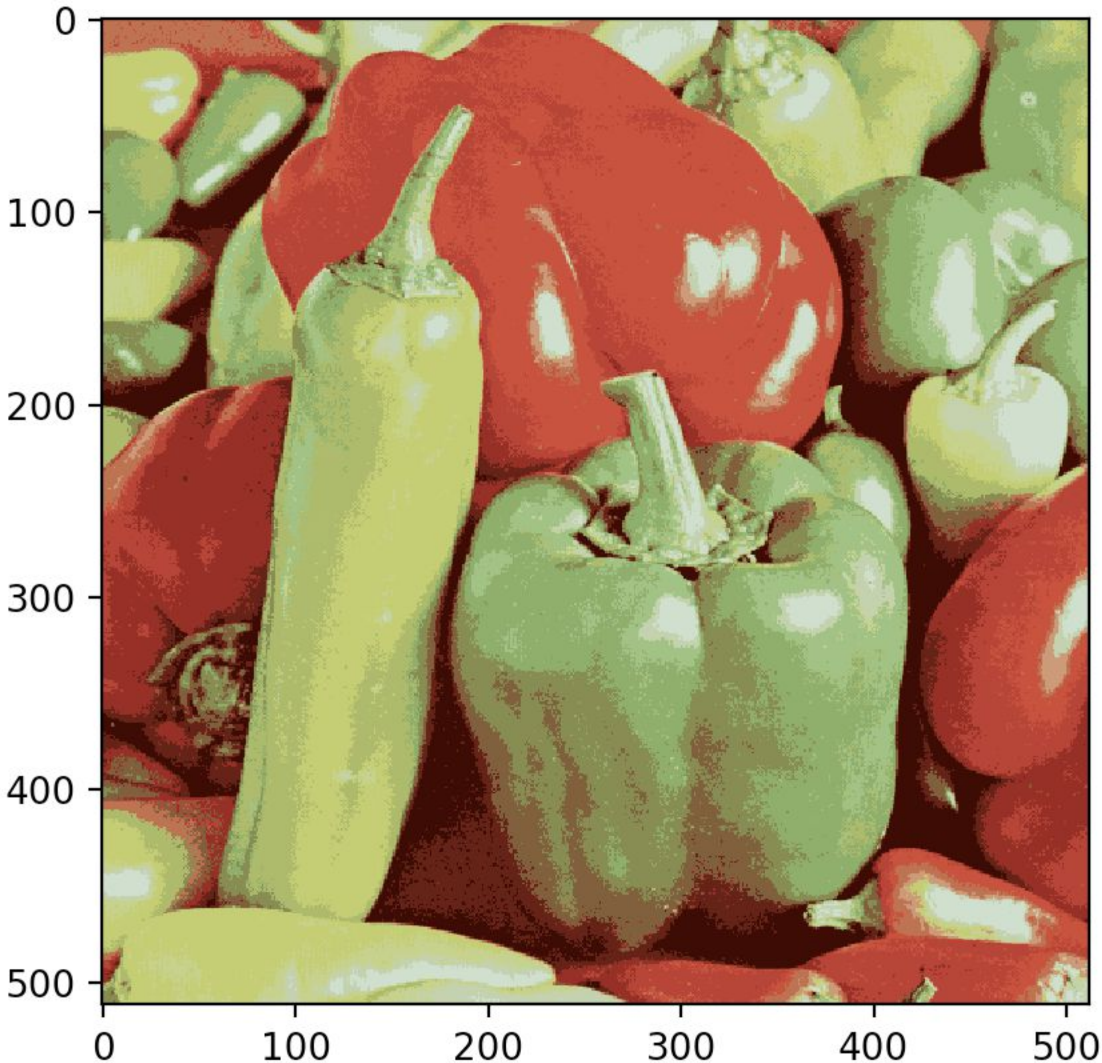
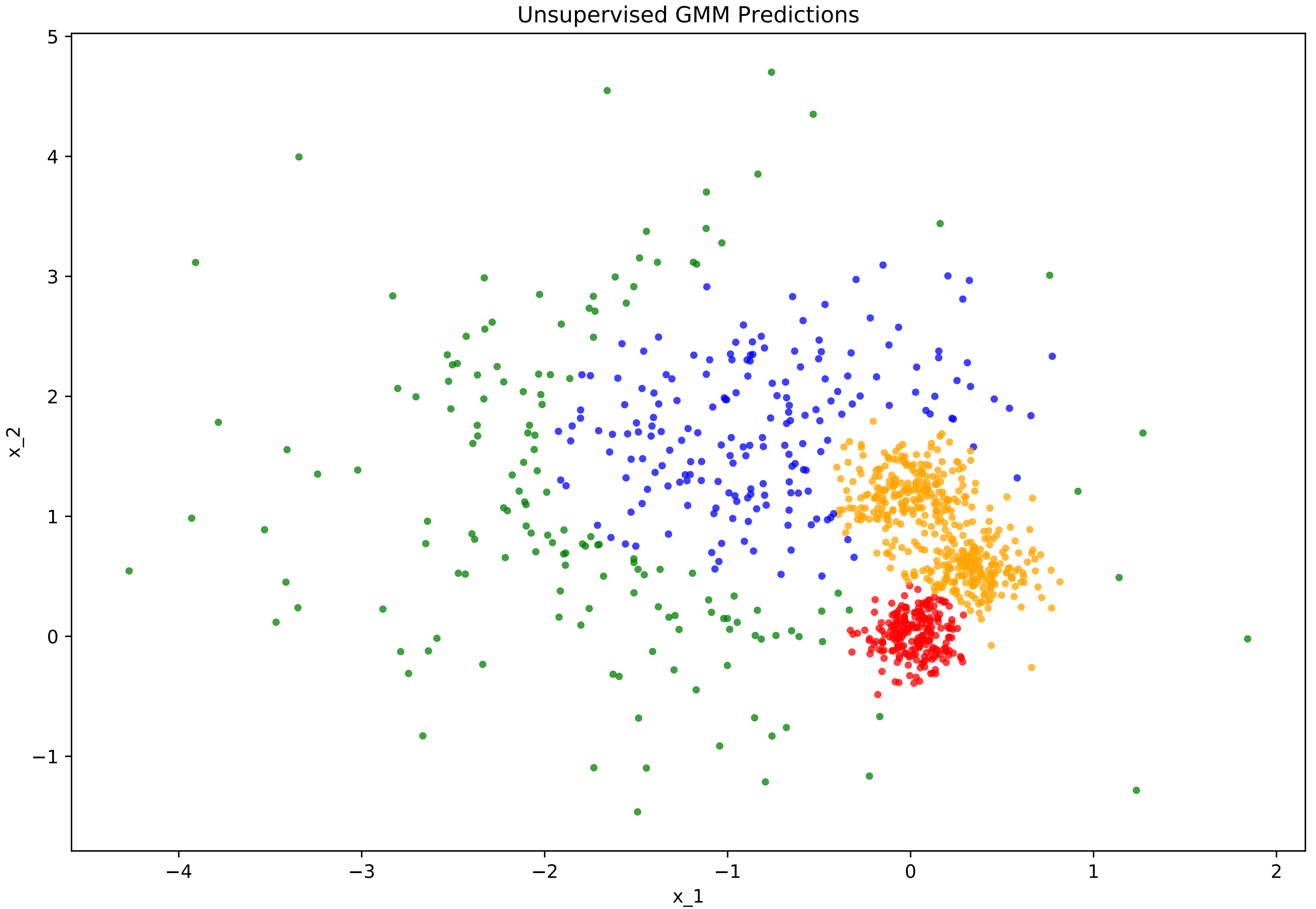
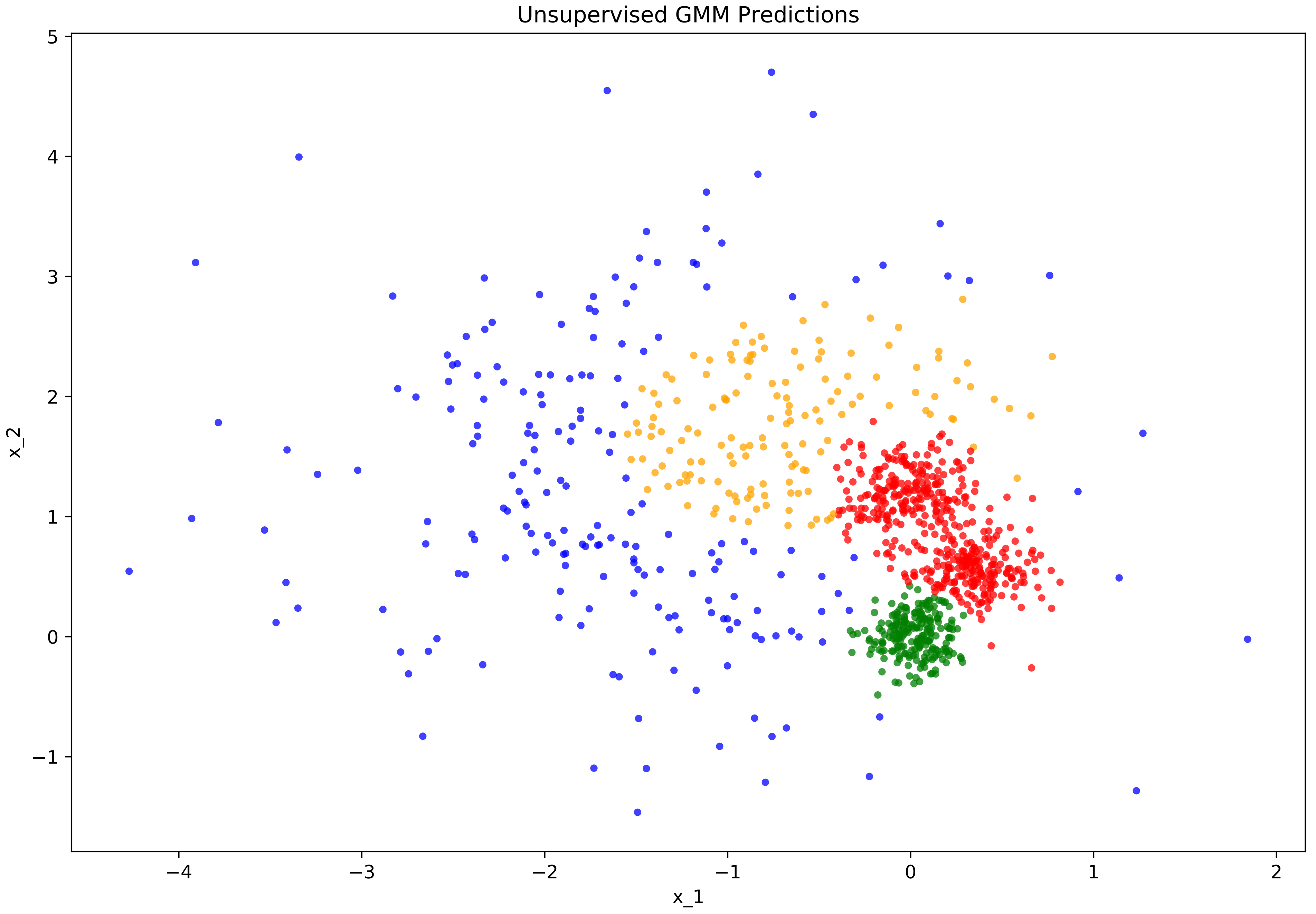

0
100
200
300
400
500
0
100
200
300
400
500

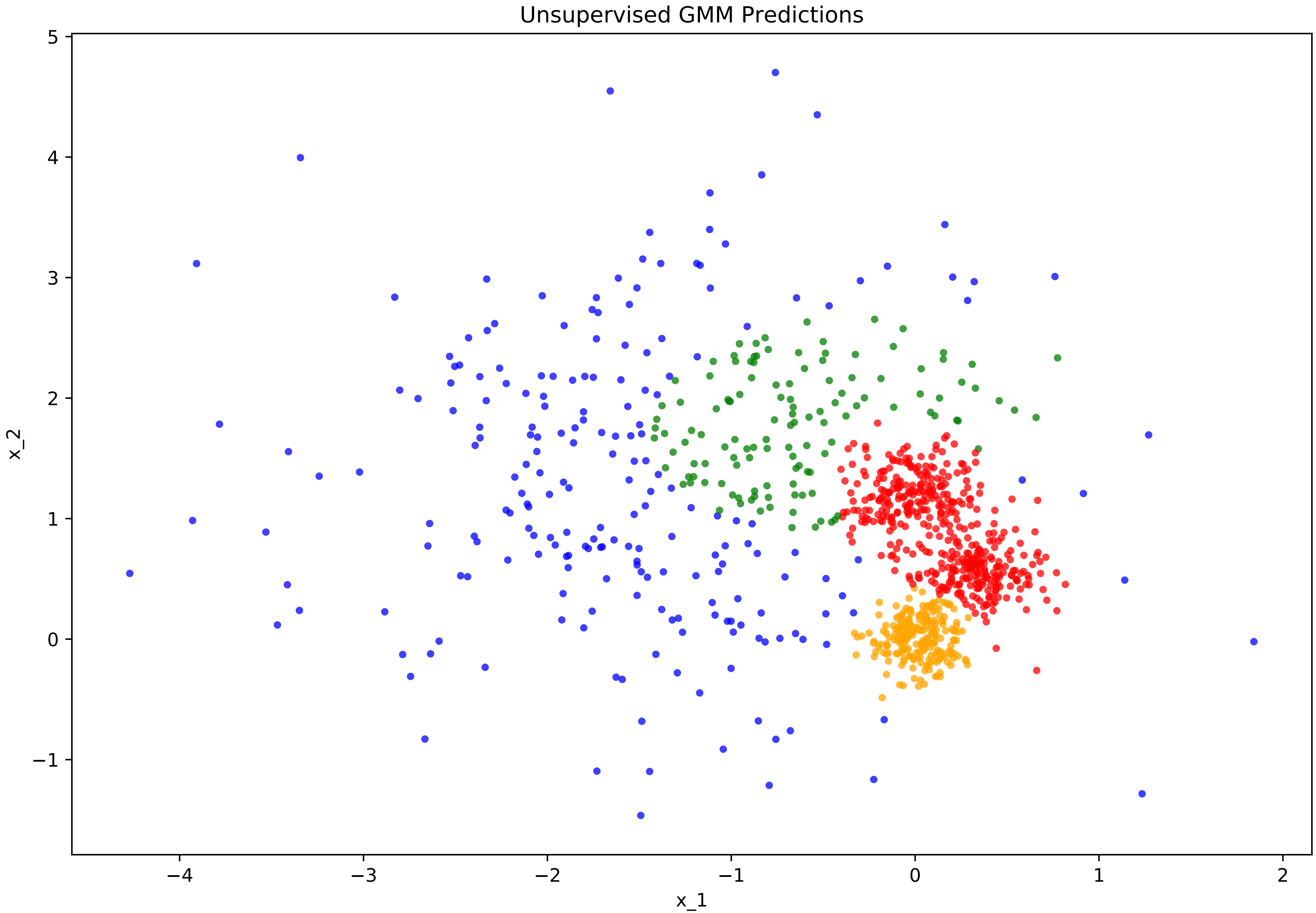
```python
import numpy as np
from matplotlib.image import imread
import matplotlib.pyplot as plt


def main():
    A = imread('peppers-large.tiff')
    plt.imshow(A)
    plt.show()

    im_small = imread('peppers-small.tiff')
    plt.imshow(im_small)
    plt.show()

    k = 16
    centroid = kmeans(im_small, k)

    # assign each example in the large image to the closest cluster usi
    dim = A.shape[0]
    A = np.reshape(A, (-1, 3))
    diffs = []
    for c in centroid:
        diff = np.linalg.norm(A - c, axis=1)
        diffs.append(diff)

    # Join the array "diff" along a new axis
    c_i = np.argmin(diffs, axis=0)

    # Compress the large image A
    compress_A = np.zeros((A.shape[0], A.shape[1]), dtype=int)
    for j in range(k):
        ind_j = np.where(c_i == j)
        compress_A[ind_j] = centroid[j]

    compress_A = compress_A.reshape(dim, dim, 3)
    plt.imshow((compress_A))
    plt.show()
```

```python
def kmeans(A, k):
    # initialize centroid by randomly picking k training examples,
    # and set the cluster centroids to be equal to the values of these k examples
    A = np.reshape(A, (-1, 3))
    m = A.shape[0]
    ind = np.random.choice(np.arange(m), size=k, replace=False)
    centroid = A[ind]

    iter = 0
    centroid = np.array(centroid)
    c_i = c_i_old = None
    while c_i_old is None or not np.array_equal(c_i, c_i_old):
        iter += 1
        c_i_old = c_i
        # Assigning each training example x_i to the closest cluster centroid miu_j
        diffs = []
        for c in centroid:
            diff = np.linalg.norm(A - c, axis=1)
            diffs.append(diff)

        c_i = np.argmin(diffs, axis=0)
        #print("c_i_old: ", c_i_old)
        #print("c_i: ", c_i)

        # Moving each cluster centroid miu_j to the mean of the points assigned to it
        miu_js = []
        for j in range(k):
            ind_j = np.where(c_i == j)
            miu_j = A[ind_j].mean(axis=0)
            miu_js.append(miu_j)

        centroid = np.array(miu_js)
        print("iteration: ", iter)
    return centroid


if __name__ == "__main__":
    main()
```

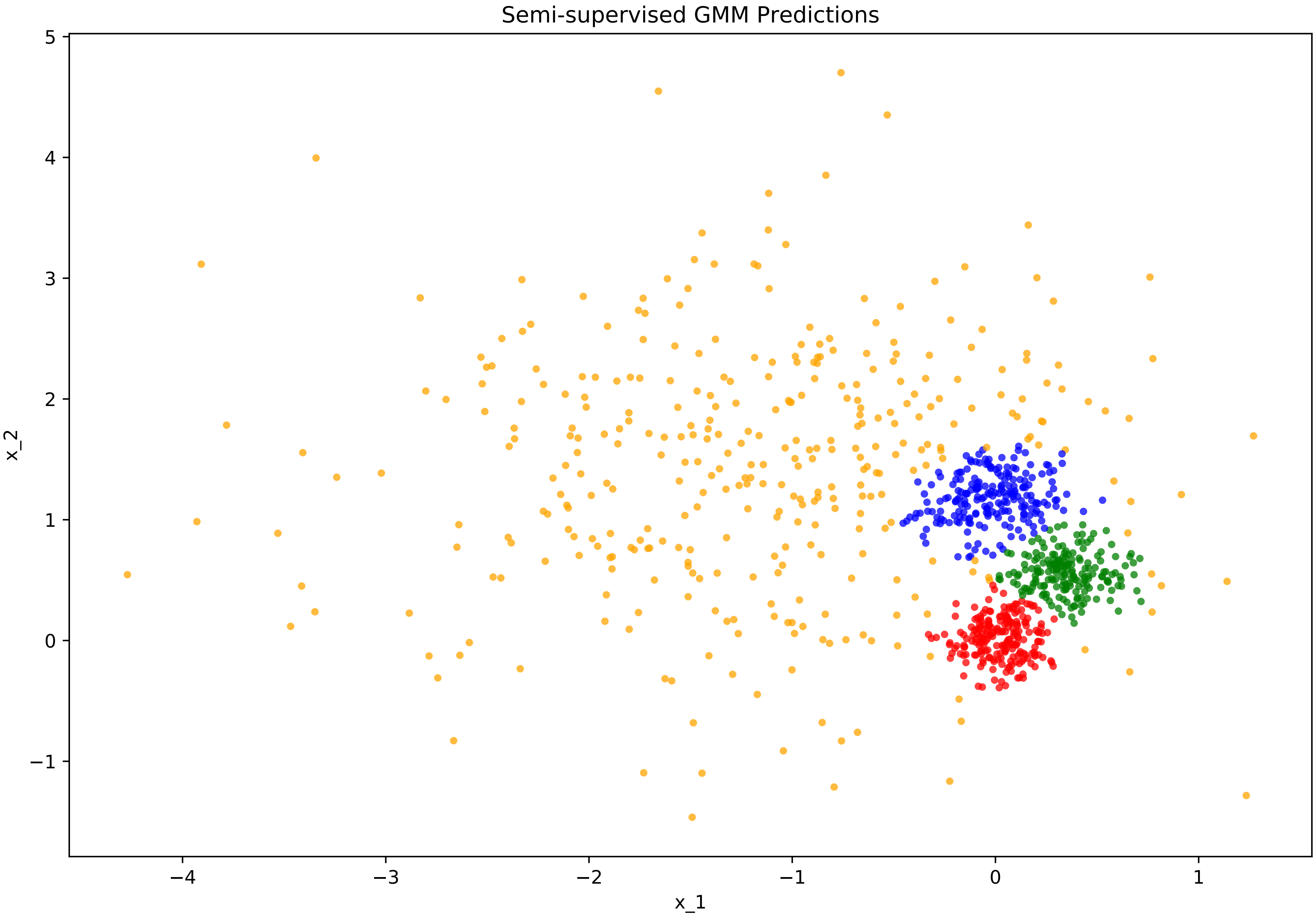

Unsupervised GMM Predictions
x_1
x_2
−4
−3
−2
−1
0
1
2
−1
0
1
2
3
4
5

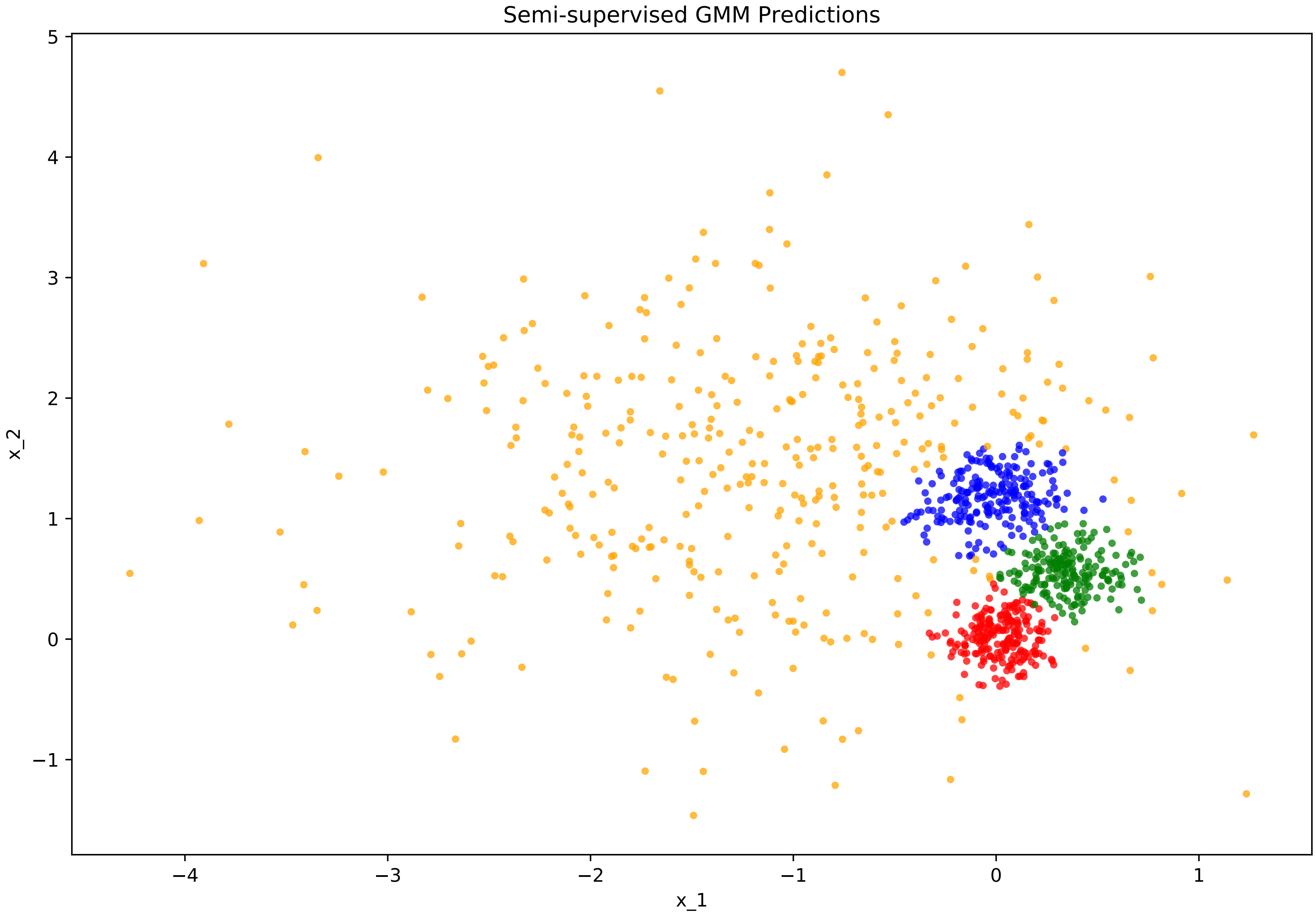


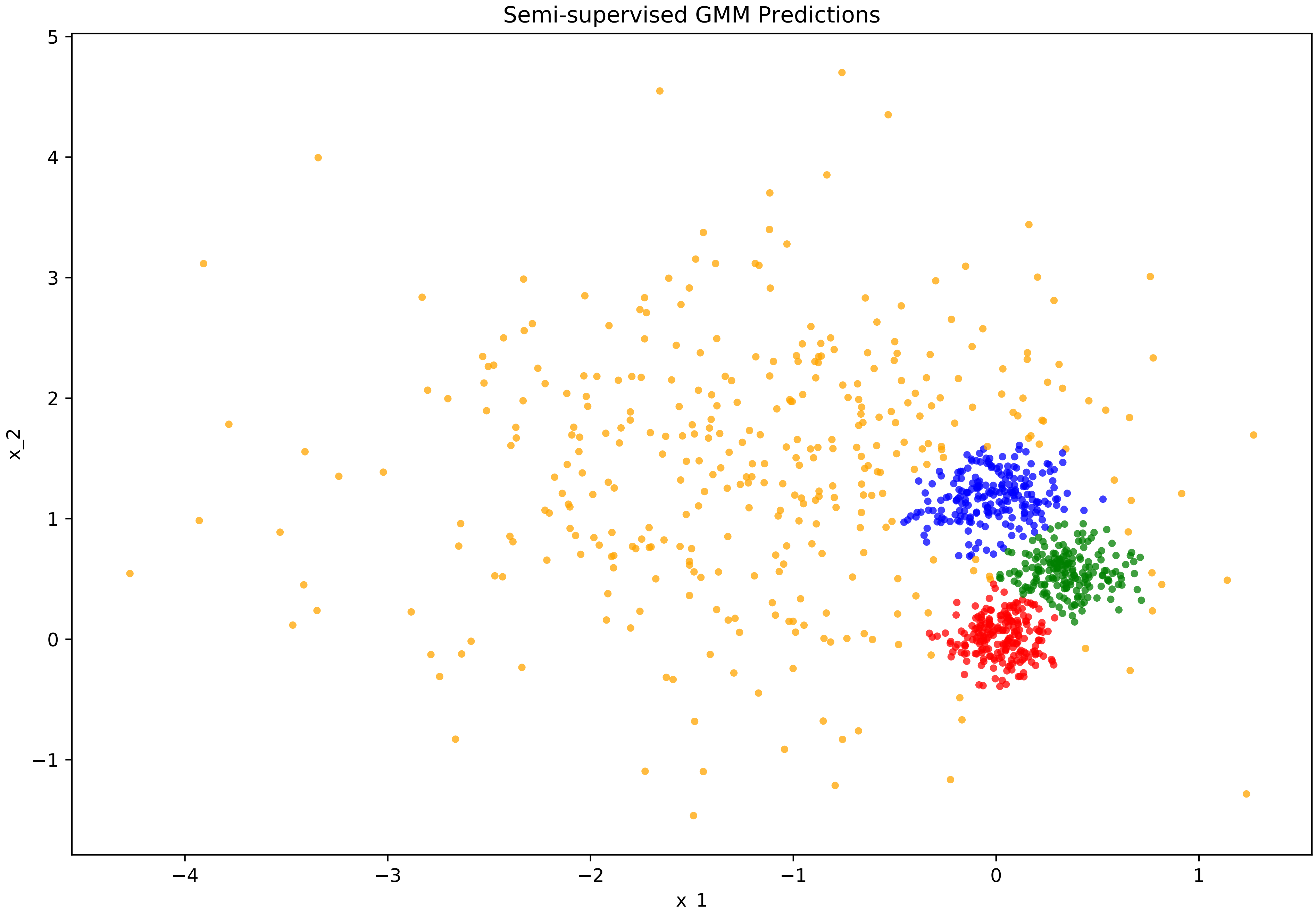
Unsupervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2



Unsupervised GMM Predictions
x_1
x_2
−4
−3
−2
−1
0
1
2
−1
0
1
2
3
4
5


Semi-supervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1


Semi-supervised GMM Predictions
x_1
x_2
−4
−3
−2
−1
0
1
5
4
3
2
1
0
−1

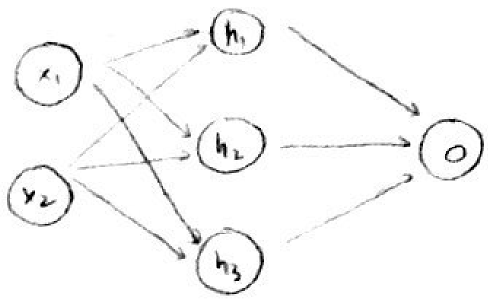

Semi-supervised GMM Predictions
x_1
x_2
−4
−3
−2
−1
0
1
−1
0
1
2
3
4
5

CS 229 HW3

Mengxuan Jin

Problem 1

m : sample size. $n = 2$ feature



$W_{1,1}^{[1]}$, $W_{2,1}^{[1]}$ $\quad W_{0,1}^{[1]}$

$W_{1,2}^{[1]}$, $W_{2,2}^{[1]}$ $\quad W_{0,2}^{[1]}$

$W_{1,3}^{[1]}$, $W_{2,3}^{[1]}$ $\quad W_{0,3}^{[1]}$

$W_1^{[2]}$ $W_2^{[2]}$ $W_3^{[2]}$ $\quad W_0^{[2]}$

(a)

$$z^{[1]} = \begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \\ z_3^{[1]} \end{bmatrix} = \begin{bmatrix} W_{1,1}^{[1]} & W_{2,1}^{[1]} \\ W_{1,2}^{[1]} & W_{2,2}^{[1]} \\ W_{1,3}^{[1]} & W_{2,3}^{[1]} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} W_{0,1}^{[1]} \\ W_{0,2}^{[1]} \\ W_{0,3}^{[1]} \end{bmatrix}, \quad \sigma(z^{[1]}) = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}$$

$$z^{[2]} = \begin{bmatrix} W_1^{[2]} & W_2^{[2]} & W_3^{[2]} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} + W_0^{[2]}, \quad \sigma(z^{[2]}) = \text{Output}$$

then $z_2^{[1](i)} = W_{1,2}^{[1]} x_1^{(i)} + W_{2,2}^{[1]} x_2^{(i)} + W_{0,2}^{[1]}$

$h_2^{(i)} = \sigma(z_2^{[1](i)})$

$z^{[2](i)} = W_1^{[2]} h_1^{(i)} + W_2^{[2]} h_2^{(i)} + W_3^{[2]} h_3^{(i)} + W_0^{[2]}$

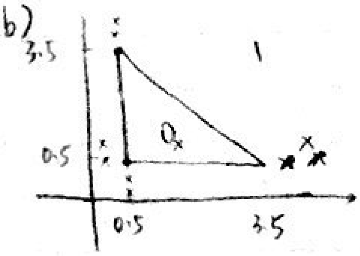
$\text{Output}^{(i)} = \sigma(z^{[2](i)})$

$$\frac{\partial l}{\partial W_{1,2}^{[1]}} = \frac{\partial l}{\partial o^{(i)}} \cdot \frac{\partial o^{(i)}}{\partial z^{[2](i)}} \cdot \frac{\partial z^{[2](i)}}{\partial h_2^{(i)}} \cdot \frac{\partial h_2^{(i)}}{\partial z_2^{[1](i)}} \cdot \frac{\partial z_2^{[1]}}{\partial W_{1,2}^{[1]}}$$

$$= \left(\frac{1}{m}\sum_{i=1}^{m} 2o^{(i)}\right) \cdot o^{(i)}(1-o^{(i)}) \cdot W_2^{[2]} \cdot h_2^{(i)}(1-h_2^{(i)}) \cdot x_1^{(i)}$$

$W_{1,2}^{[1]} = W_{1,2}^{[1]} - \alpha \frac{\partial l}{\partial W_{1,2}^{[1]}}$, where $\frac{\partial l}{\partial W_{1,2}^{[1]}}$ is given above

(b)



$$f(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases}, \quad z = w^T x + b$$

$$z^{[1]} = \begin{bmatrix} W_{0,1}^{[1]} & W_{1,1}^{[1]} & W_{2,1}^{[1]} \\ W_{0,2}^{[1]} & W_{1,2}^{[1]} & W_{2,2}^{[1]} \\ W_{0,3}^{[1]} & W_{1,3}^{[1]} & W_{2,3}^{[1]} \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

we can construct this matrix s.t. $f(z^{[1]}) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ for $x^{(i)}$ within triangle.

(0.5, 0.5) (3.5, 0.5) (0.5 3.5) are critical points

$$\begin{bmatrix} -1 & 2 & 0 \\ -1 & 0 & 2 \\ 4 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

$4 = 3.5 + 0.5$; $-1, -1$: slope

$2 = \frac{1}{0.5}$

then for the second matrix we can write as

$$z^{[2]} = [W_0^{[2]} \; W_1^{[2]} \; W_2^{[2]} \; W_3^{[2]}] \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

, want $f(z^{[2]})$ be negative for each entry, but nonnega if any of the entry in last 3 is -1.

choose $[W_0^{[2]} \; W_1^{[2]} \; W_2^{[2]} \; W_3^{[2]}] = [2.5 \; -1 \; -1 \; -1]$ satisfy the conditions desired.

(c) It's not possible Because data set B not linearly separable. if using $f(x) = x$ on h_1, h_2, h_3, it's just linear regression on $x^{(i)}$ and can't seperate the two classes. Then no matter what activation fn used in O, we can't achieve 100% accuracy

Problem 2

(a) $D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{Q(x)}{P(x)}\right)^{-1} = -\sum_{x\in\mathcal{X}} P(x)\log\left(\frac{Q(x)}{P(x)}\right)$

By def of expected value : $= -E\left(\log\frac{Q(x)}{P(x)}\right) = E\left(-\log\frac{Q(x)}{P(x)}\right)$

$-\log$ function is strictly convex : $\geq -\log E\left(\frac{Q(x)}{P(x)}\right) = -\log\sum_{x\in\mathcal{X}} P(x)\frac{Q(x)}{P(x)} = -\log\sum_{x\in\mathcal{X}} Q(x) = -\log 1 = 0$

$\therefore D_{KL}(P\|Q) \geq 0$

~~$\frac{Q(x)}{P(x)} = \log E\left(\frac{Q(x)}{P(x)}\right)$ $\frac{Q(x)}{P(x)}$ is constant by Jensen's inequality~~

- if $D_{KL}(P\|Q)=0$, then the equality holds above, which means

$$\frac{Q(x)}{P(x)} = E\left(\frac{Q(x)}{P(x)}\right) = \sum_x P(x)\frac{Q(x)}{P(x)} = \sum_x Q(x) = 1 \Rightarrow Q(x) = P(x)$$

- If $P=Q$, then $D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\cdot\underbrace{\log 1}_{0} = 0$

(b) $D_{KL}(P(X|Y)\|Q(X|Y)) = \sum_y P(y)\left(\sum_x P(x|y)\log\frac{P(x|y)}{Q(x|y)}\right)$

$D_{KL}(P(x)\|Q(x)) = \underbrace{\sum_x P(x)\log\frac{P(x)}{Q(x)}}_{\text{Eq}①}$

$D_{KL}(P(Y|X)\|Q(Y|X)) = \underbrace{\sum_x P(x)\left(\sum_y P(y|x)\log\frac{P(y|x)}{Q(y|x)}\right)}_{\text{Eq}②}$

$D_{KL}(P(X,Y)\|Q(X,Y)) = \sum_{x,y} P(x,y)\log\frac{P(x,y)}{Q(x,y)} = \sum_{x,y} P(x,y)\log\frac{P(x)P(y|x)}{Q(x)Q(y|x)}$

$= \sum_{x,y} P(x,y)\log\frac{P(x)}{Q(x)} + \sum_{x,y} P(x,y)\log\frac{P(y|x)}{Q(y|x)} = \sum_{x,y} P(x)P(y|x)\log\frac{P(x)}{Q(x)} + \sum_{x,y} P(x)P(y|x)\log\frac{P(y|x)}{Q(y|x)}$

$= \sum_x P(x)\log\frac{P(x)}{Q(x)} + \sum_x P(x)\sum_y P(y|x)\log\frac{P(y|x)}{Q(y|x)}$

$= \text{Eq}① + \text{Eq}②$

$\therefore D_{KL}(P(X,Y)\|Q(X,Y)) = D_{KL}(P(X)\|Q(X)) + D_{KL}(P(Y|X)\|Q(Y|X))$

(c) $D_{KL}(\hat{P} \| P_\theta) = \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)} = \sum_x \frac{1}{m} \sum_{i=1}^m 1\{x^{(i)} = x\} \log \frac{\frac{1}{m}\sum_{i=1}^m 1\{x^{(i)}=x\}}{P_\theta(x)}$

$= \frac{1}{m}\sum_{i=1}^m \sum_x 1\{x^{(i)}=x\} \log \frac{\frac{1}{m}\sum_{i=1}^m 1}{P_\theta(x^{(i)})} = \frac{1}{m}\sum_{i=1}^m \log \frac{1}{P_\theta(x^{(i)})} = -\frac{1}{m}\sum_{i=1}^m \log P_\theta(x^{(i)})$

so $\arg\min_\theta D_{KL}(\hat{P}\|P_\theta) = \arg\min_\theta -\frac{1}{m}\sum_{i=1}^m \log P_\theta(x^{(i)})$

$= \arg\max_\theta \frac{1}{m}\sum_{i=1}^m \log P_\theta(x^{(i)})$

$= \arg\max_\theta \sum_{i=1}^m \log P_\theta(x^{(i)})$

Problem 3

(a) $E_{y\sim p(y;\theta)}\left[\nabla_{\theta'} \log p(y;\theta')\big|_{\theta'=\theta}\right]$

$= \int_{-\infty}^{\infty} p(y;\theta)\, \nabla_\theta \log p(y;\theta)\, dy$

$= \int_{-\infty}^{\infty} \cancel{p(y;\theta)} \frac{1}{\cancel{p(y;\theta)}} \nabla_\theta p(y;\theta)\, dy$

$= \nabla_\theta \int_{-\infty}^{\infty} p(y;\theta)\, dy = \nabla_\theta 1 = 0$

) By definition of Covariance and covariance matrix
then $Cov_{y\sim p(y;\theta)}$ is a covariance matrix

~~cov(X,X) = Var(X) = E(X^2) - E(X)^2~~

$$Cov_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}\right] = E\left(\nabla_{\theta'}\log p(y;\theta')\,\nabla_{\theta'}\log p(y;\theta')^T|_{\theta'=\theta}\right) - \underbrace{E\left(\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}\right)}_{=0 \text{ by part (a)}}\underbrace{E\left(\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}\right)^T}_{0}$$

$$\therefore Cov_{y\sim p(y;\theta)}\left[\nabla_{\theta'}\log p(y;\theta')|_{\theta'=\theta}\right] = E\left(\nabla_{\theta'}\log p(y;\theta')\,\nabla_{\theta'}\log p(y;\theta')^T|_{\theta'=\theta}\right)$$

c) $\nabla^2_{\theta'}\log p(y;\theta')|_{\theta'=\theta} = H_{\log p(y;\theta)} = J\left(\frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\right)$, where J is Jacobian operator

$$= -p(y;\theta)^{-2}\cdot\nabla_\theta p(y;\theta)\,\nabla_\theta p(y;\theta)^T - \frac{\nabla^2_\theta p(y;\theta)}{p(y;\theta)}$$

$$= \frac{\nabla^2_\theta p(y;\theta)}{p(y;\theta)} - \frac{\nabla_\theta p(y;\theta)\,\nabla_\theta p(y;\theta)^T}{p(y;\theta)\,p(y;\theta)}$$

$$= \frac{H_{p(y;\theta)}}{p(y;\theta)} - \frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\left(\frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\right)^T$$

then $E(\nabla^2_\theta\log p(y;\theta)) = E\left[\frac{H_{p(y;\theta)}}{p(y;\theta)} - \frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\left(\frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\right)^T\right]$

$$= E\left[\frac{H_{p(y;\theta)}}{p(y;\theta)}\right] - E\left[\frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\left(\frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\right)^T\right]$$

$$= \int \cancel{p(y;\theta)}\frac{H_{p(y;\theta)}}{\cancel{p(y;\theta)}}dy - E\underbrace{\left[\nabla_\theta\log p(y;\theta)\;\nabla_\theta\log p(y;\theta)^T\right]}_{I(\theta)}$$

$$= \int H_{p(y;\theta)}\,dy - I(\theta)$$

$$= H_{\int p(y;\theta)dy} - I(\theta) = 0 - I(\theta)$$

$$\therefore E(\nabla^2_\theta\log p(y;\theta)) = -I(\theta) \Rightarrow E(-\nabla^2_\theta\log p(y;\theta)) = I(\theta)$$

(d) $D_{KL}(P_\theta \| P_{\theta+d}) = \int_{-\infty}^{\infty} P_\theta(x) \log \frac{P_\theta(x)}{P_{\theta+d}(x)} dx = -\int_{-\infty}^{\infty} P_\theta(x) \log \frac{P_{\theta+d}(x)}{P_\theta(x)} dx$

$$\approx -\int_{-\infty}^{\infty} P_\theta(x) \left(\cancel{\log \frac{P_\theta(x)}{P_\theta(x)}}^{\to 0} + d^T \nabla_\theta \log P_\theta(x) + \frac{1}{2} d^T \nabla_\theta^2 \log P_\theta(x)\, d \right) dx$$

$$= -\int_{-\infty}^{\infty} P_\theta(x) \left(d^T \nabla_\theta \log P_\theta(x) + \frac{1}{2} d^T \nabla_\theta^2 \log P_\theta(x)\, d \right) dx$$

$$= -\int_{-\infty}^{\infty} P_\theta(x) \left(\frac{\nabla_\theta P_\theta(x)}{P_\theta(x)} \right)^T d\, dx - \frac{1}{2} d^T \int_{-\infty}^{\infty} P_\theta(x) \frac{P_\theta(x) \nabla_\theta^2 P_\theta(x) - (\nabla_\theta P_\theta(x))(\nabla_\theta P_\theta(x))^T}{P_\theta(x)^2} d\, dx$$

$$= -\int_{-\infty}^{\infty} \nabla_\theta P_\theta(x)^T d\, dx - \frac{1}{2} d^T \int_{-\infty}^{\infty} \nabla_\theta^2 P_\theta(x)\, d\, dx + \frac{1}{2} d^T \int_{-\infty}^{\infty} P_\theta(x) \left(\frac{\nabla_\theta P_\theta(x)}{P_\theta(x)} \right) \left(\frac{\nabla_\theta P_\theta(x)}{P_\theta(x)} \right)^T d\, dx$$

$$= -\left(\frac{d}{d\theta} \underbrace{\int_{-\infty}^{\infty} P_\theta(x)\, dx}_{1} \right)^T d - \frac{1}{2} d^T \left(\nabla_\theta^2 \underbrace{\int_{-\infty}^{\infty} P_\theta(x)\, dx}_{1} \right) d + \frac{1}{2} d^T \left(\int_{-\infty}^{\infty} P_\theta(x) (\nabla_\theta \log P_\theta(x)) (\nabla_\theta \log P_\theta(x))^T dx \right) d$$

$$= -0 - 0 + \frac{1}{2} d^T \left(\underbrace{\int_{-\infty}^{\infty} P_\theta(x) (\nabla_\theta \log P_\theta(x)) (\nabla_\theta \log P_\theta(x))^T dx}_{E(\nabla_\theta \log P_\theta(x)\, \nabla_\theta \log P_\theta(x)^T) = I(\theta)} \right) d$$

$$= \frac{1}{2} d^T I(\theta)\, d$$

(e) $d^* = \arg\max_d l(\theta + d) \quad s.t. \quad D_{KL}(P_\theta \| P_{\theta+d}) = c$

$$\Rightarrow \quad \mathcal{L}(d, \lambda) = l(\theta + d) - \lambda (D_{KL}(P_\theta \| P_{\theta+d}) - c)$$

$$\approx l(\theta) + d^T \nabla_{\theta'} l(\theta')|_{\theta'=\theta} - \frac{1}{2} \lambda d^T I(\theta)\, d + \lambda c$$

$$\nabla_d \mathcal{L}(d, \lambda) = \nabla_{\theta'} l(\theta')|_{\theta'=\theta} - \lambda I(\theta)\, d = 0$$

$$\lambda I(\theta)\, d = \nabla_{\theta'} l(\theta')|_{\theta'=\theta}$$

$$d = \frac{1}{\lambda} (I(\theta))^{-1} \nabla_\theta l(\theta)$$

Problem 4

(a) $l_{sup}(\theta) = \sum_{i=1}^{\tilde{m}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta)$

$$l_{unsup}(\theta) = \sum_{i=1}^{m} \log P(x^{(i)}; \theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta)$$

let $\mathcal{L}$ be $\sum_{i=1}^{m}\left(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i^{(t)}(z^{(i)})}\right) + \alpha\left(\sum_{i=1}^{\tilde{m}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta)\right)$

then $\mathcal{L}(\theta^{(t+1)}) \geq \underbrace{\sum_{i=1}^{m}\left(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})}\right)}_{\text{part ①}} + \alpha \underbrace{\sum_{i=1}^{\tilde{m}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)})}_{\text{part ②}}$

because $\theta^{(t+1)}$ maximize $\mathcal{L}$, then $\mathcal{L}(\theta^{(t+1)}) \geq \mathcal{L}(\theta)\ \forall \theta$, including $\theta^{(t)}$

then part ① is just $\sum_{i=1}^{m} E_{z^{(i)} \sim Q_i^{(t)}} \log \frac{P(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})} \geq \sum_{i=1}^{m} \log E_{z^{(i)} \sim Q_i^{(t)}} \frac{P(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})}$

By Jesen's Inequality

$$= \sum_{i=1}^{m} \log \sum_{z^{(i)}} \cancel{Q_i^{(t)}(z^{(i)})} \frac{P(x^{(i)}, z^{(i)}; \theta^{(t)})}{\cancel{Q_i^{(t)}(z^{(i)})}}$$

$$= \sum_{i=1}^{m} \log P(x^{(i)}; \theta^{(t)}) = l_{unsup}(\theta^{(t)})$$

and part ② is $l_{sup}(\theta^{(t)})$, so $\mathcal{L}(\theta^{(t+1)}) \geq l_{sup}(\theta^{(t)}) + \alpha\, l_{unsup}(\theta^{(t)}) = \mathcal{L}(\theta^{(t)})$

(i.e. $l_{semi\text{-}sup}(\theta^{(t+1)}) \geq l_{semi\text{-}sup}(\theta^{(t)})$)

(b) μ, Σ, ϕ

$$w_j^{(i)} = P(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{\sum_{l=1}^{k} p(x^{(i)} \mid z^{(i)} = l) p(z^{(i)} = l)}$$

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

only $z^{(i)}$'s are the latent variables that need to be re-estimated (unlabelled)

(c) $$\ell = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j)\right) \cdot \phi_j}{w_j^{(i)}} +$$

$$\alpha \sum_{i=1}^{\tilde{m}} \sum_{j=1}^{k} \log\left(\frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(\tilde{x}^{(i)}-\mu_j)^T \Sigma_j^{-1} (\tilde{x}^{(i)}-\mu_j)\right) \cdot \mathbb{I}\{\tilde{z}^{(i)} = j\} \cdot \phi_j\right)$$

$$\nabla_{\mu_l} \ell = \underbrace{\sum_{i=1}^{m} w_l^{(i)} \left(\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l\right)}_{\text{from lecture note}} + \nabla_{\mu_l} \alpha \sum_{i=1}^{\tilde{m}} -\frac{1}{2} (\tilde{x}^{(i)} - \mu_l)^T \Sigma_l^{-1} (\tilde{x}^{(i)} - \mu_l) \cdot \mathbb{I}\{\tilde{z}^{(i)} = l\}$$

$$\Downarrow$$

$$\frac{1}{2}\alpha \sum_{i=1}^{\tilde{m}} \nabla_{\mu_l} \left(2\mu_l^T \Sigma_l^{-1} \tilde{x}^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l\right) \cdot \mathbb{I}\{\tilde{z}^{(i)} = l\}$$

$$= \alpha \sum_{i=1}^{\tilde{m}} \left(\Sigma_l^{-1} \tilde{x}^{(i)} - \Sigma_l^{-1} \mu_l\right) \cdot \mathbb{I}\{\tilde{z}^{(i)} = l\}$$

then $\nabla_{\mu_l} \ell = 0 \Rightarrow \sum_{i=1}^{m} w_l^{(i)} \left(\cancel{\Sigma_l^{-1}} x^{(i)} - \cancel{\Sigma_l^{-1}} \mu_l\right) + \alpha \sum_{i=1}^{\tilde{m}} \left(\cancel{\Sigma_l^{-1}} \tilde{x}^{(i)} - \cancel{\Sigma_l^{-1}} \mu_l\right) \cdot \mathbb{I}\{\tilde{z}^{(i)} = l\} = 0$

$$\mu_l := \frac{\sum_{i=1}^{m} w_l^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \mathbb{I}\{\tilde{z}^{(i)} = l\} \tilde{x}^{(i)}}{\left(\sum_{i=1}^{m} w_l^{(i)}\right) + \alpha \sum_{i=1}^{\tilde{m}} \mathbb{I}\{\tilde{z}^{(i)} = l\}}$$

$$\nabla_{\Sigma_l} \mathcal{I} = -\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)} \left(\Sigma_l^{-1} - \Sigma_l^{-1}(x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T \Sigma_l^{-1} \right) +$$

$$\underbrace{\nabla_{\Sigma_l} \alpha \sum_{i=1}^{\tilde{m}} \left(\log \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} - \frac{1}{2}(\tilde{x}^{(i)}-\mu_l)^T \Sigma_j^{-1} (\tilde{x}^{(i)}-\mu_l) \right) \cdot \mathbb{I}\{\tilde{z}^{(i)}=l\}}_{-\frac{1}{2}\alpha \sum_{i=1}^{\tilde{m}} \left(\Sigma_l^{-1} - \Sigma_l^{-1}(\tilde{x}^{(i)}-\mu_l)(\tilde{x}^{(i)}-\mu_l)^T \Sigma_l^{-1} \right) \cdot \mathbb{I}\{\tilde{z}^{(i)}=l\}} = 0.$$

$$\therefore \nabla_{\Sigma} \mathcal{I} = +\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)} \left(\Sigma_l - (x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T \right) + \frac{\alpha}{2}\sum_{i=1}^{\tilde{m}} \mathbb{I}\{\tilde{z}^{(i)}=l\} \left(\Sigma_l - (\tilde{x}^{(i)}-\mu_l)(\tilde{x}^{(i)}-\mu_l)^T \right) = 0$$

$$\Sigma_l := \frac{\sum_{i=1}^{m} w_l^{(i)} (x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T + \alpha \sum_{i=1}^{\tilde{m}} (\tilde{x}^{(i)}-\mu_l)(\tilde{x}^{(i)}-\mu_l)^T \cdot \mathbb{I}\{\tilde{z}^{(i)}=l\}}{\sum_{i=1}^{m} w_l^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \mathbb{I}\{\tilde{z}^{(i)}=l\}}$$

For ϕ_l: $\qquad (\because \sum \phi_j = 1)$

maximize $\mathcal{I}' = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta_1\left(\sum_{j=1}^{k} \phi_j - 1\right) + \alpha \left(\sum_{i=1}^{\tilde{m}}\sum_{j=1}^{k} \mathbb{I}\{\tilde{z}^{(i)}=j\} \log \phi_j + \beta_2\left(\sum_{j=1}^{k}\phi_j - 1\right) \right)$

$$\nabla_{\phi_l} \mathcal{I}' = \sum_{i=1}^{m} w_l^{(i)}/\phi_l + \beta_1 + \alpha\left(\sum_{i=1}^{\tilde{m}} \mathbb{I}\{\tilde{z}^{(i)}=l\}/\phi_l + \beta_2 \right) = 0$$

$$\Rightarrow \phi_l = \frac{\sum_{i=1}^{m} w_l^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \mathbb{I}\{\tilde{z}^{(i)}=l\}}{-\beta_1 - \alpha\beta_2} \quad \text{and} \quad -\beta_1 - \alpha\beta_2 = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{m}}\sum_{j=1}^{k} w_j^{(i)}$$

$$= m + \alpha\tilde{m}$$

$$\therefore \phi_l = \frac{\sum_{i=1}^{m} w_l^{(i)} + \alpha \sum_{i=1}^{\tilde{m}} \mathbb{I}\{\tilde{z}^{(i)}=l\}}{m + \alpha\tilde{m}}$$

(f) [i] unsupervised EM takes much more iterations to converge (over 100 iterations) than SS EM (SS EM takes about 20-30 iterations)

[ii] Based on the plot, SS EM is more stable since 3 plots are almost the same, but unsupervised EM's plots differ each time (the assignment is different, in particular, the boundaries on the left 2 groups changes each time)

[iii] SS EM has better quality from plots, it clearly has 3 Gaussian dist. w/ low variance and 1 w/ high variance

Scanned with CamScanner

$\nabla_\lambda \mathcal{L}(d,\lambda) = -\frac{1}{2} d^T I(\theta) d + c$

with $d = \frac{1}{\lambda}(I(\theta))^{-1} \nabla_\theta l(\theta)$, $\nabla_\lambda \mathcal{L}(d,\lambda) = -\frac{1}{2}\left(\frac{1}{\lambda^2}(\nabla_\theta l(\theta))^T (I(\theta))^{-T} \cancel{(I(\theta))}\cancel{(I(\theta))}^{-1} \nabla_\theta l(\theta)\right) + c = 0$

$\Rightarrow 2c = \cancel{\lambda} \frac{1}{\lambda^2} (\nabla_\theta l(\theta))^T (I(\theta))^{-T} \nabla_\theta l(\theta)$

$\lambda = \left(\frac{1}{2c}(\nabla_\theta l(\theta))^T (I(\theta))^{-T} \nabla_\theta l(\theta)\right)^{1/2}$

then $d^* = \left(\frac{1}{2c} \nabla_\theta l(\theta))^T (I(\theta))^{-T} \nabla_\theta l(\theta)\right)^{-\frac{1}{2}} (I(\theta))^{-1} \nabla_\theta l(\theta)$

(f) Natural Gradient direction: $(I(\theta))^{-1} \nabla_\theta l(\theta)$

Newton's method direction: $-(H(\theta))^{-1} \nabla_\theta l(\theta)$

$I(\theta) = \underset{y \sim p(y;\theta)}{E}[-\nabla_\theta^2 \log P(y;\theta)]$

$H(\theta) = \frac{\partial^2}{\partial\theta^2} a(\theta^T x) x X^T$ (from HW1)

$= Var(Y|X,\theta) X X^T$

$\nabla_\theta^2 l(\theta)$, $l(\theta) = \log[p(y;\theta)]$

$(I(\theta))^{-1} \nabla l(\theta) = \left(\underset{y}{E}[-\nabla_\theta^2 l(\theta)]\right)^{-1} \nabla_\theta l(\theta)$

$= \left(-\underset{y}{E}[Var(y|x;\theta) x x^T]\right)^{-1} \nabla_\theta l(\theta)$ ← from PS1, question 4

$= -\left(\underset{y}{E}[Var(y|x;\theta)] x x^T\right)^{-1} \nabla_\theta l(\theta)$

$= -\left(Var(y|x;\theta) x x^T\right)^{-1} \nabla_\theta l(\theta)$

$= -\left(\nabla_\theta^2 l(\theta)\right)^{-1} \nabla_\theta l(\theta)$

$= -(H(\theta))^{-1} \nabla_\theta l(\theta)$

Problem 5

(b) Compression factor = $\frac{\text{previous \# of bytes to store image}}{\text{new \# of bytes to store image}}$

Previous: $512 \times 512 \times 3 \times 8 = 6{,}291{,}456$ bits $= 786{,}432$ bytes

Compressed. $512 \times 512 \times 4 = 1{,}048{,}576$ bits $= 131{,}072$ bytes

~~the number "8" comes from $2^8 = 256$ since we have 256 colors, then each byte has 8 bits~~

~~the number "4" comes from $2^4 = 16$ since we now have 16 colors, then each byte has 4 bits~~

The original image has 24 bits/pixel, but now for 16 colors, we only need $\log_2 16 = 4$ bits per pixel

So compression rate $= \frac{786{,}432}{131{,}072} = \frac{24}{4} = 6$.

Scanned with CamScanner