TDDD41 - Lab 1

February 14, 2017

1 Simple k-means

1. Choose a set of attributes for clustering and give a motivation. (Hint: always ignore attribute "name". Why does the name attribute need to be ignored?) The *name* attribute is basically metadata, in the sense that it does not carry meaning during a data analysis and is only used to label the data points. For what concerns the choice of attributes, we tried to choose attributes that do not seem to be directly correlated. For instance, we picked protein and fat over energy, because energy in meat is largely depending on those two attributes, and it could make the analysis more complex without providing a lot of extra information. Then we chose to pick calcium.

2. Experiment with at least two different numbers of clusters, e.g. 2 and 5, but with the same seed value 10. The more clusters we have, the more specific subdivisions we get. Because of the nature of k-means, that is partitioning and non-overlapping, it might not carry a lot of significance to cluster data in very small sets. Mostly the k parameter should be tuned according to which information we are looking for. We experimented with $k = 2, 3, 8$.

3. Then try with a different seed value, i.e. different initial cluster centers. Compare the results with the previous results. Explain what the seed value controls. The seed value controls the starting points of the cluster centroids. If we change the value of the seed, the algorithm seems much more unstable for larger values of k, producing quite different clusters, which is likely also due to the fact that we have such a small dataset. However, if k is small, the clusters produced are quite similar for different seeds. With $k = 2$ and $k = 3$, for instance, the results are quite similar, evaluating by the centroids of the clusters. For $k = 8$, much less so. We chose these values in order to compare how meaningful the clusters when there are many vs when there are few. The number of data points in the clusters may vary too with the seed. This, along with the fact that the algorithms converges after many less iterations that the maximum we supply of 500, shows that k-means does not generate solutions corresponding to the global optimum, because otherwise starting by only changing the seed values would result in the same solutions. Instead, local optima are reached and then the algorithm stops. The choice of the seed values is somewhat arbitrary as they just influence a random starting point.

4. Do you think the clusters are "good" clusters? (Are all of its members "similar" to each other? Are members from different clusters dissimilar?) As we mentioned in question 2., how good our clusters are essentially depends on what attributes we are looking at. Since there are many variables, it is hard to visualize them all at once, visualization of each variable against itself to see if similar elements get into similar clusters is a helpful way to examine different clusters. In the plots below we can see that elements of a single cluster tend to be quite similar among each other in terms of fat and calcium—while they are dissimilar to members of the other cluster—, but dissimilar in terms of protein content—and conversely tend to be similar to members of another cluster.

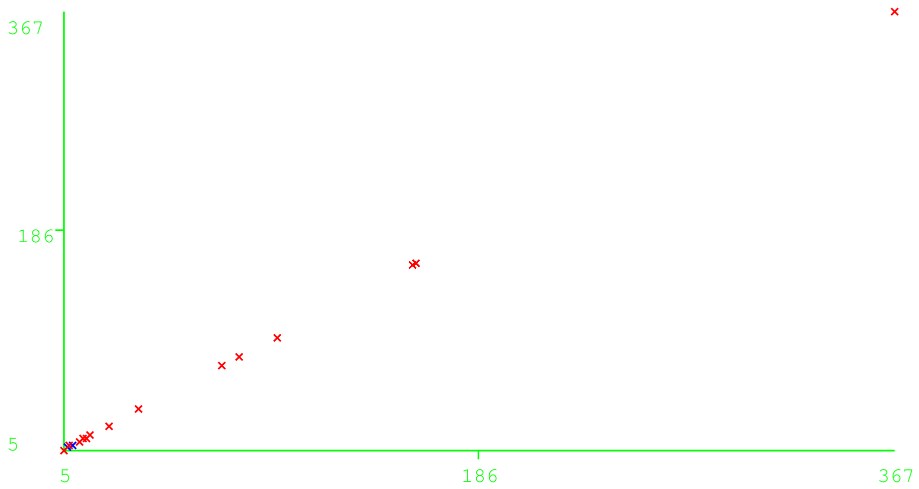

Figure 1: The blue cluster has very low levels of calcium, the red cluster covers all other amounts. There seems to be little overlap.

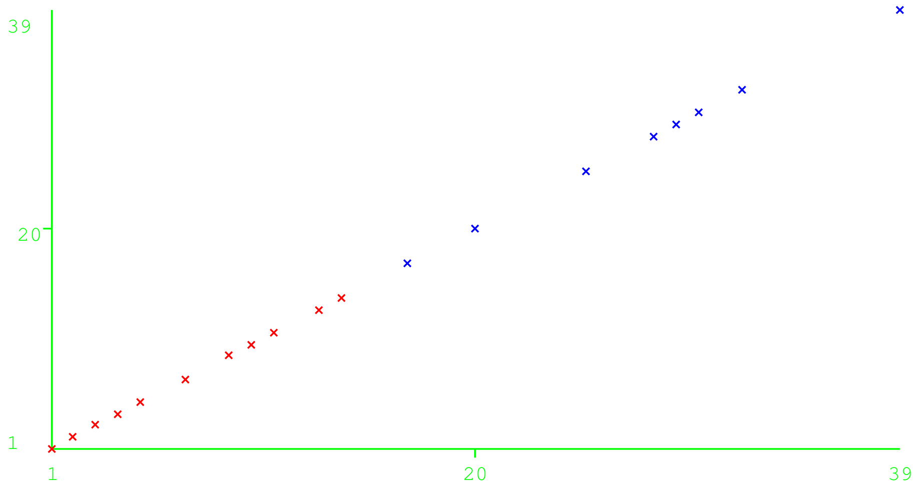

Figure 2: The blue cluster has higher amounts of fat, the red cluster has smaller amounts. Within-cluster elements have high similarity.

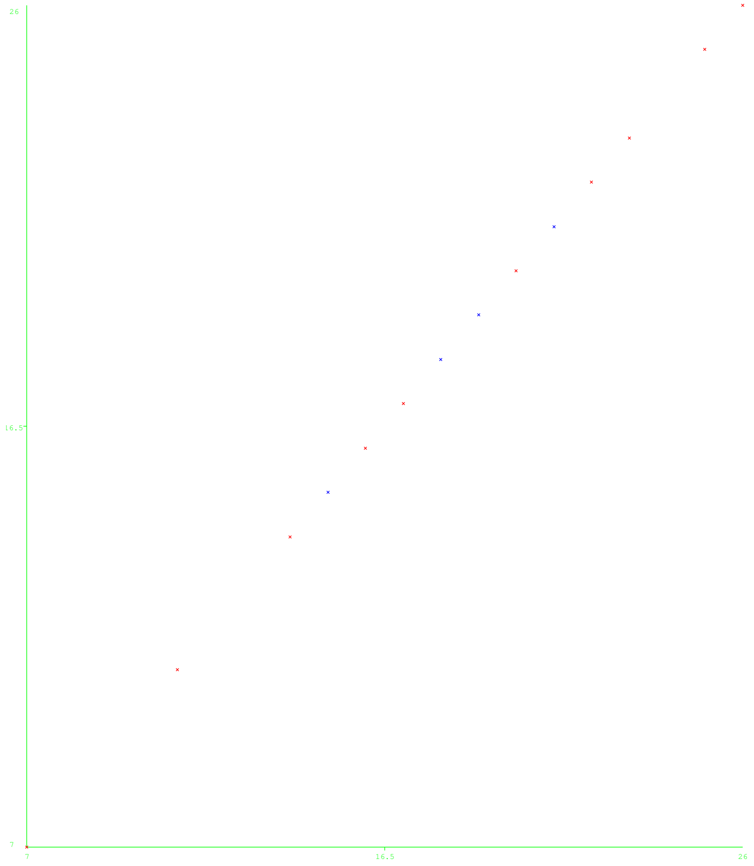

Figure 3: The clustering is not good w.r.t. protein. We see that the two clusters overlap in terms of protein content, and within-cluster similarity is low.

5. What does each cluster represent? Choose one of the results. Make up labels (words or phrases in English) which characterize each cluster. We chose $k = 2$ with seed 10. As can be seen in the plots above, the clearest separation is in terms of fat content. In particular, we can see that very fatty foods contain very little calcium. The blue cluster contains meats that are high in fat and contain very little calcium, while the red cluster contains meats that are low in fat.

2 MakeDensityBasedClusters

2. Experiment with at least two different standard deviations. Compare the results. (Hint: Increasing the standard deviation to higher values will make the differences in different runs more obvious and thus it will be easier to conclude what the parameter does). The standard deviation parameter influences the size of the final clusters. First, the algorithm produces clusters according to k-means, and then adjusts them with the minimum standard deviation. If we have a sufficiently small minimum standard deviation, we are guaranteed that we will have as many clusters as we originally asked for, in our case $k = 2$. However, if the standard deviation is too large we will end up including points that are dissimilar among each other, or even reducing the number of clusters, as more and more points need to be included to satisfy the minimum standard deviation requirement. The larger the standard deviation of an attribute for a given cluster in relation to the other attributes, the less that attribute serves to characterize the cluster. Taking cluster-means into account we need to consider the inter-cluster distance between the means for a given attribute in relation to the standard deviation of said attribute in order to evaluate its significance in characterizing a cluster.

With $k = 2$ we first used the default minimum of $\sigma = 1.0E - 6$, and we ended up with two clusters with the same data distribution as with the *k-means algorithm*. For clustering with $\sigma = 10$ we, as with $\sigma = 1.0E - 6$ got two clusters, even if two data points shifted cluster assignment. With $\sigma = 1000$ all elements end up being in one cluster as a result of no data points being outside this constraint on the minimum standard deviation.