^{-1} \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}} \mathcal{I}(\theta)^{-1} \nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}.$$

(f) As for the natural gradient, the direction should be

$$\theta := \theta + \mathcal{I}(\theta)^{-1} \nabla_\theta L(\theta).$$

As for GLM,

$$\theta := \theta - (\nabla^2_\theta L(\theta))^{-1} \nabla_\theta L(\theta).$$

They are the same, since $\mathcal{I}(\theta) = \mathbb{E}_{y\sim p(y;\theta)}[-\nabla^2_{\theta'} \log p(y;\theta')|_{\theta'=\theta}] = -\nabla^2_\theta L(\theta)$.

4 Semi-supervised EM

(a) We can derive it the same way as the lecture note.

$$\begin{aligned}
\ell_{\text{semi-sup}}(\theta^{(t+1)}) =&[\sum_{i=1}^{m}(\sum_{z^{(i)}} Q_i^{(t+1)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t+1)}(z^{(i)})}) + \alpha(\sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)}))] \\
\geq&[\sum_{i=1}^{m}(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})}) + \alpha(\sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)}))] \\
\geq&[\sum_{i=1}^{m}(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})}) + \alpha(\sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)}))] \\
=&\ell_{\text{semi-sup}}(\theta^{(t)}).
\end{aligned}$$

(b) The latent variable re-estimated is $z^{(i)}$ (actually the same as unsupervised case).

$$\begin{aligned}
w_j^{(i)} :=&p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) \\
=&\frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)}|z^{(i)} = l; \mu, \Sigma)p(z^{(i)} = l; \phi)} \\
=&\frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j))\phi_j}{\sum_{l=1}^{k} \frac{1}{(2\pi)^{d/2}|\Sigma_l|^{1/2}} \exp(-\frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1}(x^{(i)} - \mu_l))\phi_l}.
\end{aligned}$$

(c) The parameters that should be re-estimated are ϕ, μ, Σ. As for ϕ, since $\sum_{l=1}^{k} \phi_l = 1$, consider Lagrangian

$$\mathcal{L}(\phi, \lambda) = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \alpha \sum_{i=1}^{\tilde{m}}\sum_{j=1}^{k} 1\{\tilde{z}^{(i)} = j\} \log \phi_j + \lambda(\sum_{j=1}^{k} \phi_j - 1).$$

Then we have that

$$\phi_j = -\frac{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)} = j\}}{\lambda} \Rightarrow -\frac{\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}}\sum_{j=1}^{k} 1\{\tilde{z}^{(i)} = j\}}{\lambda} = 1.$$

So

$$\lambda = -(\sum_{i=1}^{m} 1 + \alpha \sum_{i=1}^{\tilde{m}} 1) = -(m + \alpha\tilde{m}), \quad \phi_j := \frac{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)} = j\}}{m + \alpha\tilde{m}}.$$

As for μ,

$$\nabla_{\mu_j} \ell_{\text{semi-sup}} = \sum_{i=1}^{m} w_j^{(i)} \Sigma_j^{-1}(x^{(i)} - \mu_j) + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)} = j\} \Sigma_j^{-1}(\tilde{x}^{(i)} - \mu_j) = 0,$$

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

As for Σ,

$$\nabla_{\Sigma_j^{-1}} \ell_{\text{semi-sup}} = \sum_{i=1}^{m} \frac{1}{2} w_j^{(i)} (\Sigma_j - (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T) + \alpha \sum_{i=1}^{\tilde{m}} \frac{1}{2} 1\{\tilde{z}^{(i)} = j\} (\Sigma_j - (\tilde{x}^{(i)} - \mu_j)(\tilde{x}^{(i)} - \mu_j)^T) = 0,$$

$$\Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)} = j\} (\tilde{x}^{(i)} - \mu_j)(\tilde{x}^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} 1\{\tilde{z}^{(i)} = j\}}.$$

Here we omit the superscript (t) to align with the lecture notes.
(d) Three plots using classical unsupervised EM implementation:

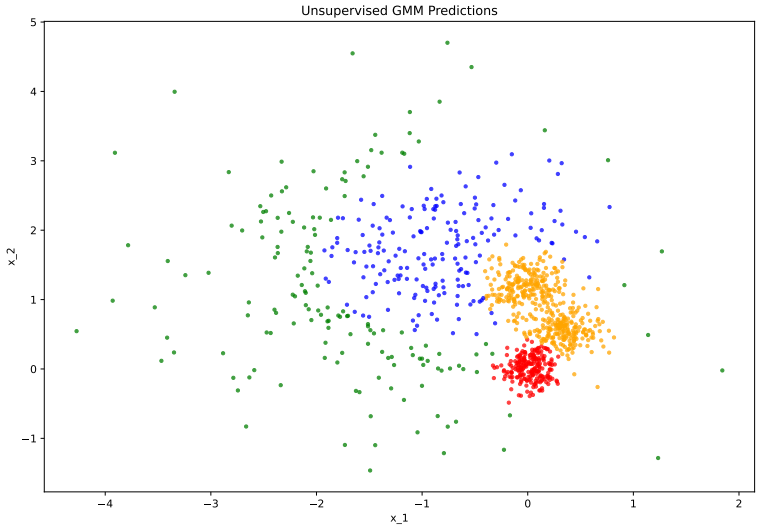

Figure 1: Unsupervised learning of 1st trial

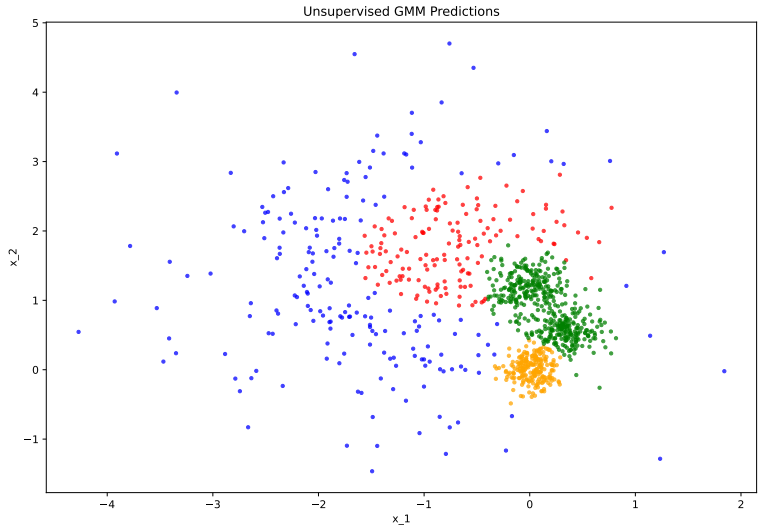

Figure 2: Unsupervised learning of 2nd trial

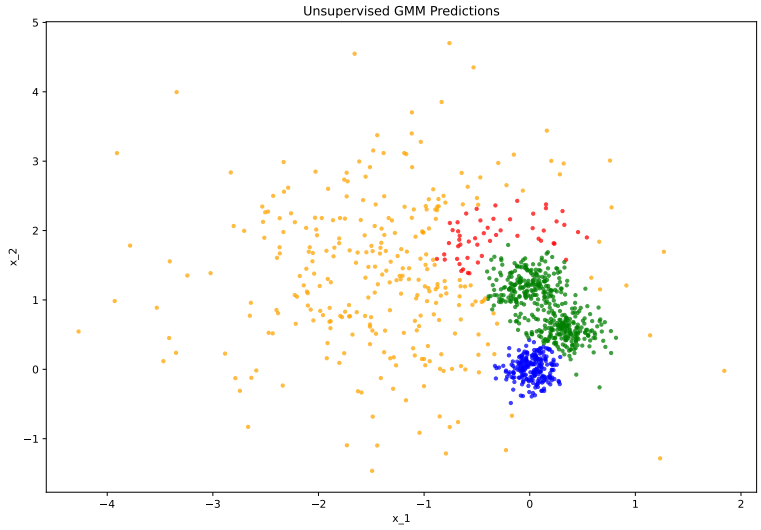

Figure 3: Unsupervised learning of 3rd trial

The code is omitted here.

(e) Three plots using classical semi-supervised EM implementation:

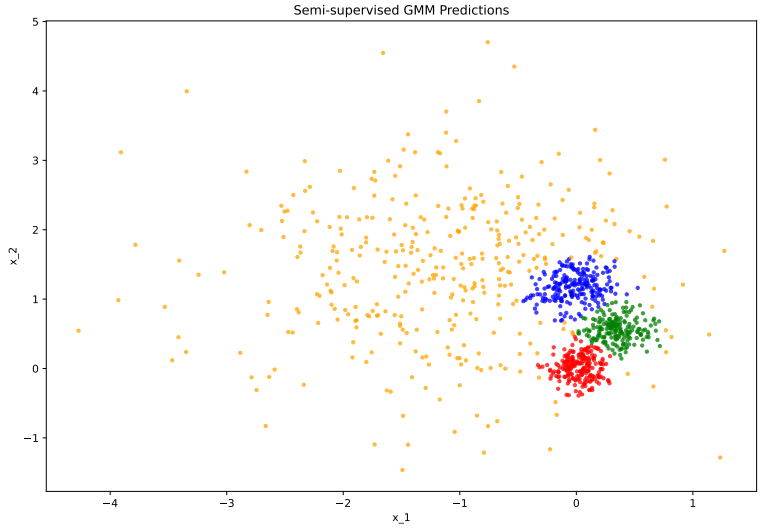

Figure 4: Semi-supervised learning of 1st trial

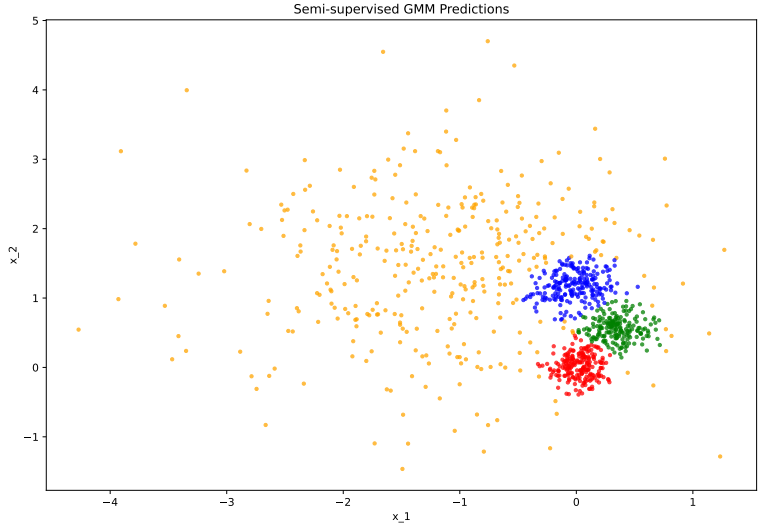

Figure 5: Semi-supervised learning of 2nd trial

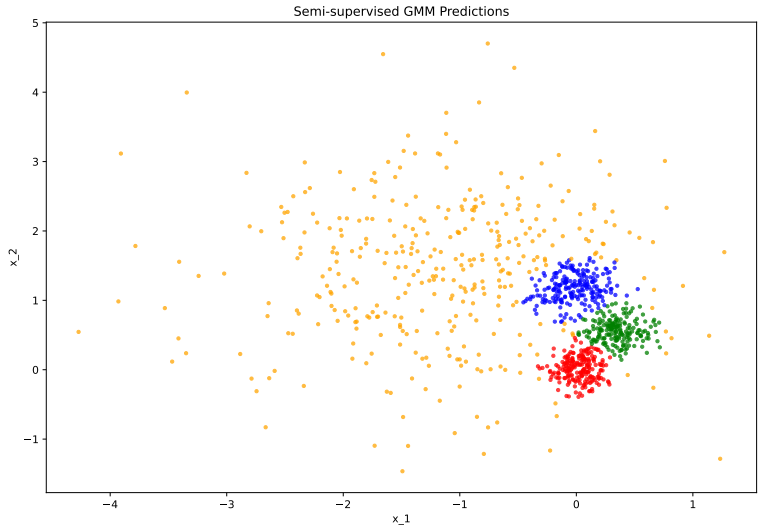

Figure 6: Semi-supervised learning of 3rd trial

The code is omitted here.
(f)

i. It takes more iterations for unsupervised EM algorithm to converge (respectively 117, 118, 101) than that for semi-supervised (respectively 19, 23, 19).

ii. Semi-supervised EM algorithm is more stable than unsupervised EM algorithm, since the assignment by unsupervised EM algorithm is random and by Semi-supervised EM algorithm is deterministic.

iii. The overall quality of assignments by semi-supervised EM algorithm are higher than unsupervised EM algorithm. There are three low-variance Gaussian distributed clusters samples and one high-variance in the plots generated by semi-supervised algorithm, the same as how they are sampled. But unsupervised EM algorithm is not as accurate as semi-supervised EM algorithm.

5 K-means for compression

(a) The following three figures are respectively original large-size image, original small-size image and compressed large image.

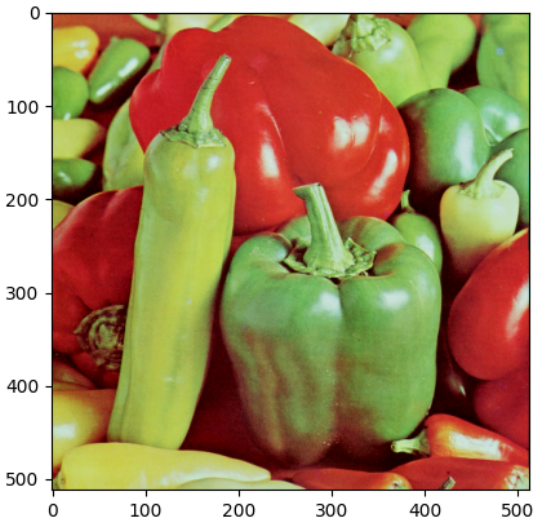

Figure 7: Original large-size image

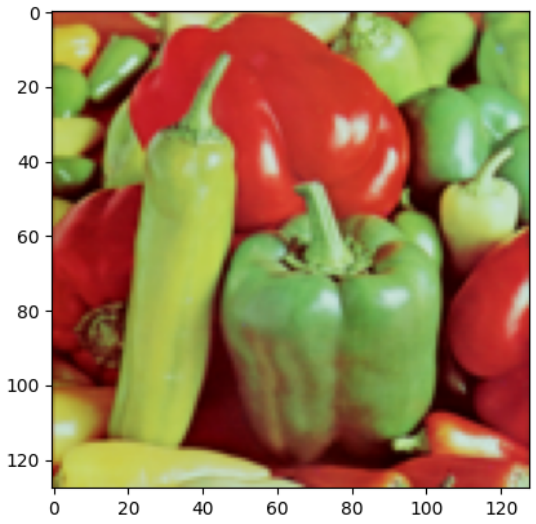

Figure 8: Original small-size image

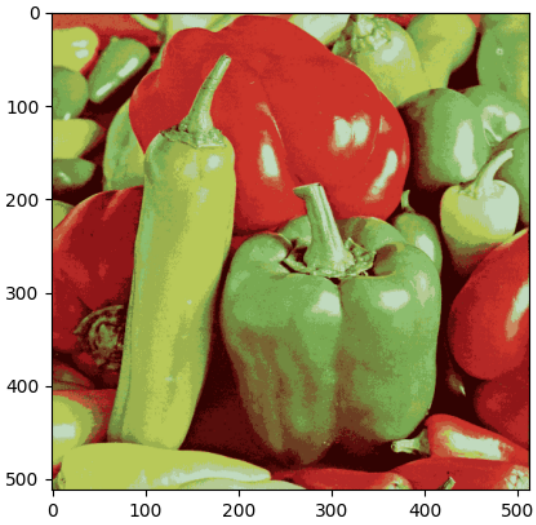

Figure 9: Large image compressed

The code is omitted here.
(b) As for the original large-size image, we use 8 bits/channel $\times$ 3 channels $= 24$ bits/pixel. After quantizing to $K = 16$ colors, we store $\log_2 16 = 4$ bits/pixel. So the compression factor should be $24/4 = 6$.