Qian Cao

1. $C_1 = \{(5,5), (8,7), (7,3)\}$

$C_2 = \{(6,5), (4,4), (9,2), (3,5), (8,4)\}$

(a) $m_1 = \left(\frac{5+8+7}{3}, \frac{5+7+3}{3}\right) = (6.667, 5)$

$m_2 = \left(\frac{6+4+9+3+8}{5}, \frac{5+4+2+5+4}{5}\right) = (6, 4)$

(b) $m = \left(\frac{6.667\times3+6\times5}{8}, \frac{5\times3+4\times5}{8}\right) = (6.25, 4.375)$

(c) $S_1 = \begin{pmatrix}-1.667\\0\end{pmatrix}(-1.667\ \ 0) + \begin{pmatrix}\frac{4}{3}\\2\end{pmatrix}(\frac{4}{3}, 2) + \begin{pmatrix}\frac{1}{3}\\-2\end{pmatrix}(\frac{1}{3}\ \ -2)$

$= \begin{pmatrix}\frac{25}{9} & 0\\0 & 0\end{pmatrix} + \begin{pmatrix}\frac{16}{9} & \frac{8}{3}\\\frac{8}{3} & 4\end{pmatrix} + \begin{pmatrix}\frac{1}{9} & -\frac{2}{3}\\-\frac{2}{3} & 4\end{pmatrix}$

$= \begin{pmatrix}4.667 & 2\\2 & 8\end{pmatrix}$

we can get $S_2 = \begin{pmatrix}26 & -9\\-9 & 6\end{pmatrix}$ with the same way.

(d) $S_W = S_1 + S_2 = \begin{pmatrix}30.667 & -7\\-7 & 14\end{pmatrix}$

(e). $S_B = 3\times\begin{pmatrix}0.417\\0.625\end{pmatrix}(0.417\ \ 0.625) + 5\begin{pmatrix}-0.25\\-0.375\end{pmatrix}(-0.25\ \ -0.375)$

$= \begin{pmatrix}0.8342 & 1.2506\\1.2506 & 1.875\end{pmatrix}$

(f) $\frac{tr(S_B)}{tr(S_W)} = \frac{0.8342+1.875}{30.667+14} = 0.06065$

2.

(a) There is one cluster (8 points):

$C = \{(0,1), (5,2), (2,3), (6,1), (3,4), (6,3), (7,2), (1,2)\}$

(b). for all $p, q \in C$:

the point p is density-connected to q

wrt. $\epsilon = 2\sqrt{2}$ and MinPts = 4.

(c). noise points: $(10,2), (0,6), (0,7)$

3.

$X = \{HHTT, HHHH, TTTT\}$

$P(A|X_n) \propto P(A)\mu_A^{heads}(1-\mu_A)^{tails}$

$P(B|X_n) \propto P(B)\mu_B^{heads}(1-\mu_B)^{tails}$

at the beginning: $P(A) = \pi = 0.6$

$P(B) = 1-\pi = 0.4$

① first iteration:

$P(A|X_1) \propto 0.6\times(0.55)^2\times(0.45)^2$

$P(B|X_1) \propto 0.4\times(0.45)^2\times(0.55)^2$

for $P(A|X_1)+P(B|X_1) = 1$, $P(A|X_1) = 0.6$

$P(B|X_1) = 0.4$

$P(A|X_2) \propto 0.6\times(0.55)^4$

$P(B|X_2) \propto 0.4\times(0.45)^4$

for $P(A|X_2)+P(B|X_2) = 1$, $P(A|X_2) = 0.77$

$P(B|X_2) = 0.23$

$P(A|X_3) \propto 0.6\times(0.45)^4$

$P(B|X_3) \propto 0.4\times(0.55)^4$

for $P(A|X_3)+P(B|X_3) = 1$, $P(A|X_3) = 0.40$

$P(B|X_3) = 0.60$

write in a table:

	coin A	coin B
	1.2H, 1.2T	0.8H, 0.8T
	3.08H	0.92H
	1.6T	2.4T
total	4.28H, 2.8T	1.72H, 3.2T

$\mu_A' = \frac{4.28}{4.28+2.8} = 0.60$

$\mu_B' = \frac{1.72}{1.72+3.2} = 0.35$

$\pi' = \frac{4.28+2.8}{4\times3} = 0.59.$

Here, we finish the first iteration.

② Second iteration:

$P(A|X_1) = 0.615$ $P(B|X_1) = 0.385$

$P(A|X_2) = 0.9255$ $P(B|X_2) = 0.0745$

$P(A|X_3) = 0.171$ $P(B|X_3) = 0.829$

Qian Cao

write in a table:

	coin A	coin B
	1.23H, 1.23T	0.77H, 0.77T
	3.702H	0.298H
	0.684T	3.316T
total	4.932H, 1.914T	1.068H, 4.086T

$$\mu_A'' = \frac{4.932}{4.932+1.914} = 0.72$$

$$\mu_B'' = \frac{1.068}{1.068+4.086} = 0.207$$

$$\pi'' = \frac{4.932+1.914}{12} = 0.57$$

③ third iteraction:

$P(A|X_1) = 0.665$ $P(B|X_1) = 0.335$

$P(A|X_2) = 0.995$ $P(B|X_2) = 0.005$

$P(A|X_3) = 0.02$ $P(B|X_3) = 0.98$

Write in a table:

	coin A	coin B
	1.33H, 1.33T	0.67H, 0.67T
	3.98H	0.02H
	0.08T	3.92T
total	5.31H, 1.41T	0.69H, 4.59T

$$\mu_A''' = \frac{5.31}{5.31+1.41} = 0.79$$

$$\mu_B''' = \frac{0.69}{0.69+4.59} = 0.13$$

$$\pi''' = \frac{5.31+1.41}{12} = 0.56$$

4.

(a) GMM > DBSCAN = Agnes > K-means

for different density of each cluster, GMM is the first choice.

for the nonconvex shape, DBSCAN and Agnes will be better than k-means.

DBSCAN and Agnes get the similar result.

(b). DBSCAN > Agnes > GMM > K-means

for clearly seeing high density area (cluster), DBSCAN is the best choice.

Agnes may have trouble in dividing clusters nearby, so it's the second choice.

It may have overlapping clusters, so GMM is better than k-means.

(c). Agnes > DBSCAN > GMM > K-means

similar as (b).

Since there are too many clusters (31 clusters), I think Agnes may be better than DBSCAN.

(d) DBSCAN > Agnes = K-means > GMM

even density for each cluster makes DBSCAN the best choice.

Agnes and K-means may have trouble in dividing clusters nearby → second choice

No need for GMM: flat geometry;
no overlapping;
even density.

(e). GMM > DBSCAN > Agnes > K-means.

for two clusters connected, there may be overlapping then GMM is desired.

It has a nonconvex shape and the clusters are nearby, so I think DBSCAN better than Agnes, which better than k-means.

(f). GMM > K-means > DBSCAN > Agnes

for varing density in each, GMM is the best.

Also for the same reason, K-means is better than DBSCAN.

for nearby clusters, DBSCAN is better than Agnes.

(g). DBSCAN > Agnes > K-means > GMM

It has nearby clusters and nonconvex shape, so DBSCAN > Agnes > k-means.

No need for GMM. (same reason as (d).

(f). DBSCAN = Agnes > k-means > GMM

It has nonconvex shape → DBSCAN and Agnes will work similarly, which are better than k-means.

No need for GMM. (also same as (d).