CS446 hw4

Junsheng Huang

March 2024

1 PCA

1.1

$X = \begin{bmatrix} 1 & 4 \\ 3 & 7 \end{bmatrix}$, $\mu = \begin{bmatrix} \frac{5}{2} \\ 5 \end{bmatrix}$, $\bar{X} = \begin{bmatrix} -\frac{3}{2} & \frac{3}{2} \\ -2 & 2 \end{bmatrix}$, so we have:

$$\Sigma = \frac{1}{2}\bar{X}\bar{X}^T = \begin{bmatrix} \frac{9}{4} & 3 \\ 3 & 4 \end{bmatrix}$$

Solving the eigenvalue of Σ, we have $\lambda_1 = \frac{25}{4}$ with $w = [\frac{3}{5}, \frac{4}{5}]^T$

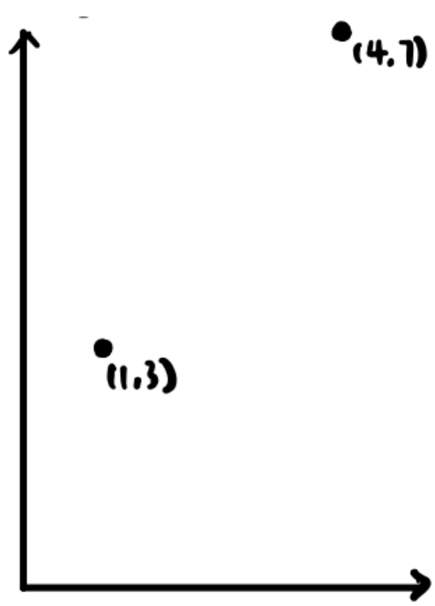

Figure 1: Q1.1

1.2

$X = \begin{bmatrix} 2 & 2 & 6 & 6 \\ 0 & 2 & 0 & 2 \end{bmatrix}$, $\mu = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$, $\bar{X} = \begin{bmatrix} -2 & -2 & 2 & 2 \\ -1 & 1 & -1 & 1 \end{bmatrix}$, so we have:

$$\Sigma = \frac{1}{4}\bar{X}\bar{X}^T = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

Solving the eigenvalue of Σ, we have $\lambda_1 = 4$ with $v_1 = (1,0)^T$; $\lambda_2 = 1$ with $v_2 = (0,1)^T$.
So we have $U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and centralized data should be: $\begin{bmatrix} -2 \\ -1 \end{bmatrix}$ $\begin{bmatrix} -2 \\ 1 \end{bmatrix}$ $\begin{bmatrix} 2 \\ -1 \end{bmatrix}$ $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$

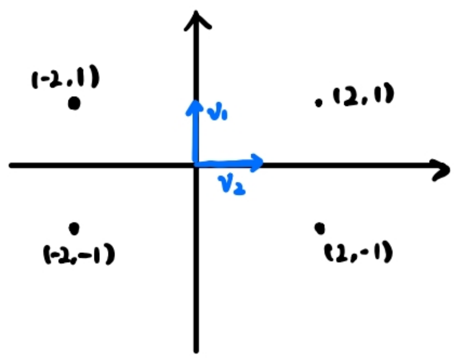

Figure 2: Q1.2

1.3

giving the Σ, we want to solve the optimal problem $max_{w:||w||^2=1}\ w^T\Sigma w$. From the lecture we know the optimal problem's solution is the largest eigenvalue and w is the corresponding eigenvector.
And we know for diagonal matrix, the eigenvalue is the diagonal itself, so the largest eigenvalue is $\lambda_1 = 20$ and the corresponding vector is $w = (0,0,1,0)^T$

2 Basics in Information Theory

2.1

$$Pr(X' = x) = Pr(X'\ P)P(x) + Pr(X'\ Q)Q(x) = \lambda P(x) + (1-\lambda)Q(x)$$

2.2

$$
\begin{aligned}
I(X';B) &= \sum_x \sum_{b=0,1} Pr(x,b)log(\frac{Pr(x,b)}{Pr(x)P(b)}) \\
&= \sum_x (Pr(x|0)Pr(0)log(\frac{Pr(x|0)Pr(0)}{Pr(x)Pr(0)}) + Pr(x|1)Pr(1)log(\frac{Pr(x|1)Pr(1)}{Pr(x)Pr(1)})) \\
&= \lambda \sum_x P(x)log(\frac{P(x)}{Pr(x)}) + (1-\lambda)\sum_x Q(x)log(\frac{Q(x)}{Pr(x)}) \qquad (1)
\end{aligned}
$$

We know from (1) that:

$$Pr(x) = \lambda P + (1-\lambda)Q$$

So:

$$I(X';B) = \lambda D_{KL}(P||\lambda P+(1-\lambda)Q)+(1-\lambda)D_{KL}(Q||\lambda P+(1-\lambda)Q) = D_\lambda(P||Q)$$

3 k-Means with Soft Assignments

3.1

for each row of A(means i is fixed): In the case of hard assignment($A \in \{0,1\}^{nxk}$), we know $A_{ik} = 1$ when

$$k = argmin_{k\in\{1,...K\}}||x^{(i)} - \mu_k||_2^2$$

We know $\{0,1\}^{nxK}$ can be seen as a subset of $[0,1]^{nxK}$,so the soft assignment has a bigger searching space for the optimal solution and at least have an upper bound of the hard assignment. That is:

$$min_{\mu_1,...,\mu k} min_{A\in[0,1]^{nxK}} \sum_{i=1}^n \sum_{k=1}^K A_{ik}||x^{(i)}-\mu_k||_2^2 \leq min_{\mu_1,...,\mu k} min_{A\in\{0,1\}^{nxK}} \sum_{i=1}^n \sum_{k=1}^K A_{ik}||x^{(i)}-\mu_k||_2^2$$

3.2

$$\sum_{i=1}^n \sum_{k=1}^K A_{ik}||x^{(i)} - \mu_k||_2^2 \geq \sum_{i=1}^n (\sum_{k=1}^K A_{ik}) min_l ||x_i - \mu_l||_2^2 = \sum_{i=1}^n min_l ||x^i - \mu_l||_2^2$$

which is the same as the hard assignment. So:

$$min_{\mu_1,...,\mu k} min_{A\in[0,1]^{nxK}} \sum_{i=1}^n \sum_{k=1}^K A_{ik}||x^{(i)}-\mu_k||_2^2 \geq min_{\mu_1,...,\mu k} min_{A\in\{0,1\}^{nxK}} \sum_{i=1}^n \sum_{k=1}^K A_{ik}||x^{(i)}-\mu_k||_2^2$$

3.3

From (1) and (2), we can see that soft assignment $\leq$ hard assignment and soft assignment $\geq$ hard assignment, so we have:

$$min_{\mu_1,\ldots,\mu k} min_{A\in[0,1]^{nxK}} \sum_{i=1}^{n}\sum_{k=1}^{K} A_{ik}||x^{(i)}-\mu_k||_2^2 = min_{\mu_1,\ldots,\mu k} min_{A\in\{0,1\}^{nxK}} \sum_{i=1}^{n}\sum_{k=1}^{K} A_{ik}||x^{(i)}-\mu_k||_2^2$$

and we can say that the soft assignment corresponds to a globally optimal hard assignment.

4 Bernoulli Mixture Model

4.1

$$Pr(x^{(i)}, z_i|\pi,\mu) = Pr(x^{(i)}|z_i,\pi,\mu)Pr(z_i|\pi,\mu) = \prod_{k=1}^{K} \pi_k^{z_{ik}} Pr(x^{(i)}|\mu_k)^{z_{ik}}$$

So we have:

$$\begin{aligned} log(Pr(x^{(i)}, z_i|\pi,\mu)) &= \sum_{k=1}^{K} log(\pi_k^{z_{ik}} Pr(x^{(i)}|\mu_k)^{z_{ik}}) \\ &= \sum_{k=1}^{K} z_{ik}(log(\pi_k) + log(\prod_{j=1}^{d} \mu_k^{x_j^{(i)}}(1-\mu_k)^{(1-x_j^{(i)})}) \\ &= \sum_{k=1}^{K} z_{ik}(log(\pi_k) + \sum_{j=1}^{d}(x_j^{(i)} log(\mu_k) + (1-x_j^{(i)})log(1-\mu_k))) \end{aligned} \tag{2}$$

4.2

$$\begin{aligned} z_{ik}^{new} &= Pr(z_{ik}=1|x^{(i)}) \\ &= \frac{P(z_{ik}=1)P(x^{(i)}|z_{ik}=1)}{\sum_k P(z_{ik}=1)P(x^{(i)}|z_{ik}=1)} \\ &= \frac{\pi_k \prod_{j=1}^{d} \mu_k^{x_j^{(i)}}(1-\mu_k)^{(1-x_j^{(i)})}}{\sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \mu_k^{x_j^{(i)}}(1-\mu_k)^{(1-x_j^{(i)})}} \end{aligned} \tag{3}$$

4.3

First, our optimal problem is:

$$min_{\pi,\mu,\sigma} = -log\prod_{i\in D} Pr(x^{(i)}|\pi,\mu,\sigma) = -\sum_{i\in D} log(\sum_{k=1}^{K} \pi_k \mu_k^{\sum_{j=1}^{d} x_j^{(i)}}(1-\mu)_k^{\sum_{j=1}^{d}(1-x_j^{(i)})})$$

Thus,for μ_k^{new}:

$$
\begin{aligned}
\frac{\partial}{\partial \mu_k} &= -\sum_{i \in D} z_{ik}^{new}[\sum_{j=1}^{d} x_j^{(i)} \frac{1}{\mu_k} - \sum_{j=1}^{d}(1-x_j^{(i)})\frac{1}{1-\mu_k}] \\
&= -\frac{1}{\mu_k(1-\mu_k)} \sum_{i \in D} z_{ik}^{new}[\sum_{j=1}^{d} x_j^{(i)} - d\mu_k]
\end{aligned} \tag{4}
$$

$$
\mu_k^{new} = \frac{\sum_{i \in D}(z_{ik}^{new} \sum_{j=1}^{d} x_j^{(i)})}{d \sum_{i \in D} z_{ik}^{new}}
$$

for π_k^{new},using Lagrange multiplier, we have:

$$
L(\pi_k, \lambda) = -\sum_{i \in D} log \sum_{k=1}^{K} \pi_k Pr(x^{(i)}|\mu_k) + \lambda(\sum_{k=1}^{K} \pi_k - 1)
$$

Thus:

$$
\frac{\partial}{\partial \pi_k} = -\sum_{i \in D} \frac{Pr(x^{(i)}|\mu_k)}{\sum_{k=1}^{K} \pi_k Pr(x^{(i)}|\mu_k)} + \lambda = 0
$$

$$
\lambda = \sum_{i \in D} \frac{z_{ik}^{new}}{\pi_k}
$$

$$
\pi_k = \sum_{i \in D} \frac{z_{ik}^{new}}{\lambda}
$$

$$
\lambda = \sum_{i \in D} \sum_{k=1}^{K} z_{ik}^{new} = N
$$

so:

$$
\pi_k^{new} = \frac{\sum_{i \in D} z_{ik}^{new}}{N}
$$

5 Variational Autoencoder(VAE)

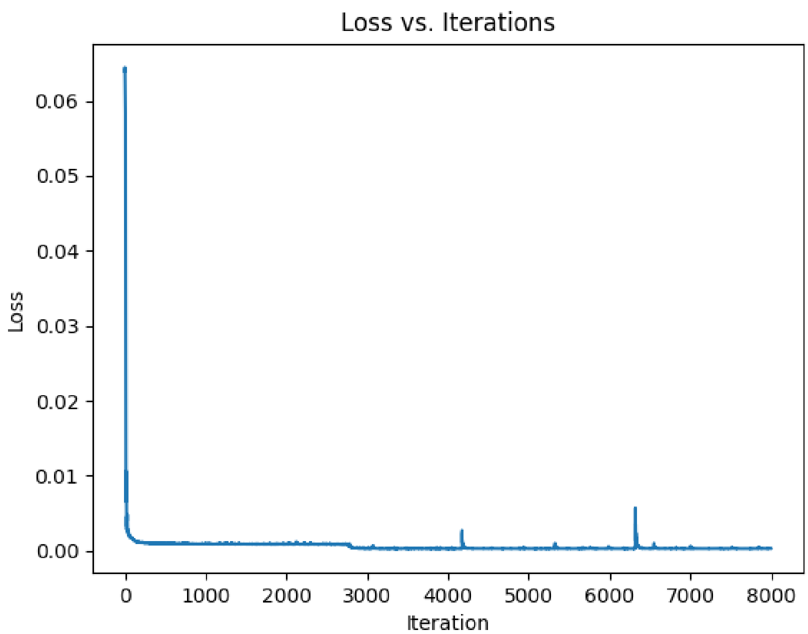

Figure 3: loss

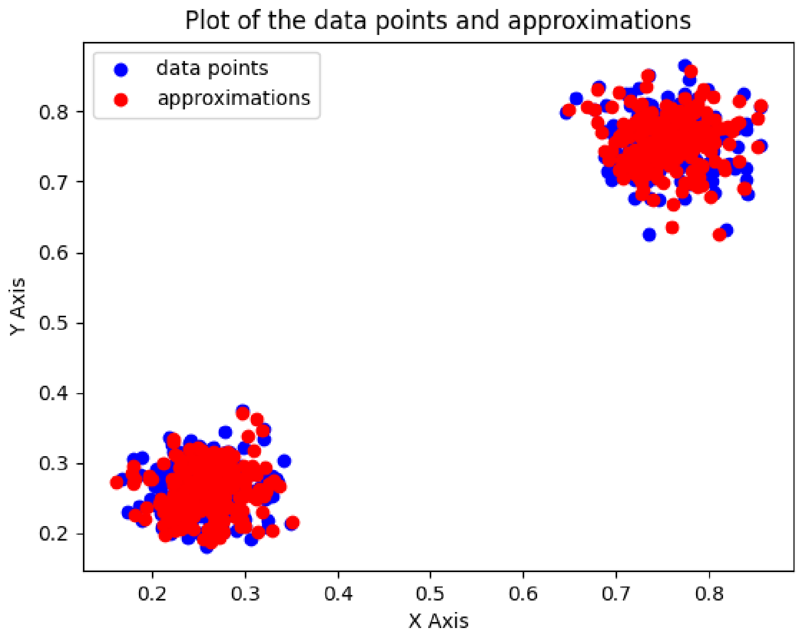

Figure 4: data vs. approximation

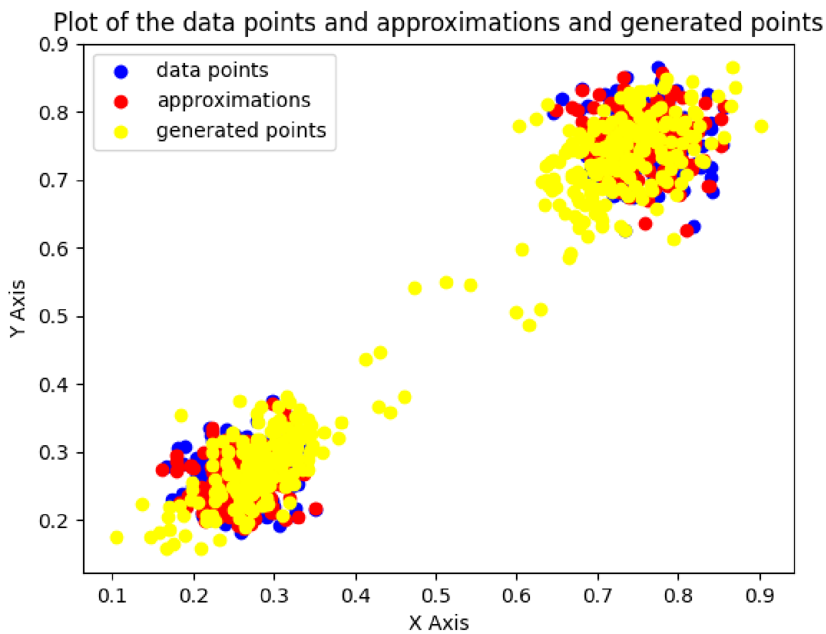

Figure 5: data, approximation and generated points