Queen Mary
University of London

Data Mining

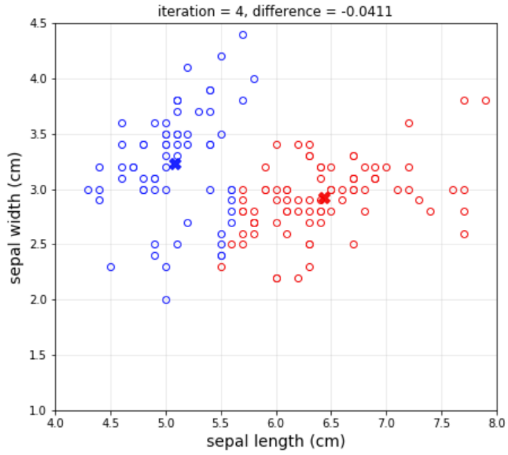
Assignment5

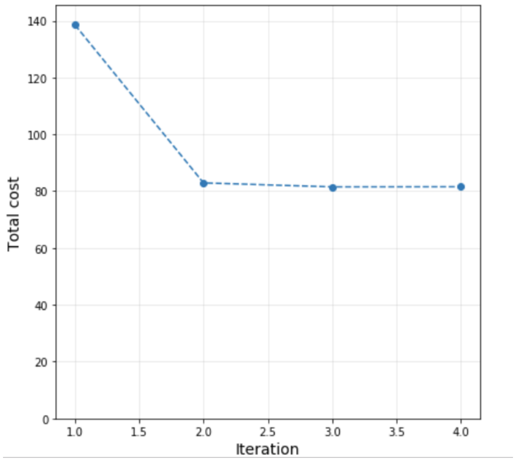
Spyridon Roumpis

181004877

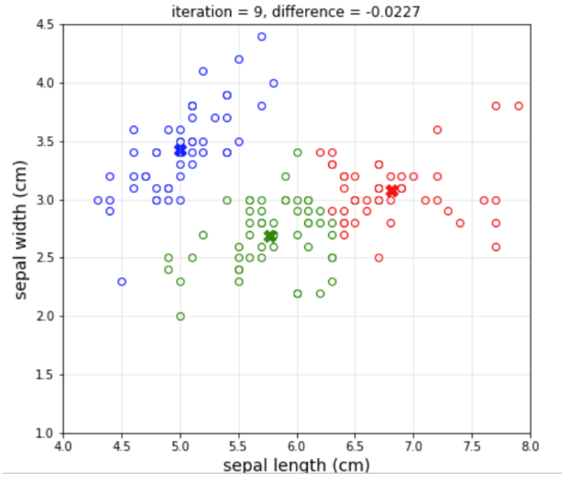
Exercise 0: What do you observe about the dependence of the final cluster quality in terms of total distance on the number of clusters K used? Why?

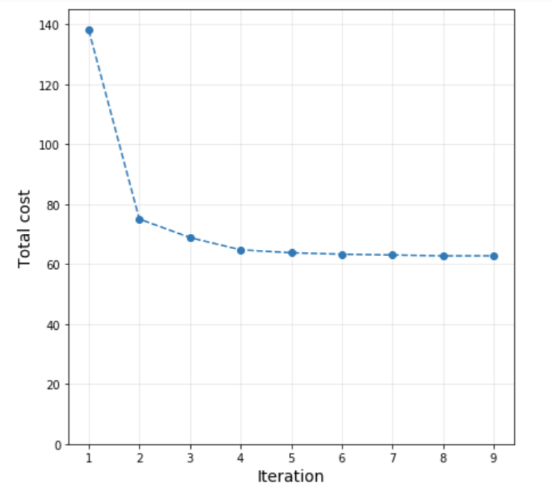
- K=2



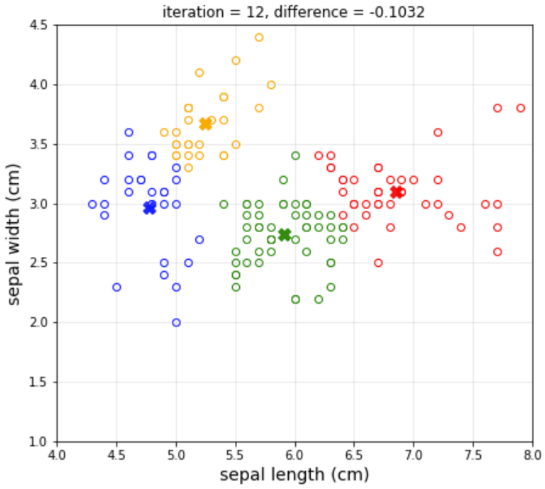


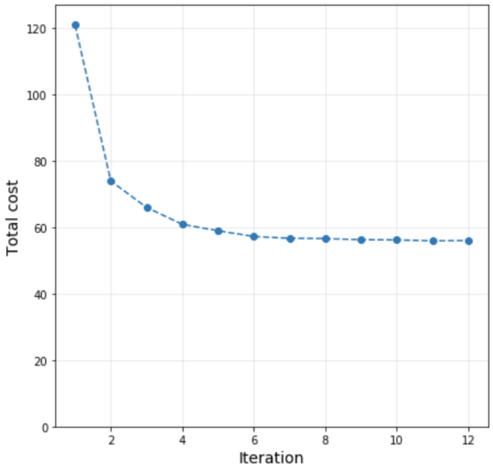
- K=3



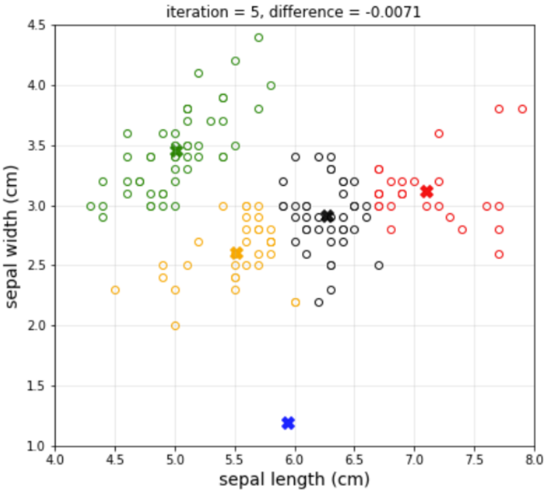


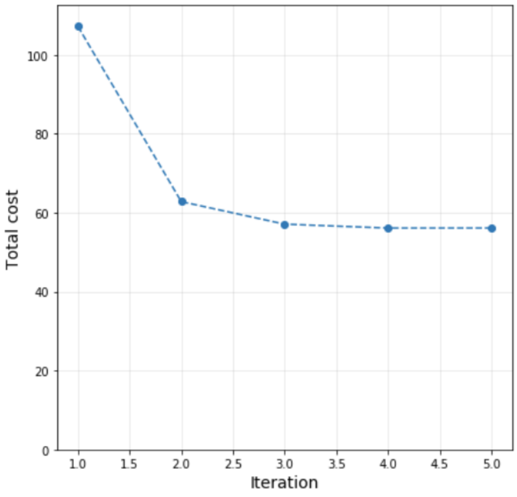
- K=4


iteration = 12, difference = -0.1032
sepal width (cm)
sepal length (cm)
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

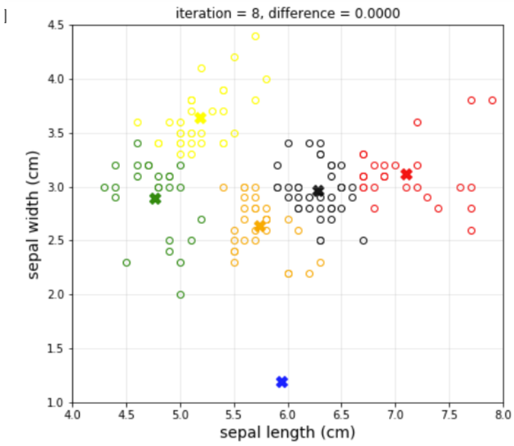

Total cost
Iteration
120
100
80
60
40
20
0
2
4
6
8
10
12

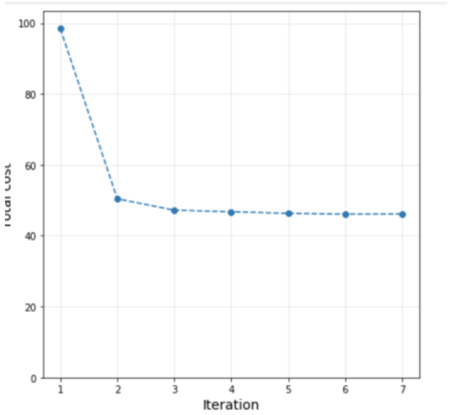
- K=5


iteration = 5, difference = -0.0071
sepal width (cm)
sepal length (cm)
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

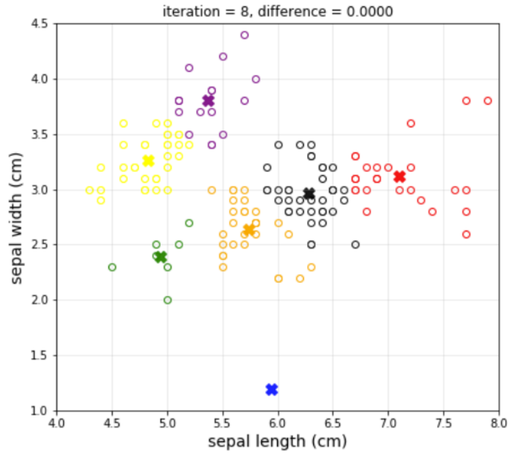

Total cost
Iteration
100
80
60
40
20
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

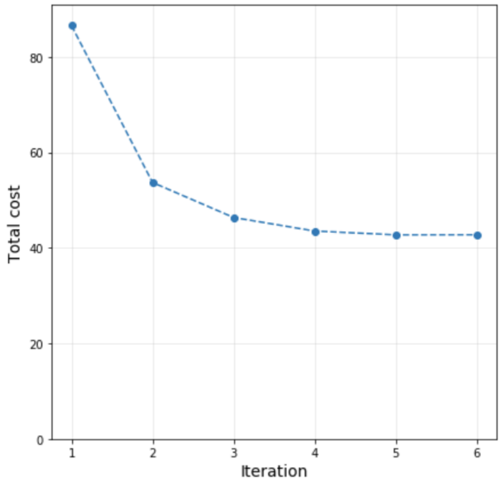
- K=6





- K=7

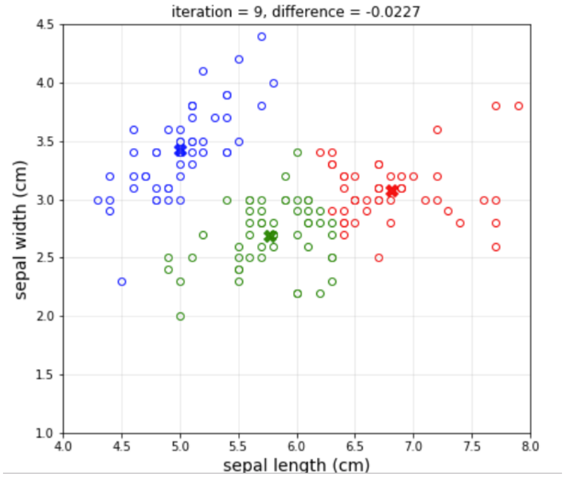


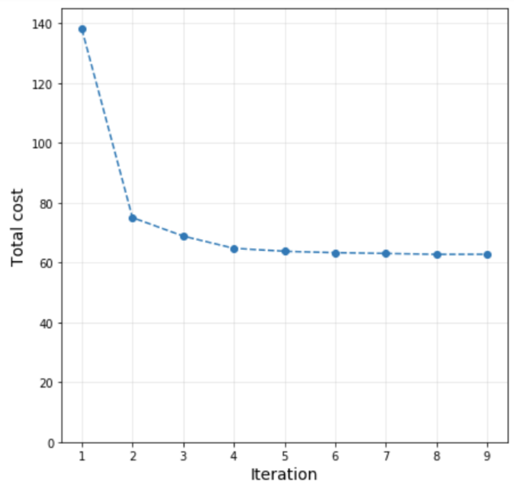


The more the iterations are the better the quality of the cluster we receive, which means that the cluster centres distances are decreased. Another observation here is that for both k=6 and k=7 the total distance in the end is zero although too many clusters makes our model overfitting and not generalize the data.

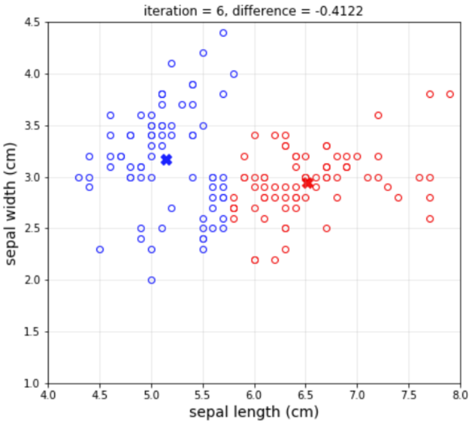
Exercise 1: Find a seed that gives a different final quality of clusters (in terms of total distance). Include both the values of the seed, the final distance and the picture of the cluster with your answer

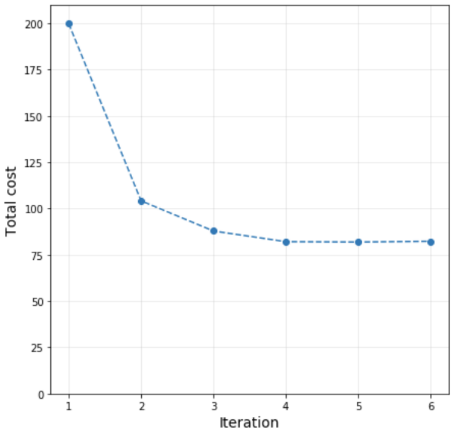
- K=3, seed=3



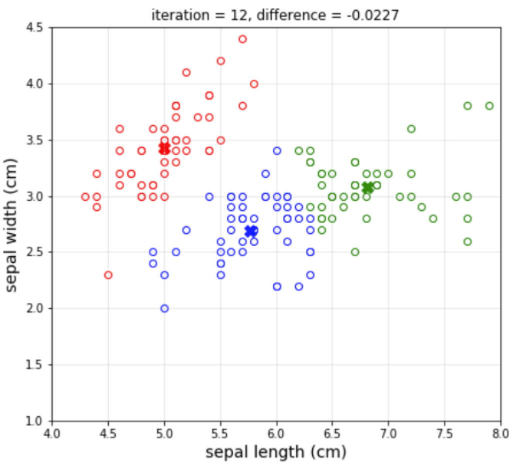


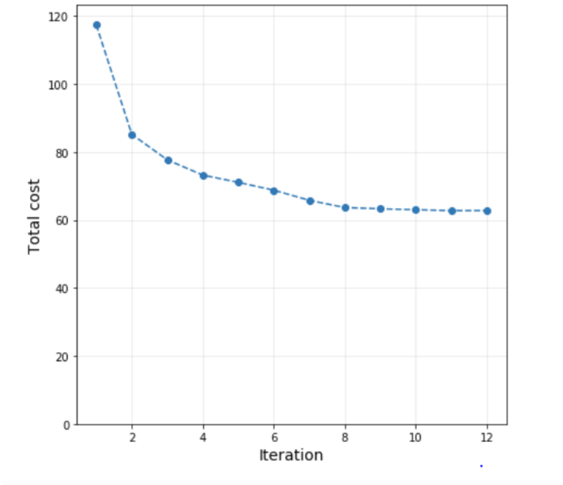
- K=3, seed=8





- K=3, seed=18





Every time we select a different value for the seed the initialization of it is random. It can be seen for the second try (when seed =8) that even we have 3 clusters only 2 can be seen on the plot and that's because the random initialization.

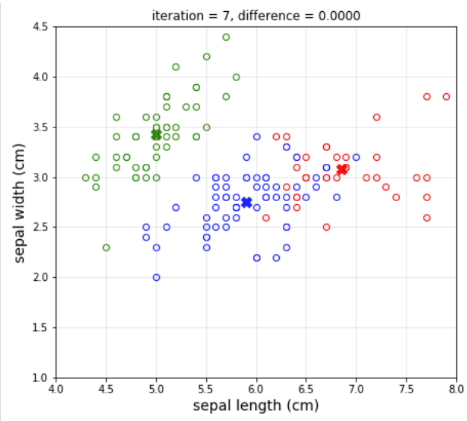
Exercise 2: Has the clustering accuracy improved from before? Why?

Accuracy before = 0.82

Accuracy after = 0.8933333333333333

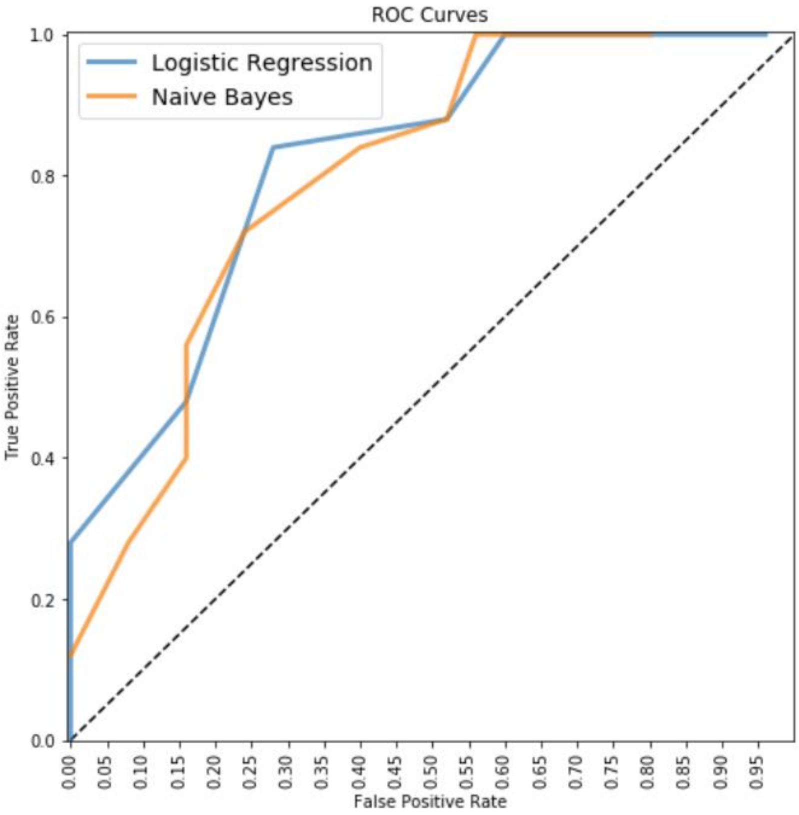
It can be seen from the values of the accuracy above that it has been improved since we already have the labels and used that as a way to evaluate the quality of clustering!

Bonus: Edit the visualization to visualize three of the K-means clustering dimensions instead of just 2



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.

For this exercise 10 different values of the thresholds were used, from 0.1 till 1. Below it can be seen the plot of the ROC curve for the Logistic Regression model and the Naïve Bayes model.



Code:

```
FPR_LR=[]
TPR_LR=[]
FPR_NB=[]
TPR_NB=[]

for i in range(1, 10, 1):
  threshold = i/10
  TPRNB, FPRNB = compute_tpr_fpr(NB_classifier, threshold, Xte, Yte)
  TPRLR, FPRLR = compute_tpr_fpr(LR_classifier, threshold, Xte, Yte)

  FPR_LR.append(FPRLR)
  TPR_LR.append(TPRLR)
  FPR_NB.append(FPRNB)
  TPR_NB.append(TPRNB)
```

For the plot the code was given in the exercise.

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

From theory we know that the AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much the model is capable of distinguishing between classes, higher the AUC, better the model is at distinguishing between patients with disease and no disease.

The AUC of the ROC for the Logistic Regression classifier is 0.7816 while for the Naïve Bayes is 0.6000000000000001. As we mentioned before the higher the AUC, better the model is working so the Logistic Regression classifier is the preferable one.

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?

Given the following formulas $\mathrm{AUC} = \mathrm{SE}/2 + \mathrm{SP}/2$ and $\mathrm{SE} = \mathrm{TPR}$ and $\mathrm{FPR} = 1 - \mathrm{SP}$ we can compute for which model the TPR-Sensitivity is the best.

- Naïve Bayes
  0.60=SE/2+(1-0.16)/2=>1.2=SE+0.84=>SE=0.36
  For this model the maximum TPR, given that the maximum allowed FPR is 0.16, is 0.36

- Logistic Regression
  0.7816=SE/2+(1-0.16)/2=>1.5632=SE+0.84=>SE=0.7232
  For this model the maximum TPR, given that the maximum allowed FPR is 0.16, is 0.7232

The maximum TPR is given for the Logistic Regression classifier, so that is the most preferable.