2.

1.

$$
\begin{aligned}
& H(X,Y) \\
=& -\sum_i \sum_j P(X=x_i, Y=y_i) \log P(X=x_i, Y=y_i) \\
=& -\sum_i \sum_j P(X=x_i|Y=y_j) P(Y=y_j) \left(\log P(X=x_i|Y=y_j) + \log P(Y=y_j)\right) \\
=& -\sum_j P(Y=y_j) \left(\sum_i P(X=x_i|Y=y_j) \log P(X=x_i|Y=y_j)\right) \\
& -\sum_j P(Y=y_j) \log P(Y=y_j) \left(\sum_i P(X=x_i|Y=y_j)\right) \\
=& H(X|Y) - \sum_j P(Y=y_j) \log P(Y=y_j) \\
=& H(X|Y) + H(Y).
\end{aligned}
$$

Similarily,

$$H(X,Y) = H(Y|X) + H(X).$$

2. If X and Y are independent, then we have

$$P(X=x_i, Y=y_i) = P(X=x_i)P(Y=y_j),$$

therefore,

$$
\begin{aligned}
H(X,Y) &= -\sum_i \sum_j P(X=x_i, Y=y_i) \log P(X=x_i, Y=y_i) \\
&= -\sum_i \sum_j P(X=x_i)P(Y=y_j) \left(\log P(X=x_i) + \log P(Y=y_j)\right) \\
&= -\sum_i P(X=x_i) \log P(X=x_i) \left(\sum_j P(Y=y_j)\right) \\
&+ -\sum_j P(Y=y_j) \log P(Y=y_j) \left(\sum_i P(X=x_i)\right) \\
&= -\sum_i P(X=x_i) \log P(X=x_i) - \sum_j P(Y=y_j) \log P(Y=y_j) \\
&= H(X) + H(Y).
\end{aligned}
$$

According to

$$\begin{cases} H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X) \\ H(X,Y) = H(X) + H(Y) \end{cases}$$

we finally get $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = 0$.

3.

$$\begin{aligned}
& D_{KL}(p(X,Y)||p(X)p(Y)) \\
=& -\sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i)p(y_j)}{p(x_i, y_j)} \\
=& -\sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i)}{p(x_i, y_j)} - \sum_i \sum_j p(x_i, y_j) \log p(y_i) \\
=& \sum_i \sum_j p(y_j|x_i)p(x_i) \log p(y_j|x_i) - \sum_i \sum_j p(x_i|y_j)p(y_j) \log p(y_i) \\
=& \sum_i p(x_i) \left(\sum_j p(y_j|x_i) \log p(y_j|x_i) \right) - \sum_j p(y_j) \log p(y_i) \left(\sum_i p(x_i|y_j) \right) \\
=& - H(Y|X) - \sum_j p(y_j) \log p(y_i) \\
=& - H(Y|X) + H(Y) = H(Y) - H(Y|X) = I(X;Y).
\end{aligned}$$

4. Knowing $-\log(x)$ is a convex function, by Jensen's inequality, we get

$$\sum_i p_i \cdot \log\left(\frac{q_i}{p_i}\right) \leq \log\left(\sum_i p_i \cdot \frac{q_i}{p_i}\right) = \log\left(\sum_i q_i\right) = \log(1) = 0.$$

Therefore, $D_{KL}(P||Q) = -\sum_i p_i \cdot \log\left(\frac{q_i}{p_i}\right) \geq 0$ for any P and Q. (Note that P and Q are probability distributions, which means p_i and q_i are between zero and one)
As a result, $I(X;Y) \geq 0$.

3.

1. The likelihood function is

$$L(\mu, a, b) = \left(\frac{1}{2}\right)^a (\mu)^b (2\mu)^c \left(\frac{1}{2} - 3\mu\right)^d$$

It follows that the log-likelihood function is

$$l(\mu, a, b) = \log L(\mu, a, b) = a\log(1/2) + b\log(\mu) + c\log(2\mu) + d\log(1/2 - 3\mu).$$

2. Since $\mathbb{P}(A) = 1/2$ and $\hat{\mathbb{P}}(B) = \hat{\mu}^{(m)}$, then we get

$$\hat{a}^{(m)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A) + \hat{\mathbb{P}}(B)} h = \frac{1/2}{1/2 + \hat{\mu}^{(m)}} h \quad \text{and} \quad \hat{b}^{(m)} = \frac{\hat{\mathbb{P}}(B)}{\mathbb{P}(A) + \hat{\mathbb{P}}(B)} h = \frac{\hat{\mu}^{(m)}}{1/2 + \hat{\mu}^{(m)}} h.$$

3. By differentiate l with respect to μ and let it be zero, we have

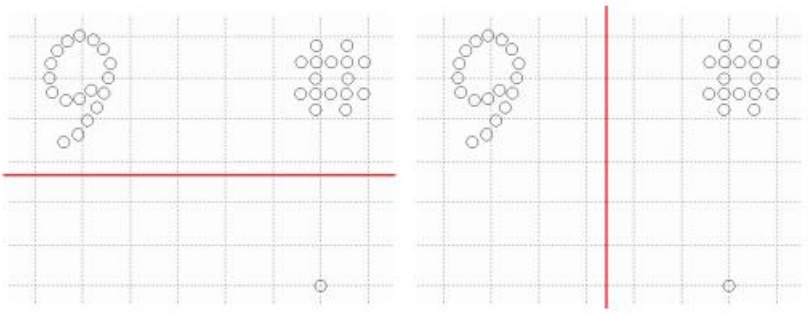
$$\frac{\partial l}{\partial \mu} = \frac{b}{\mu} + \frac{c}{\mu} + \frac{-3d}{1/2 - 3\mu} = 0.$$

This gives that

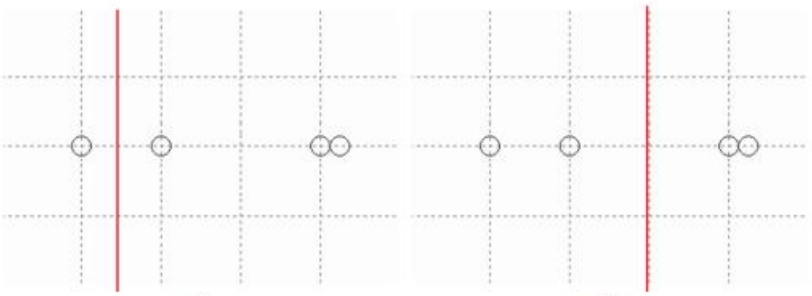
$$\hat{\mu}^{(m+1)} = \frac{\hat{b}^{(m)} + c}{6(\hat{b}^{(m)} + c + d)} = \frac{c + 2\hat{\mu}^{(m)}(h + c)}{6\Big((c + d) + 2\hat{\mu}^{(m)}(h + c + d)\Big)}.$$

4. **True**. Because the value of the log-likelihood function (lower bound) increases at each iteration.

4.



(a) Min-cut (b) Normalized cut



(a) $\sigma^2 = 50$ (b) $\sigma^2 = 0.5$

(a) The data points are shown in Figure (a) above. The grid unit is 1. Let $W_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}$, give the clustering results of min-cut and normalized cut respectively (Please draw a rough sketch and give the separation boundary in the answer book).
Sol: Shown in the figures, since min-cut allows the isolation of singlet while normalized cut does not.

(b)

Sol: Shown in the figures, since $\sigma^2 = 50$ treats all pairs of data points almost equidistant and thus min-cut isolates singlet.

2. Now back to the setting of the 2-clustering problem shown in Figure (a). The grid unit is 1.

(a) If we use Euclidean distance to construct the affinity matrix W as follows:

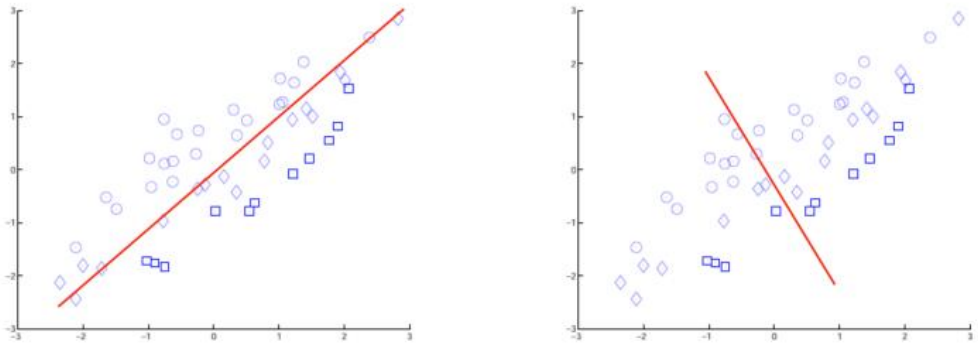
$$W_{ij} = \begin{cases} 1, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leqslant \sigma^2; \\ 0, & \text{otherwise.} \end{cases}$$

What σ^2 value would you choose? Briefly explain.
Sol: $\sigma^2 = 9 \sim 16$. Since the distance of each pair of data points within each cluster is not greater than 3, and the between-cluster distance is almost 4, it is obvious to separate these clusters using a distance between 3 and 4.

(b) The next step is to compute the first $k = 2$ dominant eigenvectors of the affinity matrix W. For the value of σ^2 you chose in the previous question, can you compute analytically the eigenvalues corresponding to the first two eigenvectors? If yes, compute and report the eigenvalues. If not, briefly explain.
Sol: 17 and 15. Note that the cluster "9" has 18 points and the cluster "#" has 16 points. The affinity matrix has the block diagonal form $W = \text{diag}(\mathbf{J}_{18} - \mathbf{I}_{18}, \mathbf{J}_{16} - \mathbf{I}_{16}, 0)$, where $\mathbf{J}$ is the matrix with entries all ones. The degree matrix is $D = \text{diag}(17\mathbf{I}_{18}, 15\mathbf{I}_{16}, 0)$. The Laplacian $L = D - W$ has two zero principal eigenvalues corresponding to the eigenvectors $\mathbf{1}_{18}$ and $\mathbf{1}_{16}$ (there are 3 connected components in the graph). Simple calculation shows that W has the two principal eigenvalues 17 and 15 corresponding to the same eigenvectors $\mathbf{1}_{18}$ and $\mathbf{1}_{16}$.

5.

1.



(a) (b)

2. a) $\mu = (0,0,0,0,0,0)^T$.

b)

$$XX^T = \begin{pmatrix} 3 & -9 & 6 & 0 & 0 & 0 \\ -9 & 27 & -18 & 0 & 0 & 0 \\ 6 & -18 & 12 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -4 & 2 \\ 0 & 0 & 0 & -4 & 8 & -4 \\ 0 & 0 & 0 & 2 & -4 & 2 \end{pmatrix}$$

$$\Rightarrow \begin{cases} \sigma_1^2 = 42 & u_1^T = \left(-\frac{\sqrt{14}}{14}, \frac{3\sqrt{14}}{14}, -\frac{\sqrt{14}}{7}, 0, 0, 0\right) \\ \sigma_2^2 = 12 & u_2^T = \left(0, 0, 0, -\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{3}, -\frac{\sqrt{6}}{6}\right) \end{cases}$$

$$X^TX = \begin{pmatrix} 14 & 14 & 14 & 0 & 0 \\ 14 & 14 & 14 & 0 & 0 \\ 14 & 14 & 14 & 0 & 0 \\ 0 & 0 & 0 & 6 & 6 \\ 0 & 0 & 0 & 6 & 6 \end{pmatrix} \Rightarrow \begin{cases} \sigma_1^2 = 42 & v_1 = \left(-\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, 0, 0\right) \\ \sigma_2^2 = 12 & v_2 = \left(0, 0, 0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \end{cases}$$

Therefore,

$$X = \begin{pmatrix} -\frac{\sqrt{14}}{14} & 0 \\ \frac{3\sqrt{14}}{14} & 0 \\ -\frac{\sqrt{14}}{7} & 0 \\ 0 & -\frac{\sqrt{6}}{6} \\ 0 & \frac{\sqrt{6}}{3} \\ 0 & -\frac{\sqrt{6}}{6} \end{pmatrix} \begin{pmatrix} \sqrt{42} & 0 \\ 0 & 2\sqrt{3} \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}.$$

Note that this (a, b, c, d) pair is not the only solution, the result is considered to be correct when the corresponding sign of (a, b, c, d) pair is correct. Similar situation for the following questions.

c)

$$\Sigma = Cov(X) = \frac{1}{6-1}\sum_{i=1}^{6}(x_i - \mu)(x_i - \mu)^T = \frac{1}{5}X^TX$$

The maximum eigenvalue of X^TX is 42, and the corresponding eigenvector is $w_1 = \left(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, 0, 0\right)^T$, thus, the first principle component is $\frac{\sqrt{3}}{3}x_1 + \frac{\sqrt{3}}{3}x_2 + \frac{\sqrt{3}}{3}x_3$.

d)

$$\{\hat{x}_i\}_{i=1}^6 = \{\sqrt{3}, -3\sqrt{3}, 2\sqrt{3}, 0, 0, 0\}$$

Then,

$$Var(\{\hat{x}_i\}_{i=1}^6) = \frac{\sum_{i=1}^6(\hat{x}_i - \bar{\hat{x}})^2}{6-1} = \frac{42}{5} = 8.4.$$

or

$$Var(\{\hat{x}_i\}_{i=1}^6) = \frac{\sum_{i=1}^6(\hat{x}_i - \bar{\hat{x}})^2}{6} = \frac{42}{6} = 7.$$

e)

$$\frac{1}{6}\sum_{i=1}^{6}\|x_i - \hat{x}_i\|_2^2 = \frac{1}{6}\sum_{i=1}^{6}\|x_i - \hat{x}_i'w_1\|_2^2 = 2,$$

where $\hat{x}_i'$ represents the projected data in (d).

6.

1.

$$\begin{cases} \dfrac{\partial \sum_{i=1}^{N} \|\mathbf{x}_i - \mu - \mathbf{V}_q \alpha_i\|^2}{\partial \mu} = -2 \sum_{i=1}^{N} (\mathbf{x}_i - \mu - \mathbf{V}_q \alpha_i) = 0 \\ \dfrac{\partial \sum_{i=1}^{N} \|\mathbf{x}_i - \mu - \mathbf{V}_q \alpha_i\|^2}{\partial \alpha_i} = -2 \mathbf{V}_q^T (\mathbf{x}_i - \mu - \mathbf{V}_q \alpha_i) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \hat{\mu} = \dfrac{1}{N} \left(\sum_{i=1}^{N} \mathbf{x}_i - \mathbf{V}_q \sum_{i=1}^{N} \alpha_i \right) \\ \hat{\alpha}_i = V_q^T (\mathbf{x}_i - \mu) \end{cases}.$$

Therefore, we can see that $\hat{\mu}$ is not unique, it depends on $\sum_{i=1}^{N} \alpha_i$.
The family of solutions for $\hat{\mu}$ is

$$\hat{\mu} = \frac{1}{N} \left(\sum_{i=1}^{N} \mathbf{x}_i - \mathbf{V}_q \sum_{i=1}^{N} \alpha_i \right).$$

If $\sum_{i=1}^{N} \alpha_i = 0$, then we get

$$\begin{cases} \hat{\mu} = \dfrac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i = \bar{\mathbf{x}} \\ \hat{\alpha}_i = \mathbf{V}_q^T (\mathbf{x}_i - \mu) = \mathbf{V}_q^T (\mathbf{x}_i - \bar{\mathbf{x}}) \end{cases}.$$

2.

$$\begin{aligned} &\min_{\mathbf{V}_q} \sum_{i=1}^{N} \|(\mathbf{x}_i - \bar{\mathbf{x}}) - \mathbf{V}_q \mathbf{V}_q^T (\mathbf{x}_i - \bar{\mathbf{x}})\|^2 \\ =& \min_{\mathbf{V}_q} \operatorname{Tr} \left(\tilde{\mathbf{X}} (\mathbf{I}_p - \mathbf{V}_q \mathbf{V}_q^T) \tilde{\mathbf{X}}^T \right) \\ =& \min_{\mathbf{V}_q} \operatorname{Tr} \left(\mathbf{U} \mathbf{D} \mathbf{V}^T (\mathbf{I}_p - \mathbf{V}_q \mathbf{V}_q^T) \mathbf{V} \mathbf{D} \mathbf{U}^T \right) \\ =& \min_{\mathbf{V}_q} \operatorname{Tr} \left(\mathbf{D} \mathbf{V}^T (\mathbf{I}_p - \mathbf{V}_q \mathbf{V}_q^T) \mathbf{V} \mathbf{D} \right) \\ =& \min_{\mathbf{V}_q} \operatorname{Tr} \left(\mathbf{D}^2 (\mathbf{I}_p - (\mathbf{V}^T \mathbf{V}_q)(\mathbf{V}^T \mathbf{V}_q)^T) \right) \\ =& \max_{\mathbf{V}_q} \operatorname{Tr} \left(\mathbf{D}^2 (\mathbf{V}^T \mathbf{V}_q)(\mathbf{V}^T \mathbf{V}_q)^T \right) \end{aligned}$$

Assume $\mathbf{D} = diag(d_1, \cdots, d_p)$, $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_p)$, and $\mathbf{V}_q = (\mathbf{x}_1, \cdots, \mathbf{x}_q)$, since $\mathbf{V}$ and

$\mathbf{V}_q$ are orthogonal, we can easily see that

$$
\begin{aligned}
&\operatorname{Tr}\left(\mathbf{D}^2(\mathbf{V}^T\mathbf{V}_q)(\mathbf{V}^T\mathbf{V}_q)^T\right) \\
=&\sum_{i=1}^{p}\sum_{j=1}^{q}(d_i\mathbf{v}_i^T\mathbf{x}_j)^2 \\
=&\sum_{i=1}^{p}d_i^2\sum_{j=1}^{q}(\mathbf{v}_i^T\mathbf{x}_j)^2 \\
&\left(\mathbf{V}\text{ and }\mathbf{V}_q\text{ are orthogonal }\Rightarrow\begin{cases}\sum_{j=1}^{q}(\mathbf{v}_i^T\mathbf{x}_j)^2\le 1\\ \sum_{i=1}^{p}(\mathbf{v}_i^T\mathbf{x}_j)^2=1\Rightarrow\sum_{i=1}^{p}\sum_{j=1}^{q}(\mathbf{v}_i^T\mathbf{x}_j)^2=q\end{cases}\right) \\
=&\sum_{i=1}^{p}y_id_i^2 \qquad \left(y_i=\sum_{j=1}^{q}(\mathbf{v}_i^T\mathbf{x}_j)^2\le 1,\sum_{i=1}^{p}y_i=q\right) \\
\le&\sum_{j=1}^{q}d_j^2 \qquad (\text{because } d_1\ge d_2\ge\cdots\ge d_p\ge 0)\,.
\end{aligned}
$$

When $\mathbf{V}_q$ consists of the first q columns of $\mathbf{V}$ in order, we can get that

$$\mathbf{V}^T\mathbf{V}_q=\begin{pmatrix}\mathbf{I}_q\\ \mathbf{0}_{(p-q)\times q}\end{pmatrix}_{p\times q},$$

which means that under this condition, we have

$$\operatorname{Tr}\left(\mathbf{D}^2(\mathbf{V}^T\mathbf{V}_q)(\mathbf{V}^T\mathbf{V}_q)^T\right)=\sum_{j=1}^{q}d_j^2.$$

Therefore, we get that the solution $\mathbf{V}_q$ to problem (4) consists of the first q columns of $\mathbf{V}$.