Lab 1 - 732A75 - Clustering Algorithms

Joris van Doorn - jorva845 | Bayu Brahmantio - baybr878 | Ismail Khalil - ismkh208

11 February 2020

1. 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?)

```
Number of iterations: 6
Within cluster sum of squared errors: 5.9104434390998226
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
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) : 0.01 seconds

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

Clustered Instances

0      11 ( 41%)
1      16 ( 59%)
```

The attributes for the simpleKmeans clustering algorithm we choose are 2 clusters, with seed = 1. Two clusters is the minimum that we can choose, and it is a good start to see what happens if we simply want to divide the data into two groups. For the seed we started with 1, just so we can easily compare to other seeds. The output is depicted in the image.

The name attribute needs to be ignored because it is a string with labels, which is already a categorization of different groups.

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

Cluster centroids:
                          Cluster#
Attribute      Full Data        0          1
                    (27)       (9)        (18)
===============================================
Energy          207.4074  331.1111   145.5556
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) : 0.01 seconds

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

Clustered Instances

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


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

Cluster centroids:
                          Cluster#
Attribute      Full Data       0          1          2         3          4
                    (27)      (7)        (8)        (6)       (1)        (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%)
```

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.

```
Number of iterations: 4
Within cluster sum of squared errors: 5.082974846131301
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute     Full Data        0            1
                   (27)      (19)          (8)
==============================================
Energy          207.4074   150.7895     341.875
Protein               19   19.1053       18.75
Fat              13.4815         7      28.875
Calcium          43.963    58.7895        8.75
Iron             2.3815     2.3579      2.4375




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

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

Clustered Instances

0      19 ( 70%)
1       8 ( 30%)

Number of iterations: 10
Within cluster sum of squared errors: 3.413323983598554
Missing values globally replaced with mean/mode

Cluster centroids:
                    Cluster#
Attribute    Full Data       0         1         2         3         4
                  (27)     (11)       (6)       (3)       (1)       (6)
======================================================================
Energy         207.4074  159.5455  341.6667  238.3333       420       110
Protein              19   21.5455   19.1667   19.6667        15      14.5
Fat             13.4815    6.8182   28.6667        17        39       4.5
Calcium          43.963   20.3636         9    8.3333         7  146.1667
Iron             2.3815    2.0364    2.4833       2.6         2    2.8667




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

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

Clustered Instances

0      11 ( 41%)
1       6 ( 22%)
2       3 ( 11%)
3       1 (  4%)
4       6 ( 22%)
```

The seed value controls for different initial centroids. Different initial centroids could lead to different clustering outcomes.

4.

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

5.

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

2. MakeDensityBasedClusters

Now with MakeDensityBasedClusters, SimpleKMeans is turned into a denstiy-based clusterer. 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).

```
Number of iterations: 10
Within cluster sum of squared errors: 3.413323983598554
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data        0          1          2          3          4
                  (27)       (11)        (6)        (3)        (1)        (6)
==============================================================================
Energy        207.4074   159.5455   341.6667   238.3333        420        110
Protein             19    21.5455    19.1667    19.6667         15       14.5
Fat            13.4815     6.8182    28.6667         17         39        4.5
Calcium         43.963    20.3636          9     8.3333          7   146.1667
Iron            2.3815     2.0364     2.4833        2.6          2     2.8667
```

```
Clustered Instances


0        10 ( 37%)
1         6 ( 22%)
2         3 ( 11%)
3         1 (  4%)
4         7 ( 26%)



Log likelihood: -18.1147
```

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)

```
Number of iterations: 10
Within cluster sum of squared errors: 3.413323983598554
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data        0          1          2          3          4
                  (27)       (11)        (6)        (3)        (1)        (6)
==============================================================================
Energy        207.4074   159.5455   341.6667   238.3333        420        110
Protein             19    21.5455    19.1667    19.6667         15       14.5
Fat            13.4815     6.8182    28.6667         17         39        4.5
Calcium         43.963    20.3636          9     8.3333          7   146.1667
Iron            2.3815     2.0364     2.4833        2.6          2     2.8667
```

```
Clustered Instances

0        10 ( 37%)
1         6 ( 22%)
2         3 ( 11%)
3         1 (  4%)
4         7 ( 26%)



Log likelihood: -18.33321
```