1. (1) $p(X;\mu) = \sum_z p(X|z;\mu)p(z;\mu)$

$p(z=k;\mu) = \pi_k$

$p(x_i|\mu_k) = \prod_{j=1}^{D} \mu_{kj}^{x_{ij}}(1-\mu_{kj})^{(1-x_{ij})}$

$$\log p(z_i, x_i|\mu_k) = \sum_k 1_{\{z_i=k\}}\left[\log\pi_k + \sum_{j=1}^{D}\log(\mu_{kj}^{x_{ij}}(1-\mu_{kj})^{(1-x_{ij})})\right]$$

$$= \sum_k 1_{\{z_i=k\}}\left[\log\pi_k + \sum_{j=1}^{D}(x_{ij}\log\mu_{kj} + (1-x_{ij})\log(1-\mu_{kj}))\right]$$

$q_i(z_i=k) = p(z_i=k|x_i;\mu^{old}) = r_{ik}$

$$\mu_{new} = \arg\max_\mu \sum_i E\left[\sum_k 1_{\{z_i=k\}}(\log\pi_k + \sum_{j=1}^{D}\log\mu_{kj}^{x_{ij}}(1-\mu_{kj})^{(1-x_{ij})})\right]$$

$$= \arg\max_\mu \sum_i\sum_k r_{ik}(\log\pi_k + \sum_{j=1}^{D}(x_{ij}\log\mu_{kj} + (1-x_{ij})\log(1-\mu_{kj})))$$

Take derivatives and set it to zero:

$$\frac{\partial}{\partial\mu_{kj}}\sum_i\sum_k r_{ik}(\log\pi_k + \sum_{j=1}^{D}(x_{ij}\log\mu_{kj} + (1-x_{ij})\log(1-\mu_{kj}))) = 0$$

$$\sum_i r_{ik}\left(\frac{x_{ij}}{\mu_{kj}} - \frac{1-x_{ij}}{1-\mu_{kj}}\right) = 0$$

$$(1-\mu_{kj})\sum_i r_{ik}x_{ij} = \mu_{kj}\sum_i r_{ik}(1-x_{ij})$$

$$\mu_{kj} = \frac{\sum_i r_{ik}x_{ij}}{\sum_i r_{ik}}$$

(2) $\beta(\alpha,\beta,\mu_{kj}) = A\mu_{kj}^{\alpha-1}(1-\mu_{kj})^{\beta-1}$ where $A = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}$

$\mu_{kj} = \arg\max_{\mu_{kj}}[L(q;\mu) + \sum_j\sum_k \log\beta(\alpha,\beta,\mu_{kj})]$

Take derivatives and set it to zero

$$\frac{\partial}{\partial\mu_{kj}}L(q;\mu) + \frac{\partial}{\partial\mu_{kj}}\sum_j\sum_k\log\beta(\alpha,\beta,\mu_{kj}) = 0$$

$$\sum_i r_{ik}\left(\frac{x_{ij}}{\mu_{kj}} - \frac{1-x_{ij}}{1-\mu_{kj}}\right) + \frac{\partial}{\partial\mu_{kj}}\sum_j\sum_k[\log A + (\alpha-1)\log\mu_{kj} + (\beta-1)\log(1-\mu_{kj})] = 0$$

$$\sum_i r_{ik}\left(\frac{x_{ij}}{\mu_{kj}} - \frac{1-x_{ij}}{1-\mu_{kj}}\right) + \frac{\alpha-1}{\mu_{kj}} - \frac{\beta-1}{1-\mu_{kj}} = 0$$

$$(1-\mu_{kj})(\alpha-1+\sum_i r_{ik}x_{ij}) = \mu_{kj}(\beta-1+\sum_i r_{ik}(1-x_{ij}))$$

$$\mu_{kj} = \frac{\alpha-1+\sum_i r_{ik}x_{ij}}{\alpha+\beta-2+\sum_i r_{ik}}$$

2. (1) T_1:

squared distance / data / Centroid	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
(2,10)	0 ✓	25	72	13	50	52	65	5
(5,8)	13	18	25 ✓	0 ✓	13 ✓	17 ✓	52	2 ✓
(1,2)	65	10 ✓	53	52	45	29	0 ✓	58

"✓" means optimal of the column.

		Center
Cluster1	(2,10)	(2,10)
Cluster2	(8,4) (5,8) (7,5) (6,4) (4,9)	(6,6)
Cluster3	(2,5) (1,2)	(1.5, 3.5)

T_2:

squared distance / data / Centroid	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
(2,10)	0 ✓	25	72	13	50	52	65	5 ✓
(6,6)	32	17	8 ✓	5 ✓	2 ✓	4 ✓	41	13
(1.5,3.5)	42.5	2.5 ✓	42.5	32.5	32.5	20.5	2.5 ✓	36.5

		Center
Cluster1	(2,10) (4,9)	(3, 9.5)
Cluster2	(8,4) (5,8) (7,5) (6,4)	(6.5, 5.25)
Cluster3	(2,5) (1,2)	(1.5, 3.5)

T_3:

squared distance / data / Centroid	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
(3,9.5)	1.25 ✓	21.25	55.25	6.25 ✓	36.25	39.25	60.25	1.25 ✓
(6.5,5.25)	41.81	20.31	3.81 ✓	9.81	0.31 ✓	1.81 ✓	40.81	20.31
(1.5,3.5)	42.5	2.5 ✓	42.5	32.5	32.5	20.5	2.5 ✓	36.5

		center
Cluster1	(2,10) (5,8) (4,9)	$(\frac{11}{3}, 9)$
Cluster2	(8,4) (7,5) (6,4)	$(7, \frac{13}{3})$
Cluster3	(2,5) (1,2)	(1.5, 3.5)

T_4:

squared distance / data / Centroid	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
$(\frac{11}{3}, 9)$	$\frac{34}{9}$ ✓	$\frac{169}{9}$	$\frac{394}{9}$	$\frac{25}{9}$ ✓	$\frac{224}{9}$	$\frac{274}{9}$	$\frac{505}{9}$	$\frac{1}{9}$ ✓
$(7, \frac{13}{3})$	$\frac{514}{9}$	$\frac{229}{9}$	$\frac{10}{9}$ ✓	$\frac{157}{9}$	$\frac{4}{9}$ ✓	$\frac{10}{9}$ ✓	$\frac{373}{9}$	$\frac{277}{9}$
(1.5, 3.5)	$\frac{85}{2}$	$\frac{5}{2}$ ✓	$\frac{85}{2}$	$\frac{65}{2}$	$\frac{65}{2}$	$\frac{41}{2}$	$\frac{5}{2}$ ✓	$\frac{73}{2}$

assignments don't change

converge

Cluster1	A_1 A_4 A_8
Cluster2	A_3 A_5 A_6
Cluster3	A_2 A_7

(2) As the figure, A_1 must link A_8, A_2 must link A_6, A_4 can't link A_5.

squared distance / data / Centroid	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
(2,10)	0 ✓	25	72	13	50	52	65	5 ($\rightarrow A_1$) ✓
(5,8)	13	18	25 ✓	0 ✓	13 ($\not\rightarrow A_4$)	17	52	2
(1,2)	65	10 ✓	53	52	45 ✓	29 ($\rightarrow A_2$) ✓	0 ✓	58

		center
Cluster1	(2,10) (4,9)	(3, 9.5)
Cluster2	(8,4) (5,8)	(6.5, 6)
Cluster3	(2,5) (7,5) (6,4) (1,2)	(4,4)

Centroid \ squared distance \ data	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
(3, 9.5)	1.25 ✓	21.25	55.25	6.25	36.25	33.25	60.25	1.25 ($\to A_1$) ✓
(6.5, 6)	36.25	21.25	6.25 ✓	6.25 ✓	1.25 ($\not\to A_4$)	4.25	46.25	15.25
(4,4)	40	5 ✓	16	17	10 ✓	4 ($\to A_2$) ✓	13 ✓	25

converge

Cluster1	A_1 A_8
Cluster2	A_3 A_4
Cluster3	A_2 A_5 A_6 A_7

3. Given a dataset $D = \{x^{(1)}, \cdots, x^{(N)}\} \subset R^D$

Let the mean be $\mu := \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$

K-dimensional subspace S is spanned by an orthonormal basis $\{u_k\}_{k=1}^{K}$ where $u_k \in R^D$, $u_i^T u_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$

Approximate data x : $\tilde{x} = \mu + Proj_S(x-\mu) = \mu + \sum_{k=1}^{K} z_k u_k$ where $z_k = u_k^T(x-\mu)$

Let $U \in R^{D \times K}$ be a matrix with columns $\{u_k\}_{k=1}^{K}$

reconstruction of x : $\tilde{x} = \mu + Uz$

representation of x : $z = U^T(x-\mu)$

the mean of the reconstructions $\tilde{\mu} = \frac{1}{N}\sum_{i=1}^{N}\tilde{x}^{(i)}$

we need to $\max_{U,\, U^TU=I} \frac{1}{N}\sum_{i=1}^{N} \|\tilde{x}^{(i)} - \tilde{\mu}\|^2$

$$
\begin{aligned}
\tilde{\mu} &= \frac{1}{N}\sum_{i=1}^{N}\tilde{x}^{(i)} \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mu + U z^{(i)}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mu + UU^T(x^{(i)}-\mu)) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mu + x^{(i)} - \mu) \\
&= \frac{1}{N}\sum_{i=1}^{N} x^{(i)} \\
&= \mu
\end{aligned}
$$

$\therefore \tilde{x}^{(i)} - \tilde{\mu} = (\mu + U z^{(i)}) - \mu = U z^{(i)}$

$\|U z^{(i)}\|^2 = (U z^{(i)})^T (U z^{(i)}) = z^{(i)T} U^T U z^{(i)} = z^{(i)T} z^{(i)} = \|z^{(i)}\|^2$

$\max_{U,\, U^TU=I} \frac{1}{N}\sum_{i=1}^{N}\|\tilde{x}^{(i)} - \tilde{\mu}\|^2 = \max_{U,\, U^TU=I} \frac{1}{N}\sum_{i=1}^{N}\|U z^{(i)}\|^2 = \max_{U,\, U^TU=I} \frac{1}{N}\sum_{i=1}^{N}\|z^{(i)}\|^2$

$\because z^{(i)} = U^T(x^{(i)} - \mu)$

$$\max_{U^TU=I} \frac{1}{N}\sum_{i=1}^{N} \| z^{(i)} \|^2 = \max_{U^TU=I} \frac{1}{N}\sum_{i=1}^{N} \| U^T(x^{(i)}-\mu) \|^2$$

$$= \max_{U^TU=I} \frac{1}{N}\sum_{i=1}^{N} \mathrm{Trace}(U^T(x^{(i)}-\mu)(x^{(i)}-\mu)^T U)$$

The empirical covariance matrix $\Sigma = \frac{1}{N}\sum_{i=1}^{N} (x^{(i)}-\mu)(x^{(i)}-\mu)^T$

$$\max_{U^TU=I} \frac{1}{N}\sum_{i=1}^{N} \mathrm{Trace}(U^T(x^{(i)}-\mu)(x^{(i)}-\mu)^T U) = \max_{U^TU=I} \mathrm{Trace}(U^T\Sigma U) = \max_{U^TU=I} \sum_{k=1}^{K} u_k^T \Sigma u_k$$

The Lagrangian function $L(U,\Lambda_K) = \mathrm{Trace}(U^T\Sigma U) + \mathrm{Trace}(\Lambda_K^T(I-U^TU))$ where $\Lambda_K = \mathrm{diag}([\hat\lambda_1,\cdots,\hat\lambda_K])$

Let $\frac{\partial L(U,\Lambda_K)}{\partial U} = 2\Sigma U - 2U\Lambda_K = 0$

$\Sigma u_k = \hat\lambda_k u_k$, $k=1,\cdots,K$

u_k is eigenvector , $\hat\lambda_k$ is eigenvalue

utilizing SVD decomposition, $\Sigma = Q\Lambda_D Q^T = \sum_{i=1}^{D} \lambda_i q_i q_i^T$ where $Q=[q_1,\cdots,q_D] \in R^{D\times D}$, q_i is the eigenvector corresponding to ith largest eigenvalue λ_i , $\Lambda_D = \mathrm{diag}([\lambda_1,\cdots,\lambda_D])$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$

$$\sum_{k=1}^{K} u_k^T\Sigma u_k = \sum_{k=1}^{K}\sum_{i=1}^{D} \lambda_i (u_k^T q_i)(q_i^T u_k) = \sum_{t\in T} \lambda_t$$

$T \subset \{1,2,\cdots,D\}$, $|T| = K$

we should pick top-K eigenvalues, correspondingly, the first K columns of Q are the optimal solution to U

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \begin{bmatrix} 6.9 \\ 3.5 \\ 5.1 \end{bmatrix}$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} (x_i-\mu)(x_i-\mu)^T = \begin{bmatrix} 2.09 & 1.45 & -0.39 \\ 1.45 & 2.25 & -1.15 \\ -0.39 & -1.15 & 7.09 \end{bmatrix}$$

SVD decomposition $\Sigma = Q\Lambda_D Q^T = \begin{bmatrix} -0.1376 & 0.6990 & 0.7017 \\ -0.2505 & 0.6609 & -0.7075 \\ 0.9583 & 0.2731 & -0.0842 \end{bmatrix} \begin{bmatrix} 7.4465 & 0 & 0 \\ 0 & 3.3085 & 0 \\ 0 & 0 & 0.6749 \end{bmatrix} \begin{bmatrix} -0.1376 & -0.2505 & 0.9583 \\ 0.6990 & 0.6609 & 0.2731 \\ 0.7017 & -0.7075 & -0.0842 \end{bmatrix}$

pick top 2 eigenvalues, $U = [q_1, q_2] = \begin{bmatrix} -0.1376 & 0.6990 \\ -0.2505 & 0.6609 \\ 0.9583 & 0.2731 \end{bmatrix}$

The new representation $z = U^T(x_i - \mu)$

$z_1 = \begin{bmatrix} -2.1514 \\ -0.1731 \end{bmatrix}$ $z_2 = \begin{bmatrix} 3.8042 \\ -2.8875 \end{bmatrix}$ $z_3 = \begin{bmatrix} 0.1532 \\ -0.9869 \end{bmatrix}$ $z_4 = \begin{bmatrix} -4.7065 \\ 1.3015 \end{bmatrix}$ $z_5 = \begin{bmatrix} 1.2938 \\ 2.2791 \end{bmatrix}$

$z_6 = \begin{bmatrix} 4.0993 \\ 0.1436 \end{bmatrix}$ $z_7 = \begin{bmatrix} -1.6258 \\ -2.2321 \end{bmatrix}$ $z_8 = \begin{bmatrix} 2.1145 \\ 3.2512 \end{bmatrix}$ $z_9 = \begin{bmatrix} -0.2348 \\ 0.3730 \end{bmatrix}$ $z_{10} = \begin{bmatrix} -2.7464 \\ -1.0689 \end{bmatrix}$

Programming Problems

1. PCA

(1) partition the dataset

```
X = dataset[:,:-1] is the features
y = dataset[:,-1] is the labels
```

(2) calculate the mean, covariance, SVD decomposition

```
mu = np.mean(X, axis=0) is the mean of dataset.
```

```
sigma = np.cov(X.T) is the covariancee matrix
```

```
u, s, v = np.linalg.svd(sigma) this is the process of SVD decompostion
```

```
U = u[:,:2] is the top 2 eigenvectors
```

(3) calculate the new representation

```
z = U.T @ (X-mu).T is the new representation of dataset X
```

2. K-means

(1) randomly pick 3 points as centers

```
initialCenter = np.random.choice(dataset.shape[0], size=3, replace=False)
x = z.T
center = (x[initialCenter[0]],x[initialCenter[1]],x[initialCenter[2]])
```

(2) repeat “assignment” and “refitting” until convergence

```
while True:
    center = refitting(x,labelList)
    if all(assignment(x,center)==labelList):
        break # converge
    labelList = assignment(x,center) every element corresponds to a label
```

3. Silhouette Coefficient

(1) For each data point i, let a_i be the mean distance between i and all other points in the same cluster, and let b_i be the mean distance between i and all other points in the next nearest cluster. The Silhouette Coefficient for data point i is defined as $s_i = (b_i - a_i)/\max(a_i, b_i)$. The overall silhouette coefficient is the average of s_i overall data points.

(2) silhouette coefficient: 0.4802142699427175

4. Rand Index

(1) Given two clusterings of these data points, let "a" be the number of pairs of elements that are in the same cluster in both clusterings, and let "b" be the number of pairs of elements that are in different clusters in both clusterings. The Rand Index is defined as $RI = 2(a+b)/(n(n-1))$ where n is the number of data.

(2) rand index: 0.8743677375256322