**SimpleKMeans**

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?)

Chose ’Fat’ and ‘Energy’.

Motivation: Higher fat and energy means taking this food will easily cause fat. And, for the attribute ’name’, the distance within this feature has no physical meaning which means take this into consideration will add a unnecessary bias into the model.

图形用户界面, 图表

描述已自动生成

We can see this clustering method can work well.Cluster1 (more healthy ) means lower fat and energy where as cluster0 means higher fat and energy. The Within cluster sum of squared error is 0,84819.

2. Experiment with at least two different numbers of clusters, e.g. 2 and 5, but with the same seed value 10.

图表, 散点图

描述已自动生成

This is the result of 4 clusters. Compare with 2 clusters(can be seen in the q1). The Within cluster sum of squared error is 0.29069.

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.

图表, 散点图

描述已自动生成

The seed in this question is 18. Compared with picture in q1, the point circled red now belongs to cluster0(less healthy). Different seeds will have different initial centers, which will have great impact on the following iterations. AS a result, some instances located in the ‘middle area’ probably change their group.

4. Do you think the clusters are "good" clusters? (Are all of its members "similar" to each other? Are members from different clusters dissimilar?)

In the question2 and question3 we have the picture for 2 clusters and 4 clusters situation. As the pictures show, we can find a clear boundary of every cluster. In this case, we would say these clusters are good clusters(All of its members similar to each other).

5. What does each cluster represent? Choose one of the results. Make up labels (words or phrases in English) which characterize each cluster.

This is the result of question3 , where we have 2 clusters. The x-axis is fat and y-axis is energy. We can see that red cluster means lower fat and energy while blue cluster means higher fat and energy. In this way, we can set labels like:

red cluster: Food wouldn’t cause fat.

blue cluster: Food more likely cause fat.

图表, 散点图

描述已自动生成

**MakeDensityBasedClusters**

1. Use the SimpleKMeans clusterer which gave the result you haven chosen in 5).

图表, 散点图

描述已自动生成

This is the same result as 5).

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).

minstddev=100

图表, 散点图

描述已自动生成

minstddev=1e-6

图表, 散点图

描述已自动生成

When we increase the Minstddev, the number of clusters decrease from 4 to 2 although we set there should be 4 clusters. The deviation controls the width of gaussian distribution. When we increase the deviation, we will get wider gaussian distribution which means more points will be included into one cluster.