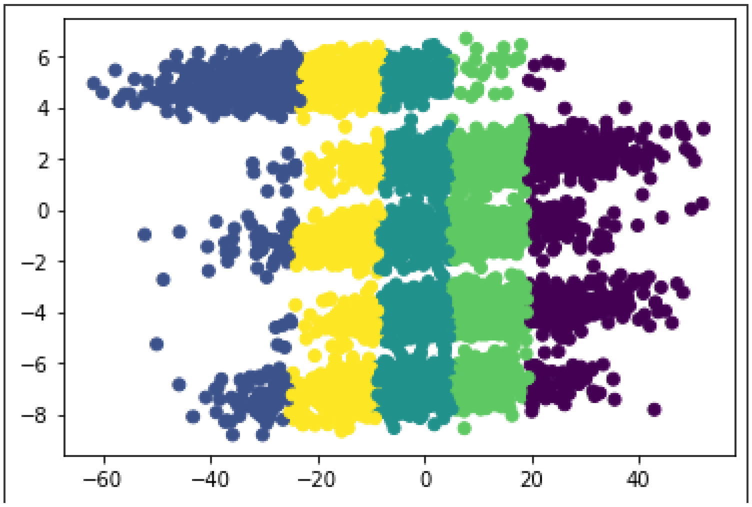
CS 6220 Homework 4

Haoping Lin

lin.haop@northeastern.edu

Question 1

1.b



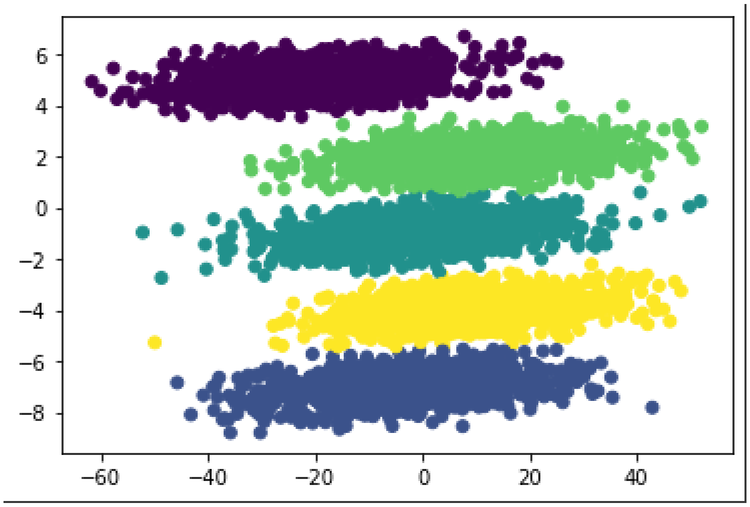
Classification Plot by K-means

1.c

From my observation on scatter plot, there are five groups of data horizontally lying on the graph. In an ideal case, each horizontal group should be classified as one class. So, an initialization of 5 points make sense here. However, it did not cluster very well, each cluster is vertical separated, not horizontal. The reason behind this is about variance. From comparison between axis, we can see that x-axis has a range from -60 to 50, while y-axis only has a range from -9 to 6. There are more variance along x-axis, in which case, classification based on normal distance is not reasonable. Therefore, even if we change initialization points, the final result will not be an ideal case as we want.

Quesiton 2

2.a

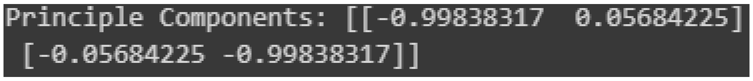


Classification Plot by K-means (Mahalanobis Version)

This time the cluster is in ideal condition, where five groups of data are classified horizontally. This is different from previous one because we use Mahalanobis Distance to find relationship between data.

2.b

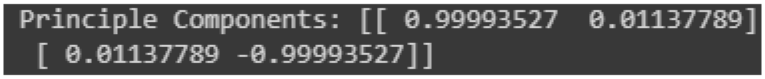
Principle components of aggregate data is:



```
Principle Components: [[-0.99838317  0.05684225]
 [-0.05684225 -0.99838317]]
```

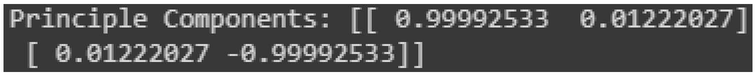
2.c

Principle components of aggregate data is:



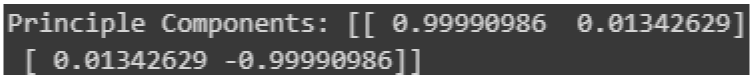
```
Principle Components: [[ 0.99993527  0.01137789]
 [ 0.01137789 -0.99993527]]
```

Cluster 1



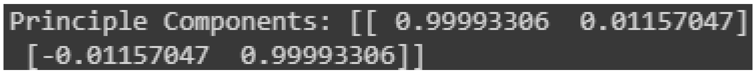
```
Principle Components: [[ 0.99992533  0.01222027]
 [ 0.01222027 -0.99992533]]
```

Cluster 2



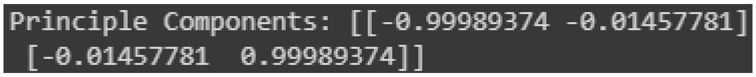
```
Principle Components: [[ 0.99990986  0.01342629]
 [ 0.01342629 -0.99990986]]
```

Cluster 3



```
Principle Components: [[ 0.99993306  0.01157047]
 [-0.01157047  0.99993306]]
```

Cluster 4

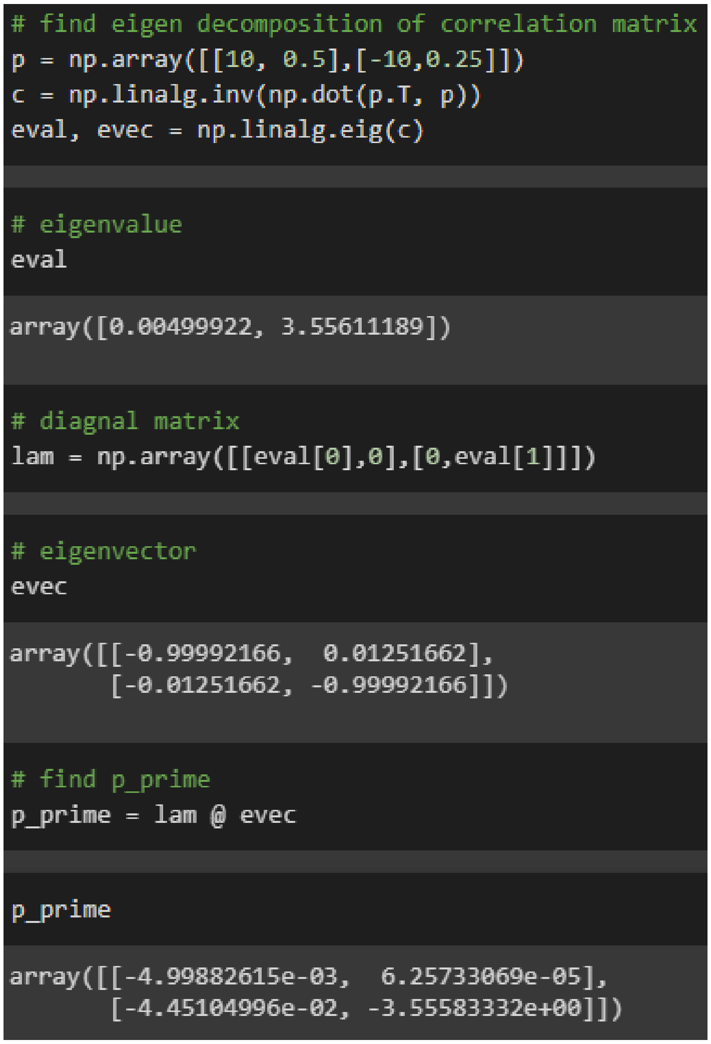


```
Principle Components: [[-0.99989374 -0.01457781]
 [-0.01457781  0.99989374]]
```

Cluster 5

There are not the exactly same, but they are almost same. Most of them share same direction and eigenvalues. However, their origins, or in other words, centroids are different.

2.d



```
# find eigen decomposition of correlation matrix
p = np.array([[10, 0.5],[-10,0.25]])
c = np.linalg.inv(np.dot(p.T, p))
eval, evec = np.linalg.eig(c)
```

```
# eigenvalue
eval
```

```
array([0.00499922, 3.55611189])
```

```
# diagnal matrix
lam = np.array([[eval[0],0],[0,eval[1]]])
```

```
# eigenvector
evec
```

```
array([[-0.99992166,  0.01251662],
       [-0.01251662, -0.99992166]])
```

```
# find p_prime
p_prime = lam @ evec
```

```
p_prime
```

```
array([[-4.99882615e-03,  6.25733069e-05],
       [-4.45104996e-02, -3.55583332e+00]])
```

I find p_prime by above steps. p_prime is the way we transfer data to an coordinate along principle component of data. This is the core of Mahalanobis distance, since we are calculating distance in an coordinate that x and y coordinate have same impact.

Q3

3a

F_3	→	C_1	→	C_4
(1, 2, 3)		(1) : 4		(1, 2, 3, 4) : 4
(1, 2, 4)		(2) : 5		(1, 2, 3, 5) : 3
(1, 2, 5)		(3) : 5		(1, 2, 4, 5) : 2
(1, 3, 4)		(4) : 4		(1, 3, 4, 5) : 2
(2, 3, 4)		(5) : 3		(2, 3, 4, 5) : 3
(2, 3, 5)				
(3, 4, 5)				

3b

F_3	→	C_4
(1, 2, 3)		(1, 2, 3, 4)
(1, 2, 4)		(1, 2, 3, 5)
(1, 2, 5)		(1, 2, 4, 5)
(1, 3, 4)		(2, 3, 4, 5)
(2, 3, 4)		
(2, 3, 5)		
(3, 4, 5)		

3c (1, 2, 3, 4)

Q4

4a. $R = 3^7 - 2^8 + 1 = 1932$

4b. $\text{Confidence} = \frac{\sigma(\{\text{Milk}, \text{Diapers}, \text{Butter}\})}{\sigma(\{\text{Milk}, \text{Diapers}\})} = \frac{2}{4} = 0.5$

4c. $\text{Support} = \frac{\sigma(\{\text{Milk}, \text{Diapers}, \text{Butter}\})}{|T|} = \frac{2}{10} = 0.2$

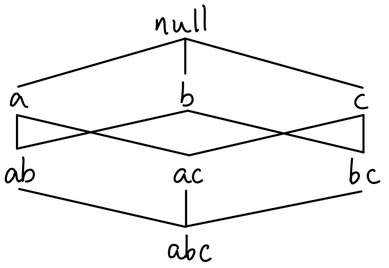
4d. True According to Apriori Principle, if an itemset is frequent, then its subset is too.

4e. False Apriori Principle cannot be used backward.

4f. False Imagine a case where {a, b} shows, {b, c} does not show. So, there could be at most 50 support for {b}

4g. False There can at most be 10 of size-2 frequent itemset. 4+3+2+1 = 10

4h.


null
a
b
c
ab
ac
bc
abc