Lab 1 Clustering Group 10

Liuxi Mei Xiaochen Liu

2025-02-03

SimpleKmeans

Apply "SimpleKMeans" to your data. In Weka euclidian distance is implemented in SimpleKmeans. You can set the number of clusters and seed of a random algorithm for generating initial cluster centers. Experiment with the algorithm as follows:

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 set is chosen is the attributes excluding attributes 'Names', 'Protein', and 'Fat'. The reason for excluding 'Protein' and 'Iron' data is that they both have quite similar mean values across clusters, see picture below , which means that the differences among cluster are insignificant and contribute quite less to clustering results.

'Name' is also ignored as it is a string attribute. For clustering, Euclidean distance can not be calculated for strings and Weka will also report a mistake.

```
Number of iterations: 3
Within cluster sum of squared errors: 3.2290308976556164
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3
                  (27)        (8)       (11)        (7)        (1)
=============================================================================
Energy         207.4074   341.875   170.4545   115.7143        180
Protein              19      18.75    22.1818    13.8571         22
Fat             13.4815     28.875     8.1818     4.8571          9
Calcium          43.963       8.75    19.1818         77        367
Iron             2.3815     2.4375     2.3455     2.3571        2.5
```

Figure 1: Clustering result with all numeric attributes

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

Number of iterations: 3
Within cluster sum of squared errors: 1.2005472547937988
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(27)	(8)	(12)	(7)
Energy	207.4074	341.875	174.5833	110
	+/-101.2078	+/-46.3633	+/-35.7045	+/-46.9929
Fat	13.4815	28.875	8.75	4
	+/-11.257	+/-5.4625	+/-4.3719	+/-3.6968
Calcium	43.963	8.75	11.8333	139.2857
	+/-78.0343	+/-0.7071	+/-5.5895	+/-109.5928

Figure 2: Calculated results with 3 clusters, seed set to 10

Number of iterations: 4
Within cluster sum of squared errors: 0.46682463309541183
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(27)	(6)	(11)	(6)	(1)	(3)
Energy	207.4074	361.6667	168.1818	98.3333	180	270
	+/-101.2078	+/-31.728	+/-29.349	+/-38.8158	+/-0	+/-27.8388
Fat	13.4815	31	8	3.1667	9	20.6667
	+/-11.257	+/-4.1952	+/-3.6878	+/-3.2506	+/-0	+/-4.0415
Calcium	43.963	8.6667	12.0909	101.3333	367	9
	+/-78.0343	+/-0.8165	+/-5.7871	+/-48.0985	+/-0	+/-0

Figure 3: Calculated results with 5 clusters, seed set to 10

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.

With seed = 20, following results are given. Both the numbers of instances in clusters and the instances belong to clusters have changed. As a result, the variance and mean value for each cluster are different from the previous results.

When initializing this algorithm, N (corresponding to cluster number) initial 'means' are generated randomly within the data domain, which is controlled by the seed number. Different seeds give different initial 'means', resulting different results in this case.

Number of iterations: 4
Within cluster sum of squared errors: 1.172520467950207
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(27)	(3)	(15)	(9)
Energy	207.4074	151.6667	144.3333	331.1111
	+/-101.2078	+/-30.1386	+/-49.5648	+/-54.0704
Fat	13.4815	7.6667	6.2	27.5556
	+/-11.257	+/-2.3094	+/-4.3948	+/-6.4636
Calcium	43.963	227.6667	28.3333	8.7778
	+/-78.0343	+/-120.6703	+/-30.647	+/-0.6667

Figure 4: Calculated results with 3 clusters, seed set to 20

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

Those are not 'good' clusters, at least not appropriate for the application of SimpleKmeans method. If they are obviously similar to each other, the clusters should include the same instances despite different seed values.

By applying 'standardize' filter to raw data and with seed set to 20, the regenerated results are shown as the figures below. When cluster number is chosen to be 3, cluster 0 and 1 different calcium values but very similar Energy and Fat values. And cluster 2 seems to be a more independent. For 5 clusters, cluster 0 and 4 are similar in Energy and Fat values but unsimilar calcium values. The similarity in calcium values is also noticed between cluster 0 and 1.

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

Take figure 6 as an example, cluster 0 is the one with moderate energy, moderate fat and high calcium level. Cluster 1 is the one with moderate energy, moderate fat and low calcium level. Cluster 3 is the one with high energy, high fat and moderate energy level.

Number of iterations: 4
Within cluster sum of squared errors: 0.5497295323397319
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(27)	(3)	(5)	(4)	(7)	(8)
Energy	207.4074	151.6667	86	228.75	352.8571	166.25
	+/-101.2078	+/-30.1386	+/-29.0259	+/-31.4576	+/-37.1772	+/-21.9984
Fat	13.4815	7.6667	1.6	16	30.1429	7.25
	+/-11.257	+/-2.3094	+/-0.8944	+/-3.1623	+/-4.4508	+/-2.5495
Calcium	43.963	227.6667	60	7.5	8.7143	14.125
	+/-78.0343	+/-120.6703	+/-36.4417	+/-1.9149	+/-0.7559	+/-5.4625

Figure 5: Calculated results with 5 clusters, seed set to 20

Number of iterations: 4
Within cluster sum of squared errors: 1.1725204679502073
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(27)	(3)	(15)	(9)
Energy	0	-0.5508	-0.6232	1.2223
	+/-1	+/-0.2978	+/-0.4897	+/-0.5343
Fat	0	-0.5165	-0.6468	1.2502
	+/-1	+/-0.2052	+/-0.3904	+/-0.5742
Calcium	0	2.3541	-0.2003	-0.4509
	+/-1	+/-1.5464	+/-0.3927	+/-0.0085

Figure 6: Calculated results with 3 clusters, seed set to 20, with standardized data

Number of iterations: 4
Within cluster sum of squared errors: 0.5497295323397318
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(27)	(3)	(5)	(4)	(7)	(8)
Energy	0	-0.5508	-1.1996	0.2109	1.4371	-0.4067
	+/-1	+/-0.2978	+/-0.2868	+/-0.3108	+/-0.3673	+/-0.2174
Fat	0	-0.5165	-1.0555	0.2237	1.4801	-0.5536
	+/-1	+/-0.2052	+/-0.0795	+/-0.2809	+/-0.3954	+/-0.2265
Calcium	0	2.3541	0.2055	-0.4673	-0.4517	-0.3824
	+/-1	+/-1.5464	+/-0.467	+/-0.0245	+/-0.0097	+/-0.07

Figure 7: Calculated results with 5 clusters, seed set to 20, with standardized data

MakeDensityBasedClusters

Now with MakeDensityBasedClusters, SimpleKMeans is turned into a density-based cluster. You can set the minimum standard deviation for normal density calculation. Experiment with the algorithm as the follows:

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 will make the differences in different runs more obvious and thus it will be easier to conclude what the parameter does)

With 0.001 and 0.5 selected for minStdDev, the same cluster centroids for 2 custers are initilized, see figure 8. However, the fitted cluster vary and cluster have 9 and 8 instances when the minStdDev is 0.001 and 0.5 respectively. The standard deviations for fitted normal distribution models are also different to fulfill the setting of minStdDev. As a result, log likelihood is -1.7 for the model with 0.001 minStdDev and 3.8 for the other.

Number of iterations: 2
Within cluster sum of squared errors: 1.9487869540091085
Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(27)	(8)	(19)
Energy	0	1.3286	-0.5594
Fat	0	1.3675	-0.5758
Calcium	0	-0.4513	0.19

Figure 8: Initial clusters with cebtroids with MakeDensityBasedClusters applied