10/4/2016

K-Means Algorithm Report

Bijie Zhu

CSCI 780 Machine Learning

Part 1: The algorithm steps and implementation:

Step 1: First read in the dataset into a 2D array, each row represent a 7 dimensional data point.

Step 2: Then I use the random function to generate the row number, which will be my initial centroid

Step 3: for each data point:

1) Calculate the distance between the data point and the three centroid using the Euclidean Distance

2) Add the point to the closest centroid

3) If the data points changed the centroid, set the change_centriod variable to true;

Step 4: for each cluster: calculate the average value for all the data points in that cluster, and set it as the new centroid for that cluster.

Step 5: if the change_centriod is still false after finishing all the data points' assignment, break the loop, else continually running.

Step 6: reset the value of change_centriod to false

Step 7: repeat step 3- step 6 till the change centroid is false

Step 8: calculate the IV and EV

Part 2: The Experiment result:

1. Table for all the pairs:

Number of Run	IV	EV	IV/EV
1	313.2168	27309.287	0.011469
2	313.7343	27255.514	0.011511
3	33.2169	27309.354	0.011469
4	313.2168	27309.287	0.011469
5	313.7342	27255.627	0.011511
Best Manually Picked Result	313.2169	27309.389	0.011469

2.Detailed result:

1. randomly generate result:

```
the randomly picked sets 1th time run :
k1 is :
12.72 13.57 0.8686 5.226 3.049 4.102 4.914
```

```
k2 is :
15.05 14.68 0.8779 5.712 3.328 2.129 5.36
k3 is :
11.34 12.87 0.8596 5.053 2.849 3.347 5.003
The Value of IV is : 313.2168
The Value of EV is : 27309.287
The Value of IV/EV is : 0.011469241
Centriod 1 :
11.964415 13.274805 0.8522002 5.2292852 2.8729222 4.7597394 5.088519
Centriod 2 :
18.721804 16.297373 0.88508695 6.208934 3.7226715 3.6035905 6.0660987
Centriod 3 :
14.648472 14.460415 0.87916666 5.5637784 3.2779036 2.6489332 5.19232
----------------------------
the randomly picked sets 2th time run :
k1 is :
14.88 14.57 0.8811 5.554 3.333 1.018 4.956
k2 is :
14.29 14.09 0.905 5.291 3.337 2.699 4.825
k3 is :
18.94 16.49 0.875 6.445 3.639 5.064 6.362
The Value of IV is : 313.73425
The Value of EV is : 27255.514
The Value of IV/EV is : 0.011510855
Centriod 1 :
14.819103 14.537164 0.88052243 5.5910153 3.299359 2.706585 5.2175374
Centriod 2 :
11.988659 13.284389 0.8527366 5.2274265 2.8800857 4.583926 5.0742435
Centriod 3 :
18.721804 16.297373 0.88508695 6.208934 3.7226715 3.6035905 6.0660987
----------------------------
the randomly picked sets 3th time run :
k1 is :
14.86 14.67 0.8676 5.678 3.258 2.129 5.351
k2 is :
15.49 14.94 0.8724 5.757 3.371 3.412 5.228
k3 is :
11.84 13.21 0.8521 5.175 2.836 3.598 5.044
The Value of IV is : 313.2169
The Value of EV is : 27309.354
The Value of IV/EV is : 0.011469216
Centriod 1 :
14.648472 14.460415 0.87916666 5.5637784 3.2779036 2.6489332 5.19232
Centriod 2 :
18.721804 16.297373 0.88508695 6.208934 3.7226715 3.6035905 6.0660987
Centriod 3 :
11.964415 13.274805 0.8522002 5.2292852 2.8729222 4.7597394 5.088519
----------------------------
the randomly picked sets 4th time run :
k1 is :
11.14 12.79 0.8558 5.011 2.794 6.388 5.049
k2 is :
16.87 15.65 0.8648 6.139 3.463 3.696 5.967
k3 is :
16.19 15.16 0.8849 5.833 3.421 0.903 5.307
The Value of IV is : 313.2168
The Value of EV is : 27309.287
The Value of IV/EV is : 0.011469241
Centriod 1 :
11.964415 13.274805 0.8522002 5.2292852 2.8729222 4.7597394 5.088519
Centriod 2 :
18.721804 16.297373 0.88508695 6.208934 3.7226715 3.6035905 6.0660987
Centriod 3 :
14.648472 14.460415 0.87916666 5.5637784 3.2779036 2.6489332 5.19232
----------------------------
the randomly picked sets 5th time run :
k1 is :
19.14 16.61 0.8722 6.259 3.737 6.682 6.053
k2 is :
15.69 14.75 0.9058 5.527 3.514 1.599 5.046
k3 is :
15.56 14.89 0.8823 5.776 3.408 4.972 5.847
The Value of IV is : 313.73422
The Value of EV is : 27255.627
The Value of IV/EV is : 0.011510805
Centriod 1 :
18.721804 16.297373 0.88508695 6.208934 3.7226715 3.6035905 6.0660987
Centriod 2 :
14.819103 14.537164 0.88052243 5.5910153 3.299359 2.706585 5.2175374
Centriod 3 :
11.988659 13.284389 0.8527366 5.2274265 2.8800857 4.583926 5.0742435
----------------------------
```

2. Manually picked result (the best one):

```
The Value of IV is : 313.21686
The Value of EV is : 27309.389
The Value of IV/EV is : 0.0114692
Centriod 1 :
18.721804 16.297373 0.88508695 6.208934 3.7226715 3.6035905 6.0660987
Centriod 2 :
14.648472 14.460415 0.87916666 5.5637784 3.2779036 2.6489332 5.19232
Centriod 3 :
11.964415 13.274805 0.8522002 5.2292852 2.8729222 4.7597394 5.088519
----------------------------
```

Part 3: The Experiment Explanation:

1. Random Generation Process:

I use the random function in java to generate the number of rows in a 2D array, which one row represent a seven dimensional data points. Then the 3 clusters is fed into the algorithm to generate the outputs.

2. The Optimal Result Range from the observation of Random Generation Process:

The best result from the random generation process is 2,3 and 4. One reason I guessed is that the three centroid points randomly picked in these groups are spread broadly (the difference between the starting centroids are relatively far from each other).

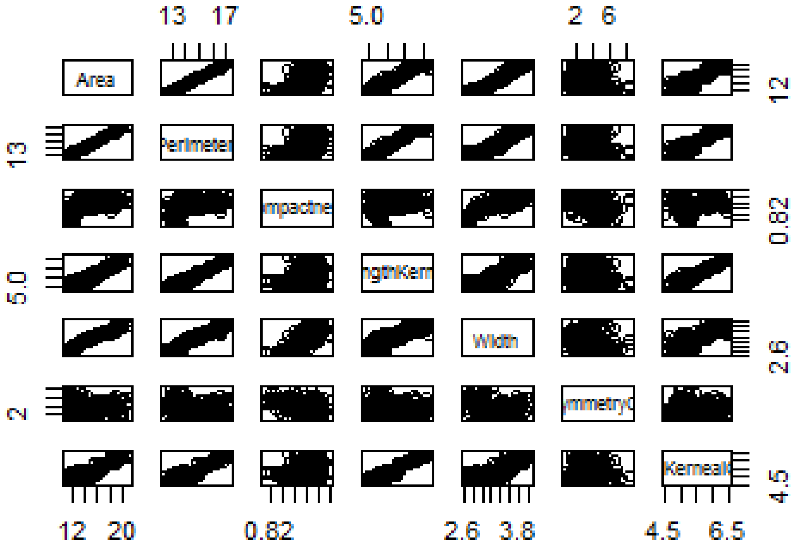
3. Important Implementation Details:

1) Data Structure: I used 2D array to represent the data, and each row is one data point and each column represent a different feature.

2) The variable change_centriod is a Boolean variable to decide when the convergence happening.

4. The Manually Picking Centroid Process:

1) First I generated a relational Matrix for all features using R:



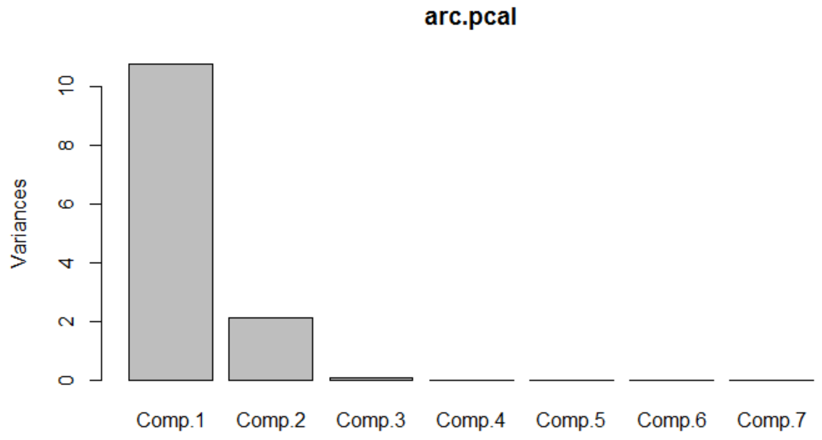
Between the different features, it shows a pretty straight forward linear relationship between most of the pairs. But two features are exceptions : asymmetry coefficient and compactness.

2) Then I did a principal component analysis in order to reduce the data dimension:

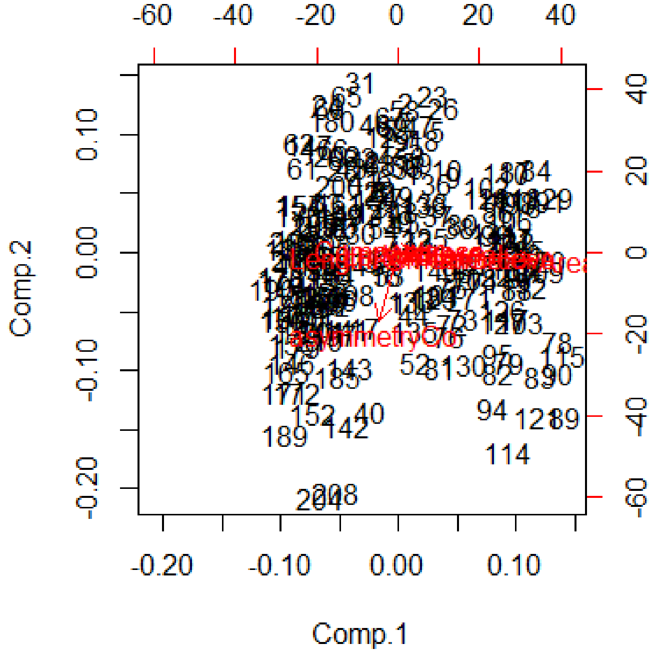
```
Loadings:
                    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Area                 0.884 -0.101  0.265 -0.199  0.137  0.281       
Perimeter            0.395        -0.283  0.579 -0.575 -0.302       
Compactness                                                   0.994
```

```
LengthKernal         0.129        -0.400  0.436  0.787 -0.113
Width                0.111         0.319 -0.234  0.145 -0.896
asymmetryCo         -0.128 -0.989                             
LengthKernealGroove  0.129        -0.762 -0.613        -0.110

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```



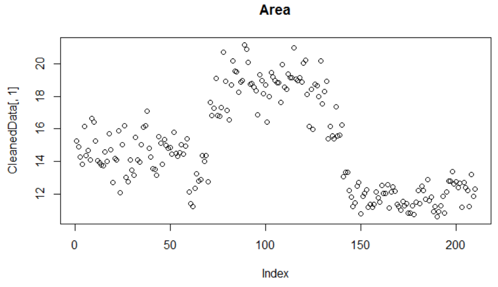
The first Component accounts for majority of the variance inside the data. Inside the first component, the Area and Perimeter accounts for two largest loadings.

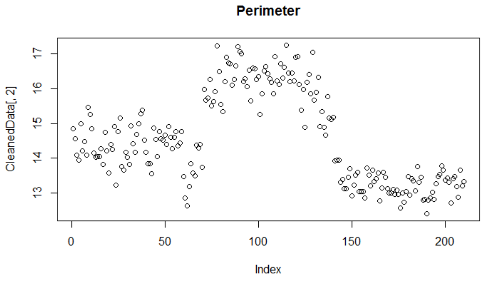


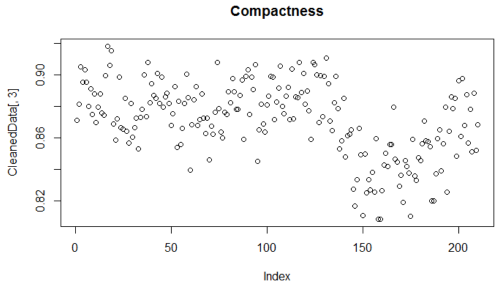
In addition, the biplot shows that asymmetry coefficient has pointed a different direction, showing a significant deviation from other features.

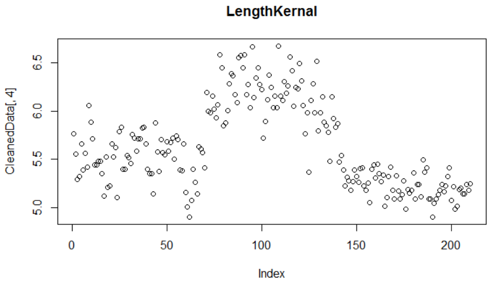
I also plotted the distribution for all seven different features:

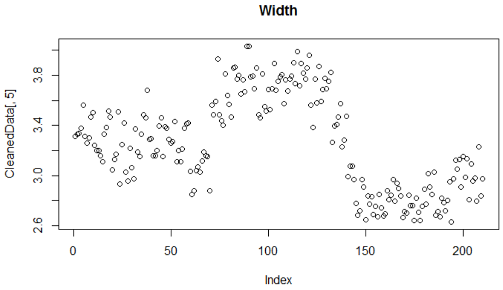
areas, perimeter, length of the kernel and length of the kernel groove shows a distribution similar to normal distribution. And the asymmetry coefficient shows a totally random distribution.

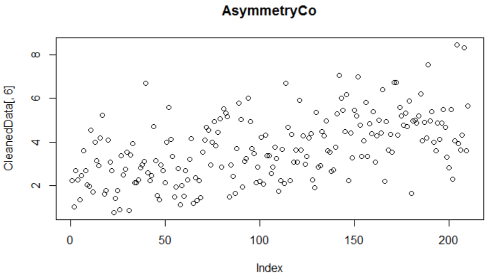

Area

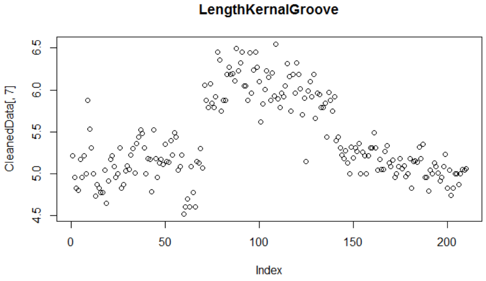

Perimeter


Compactness


LengthKernal







The ranges for concentration of the features are:

```
 Area: 14,16,20
 Perimeter: 14,15,17
 Compactness: 0.82,0.86,0.90
 LengthKernal:5.5,6.0,6.5
 Width: 3.0,3.6,3.8
 Asymmetry:2,4,6
 LengthKernalGrave: 5.0,5.5,6.0
```

3) I manually constructed three groups of centroids:

Group 1: I make the areas and perimeter as two dominate features. The way I match them is assuming a linear relationship (from the pair plot). So the lower area is matched with lower perimeter. Vice versa. For other features, I hold them constant across 3 clusters and use their averages as the centroid data.

```
float [] k1={14,14,(float) 0.87099,(float) 5.628533, (float) 3.258605,(float)
3.700201,(float) 5.408071};
float [] k2={17,15,(float) 0.87099,(float) 5.628533, (float) 3.258605,(float)
3.700201,(float) 5.408071};
float [] k3={19,17,(float) 0.87099,(float) 5.628533, (float) 3.258605,(float)
3.700201,(float) 5.408071};
```

Group 2: Since the asymmetry coefficient is deviated from other features so much, I make this group to only test asymmetry while holding others features the same(using mean).

```
float [] k1={17,15,(float) 0.87,(float) 5.628533, (float) 3.2,(float)
2,(float) 5.408071};
float [] k2={17,15,(float) 0.87,(float) 5.628533, (float) 3.2,(float)
4,(float) 5.408071};
float [] k3={17,15,(float) 0.87,(float) 5.628533, (float) 3.2,(float)
6,(float) 5.408071};
```

Group 3: Since most of the features has a positive linear correlation according to the pair plot, so for this group, I assume every feature is positively correlated.

```
float [] k1={14,14,(float) 0.82,(float) 5.5, (float) 3,(float) 2,(float) 5};
float [] k2={17,15,(float) 0.86,(float) 6.0, (float) 3.6,(float) 4,(float)
5.5};
float [] k3={19,17,(float) 0.9,(float) 6.5, (float) 3.8,(float) 6,(float) 6};
```

Result table for manually picked data points:

Number of Group	IV	EV	IV/EV
1	313.7342	27255.49	0.011511
2	313.2169	27309.39	0.011469
3	313.7342	27255.49	0.011511

Result discuss:

From the result, it shows that the second manually picked group actually has the best result. The second group is dominated by the asymmetry coefficient feature and is deviated from most other features. In addition, the asymmetry feature's distribution is relatively random compare to other features. I think this shows that a wider spread centroid starting point will generate a better result. This may due to the reason that 1) it covers more areas of the data points 2) the more areas the centroids cover, the more calculation they needs to reach convergence, which may lead to a more accurate result.