Question 1

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more similar the data points are within the same cluster.

$$J = \sum_{i=1}^{m}\sum_{k=1}^{K} w_{ik}||x^i - \mu_k||^2$$

This is a minimization problem of two parts.

1. differenciate J w.r.t w_{ik} and update cluster assigments :

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m}\sum_{k=1}^{K} ||x^i - \mu_k||^2 \Longrightarrow w_{ik} = \begin{cases} 1 & k = \text{argmin}_j ||x^i - \mu_j||^2 \\ 0 & otherwise \end{cases}$$

   In other words, assign data to x_i to the closest cluster judged by its sum of squared distance from cluster's centroid.

2. differenciate J w.r.t μ_k and recompute the centroid after the cluster assigments from previous step :

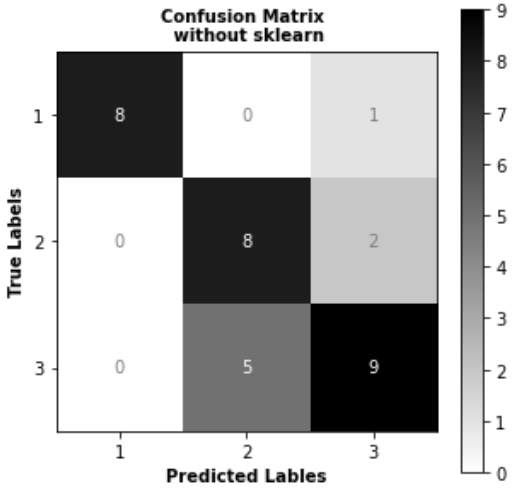
$$\frac{\partial J}{\partial \mu_k} = 2\sum_{i=1}^{m} w_{ik}(x^i - \mu_k) \Longrightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik}x^i}{\sum_{i=1}^{m} w_{ik}}$$

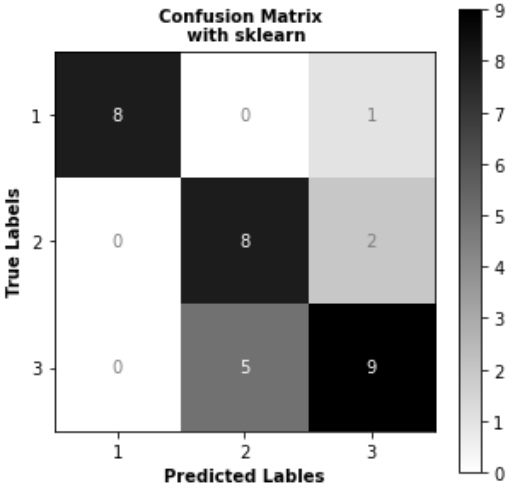
   In other words,recomputig the centroid of each cluster to reflect the new assigments.

Initial centers of clusters were random numbers scaled with variance and biased with mean of train data.

```
Accuracy of k-means with k = 3 is: % 75.76
```

```
Accuracy of k-means with k = 3 using sklearn is: % 75.76
```




Confusion Matrix
with sklearn
True Labels
Predicted Lables
1
2
3
8
0
1
0
8
2
0
5
9
9
8
7
6
5
4
3
2
1
0

Quesion 2

- Agglomerative : "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
  - Complete-Link : distance between two clusters to be the maximum distance between any single data point in the first cluster and any single data point in the second cluster.

$$cd(X,Y) = \underset{x \epsilon X, y \epsilon Y}{max} d(x,y)$$

1. step one :

$cd_{1,2} = 15.16 \quad , \quad cd_{1,3} = 11.22 \quad , \quad cd_{1,4} = 14.35 \quad , \quad cd_{1,5} = 5.47$
$cd_{2,3} = 17.72 \quad , \quad cd_{2,4} = 7.34 \quad , \quad cd_{2,5} = 13.92$
$cd_{3,4} = 12.08 \quad , \quad cd_{3,5} = 9.79$
$cd_{4,5} = 11.74$

$min(all\ \ cds) = 5.47 \Rightarrow merge : 1\ \ and\ \ 5 \rightarrow new\ \ cluster : [1,5]$

2. step two:

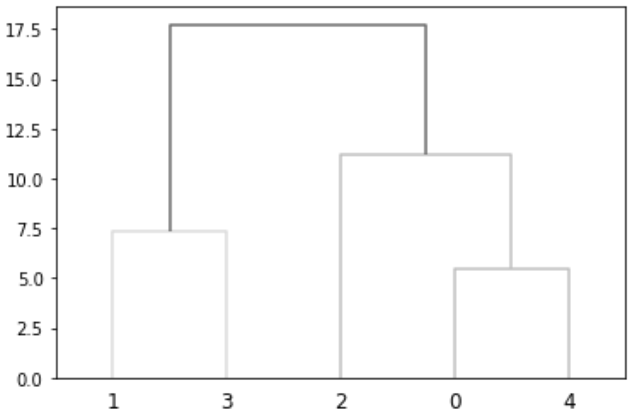
$cd_{2,3} = 17.72 \quad , \quad cd_{2,4} = 7.34 \quad , \quad cd_{3,4} = 12.08$
$cd_{2,[1,5]} = max(15.16\ \ ,\ \ 13.92) = 15.16$
$cd_{3,[1,5]} = max(11.22\ \ ,\ \ 9.79) = 11.22$
$cd_{4,[1,5]} = max(14.35\ \ ,\ \ 11.74) = 14.35$

$min(all\ \ cds) = 7.34 \Rightarrow merge : [1,5]\ \ and\ \ [2,4] \rightarrow new\ \ cluster$
$: [[1,\ \ 5],\ \ [2,\ \ 4]]$

3. step three:

$cd_{3,[1,5]} = max(11.22\ \ ,\ \ 9.79) = 11.22$
$cd_{3,[2,4]} = max(17.72\ \ ,\ \ 12.08) = 17.722$

$min(all\ \ cds) = 11.72 \Rightarrow merge : [1,5]\ \ and\ \ [3] \rightarrow new\ \ cluster$
$: [[[1,\ \ 5], [3]],\ \ [2,\ \ 4]]$



- Centroid:distance between two clusters is the distance between the two mean vectors of the clusters.

$$cd(X,Y) = d(avg(X), avg(Y))$$

1. step one :

$cd_{1,2} = 15.16 \quad , \quad cd_{1,3} = 11.22 \quad , \quad cd_{1,4} = 14.35 \quad , \quad cd_{1,5} = 5.47$
$cd_{2,3} = 17.72 \quad , \quad cd_{2,4} = 7.34 \quad , \quad cd_{2,5} = 13.92$
$cd_{3,4} = 12.08 \quad , \quad cd_{3,5} = 9.79$
$cd_{4,5} = 11.74$

$min(all\ \ cds) = 5.47 \Rightarrow merge : 1\ \ and\ \ 5 \rightarrow new\ \ cluster : [1,5]$

2. step two:

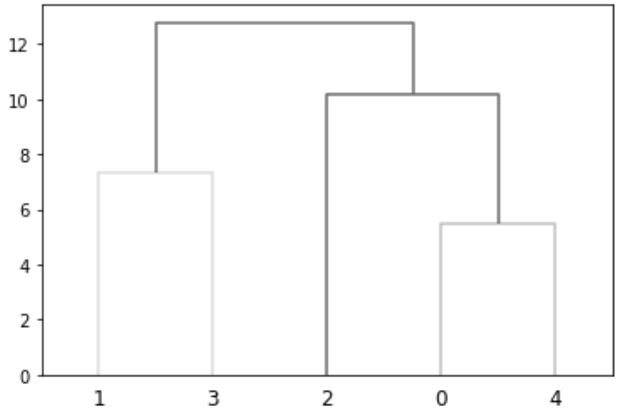
$cd_{2,3} = 17.72 \quad , \quad cd_{2,4} = 7.34 \quad , \quad cd_{3,4} = 12.08$
$cd_{2,[1,5]} = cd([-2,4,4]\ \ , \ \ avg([12,9,7]\ \ , \ \ [11,4,9]))$
$= cd([-2,4,4]\ \ , \ \ [11.5,6.5,8]) = 14.30$
$cd_{3,[1,5]} = cd([15,0,1]\ \ , \ \ [11.5,6.5,8]) = 10.17$
$cd_{4,[1,5]} = cd([15,0,1]\ \ , \ \ [11.5,6.5,8]) = 12.82$

$min(all\ \ cds) = 7.34 \Rightarrow merge : [1,5]\ \ and\ \ [2,4] \rightarrow new\ \ cluster$
$: [[1,\ \ 5],\ \ [2,\ \ 4]]$

3. step three:

$cd_{3,[1,5]} = cd([15,0,1]\ \ , \ \ [11.5,6.5,8]) = 10.17$
$cd_{3,[2,4]} = cd([15,0,1]\ \ , \ \ [11.5,6.5,8]) = 14.71$

$min(all\ \ cds) = 10.17 \Rightarrow merge : [1,5]\ \ and\ \ [3] \rightarrow new\ \ cluster$
$: [[[1,\ \ 5],[3]],\ \ [2,\ \ 4]]$



- Divisive : "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.
  - Single-link : the distance between two clusters as the minimum distance between any single data point in the first cluster and any single data point in the second cluster.

$$cd(X,Y) = \min_{x \epsilon X, y \epsilon Y} d(x,y)$$

1. step one :

$cd(1,[2,3,4,5]) = min(d_{1,2}\ ,\ d_{1,3}\ ,\ d_{1,4}\ ,\ d_{1,5}) = d_{1,5} = 5.47$

$cd(2,[1,3,4,5]) = min(d_{2,1}\ ,\ d_{2,3}\ ,\ d_{2,4}\ ,\ d_{2,5}) = d_{2,4} = 7.34$

$cd(3,[1,2,4,5]) = min(d_{3,1}\ ,\ d_{3,2}\ ,\ d_{3,4}\ ,\ d_{3,5}) = d_{3,5} = 9.79$

$cd(4,[1,2,3,5]) = min(d_{4,1}\ ,\ d_{4,2}\ ,\ d_{4,3}\ ,\ d_{4,5}) = d_{2,4} = 7.34$

$cd(5,[1,2,3,4]) = min(d_{5,1}\ ,\ d_{5,2}\ ,\ d_{5,3}\ ,\ d_{5,4}) = d_{1,5} = 5.47$

$cd([1,2],[3,4,5]) = min(d_{1,3}\ ,\ d_{1,3}\ ,\ d_{1,4}\ ,\ d_{1,5}, \dots) = d_{1,5} = 5.47$

$cd([1,3],[2,4,5]) = d_{1,5} = 5.47$

$cd([1,4],[2,3,5]) = d_{1,5} = 5.47$

$cd([1,5],[2,3,4]) = d_{3,5} = 9.79$

$cd([2,3],[1,4,5]) = d_{2,4} = 7.34$

$cd([2,4],[1,3,5]) = d_{4,5} = 11.74$

$cd([2,5],[1,3,4]) = d_{2,4} = 7.34$

$cd([3,4],[1,2,5]) = d_{2,4} = 7.34$

$cd([3,5],[1,2,4]) = d_{5,1} = 5.47$

$cd([4,5],[1,2,3]) = d_{5,1} = 5.47$

$max(all\ \ cds) = 11.74 \Rightarrow split : [2,4]\ \ and\ \ [1,3,5] \rightarrow new\ \ cluster$
$: [[2,4]\ \ ,\ \ [1,3,5]]$

2. step two :

$cd_{2,4} = 7.34$

$cd(1,[3,5]) = min(d_{1,3}\ \ ,\ \ d_{1,5}) = d_{1,5} = 5.47$
$cd(3,[1,5]) = min(d_{1,3}\ \ ,\ \ d_{3,5}) = d_{3,5} = 9.79$
$cd(5,[1,3]) = min(d_{1,5}\ \ ,\ \ d_{3,5}) = d_{1,5} = 5.47$

$max(all\ \ cds) = 9.79 \Rightarrow split : [3]\ \ and\ \ [1,5] \rightarrow new\ \ cluster$
$: [[2,4]\ \ ,\ \ [[1],[3,5]]]$

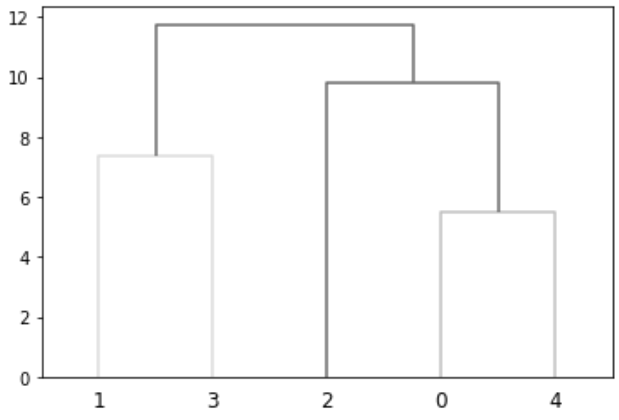
A. step three:

$cd_{2,4} = 7.34$
$cd_{1,5} = 5.47$

$max(all\ \ cds) = 7.34 \Rightarrow split : [2]\ \ and\ \ [4] \rightarrow new\ \ cluster$
$: [[[2],[4]]\ \ ,\ \ [[1],[3,5]]]$

B. step four :

$split : [3]\ \ and\ \ [5] \rightarrow new\ \ cluster : [[[2],[4]]\ \ ,\ \ [[1],[3],[5]]]$


12
10
8
6
4
2
0
1
3
2
0
4

Question 3

part 1.

$p(w_1|x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})$

$p(w_2|x) = \frac{1}{2\sqrt{2\pi}} exp(-\frac{(x-4)^2}{8})$

$p(w_1) = p(w_2) = \frac{1}{2}$

$$x_0 : 2exp(-\frac{x^2}{2}) = exp(-\frac{(x-4)^2}{8}) \xrightarrow{ln(.)} ln(2) - \frac{x^2}{2} = -\frac{(x-4)^2}{8} \rightarrow 8ln(2) - 4x^2 = \\ -x^2 + 8x - 16 \rightarrow 3x^2 + 8x - (16 + 8ln2) = 0 \\ \implies x_0 = 1.65$$

part 2.

$p(w_1|x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})$

$p(w_2|x) = \frac{1}{2\sqrt{2\pi}} exp(-\frac{(x-4)^2}{8})$

$p'(w_1) = \frac{1}{2}(\lambda_{12} - \lambda_{22}) = \frac{1}{2} \times 2 = 1$

$p'(w_2) = \frac{1}{2}(\lambda_{21} - \lambda_{11}) = \frac{1}{2} \times 1 = \frac{1}{2}$

$$\bar{x_0} : 2exp(-\frac{x^2}{2}) = exp(-\frac{(x-4)^2}{8}) \xrightarrow{ln(.)} ln(4) - \frac{x^2}{2} = -\frac{(x-4)^2}{8} \rightarrow 8ln(4) - 4x^2 = \\ -x^2 + 8x - 16 \rightarrow 3x^2 + 8x - (16 + 8ln4) = 0 \\ \implies x_0 = 1.95$$

$$\Rightarrow \textit{as expected, } \bar{x_0} \textit{ moves to the right of } x_0 = 1.65 \textit{ to minimize the risk}$$

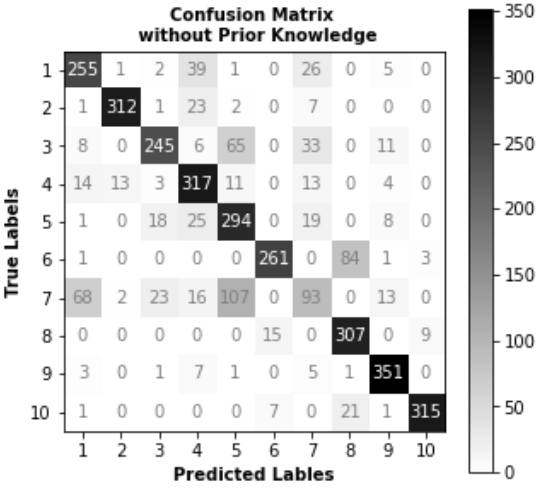
Given that the risk of choosing class 2 is higher if the correct choice is class 1, x_0 must move to the right to reduce the risk and probability of error for class 1, in other words, the area of class 1 is larger.

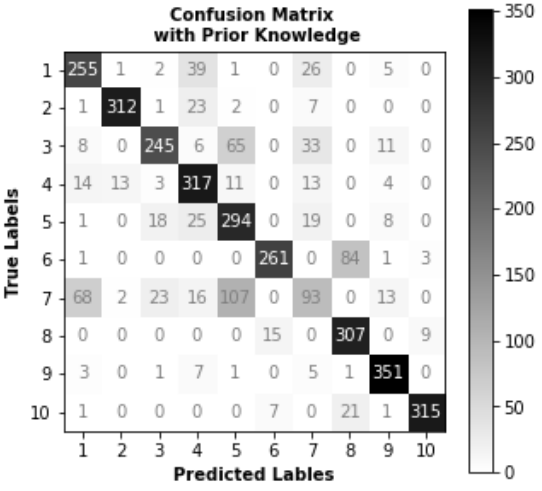
Question 4

The Optimal Bayes classifier chooses the class that has greatest a posteriori probability of occurrence (so called maximum a posteriori estimation, or MAP). It can be shown that of all classifiers, the Optimal Bayes classifier is the one that will have the lowest probability of miss classifying an observation, i.e. the lowest probability of error. So if we know the posterior distribution, then using the Bayes classifier is as good as it gets.

```
Accuracy of without Prior Knowledge is: % 78.57

Accuracy of with Prior Knowledge is: % 78.57
```





Prior probabilities are as below:

```
array([0.0998, 0.1049, 0.1024, 0.0994, 0.0975, 0.0993, 0.0946, 0.1008,
       0.1002, 0.1011])
```

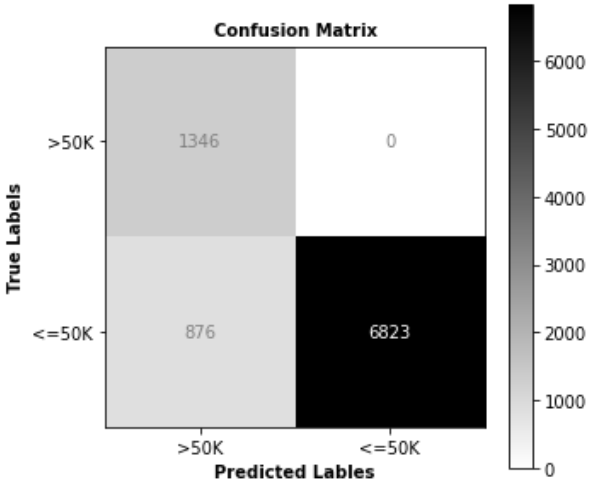
Because the probabilities of the classes in our prior knowledge are very close to each other and close to 0.1, estimation with this knowledge does not make a difference in our accuracy and confusion matrix.

Question 5

Naive Bayes is a family of algorithms based on applying Bayes theorem with a strong(naive) assumption, that every feature is independent of the others, in order to predict the category of a given sample. They are probabilistic classifiers, therefore will calculate the probability of each category using Bayes theorem, and the category with the highest probability will be output.

It is problematic when a frequency-based probability is zero, because it will wipe out all the information in the other probabilities. A solution would be Laplace smoothing , which is a technique for smoothing categorical data. A small-sample correction, or pseudo-count, will be incorporated in every probability estimate. Consequently, no probability will be zero. this is a way of regularizing Naive Bayes, I added 1 to every probability.

```
Accuracy of clasification using Naive Bayes with laplace smoothing is: % 90.32
```



Note: datas with missing attributes were omitted.