DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Assignment 2

Data preprocessing and clustering

ENGG 5103: Techniques for data mining Deadline: 5:00 pm, Nov 22th, 2016

Submission: On CUHK E-learning

Name: Lin Huangjing

SID: 1155072677

---

1. Data preprocessing

(a) After cleaning the record "Lion" that misses the value, the cleaned result is as follow:

Student id	Name	Gender	Current GPA	CS Student	Course Enrolled
1034	Jerry	Male	3.2	Yes	Data Mining
1034	Jerry	Male	3.2	Yes	Physics
2578	Tom	Male	3.8	No	Data Mining
2578	Tom	Male	3.8	No	Linear Algebra
3729	Spike	Male	2.0	No	Linear Algebra
4280	Lily	Female	4.0	No	Physics
5530	Lucy	Female	3.7	Yes	Data Mining

(b) The records without name are presented as follow:

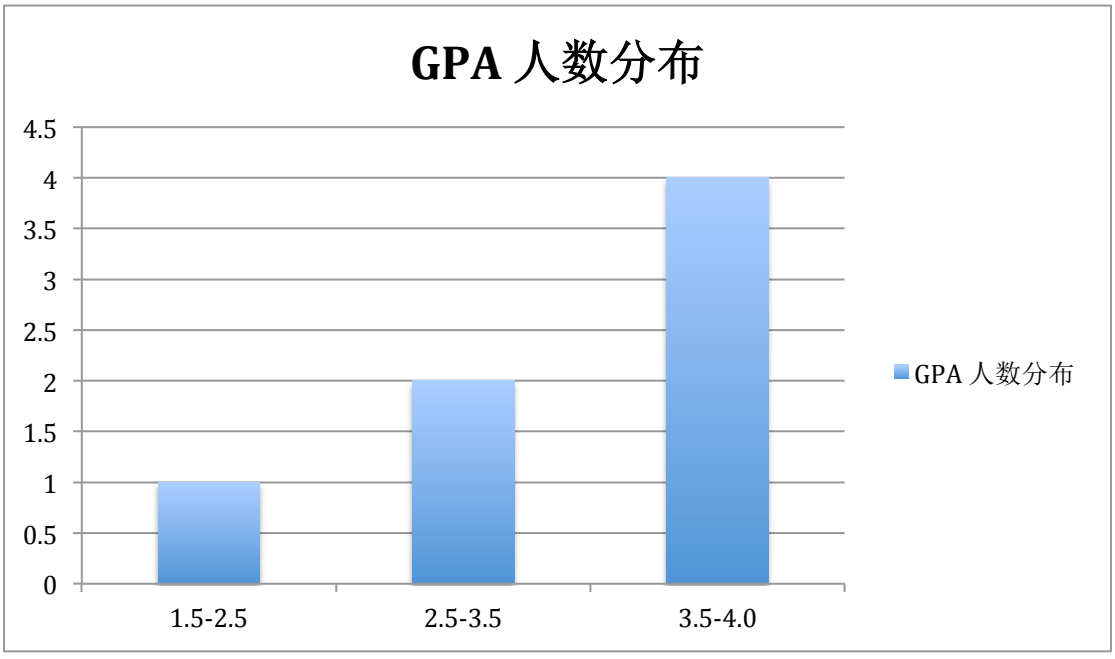
Student id	Gender	Current GPA	CS Student	Course Enrolled
1034	Male	3.2	Yes	Data Mining
1034	Male	3.2	Yes	Physics
2578	Male	3.8	No	Data Mining
2578	Male	3.8	No	Linear Algebra
3729	Male	2.0	No	Linear Algebra
4280	Female	4.0	No	Physics
5530	Female	3.7	Yes	Data Mining

(c) The coded data are presented as follow:

Student id	Gender	Current GPA	CS Student	Course Enrolled
1034	Male	3.2	1	1
1034	Male	3.2	1	2
2578	Male	3.8	0	1
2578	Male	3.8	0	3
3729	Male	2.0	0	3
4280	Female	4.0	0	2
5530	Female	3.7	1	1

(CS Student: {1:Yes, 0:No}, Course Enrolled: {1: Data Mining, 2: Physics, 3: Linear Algebra})

(d) Distribution of GPA after processing step (a),(b) and (c):



2. K-means Clustering

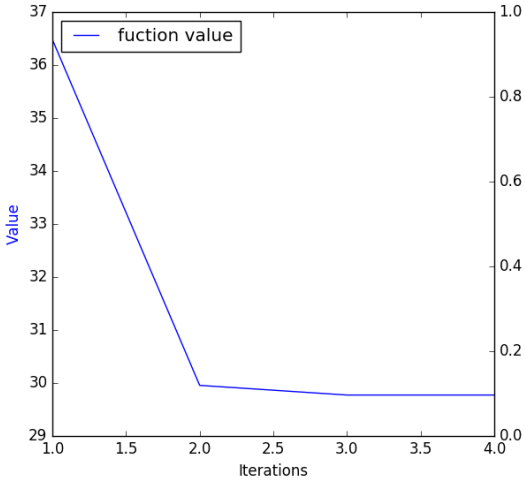
(a) Euclidean distance as distance metric

The centroids of clustering results are [[2.4, 3.7], [3.0, 1.2]].

The values of objective function are as table:

Iterations	1	2	3	4
Function Value	36.4803	29.9554	29.7729	29.7729

The line chart of object function value with iteration is presented as follow:



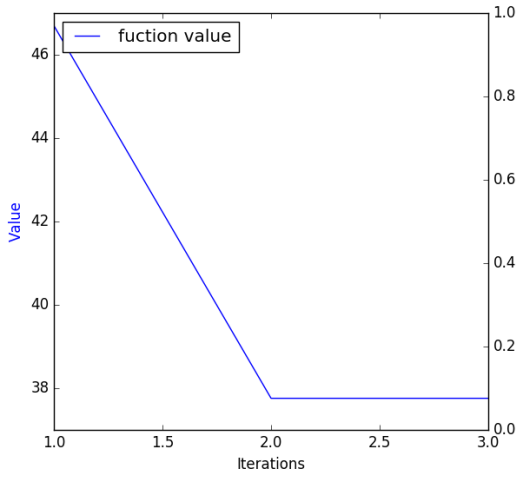
(b) Cityblock distance as distance metric

The centroids of clustering results are : [[2.6, 3.4], [2.8, 0.8]]

The values of objective function are as table:

Iterations	1	2	3
Function Value	46.7000	37.7582	37.7582

The line chart of object function value with iteration is presented as follow:



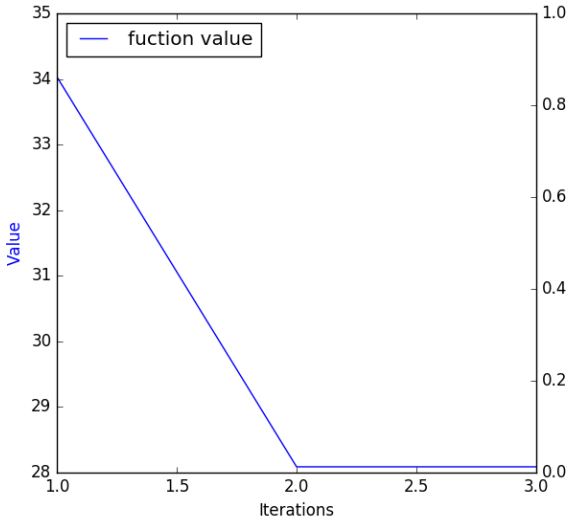
(c) Minkowski distance with lambda=3 as distance metric

The centroids of clustering results are : [[2.1, 3.7], [3.1, 1.5]]

The values of objective function are as table:

Iterations	1	2	3
Function Value	34.0333	28.0884	28.0884

The line chart of object function value with iteration is presented as follow:



(d) Answer the Questions

[1] **Answer:** The clustering results are different if given the same initialization center and different distance metric. The reason is that he distance of two points will be different if it is calculated by different metric rule.

[2] **Answer:** The Euclidean distance can be used in setting distribution center for several communities on a 2-D map. The Cityblock distance is suitable for problem of locating several bus stations in a block-structure city. The Minkowski distance is suitable for high dimension situations.

3. Hierarchical Clustering

(a) Initialization to 5 Clusters

The initial proximity matrix is as

```
0.00 2.48 4.46 2.53 4.53 3.93 2.42 1.58 1.98 4.50 3.86 2.97 0.81 4.20 2.18 2.19 1.00 2.19 2.75 1.75
2.48 0.00 1.98 1.48 2.50 2.25 2.84 2.73 0.54 2.10 1.49 0.71 3.26 2.97 1.26 0.82 1.88 2.72 2.38 2.41
4.46 1.98 0.00 2.72 2.00 2.44 4.37 4.53 2.50 0.85 1.00 1.61 5.25 3.32 2.85 2.47 3.78 4.32 3.55 4.07
2.53 1.48 2.72 0.00 3.88 3.72 3.94 3.52 1.63 3.21 2.71 1.20 3.30 4.44 2.55 0.71 1.55 3.76 0.92 1.42
4.53 2.50 2.00 3.88 0.00 0.81 3.36 3.89 2.73 1.17 1.25 2.72 5.19 1.50 2.37 3.31 4.26 3.41 4.81 4.90
3.93 2.25 2.44 3.72 0.81 0.00 2.56 3.14 2.32 1.71 1.48 2.66 4.53 0.89 1.75 3.07 3.80 2.62 4.62 4.54
2.42 2.84 4.37 3.94 3.36 2.56 0.00 0.91 2.40 3.96 3.42 3.55 2.62 2.33 1.60 3.27 2.97 0.22 4.58 3.89
1.58 2.73 4.53 3.52 3.89 3.14 0.91 0.00 2.21 4.27 3.67 3.41 1.71 3.10 1.70 2.93 2.30 0.70 4.02 3.20
1.98 0.54 2.50 1.63 2.73 2.32 2.40 2.21 0.00 2.52 1.89 1.20 2.76 2.92 0.92 0.92 1.53 2.26 2.44 2.22
4.50 2.10 0.85 3.21 1.17 1.71 3.96 4.27 2.52 0.00 0.64 2.01 5.25 2.57 2.58 2.79 3.98 3.94 4.10 4.44
3.86 1.49 1.00 2.71 1.25 1.48 3.42 3.67 1.89 0.64 0.00 1.51 4.61 2.37 1.97 2.22 3.36 3.38 3.62 3.86
2.97 0.71 1.61 1.20 2.72 2.66 3.55 3.41 1.20 2.01 1.51 0.00 3.77 3.47 1.96 0.86 2.19 3.42 2.11 2.47
0.81 3.26 5.25 3.30 5.19 4.53 2.62 1.71 2.76 5.25 4.61 3.77 0.00 4.69 2.82 3.00 1.75 2.41 3.40 2.32
4.20 2.97 3.32 4.44 1.50 0.89 2.33 3.10 2.92 2.57 2.37 3.47 4.69 0.00 2.14 3.76 4.28 2.46 5.32 5.10
2.18 1.26 2.85 2.55 2.37 1.75 1.60 1.70 0.92 2.58 1.97 1.96 2.82 2.14 0.00 1.84 2.14 1.50 3.34 2.97
2.19 0.82 2.47 0.71 3.31 3.07 3.27 2.93 0.92 2.79 2.22 0.86 3.00 3.76 1.84 0.00 1.35 3.10 1.57 1.65
1.00 1.88 3.78 1.55 4.26 3.80 2.97 2.30 1.53 3.98 3.36 2.19 1.75 4.28 2.14 1.35 0.00 2.76 1.77 0.92
2.19 2.72 4.32 3.76 3.41 2.62 0.22 0.70 2.26 3.94 3.38 3.42 2.41 2.46 1.50 3.10 2.76 0.00 4.38 3.68
2.75 2.38 3.55 0.92 4.81 4.62 4.58 4.02 2.44 4.10 3.62 2.11 3.40 5.32 3.34 1.57 1.77 4.38 0.00 1.12
1.75 2.41 4.07 1.42 4.90 4.54 3.89 3.20 2.22 4.44 3.86 2.47 2.32 5.10 2.97 1.65 0.92 3.68 1.12 0.00
```

The 5 centroids are generated by hierarchical cluster as follow:

C1= [0.567, 2.633]

C2 = [3.414, 1.900]

C3 = [1.975, 0.525]

C4 = [4.267, 4.033]

C5 = [2.467, 4.867]

(b) Complete linkage.

Step 0:

C1 = [0.567, 2.633]

C2 = [3.414, 1.900]

C3 = [1.975, 0.525]

C4 = [4.267, 4.033]

C5 = [2.467, 4.867]

The proximity metric is as follow:

```
0.00 2.94 2.54 3.96 2.93
2.94 0.00 1.99 2.30 3.11
2.54 1.99 0.00 4.19 4.37
3.96 2.30 4.19 0.00 1.98
2.93 3.11 4.37 1.98 0.00
```

Step 1:

C1 = [3.367, 4.450]

C1 = [3.414, 1.900]

C2 = [1.975, 0.525]

C3 = [0.567, 2.633]

The proximity metric is as follow:

0.00 2.55 4.16 3.34
2.55 0.00 1.99 2.94
4.16 1.99 0.00 2.54
3.34 2.94 2.54 0.00

Step 2:

C1 = [2.695, 1.215]

C2 = [0.567, 2.633]

C3 = [3.667, 4.450]

The proximity metric is as follow:

0.00 2.56 3.31
2.56 0.00 3.34
3.31 3.34 0.00

Step 3:

C1 = [1.631, 1.923]

C2 = [3.367, 4.450]

The proximity metric is as follow:

0.00 3.07
3.07 0.00

Step 4:

C1 = [2.499, 3.186]

The proximity metric is N/A.

(c) Single Linkage

Step 0:

C1 = [0.567, 2.633]

C2 = [3.414, 1.900]

C3 = [1.975, 0.525]

C4 = [4.267, 4.033]

C5 = [2.467, 4.867]

The proximity metric is as follow:

```
0.00 2.94 2.54 3.96 2.93
2.94 0.00 1.99 2.30 3.11
2.54 1.99 0.00 4.19 4.37
3.96 2.30 4.19 0.00 1.98
2.93 3.11 4.37 1.98 0.00
```

Step 1:

C1 = [3.367, 4.450]

C2 = [3.414, 1.900]

C3 = [1.975, 0.525]

C4 = [0.567, 2.633]

The proximity metric is as follow:

```
0.00 2.55 4.16 3.34
2.55 0.00 1.99 2.94
4.16 1.99 0.00 2.54
3.34 2.94 2.54 0.00
```

Step 2:

C1 = [2.695, 1.215]

C2 = [0.567, 2.633]

C3 = [3.667, 4.450]

The proximity metric is as follow:

```
0.00 2.56 3.31
2.56 0.00 3.34
3.31 3.34 0.00
```

Step 3:

C1 = [1.631, 1.923]

C2 = [3.367, 4.450]

The proximity metric is as follow:

0.00 3.07

3.07 0.00

Step 4:

C1 = [2.499, 3.186]

The proximity metric is N/A.

(d) Group Average Linkage

Step 0:

C1 = [0.567, 2.633]

C2 = [3.414, 1.900]

C3 = [1.975, 0.525]

C4 = [4.267, 4.033]

C5 = [2.467, 4.867]

The proximity metric is as follow:

```
0.00 2.94 2.54 3.96 2.93
2.94 0.00 1.99 2.30 3.11
2.54 1.99 0.00 4.19 4.37
3.96 2.30 4.19 0.00 1.98
2.93 3.11 4.37 1.98 0.00
```

Step 1:

C1 = [3.367, 4.450]

C2 = [3.414, 1.900]

C3 = [1.975, 0.525]

C4 = [0.567, 2.633]

The proximity metric is as follow:

0.00 2.55 4.16 3.34

2.55 0.00 1.99 2.94

4.16 1.99 0.00 2.54

3.34 2.94 2.54 0.00

Step 2:

C1 = [2.695, 1.215]

C2 = [0.567, 2.633]

C3 = [3.667, 4.450]

The proximity metric is as follow:

$$\begin{matrix} 0.00 & 2.56 & 3.31 \\ 2.56 & 0.00 & 3.34 \\ 3.31 & 3.34 & 0.00 \end{matrix}$$

Step 3:

C1 = [1.631, 1.923]

C2 = [3.367, 4.450]

The proximity metric is as follow:

$$\begin{matrix} 0.00 & 3.07 \\ 3.07 & 0.00 \end{matrix}$$

Step 4:

C1 = [2.499, 3.186]

The proximity metric is N/A.

(e) Answer Question

[1] Answer: The proximity matrices contain the distance information among the different clusters. In each iteration, the most proximate two clusters will be clustered together. This process is guided by the proximity matrices generated in different phases.

[2] Single linkage is sensitive to noisy point. The complete and average linkages are robust with outliers, so it can be used in some situation with noisy points.

---

4. SOM Clustering

(a)

(b)

(c)

(d) Discuss Issues

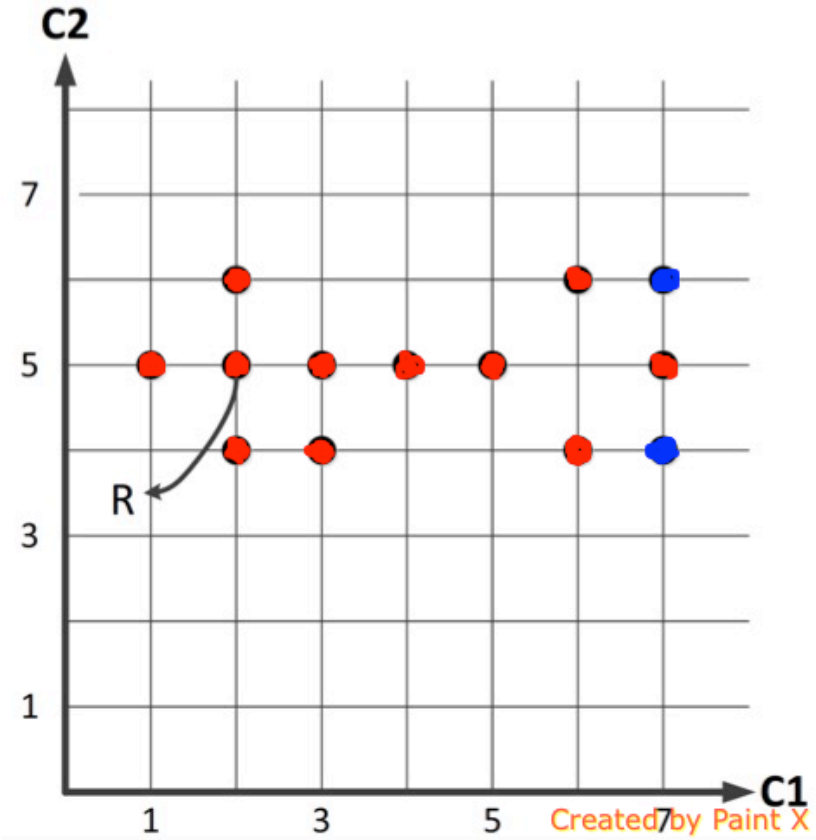
[1] Answer: If “a” is too large, the procedure will be rapid and unstable. If the “a” is too small, it will be slow to get the result.

[2] Answer: Yes , it is useful. The procedure should be fast at beginning and slow down after certain iterations

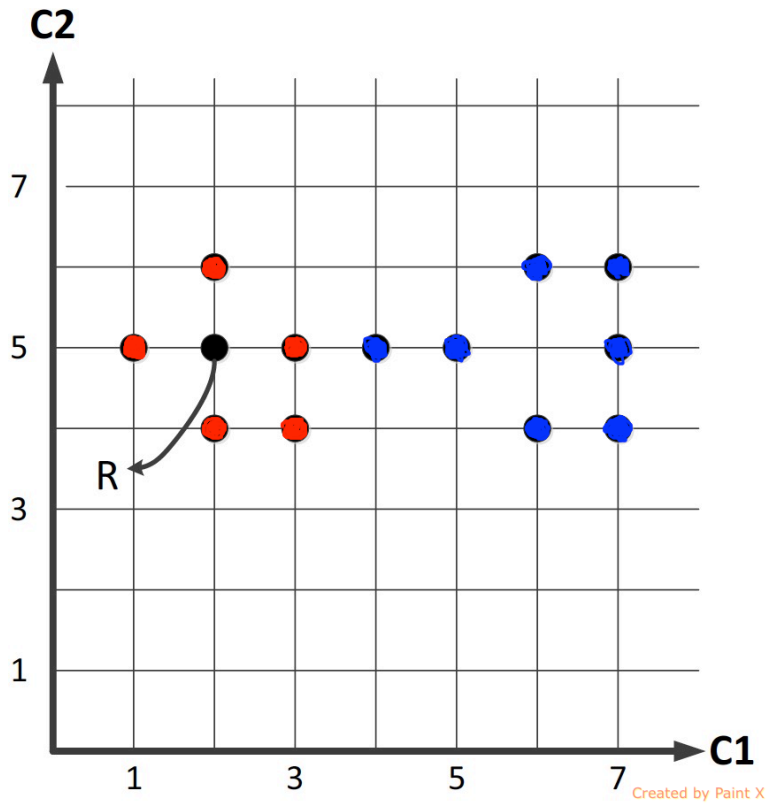
[3] Answer: If the neighborhood shrinks very rapidly, the results will be the same as one gets with the sequential k-means procedure.

5. Dense based Clustering

(a) The red points are core points. The blue points are border points



(b) The red points are directly reachable. The blue points are indirectly reachable.



(c) Answer: I would like to set Eps as 1.2 and Minpts as 4. The reason is that the Eps with 1.5 is too large so that all the points are connected together. The Eps with 1.0< Eps=1.2 < sqrt(2)=1.414 can divide the points very well.

---

6. Cluster Cohesion and Separation

(a) 2 Clusters (red points and yellow points)

The centroid of red = [2.333, 5.0]

The centroid of yellos = [5.667, 5.0]

SSE = $\sum_i \sum_{x \in C_i} (x - c_i)^2$ = 9.33

SSB = $\sum_i |C_i| (c - c_i)^2$ = 16.67

TSS = $\sum_i \sum_{x \in C_i} (x - c)^2$ = 26

(b) 3 Clusters (red points, green points and yellow points)

The centroid of red = [4.0, 6.0]

The centroid of green = [4.0, 5.0]

The centroid of yellow = [4.0, 4.0]

SSE = $\sum_i \sum_{x \in C_i} (x - c_i)^2$ = 22

SSB = $\sum_i |C_i| (c - c_i)^2$ = 4

TSS = $\sum_i \sum_{x \in C_i} (x - c)^2$ = 26

(c) General Case

SSE = $\sum_i \sum_{x \in X_i} (x - \frac{\sum X_j}{n_j \cdot e})^2$

SSB = $\sum_i |n_i| (\frac{\sum X_j}{n_j \cdot e})^2$

TSS = $\sum_i \sum_{x \in X_i} (x)^2$

(d) Objective Function

Function = 1/ ($\sum_i \sum_{x \in C_i} (x - c_i)^2$)