Lab Report: Lab1 - Clustering Lab

Dhanush Kumar Reddy Narayana Reddy (dhana004), Udaya Shanker Mohanan Nair (udamo524)

2025-03-26

Introduction

Implementation of Kmeans, SimpleKMeans and density-based method.

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?)

We choose attributes energy, fat, calcium as viewing values directly we could understand that energy, protein and fat contribute more than other attributes given. Name is always ignored because it is an identical attribute and does not contribute during clustering.

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

In first experiment, tried with cluster value as 2 and seed value as 10 for all attributes mentioned in the first questions answer.

```
Number of iterations: 3
Within cluster sum of squared errors: 0.6840965780384897
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        180   100.7143        180
Fat             13.4815     28.875     9.2727     3.1429          9
Calcium          43.963       8.75    12.1818         88        367



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

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

Clustered Instances

0       8 ( 30%)
1      11 ( 41%)
2       7 ( 26%)
3       1 (  4%)
```

In second experiment, tried with cluster value as 4 and seed value as 10 for all attributes mentioned in the first questions answer.

```
Number of iterations: 3
Within cluster sum of squared errors: 0.6840965780384897
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        180   100.7143        180
Fat             13.4815     28.875     9.2727     3.1429          9
Calcium          43.963       8.75    12.1818         88        367



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

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

Clustered Instances

0       8 ( 30%)
1      11 ( 41%)
2       7 ( 26%)
3       1 (  4%)
```

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.

Changing seed value to 20.

Results of first experiment

```
Number of iterations: 6
Within cluster sum of squared errors: 1.93522433869064
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                  (27)       (18)        (9)
=============================================
Energy         207.4074   145.5556   331.1111
Fat             13.4815     6.4444    27.5556
Calcium          43.963    61.5556     8.7778



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

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

Clustered Instances

0      18 ( 67%)
1       9 ( 33%)
```

Results of second experiment

```
Number of iterations: 6
Within cluster sum of squared errors: 1.93522433869064
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute    Full Data          0          1
                  (27)       (18)        (9)
=============================================
Energy         207.4074   145.5556   331.1111
Fat             13.4815     6.4444    27.5556
Calcium          43.963    61.5556     8.7778



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

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

Clustered Instances

0      18 ( 67%)
1       9 ( 33%)
```

Comparison of results

For Cluster 2

Cluster sum of squared error changed from 1.94879 -> 1.93522.

For Cluster 4

Cluster sum of squared error changed from 0.6841 -> 0.7243.

Seed value controls the initial cluster centers, which controls the clustering.Different seed value results in different cluster formation and squeared error 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?)

Clusters are not properly good, as for k=2, the clustering is weak where one cluster dominates the other for both cases of seed(10 or 20). When k=4, the clustering is better and clusters are distributed more fairly when compared to other k value for both the seed values.

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

Below table shows the cluster distribution when k=4 and seed value is 10.

Table 1: Cluster Labelling

Cluster	Energy	Fat	Calcium	Label
0	341.8750	28.8750	8.7500	High-Fat, Energy, calicum product
1	180.0000	9.2727	12.1818	Balanced Fat Energy Calcium product
2	100.7143	3.1429	88.0000	Rich Calcium Dairy Product

Cluster	Energy	Fat	Calcium	Label
3	180.0000	9.0000	367.0000	High Calcium Product

MakeDensityBasedClusters

1. Use the SimpleKMeans clusterer (which gave the result you haven chosen in 5).Experiment with at least two different standard deviations. Compare the results.

Results of First Experiment:

where Standard deviation is 0.1, cluster distribution when k=4 and seed value is 10.

```
=== Run information ===

Scheme:weka.clusterers.MakeDensityBasedClusterer -M 0.1 -W weka.clusterers.SimpleKMeans -- -N 4 -A "weka
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Fat
              Calcium
Ignored:
              Name
              Protein
              Iron
Test mode:evaluate on training data

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

MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 0.6840965780384897
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        180   100.7143        180
Fat             13.4815     28.875     9.2727     3.1429          9
Calcium          43.963       8.75    12.1818         88        367



Fitted estimators (with ML estimates of variance):
```

```
Cluster: 0 Prior probability: 0.2903

Attribute: Energy
Normal Distribution. Mean = 341.875 StdDev = 43.3689
Attribute: Fat
Normal Distribution. Mean = 28.875 StdDev = 5.1097
Attribute: Calcium
Normal Distribution. Mean = 8.75 StdDev = 0.6614

Cluster: 1 Prior probability: 0.3871

Attribute: Energy
Normal Distribution. Mean = 180 StdDev = 30.3764
Attribute: Fat
Normal Distribution. Mean = 9.2727 StdDev = 3.9793
Attribute: Calcium
Normal Distribution. Mean = 12.1818 StdDev = 5.4576

Cluster: 2 Prior probability: 0.2581

Attribute: Energy
Normal Distribution. Mean = 100.7143 StdDev = 33.3197
Attribute: Fat
Normal Distribution. Mean = 3.1429 StdDev = 2.7479
Attribute: Calcium
Normal Distribution. Mean = 88 StdDev = 52.1454

Cluster: 3 Prior probability: 0.0645

Attribute: Energy
Normal Distribution. Mean = 180 StdDev = 101.2078
Attribute: Fat
Normal Distribution. Mean = 9 StdDev = 11.257
Attribute: Calcium
Normal Distribution. Mean = 367 StdDev = 78.0343


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

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

Clustered Instances

0       9 ( 33%)
1       9 ( 33%)
2       8 ( 30%)
3       1 (  4%)


Log likelihood: -12.03857
```

Results of Second Experiment:

where Standard deviation is 1.0, cluster distribution when k=4 and seed value is 10.

```
=== Run information ===

Scheme:weka.clusterers.MakeDensityBasedClusterer -M 1.0 -W weka.clusterers.SimpleKMeans -- -N 4 -A "weka
Relation:     food
Instances:    27
Attributes:   6
              Energy
              Fat
              Calcium
Ignored:
              Name
              Protein
              Iron
Test mode:evaluate on training data

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

MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 0.6840965780384897
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        180   100.7143        180
Fat             13.4815     28.875     9.2727     3.1429          9
Calcium          43.963       8.75    12.1818         88        367



Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.2903

Attribute: Energy
Normal Distribution. Mean = 341.875 StdDev = 43.3689
Attribute: Fat
Normal Distribution. Mean = 28.875 StdDev = 5.1097
Attribute: Calcium
Normal Distribution. Mean = 8.75 StdDev = 78.0343

Cluster: 1 Prior probability: 0.3871

Attribute: Energy
Normal Distribution. Mean = 180 StdDev = 30.3764
```

```
Attribute: Fat
Normal Distribution. Mean = 9.2727 StdDev = 3.9793
Attribute: Calcium
Normal Distribution. Mean = 12.1818 StdDev = 5.4576

Cluster: 2 Prior probability: 0.2581

Attribute: Energy
Normal Distribution. Mean = 100.7143 StdDev = 33.3197
Attribute: Fat
Normal Distribution. Mean = 3.1429 StdDev = 2.7479
Attribute: Calcium
Normal Distribution. Mean = 88 StdDev = 52.1454

Cluster: 3 Prior probability: 0.0645

Attribute: Energy
Normal Distribution. Mean = 180 StdDev = 101.2078
Attribute: Fat
Normal Distribution. Mean = 9 StdDev = 11.257
Attribute: Calcium
Normal Distribution. Mean = 367 StdDev = 78.0343


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

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

Clustered Instances

0       8 ( 30%)
1      10 ( 37%)
2       8 ( 30%)
3       1 (  4%)


Log likelihood: -13.32433
```

Comparison of results

Metric	SD = 0.1	SD = 1.0
Log Likelihood	-12.03857	-13.32433
Cluster 0 Size	9	8
Cluster 1 Size	9	10
Cluster 2 Size	8	8
Cluster 3 Size	1	1

(1) **Lower standard deviation (0.1)** leads to tighter density estimation and clearer cluster membership.

(2) **Higher standard deviation (1.0)** allows more flexibility, leading to broader densities and potentially more overlapping clusters.

(3) The choice of standard deviation affects both model fit (log likelihood) and instance assignments.

(4) For cleaner clustering with higher certainty, a lower SD is preferable; for flexibility and noise tolerance, a higher SD can be useful.