Lab 1

February 4, 2017

1 SimpleKmeans

1. 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. 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.
3.
4. As we mentioned in question 2., how good our clusters are essentially means on what we are looking for.

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. 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 kmeans 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.

2 MakeDensityBasedClusters