MLCourse-LU

s2538334

Assignment 3

1 Visualizing clustering results

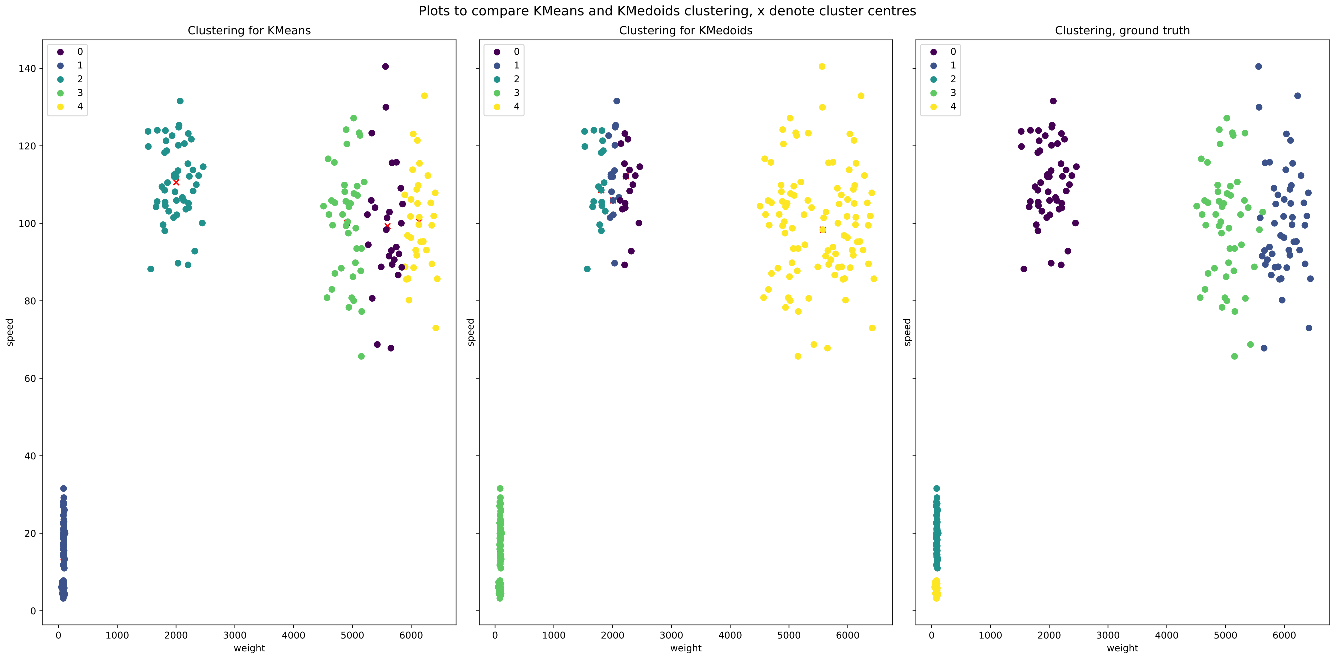

Figure 1: Plot of a dataset with labels compared to the fit of k-means and k-medoids. The colours in the plot refer to the label each datapoint has.

Analysis of figure 1. It can be seen that both models perform at best mediocre: the k-means model falsely assigns a label to the middle third set of datapoints in the right cluster, while the k-medoids model groups them all as one, and instead splits the upper left cluster into two groups. Both models fail to distinguish the difference in labels in the lower left cluster.

2 $k-$means vs. $k-$medoids

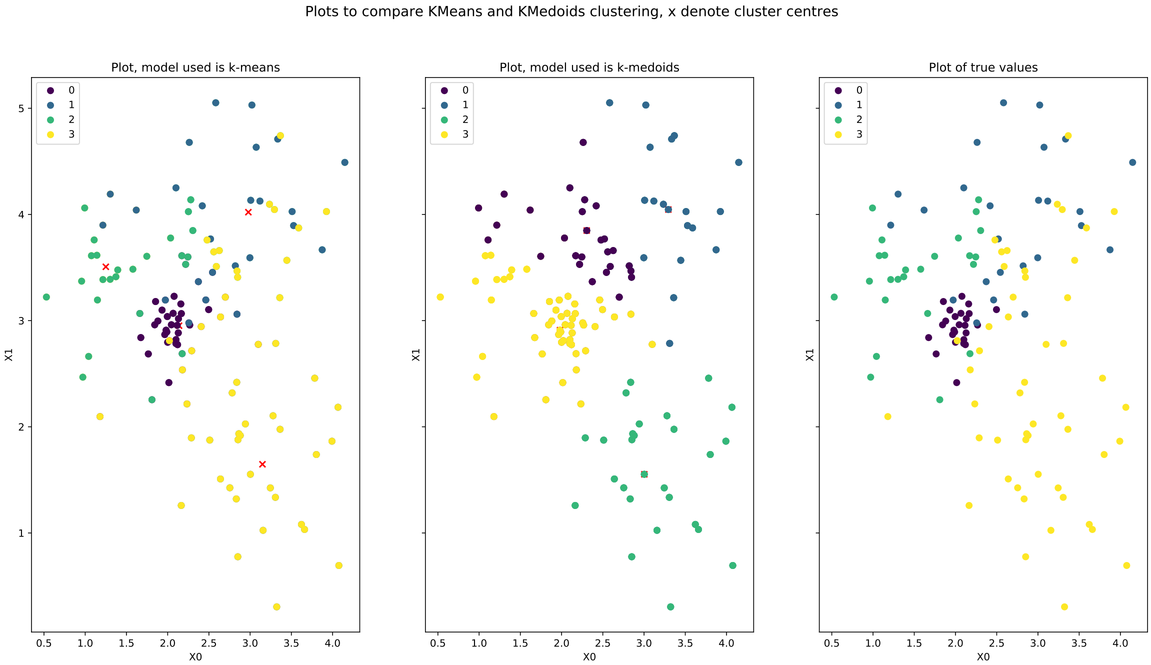

Figure 2: Plot of a another dataset with labels compared to the fit of k-means and k-medoids. We have used k=4, since in the ground truth dataset there are also four labels. The k-means model here is clearly superior: one can see that most, if not all labels correspond neatly to those of the ground truth, while the k-medoids model splits the dataset into a central cluster, surrounded by three surrounding clusters, failing to take into account the overlap in the central cluster between the labels.

Why k-means works best. The difference between k-means and k-medoids is that the centre of a cluster in k-means does not necessarily have to be an actual data point, while in k-medoids a cluster centre must also be an actual data point. Also the algortihm is slightly different: k-means builds cluster labels based upon the Euclidean distance of each point to the mean location of all the points in a cluster, while k-medoids does not necessarily take the distance but instead compare how similar two points are to each other. This means k-means works best if there is not a clearly defined centre, which is in this dataset the case for most of the clusters. This is also reflected in higher homogeneity and completeness scores, which can be seen in the table below.

A situation where k-means works better than k-medoids, and vice versa. For a dataset with spread-out points k-means would work better than k-medoids, for the aforementioned reasons. K-medoids would work better than k-means when there is a clearly defined centre with a significant amount of outliers, since k-means is more sensitive to scattered data points than k-medoids is.

Model	**Homogeneity**	**Completeness**	**k-value**
k-means	0.417	0.420	5
k-medoids	0.373	0.378	5

Table 1: Table of the homogeneity and completeness scores for each model.

3 Using the silhouette method to compare $k-$means vs. $k-$medoids on unlabeled data

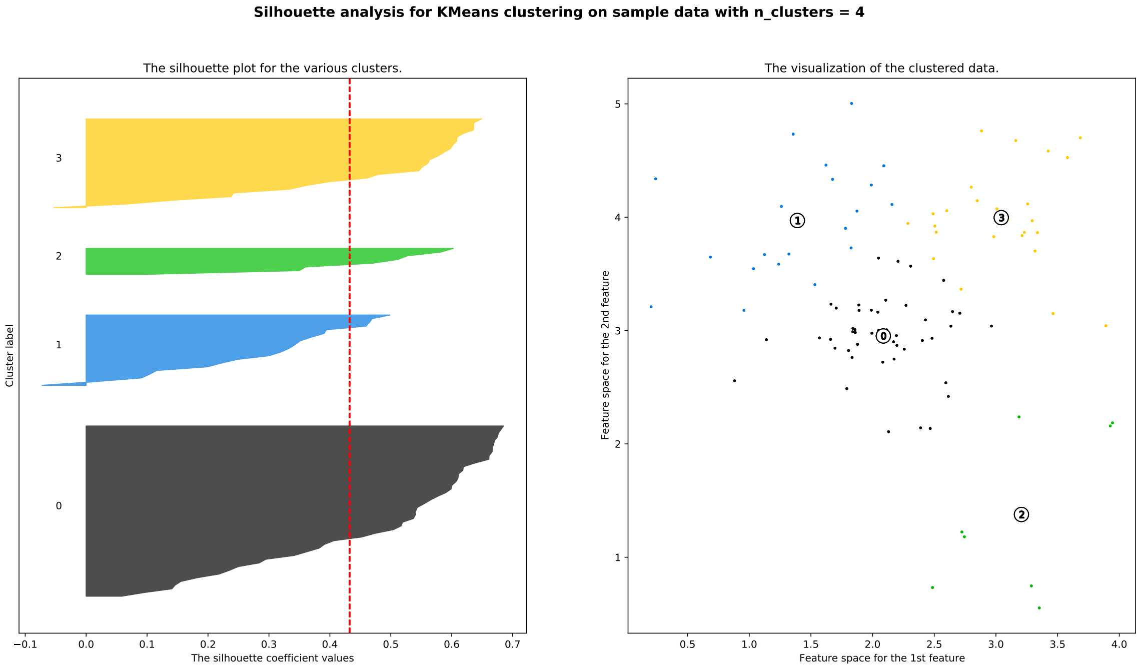

Figure 3: Silouette and scatter plot of the k-means method used to analyse dataset 3. Reported average silouette score is 0.483.

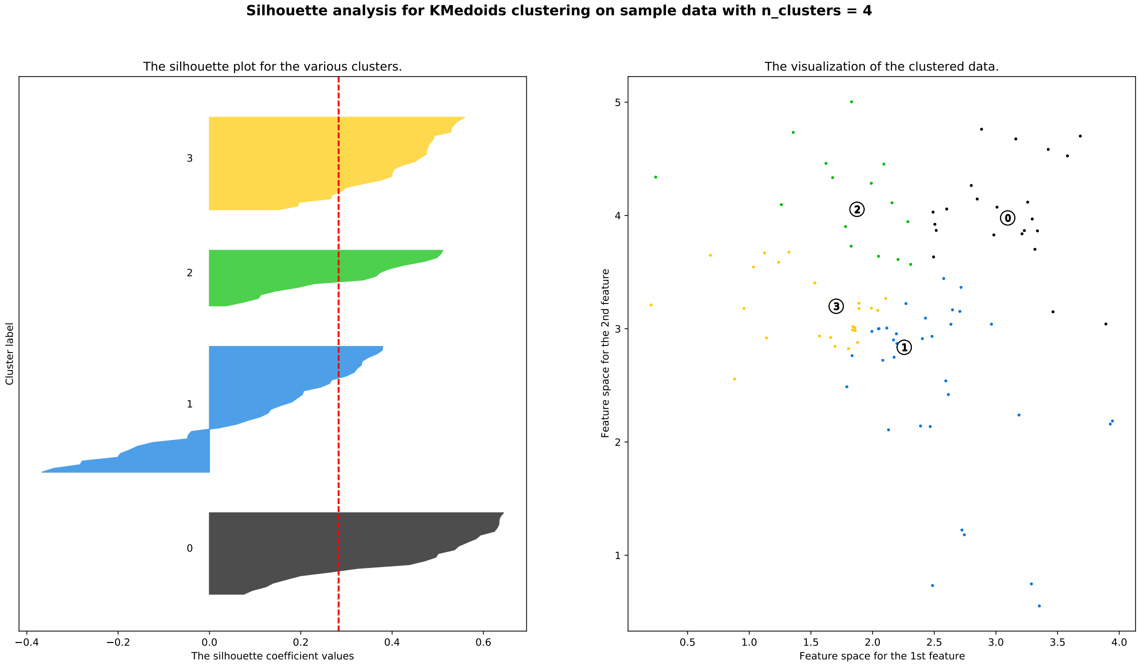

Figure 4: Silouette and scatter plot of the k-medoids method used to analyse dataset 3. Reported average silouette score is 0.283.

Analysis of the silouette plots. The silhouette score is a measure for how well-matched the points of a cluster are to that cluster. The average silhouette score for k-means is higher than for k-medoids, leading us to conclude that k-means is the better performing model here. This is supported by the scatter plot, because the lower right cluster is likely to be its own cluster (as can be seen with k-means) due to the average distance between any set of points than its part of central cluster.

4 The elbow method

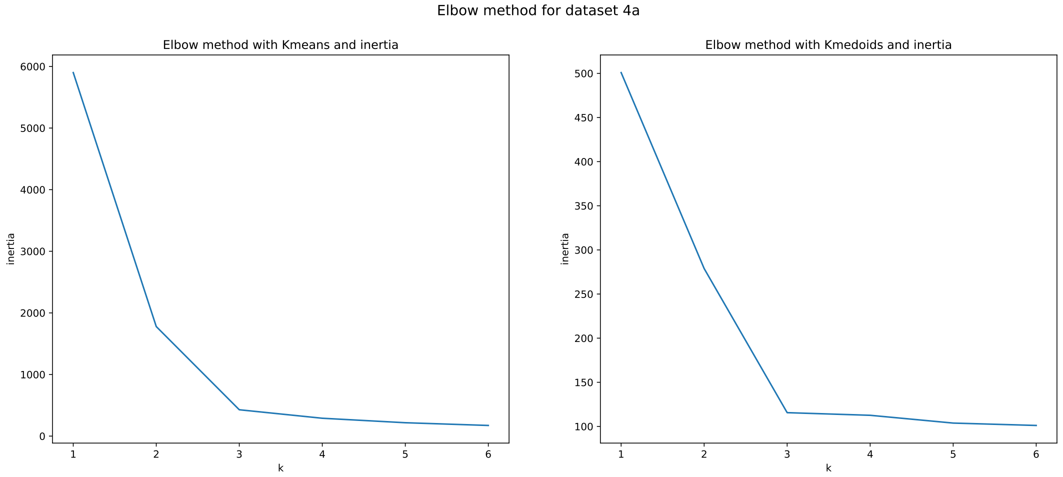

Figure 5: Elbow plot for dataset 4a for both k-means and k-medoids.

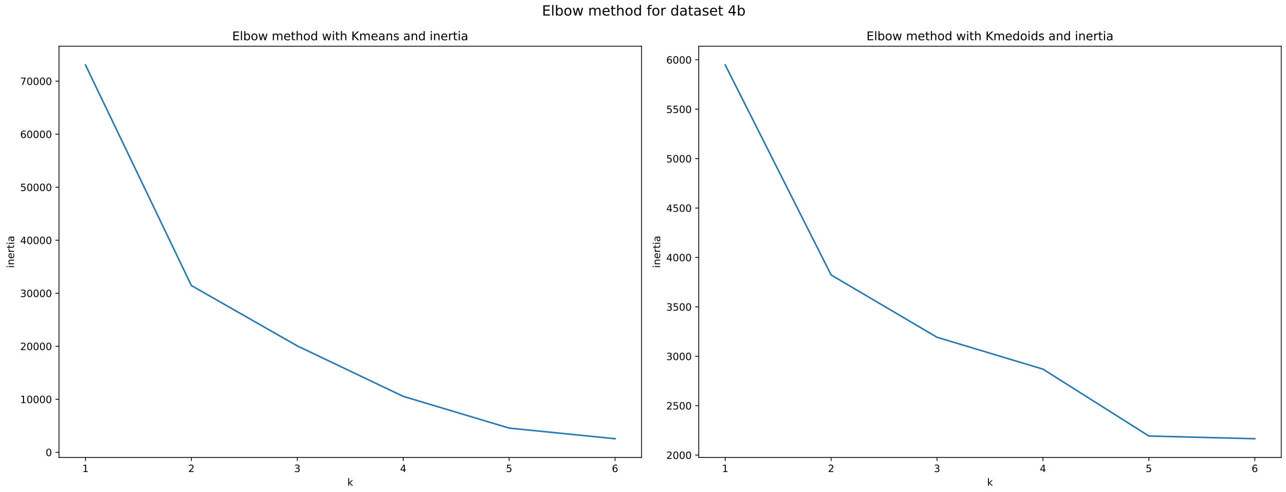

Figure 6: Elbow plot for dataset 4b for both k-means and k-medoids.

Dataset	**Min. value of k**	**max. value of k**
4a	1	6
4b	1	6

Table 2: Table of values of k. The minimal and maximal values of k were chosen because the elbow falls neatly between the first third and midway of the plot for each plot shown.

Choice of k. For dataset 4a I think k=3 to be the most likely value to be chosen to generate this dataset with, because the plot for k-medoids clearly shows an elbow at k=3. While the plot for k-means shows two elbows, the one at k=3 shows a stronger decline, and is additionally supported by the elbow in the k-medoids plot. For dataset 4b it is somewhat harder to determine, but again we think the most likely k-value is k=2. The k-means plot shows here the greatest elbow (the other curvers at k=4 and k=5 are almost smooth), while the k-medoids plots show at k=2 and k=5 the greatest elbow, but it has to be noted that the curve for k-medoids proceeds weirdly anyways.

5 Generating difficult clusters

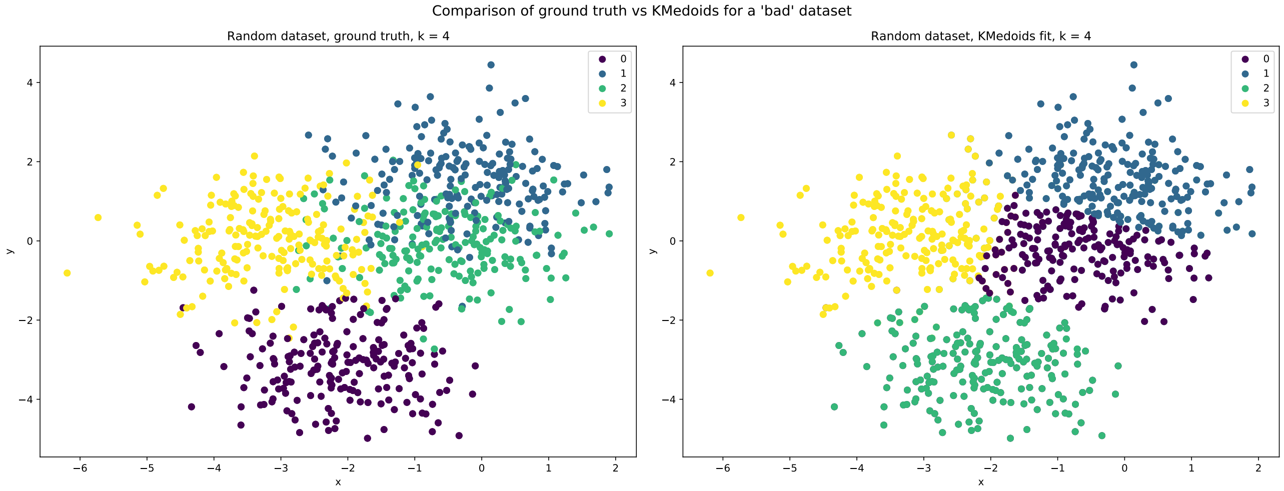

Figure 7: Ground truth and a k-medoids fit for a randomly generated dataset.

Generation of dataset. As k-medoids has issues with properly clustering a dataset of which the points are relatively evenly spaced, but some of them are still clustered, we needed to generate a dataset that has those properties. This leaded to a dataset with only two features, four centres and a relatively high point density and low cluster standard deviation. We've played some time with those values until we could reliably generate a random dataset for which k-medoids is hard to cluster.

Analysis of dataset. As noted before, k-medoids determine which point belongs to which cluster based on similarities. This means any dataset which is similar could be hard for k-medoids to cluster, so when we introduce a dataset with relatively small variations of distances between points, is it hard to cluster. This dataset is still easy for a human to cluster: there is a noticeably (though not very much higher) higher point density in the middle-left, lower right and upper right part of the upper ”island”.