Student: aksgu350 (Akshay Gurudath)

Student: kespa139 (Keshav Padiyar Manuru)

TDDD41/732A75: Clustering Lab ¶

Dataset:

The given dataset consists of nutrients levels of 27 different food products. The quantities of energy, protien, fat, calcium, and iron have been measured in a 3 ounce portion of the various foods.

Clustering using 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 beignored?)

Below are the attribute details of the given dataset.

Attribute Name	Data Type
Name	string
Energy	real
Protein	real
Fat	real
Calcium	real

The attribute *name* is a string/categorical attribute. K means is applicable only when mean/centroid is defined hence it works only on the numeric data. Therefore, we are ignoring this attribute.

Except *name*, remaining all attributes are real numbers hence we excpect them to contribute in the clustering.

So, we are considering all the attributes except *name* for clustering.

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

Experiment 1:\ k = 2\ seed = 10

Output from Weka:

Number of iterations: 2\ Within cluster sum of squared errors: 5.069321339929419\ Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (27)	0 (9)	1 (18)
Energy	207.4074	331.1111	145.5556

Attribute	Full Data (27)	0 (9)	1 (18)
Protein	19	19	19
Fat	13.4815	27.5556	6.4444
Calcium	43.963	8.7778	61.5556
Iron	2.3815	2.4667	2.3389

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 - 9 (33%)\ 1 - 18 (67%)

Obsercations of Experiment 1: Clustering of food products done based on its nutrition value using simple K means clustering, with *k=2*. That means, we have 2 clusters. From above table, its evident that attribute *Energy* and *Calcium* are main contributors to the clustering. For these 2 attributes, we see that the individual cluster centroids are farthest away from the centroid of the attribute.

Experiment 2:\ k = 5\ seed = 10

Output from Weka:

Number of iterations: 4 Within cluster sum of squared errors: 2.750432407251998\ Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (27)	0 (7)	1 (8)	2 (6)	3 (1)	4 (5)
Energy	207.4074	352.8571	153.125	102.5	180	222
Protein	19	18.5714	23.25	13.5	22	18.8
Fat	13.4815	30.1429	5.75	3.8333	9	15
Calcium	43.963	8.7143	23.75	87.5	367	8.8
Iron	2.3815	2.4143	2.45	2.5333	2.5	2.02

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 - 7 (26%)\ 1 - 8 (30%)\ 2 - 6 (22%)\ 3 - 1 (4%)\ 4 - 5 (19%)

Obsercations of Experiment 2: Clustering of food products done based on its nutrition value using simple K means clustering, with *k=5*. That means, we have 5 clusters. From above table, its evident that attribute *Energy* and *Calcium* are main contributors to the clustering. For these 2 attributes, we see that the individual cluster centroids are farthest away from the centroid of the attribute.

Overall Observations for Experiment 1 and Experiment 2: From above experiments we could see that, *Energy and Calcium* are the most contributing attributes in both the case. However, *Sum of Squared Error* within the clusters formed in Experiment 1 is greater than in Experiment 2. This metric implies that the 5 clusters are ideal ones than 2 clusters.

3. Then try with a different seed value, i.e. different initial cluster centers. Compare the results with the previous results. Explain what the seedvalue controls.

Experiment 3:\ k = 2\ seed = 12344

Output from Weka:

Number of iterations: 4 Within cluster sum of squared errors: 5.9104334390998226\ Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data	0	1
	(27)	(11)	(16)
Energy	207.4074	122.7273	265.625
Protein	19	16.7273	20.5625
Fat	13.4815	4.6364	19.5625
Calcium	43.963	94.2727	9.375
Iron	2.3815	2.1455	2.5438

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 - 11 (41%)\ 1 - 16 (59%)

Obsercations of Experiment 3:

Changing the seed value = 12344 and comparing with the output of *Experiment 1*

1. The number of iterations have been increased to 4.
2. Cluster centroids are chaged
3. Due to change in cluster centroid, the data distribution within the clusters also varied.
4. Sum of square error within the clusters also changed.

Experiment 4:\ k = 5\ seed = 1234

Output from Weka:

Number of iterations: 4\ Within cluster sum of squared errors: 3.240800908192409\ Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data	0	1	2	3	4
	(27)	(1)	(6)	(6)	(6)	(8)
Energy	207.4074	420	110	207.5	341.6667	153.125
Protein	19	15	14.5	18.3333	19.1667	23.25
Fat	13.4815	39	4.5	13.3333	28.6667	5.75
Calcium	43.963	7	146.1667	9.8333	9	23.75

Attribute	Full Data (27)	0 (1)	1 (6)	2 (6)	3 (6)	4 (8)
Iron	2.3815	2	2.8667	1.7667	2.4833	2.45

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 - 1 (4%)\ 1 - 6 (22%)\ 2 - 6 (22%)\ 3 - 6 (22%)\ 4 - 8 (30%)

Obsercations of Experiment 4: Changing the seed value = 1234 and comparing with the output of *Experiment 2*

1. Cluster centroids are chaged
2. Due to change in cluster centroid, the data distribution within the clusters also varied.
3. Sum of square error within the clusters also changed.

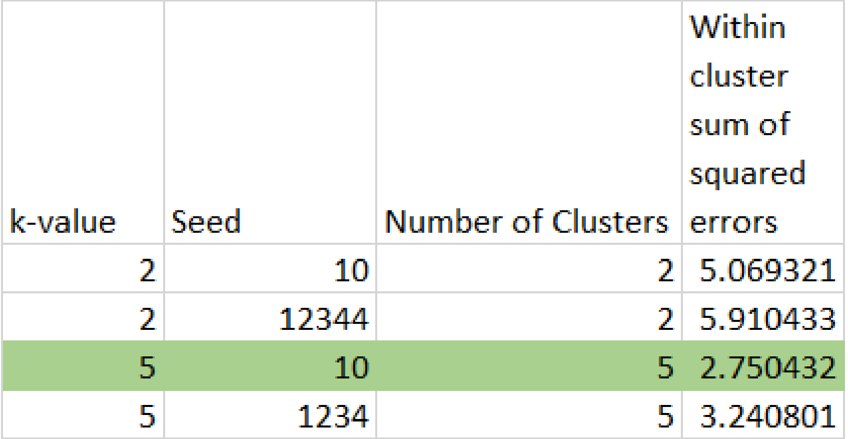
Overall Observations for Experiment 3 and Experiment 4: Seed value controls the initialization/selection of first centroid value. The final or the best centroid is dependenet on the first cluster formed, which depends on the inital centroid value. Hence, we observed variation in results for different seeds and same k values.

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

Yes, we think that these clusters are good clusters. As Sum of squared error within the cluster gives us the similarity metric, smaller the error better the cluster. we also think, members from different clusters are dissimilar.

However, we shouldn't increase the value of k too much, so as to prevent overfitting.

Below table consists of error values obtained in our Experiments.



k-value	Seed	Number of Clusters	Within cluster sum of squared errors
2	10	2	5.069321
2	12344	2	5.910433
5	10	5	2.750432
5	1234	5	3.240801

From above table, clusters formed with k = 5; seed = 10 is resulted in relatively smaller within cluster sum of squared error. Hence we consider it as a better cluster.

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

We are considering cluster formed with k= 5; seed = 10.

Attribute	Full Data (27)	0 (7)	1 (8)	2 (6)	3 (1)	4 (5)

Attribute	Full Data (27)	0 (7)	1 (8)	2 (6)	3 (1)	4 (5)
Energy	207.4074	352.8571	153.125	102.5	180	222
Protein	19	18.5714	23.25	13.5	22	18.8
Fat	13.4815	30.1429	5.75	3.8333	9	15
Calcium	43.963	8.7143	23.75	87.5	367	8.8
Iron	2.3815	2.4143	2.45	2.5333	2.5	2.02

Methedology: We looked at the cluster centroid of each attribute and compared it with the over all centroid of that attribute. Where ever we observed a high difference from the over all centroid, we marked that attribute as High or low in content.

cluster 0:\ High Energy and High Fat content\ cluster 1:\ High Protein and Low Fat content\ cluster 2:\ High Calcium and Low Fat content\ cluster 3:\ High Calcium and High Protein content\ cluster 4:\ Well Balanced Nutrient Contents

MakeDensityBasedClusters

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

2. Experiment with at least two different standard deviations. Compare the results. (Hint: Increasing the standard deviation to higher values willmake the differences in different runs more obvious and thus it will be easier to conclude what the parameter does)

Experiment 5:\ k = 5\ seed = 10

Number of iterations: 4 Within cluster sum of squared errors: 2.750432407251998\ Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (27)	0 (7)	1 (8)	2 (6)	3 (1)	4 (5)
Energy	207.4074	352.8571	153.125	102.5	180	222
Protein	19	18.5714	23.25	13.5	22	18.8
Fat	13.4815	30.1429	5.75	3.8333	9	15
Calcium	43.963	8.7143	23.75	87.5	367	8.8
Iron	2.3815	2.4143	2.45	2.5333	2.5	2.02

Time taken to build model (full training data) : 0.01 seconds

Min Standard Deviation = 0.01

=== Model and evaluation on training set ===

Clustered Instances

0 - 7 (26%)\ 1 - 8 (30%)\ 2 - 6 (22%)\ 3 - 1 (4%)\ 4 - 5 (19%)

For Analysis we shall consider Cluster 1 and Cluster 4 distributions with respect to each attribute:

Cluster: 1 Prior probability: 0.2813

Attribute: Energy Normal Distribution. Mean = 153.125 StdDev = 27.379 Attribute: Protein Normal Distribution. Mean = 23.25 StdDev = 1.854 Attribute: Fat Normal Distribution. Mean = 5.75 StdDev = 2.8614 Attribute: Calcium Normal Distribution. Mean = 23.75 StdDev = 28.5909 Attribute: Iron Normal Distribution. Mean = 2.45 StdDev = 1.6023

Cluster: 4 Prior probability: 0.1875

Attribute: Energy Normal Distribution. Mean = 222 StdDev = 27.8568 Attribute: Protein Normal Distribution. Mean = 18.8 StdDev = 1.7205 Attribute: Fat Normal Distribution. Mean = 15 StdDev = 3.1623 Attribute: Calcium Normal Distribution. Mean = 8.8 StdDev = 2.9933 Attribute: Iron Normal Distribution. Mean = 2.02 StdDev = 0.7194

Observation: We note that, in each cluster and each attribute there is a fitted normal distribution with mean = centroid and standard devivation. We observe that these standard deviations are greater than the user input *min Standard deviation = 0.01* and that is why, in the next experiment we choose a considerably higher standard deviation. In addition, there is no difference in the allocation of data points into different clusters.

Min Standard Deviation = 150

=== Model and evaluation on training set ===

Clustered Instances

0 - 7 (26%)\ 1 - 18 (67%)\ 2 - 1 (4%)\ 3 - 1 (4%)

For Analysis we shall consider Cluster 1 and Cluster 4 distributions with respect to each attribute:

Cluster: 1 Prior probability: 0.2813

Attribute: Energy\ Normal Distribution. Mean = 153.125 StdDev = 150\ Attribute: Protein\ Normal Distribution. Mean = 23.25 StdDev = 150\ Attribute: Fat\ Normal Distribution. Mean = 5.75 StdDev = 150\ Attribute: Calcium\ Normal Distribution. Mean = 23.75 StdDev = 150\ Attribute: Iron\ Normal Distribution. Mean = 2.45 StdDev = 150

Cluster: 4 Prior probability: 0.1875

Attribute: Energy\ Normal Distribution. Mean = 222 StdDev = 150\ Attribute: Protein\ Normal Distribution. Mean = 18.8 StdDev = 150\ Attribute: Fat\ Normal Distribution. Mean = 15 StdDev = 150\ Attribute: Calcium\ Normal Distribution. Mean = 8.8 StdDev = 150\ Attribute: Iron\ Normal Distribution. Mean = 2.02 StdDev = 150

Observation: Firstly we noticed that, most of the observations are being classified into cluster 1. There are no observations being allocated to cluster 4. As we increase the *min Standard Deviation*, the normal distrbution begins to widen and intersect with the normal distributions of other clusters. Therefore, a datapoint may fall in the intersection of different clusters'normal distribution. This data point will get different probabilities under different clusters'normal distributions. We believe this point is allocated to the cluster, with whose normal distributions it has the maximum probability. In this case, Cluster 1 is closer to the actual mean and will tend to dominate over other clusters as *min Standard Deviation* increases, which is what we