732A61
Lab 1

Rasmus Holm

February 10, 2017

Contents

1 k-Means

In this exercise I have used the default options for the k-means algorithm unless specified otherwise.

I have decided to use two settings, 2 and 5 clusters, and done two different cluster analysis. I have excluded the name attribute since it is just a string and it is not possible to calculate distances based on arbitrary string values.

1.1 Result 1

In the first clustering I decided the use the features: Energy, Protein, Fat. The decision was based on the scatterplot matrix where I could see that there seemed to be positive correlations between protein/fat and energy.

1.1.1 2 Clusters

The resulting clusters can be seen below.

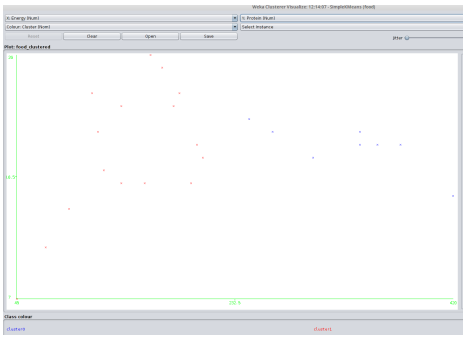

Figure 1: X-axis: Energy, Y-axis: Protein

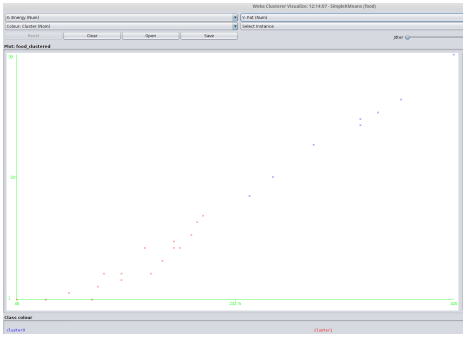

Figure 2: X-axis: Energy Y-axis: Fat

We can see from the plots that the clusters seem good and are well separated. The observations in the blue cluster have greater values in energy and fat with

kind of an average value in protein. We could say that the blue cluster contains energy rich and healthy food that are good when working out while the red cluster contains food that are low-fat and low-energy that are good for weight loss.

1.1.2 5 Clusters

The resulting clusters can be seen below.

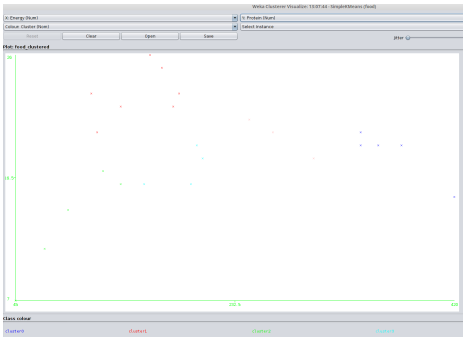

Figure 3: X-axis: Energy, Y-axis: Protein

The clusters seem reasonable except that the green and the bright blue clusters in the middle are not as well defined as they could be. The two green observations close to the blue cluster could be considered part of the blue cluster and I think that would have been better cluster assignments.

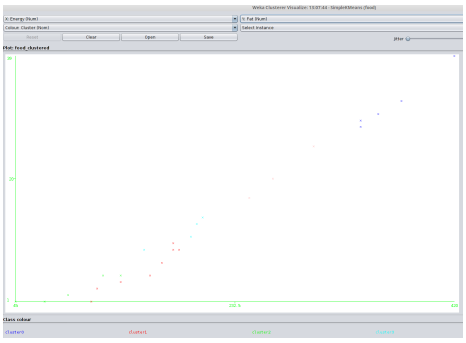

Figure 4: X-axis: Energy, Y-axis: Fat

However, this plot shows that the clusters are not as well defined as they first seemed. We can see that the green, red, and bright blue clusters overlap. Overall, I would say that it is more appropriate to use 2 clusters than 5 using these attributes.

1.2 Result 2

Here I decided the use the remaining features: iron and calcium.

1.2.1 2 Clusters

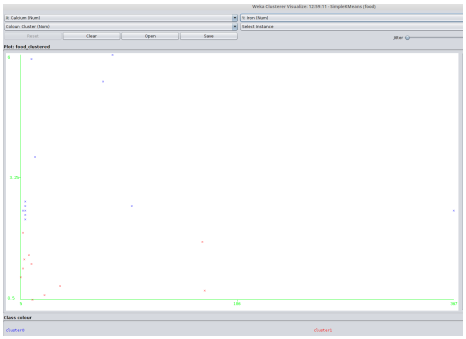

Figure 5: X-axis: Calcium, Y-axis: Iron

The clusters are not as well separated in this case since the observations are more scattared. It looks like it is suitable to find more than 2 clusters in this case.

1.2.2 5 Clusters

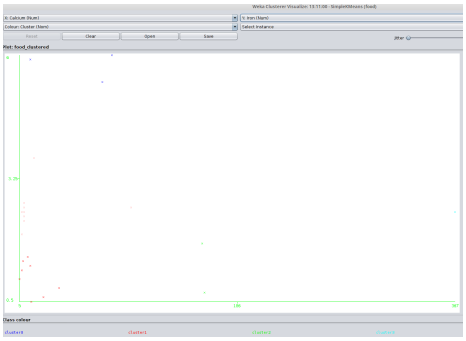

Figure 6: X-axis: Calcium, Y-axis: Iron

Using 5 clusters yield much better clusters with higher separation and lower intra-cluster distances. The within cluster sum of squared errors for 2 clusters was 2.13 and for 5 clusters it was 0.20 so we can see that it has been reduced by 10 times.

1.3 Seed

Since I am using random initialization the seed controls where the centroids are initialized which greatly influence what the final clusters will look like.

2 Density

I will be using the data I worked with in result 1 above and use 2 clusters but this time turning k-means into a density-based clustering algorithm.

The MakeDensityBasedClusters class will create Gaussian distributions at the cluster centroids produced by the k-means algorithm, i.e. the mean of the distributions are the cluster centroids, and the standard deviation (std) is given by the observations assigned to each cluster. However, we can control the distributions by specifying the minimum std which can increase the widths. The densities can then be used to calculate probabilities for each observation being in each cluster which is also influenced by the prior probabilities, the ratio of points in each cluster produced by k-means, and an observation is assigned to the cluster with the highest probability.

Below are results with different minimum standard deviations.

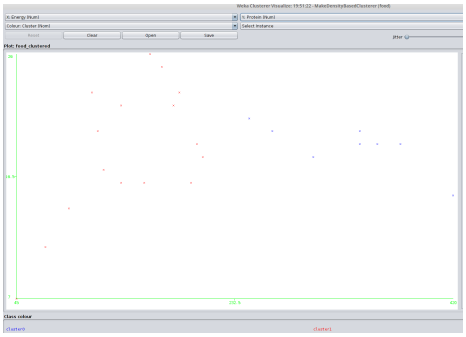

Figure 7: X-axis: Energy, Y-axis: Protein

In the above plot I have used min std = 0.1 and we can see that the clusters are exactly the same as result from the previous assignment using the regular k-means algorithm.

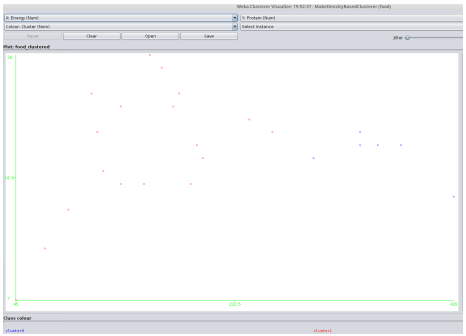

Figure 8: X-axis: Energy, Y-axis: Protein

Given min std = 100 we can see changes in the clusters and as expected the

red cluster with more observations, higher prior probability, consumes observations from the blue cluster.

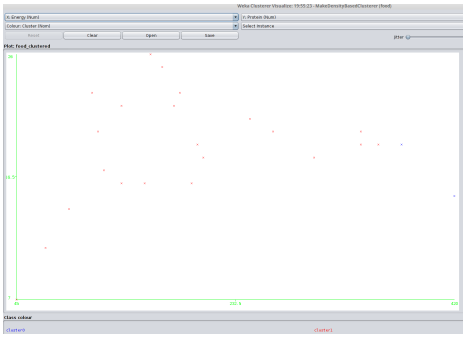

Figure 9: X-axis: Energy, Y-axis: Protein

Similarly as previously, min std = 200 means the red cluster consumes even more points from the blue cluster.