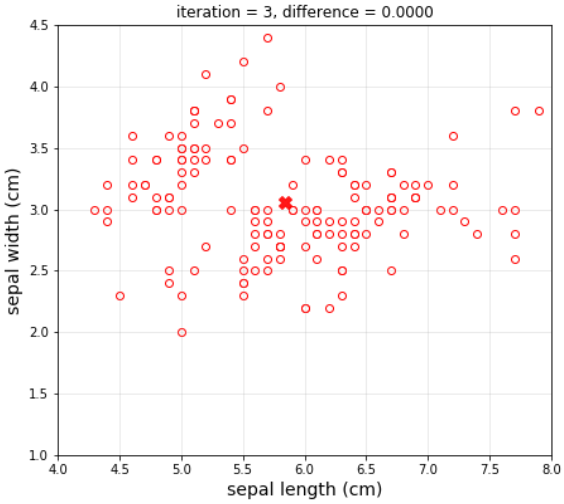
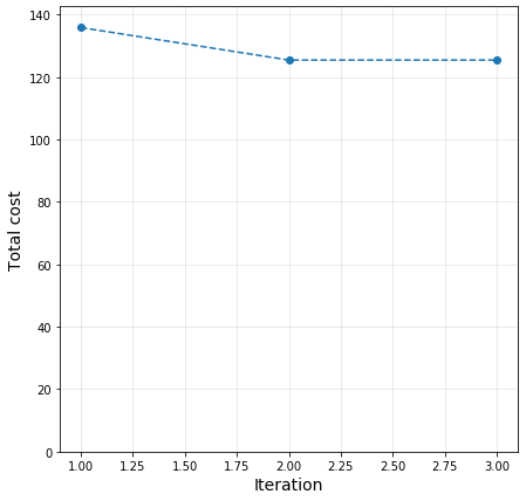
Cara Evangeline Chandra Mohan
190539421

Data Mining - Lab Assignment 5

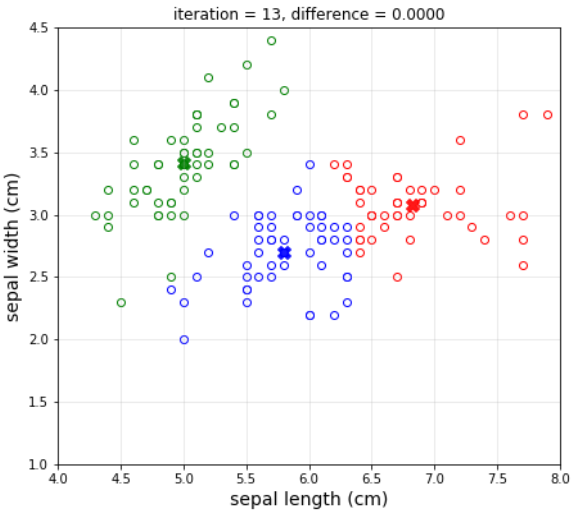
Exercise 0: What do you observe about the dependence of the final cluster quality in terms of total distance on the number of clusters K used? Why?

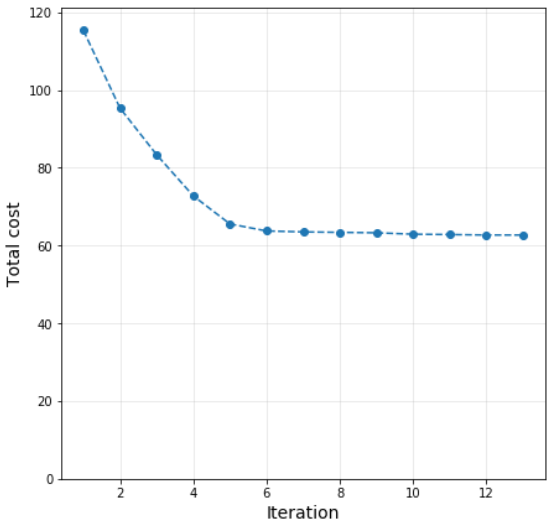
For K = 2



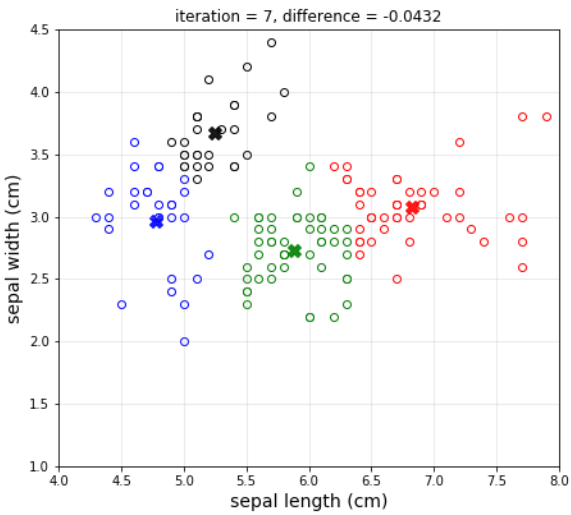


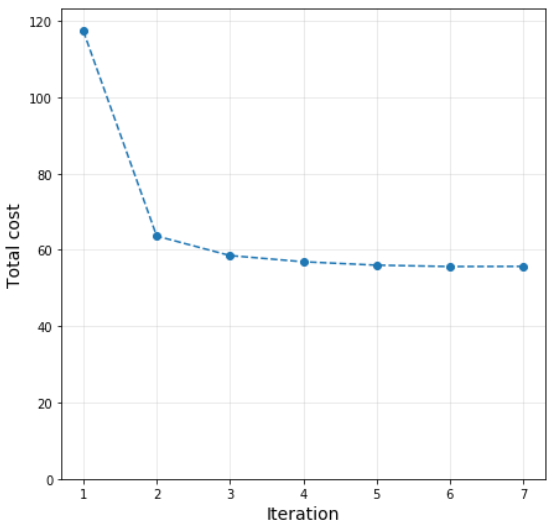
For K = 3





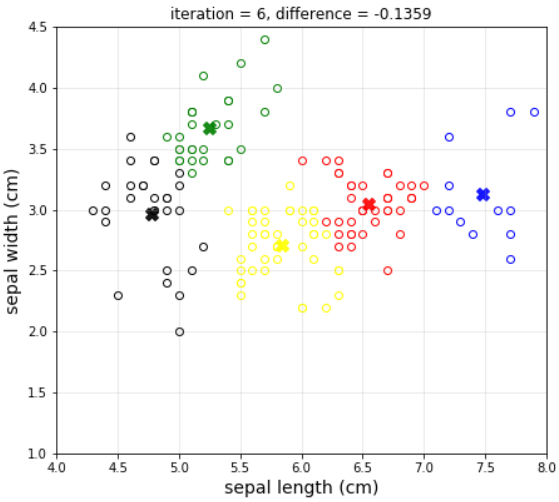
For K = 4

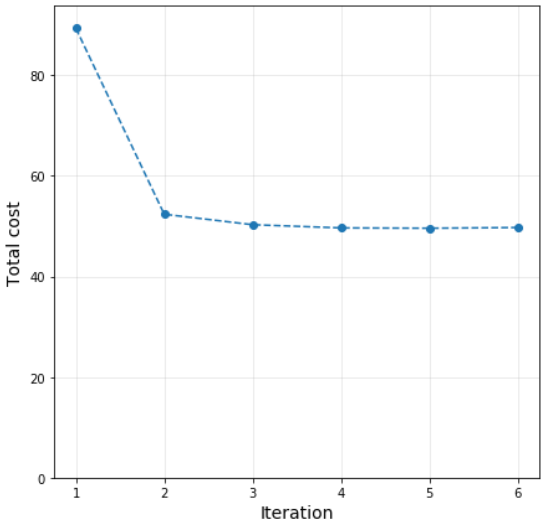




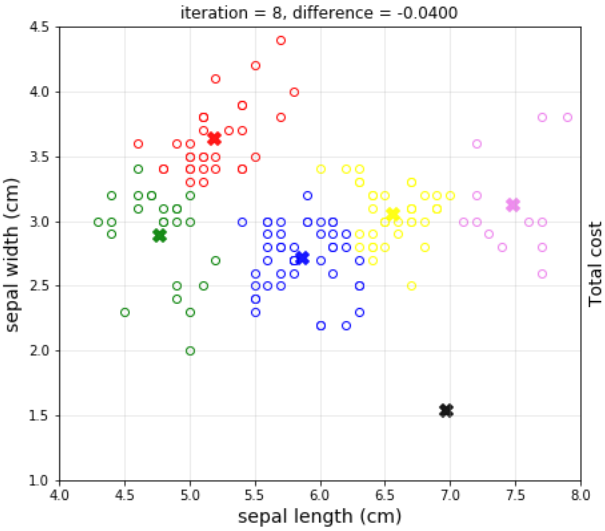
Cara Evangeline Chandra Mohan
190539421

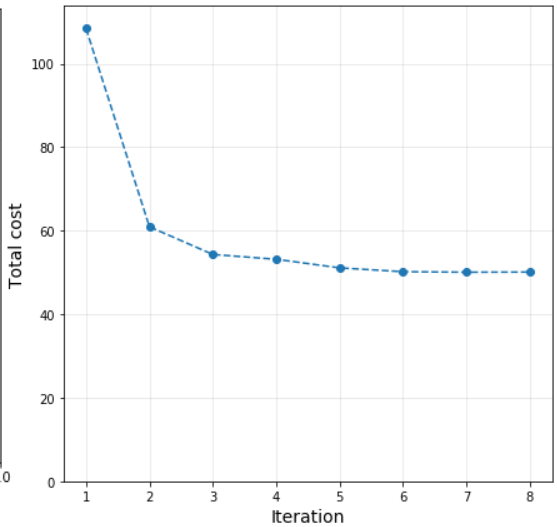
For K = 5



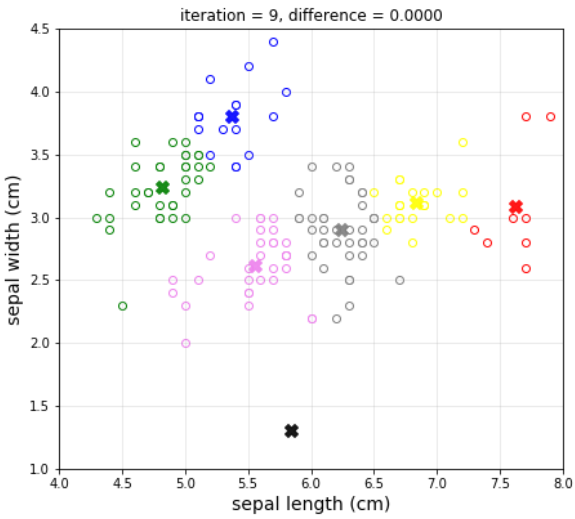


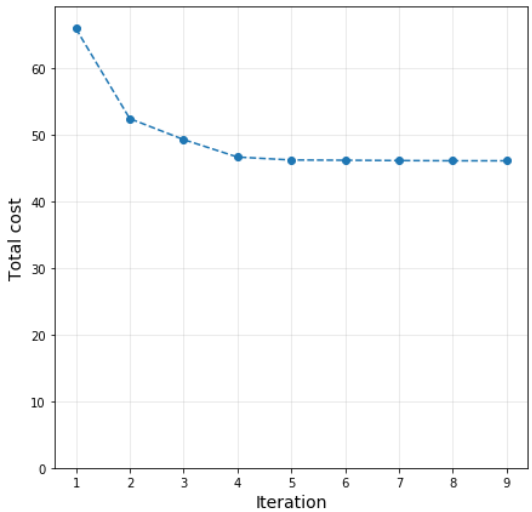
For K = 6





For K = 7



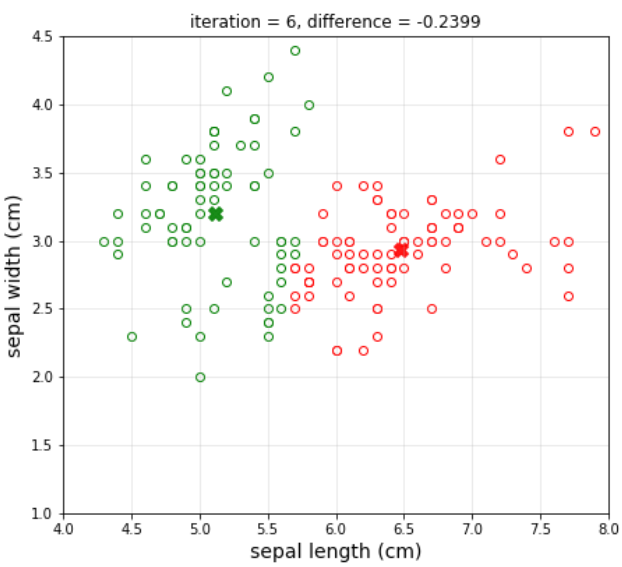


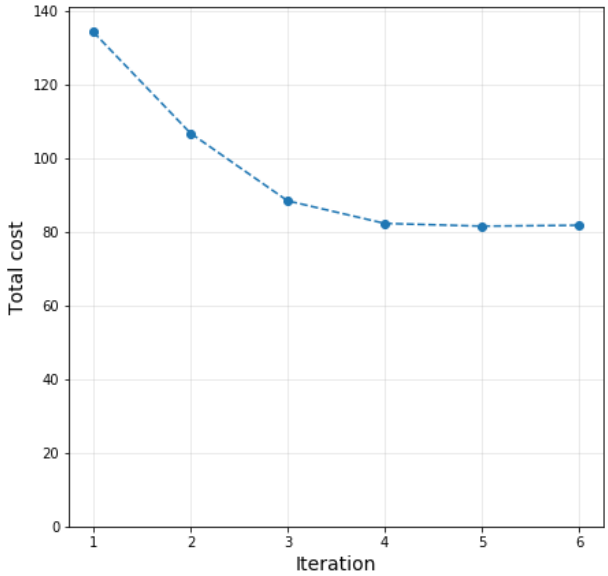
As we can see from the above figures and iteration vs cost graphs, when the number of clusters is increasing, the total distance of every point to its corresponding cluster is decreasing because the model is trying to fit as the number of clusters increases. To see how many clusters is good to choose, we can use elbow method. Here, increasing the number of clusters can give less cost

function but it may lead to overfitting of data.

Exercise 1: Find a seed that gives a different final quality of clusters (in terms of total distance). Include both the values of the seed, the final distance and the picture of the cluster with your answer.

```
np.random.seed(seed = 2)
```





The final distance for the seed value of 2 is 81.79890607855566
When seed = 2, the number of clusters formed is 2 whereas with other values of seed greater than 2 we get 3 clusters formed.

Exercise 2: Has the clustering accuracy improved from before? Why?

Yes, the accuracy has improved from before
For 2 dimensions : - Accuarcy : 0.8133333333333334
For 4 dimensions : - Accuracy : 0.8933333333333333
The accuracy when d = 4, is more because the number of features used to cluster the dataset has increased.

Exercise 3: The following cell contains a function that given a classifier and a threshold and some (test) samples, returns the TPR and FPR. Use this functoin to try a bunch of thresholds and fill in the TPR and FPR vectors. Plot TPR (y-axis) against FPR (x-axis) to visualise the resulting ROC curve. Provide both your code and the figure.

Code:

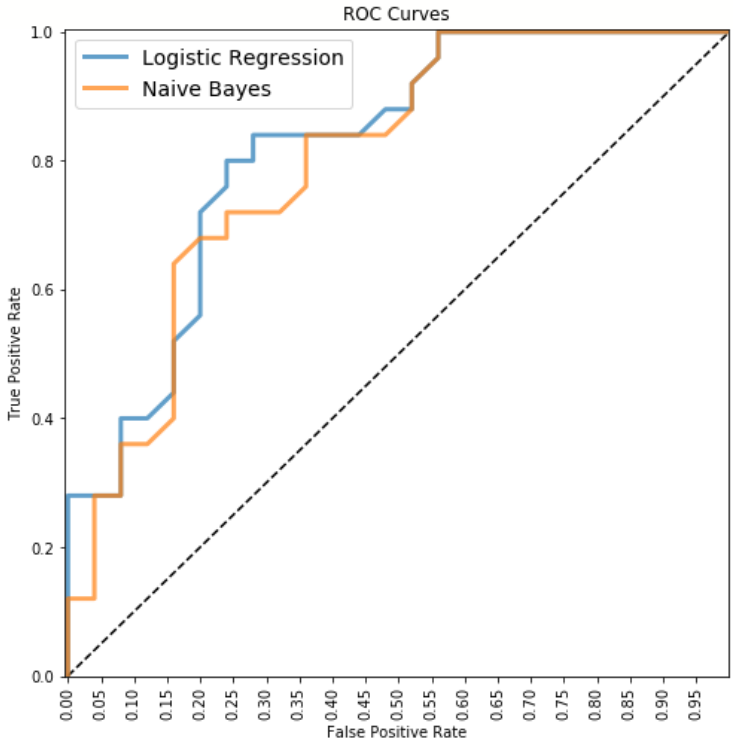
```
import numpy as np
TPR_LR = [] * 10000
```

Cara Evangeline Chandra Mohan
190539421

```
FPR_LR = [] * 10000
for i in np.linspace(0,1,10000):
 TP, FP = compute_tpr_fpr(LR_classifier, i, Xte, Yte)
 TPR_LR.append(TP)
 FPR_LR.append(FP)

TPR_NB = [] * 10000
FPR_NB = [] * 10000
for i in np.linspace(0,1,10000):
 TP1, FP1 = compute_tpr_fpr(NB_classifier, i, Xte, Yte)
 TPR_NB.append(TP1)
 FPR_NB.append(FP1)
```

Figure:



Exercise 4: Compare the AUC of the ROCs of the two classifiers. Which one is preferable by the AUC metric?

- AUC of Logistic regression is: 0.8200000000000001
- AUC of Naive Bayes is: 0.7991999999999999

According to Area Under Curve metrics, Logistic Regression is preferred to Naive bayes. It is

better to have a classifier whose AUC is tending to 1. The AUC of Logistic Regression classifier is more than the AUC of Naive bayes, it simply means that for a given FPR, in most of the cases Logistic Regression has maximum TPR than Naive Bayes and hence Logistic Regression is preferred.

Exercise 5: Suppose for a particular application, the maximum allowed FPR is 0.16. Which classifier is preferable? Obtains the maximum TPR given this FPR constraint?

Naive Bayes classifier is preferred in comparison to Logistic classifier.
For FPR < = 16,
Naive Bayes --> MaximumTPR = 0.64
Logistic Regression --> Maximum TPR = 0.52
So, here the classifier which can give more true positives is preferred to the one that gives less true positives and hence Naive Bayes classifier is preferred.