Seoul National University

M1522.001400-001 Introduction to Data Mining

Homework 4: Frequent Itemsets (Chapter 6) & Clustering (Chapter 7)

2017-18538 Hwang SunYoung

1.

Basket	Items
1	a, b, c
2	a, b, d
3	b, c, d, e
4	c, e, f
5	b, c, f
6	a, c, e

(1)

Support of {a} = 3

Support of {b} = 4

Support of {c} = 5

Support of {b, c} = 3

Support of {a, c, e} = 1

(2) minimum support = 3

Frequent itemsets: {a}, {b}, {c}, {e}, {b, c}, {c, e}

(3)

Confidence of {a, b} -> c is

conf({a, b}→c) = $\frac{support(\{a,b\} \cup c)}{support(\{a,b\})}$ = 1/2 = 0.5

(4)

Interest of {a, c} -> e is

Interest({a, c}→e)= conf({a, c}→e) - Pr[e] (Pr[e] is fraction of baskets that contain e)

$= \frac{support(\{a,c\} \cup e)}{support(\{a,c\})} - 3/6$

$= 1/2 - 1/2 = 0$

2.

(1)

Basket	Itemsets
1	1
2	1, 2
3	1, 3
4	1, 2, 4
5	1, 5
6	1, 2, 3, 6
7	1, 7
8	1, 2, 4, 8
9	1, 3, 9
10	1, 2, 5, 10

(2) minimum support = 2

By A-Priori algorithm,

C_1 = {{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}}

Count the support of itemsets in C_1 and prune non-frequent. (save itemset that is support≥2)

L_1 = {{1}, {2}, {3}, {4}, {5}}

Generate C_2 = {{1, 2}, {1, 3}, {1, 4}, {1, 5}, {2, 3}, {2, 4}, {2, 5}, {3, 4}, {3, 5}, {4, 5}}

Count the support of itemsets in C_2 and prune non-frequent. (save itemset that is support≥2)

L_2 = {{1, 2}, {1, 3}, {1, 4}, {1, 5}, {2, 4}}

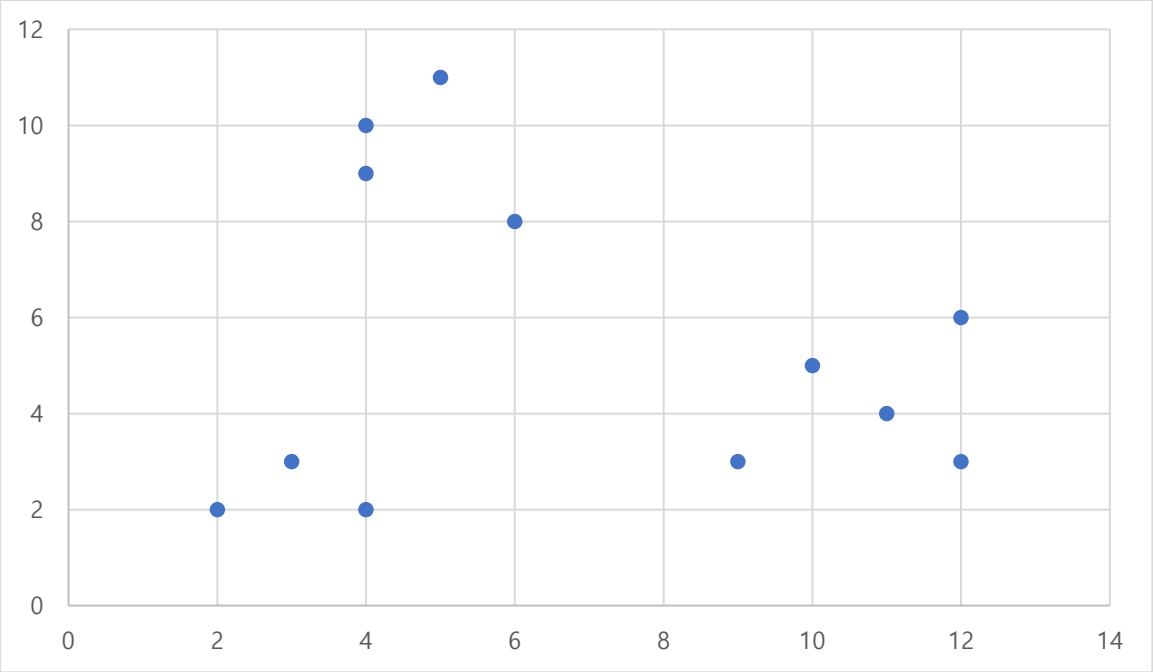
Generate C_3 = {{1, 2, 4}} ←{1, 2}, {1, 4}, {2, 4} are frequent.

Count the support of itemsets in C_3 and prune non-frequent. (save itemset that is support≥2)

L_3 = {{1, 2, 4}}

Frequent itemsets are {1}, {2}, {3}, {4}, {5}, {1, 2}, {1, 3}, {1, 4}, {1, 5}, {2, 4}, {1, 2, 4}

3.



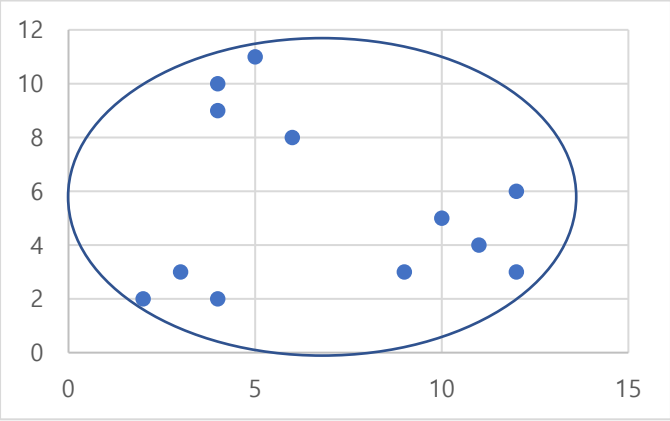
(1)

Best K is 3.

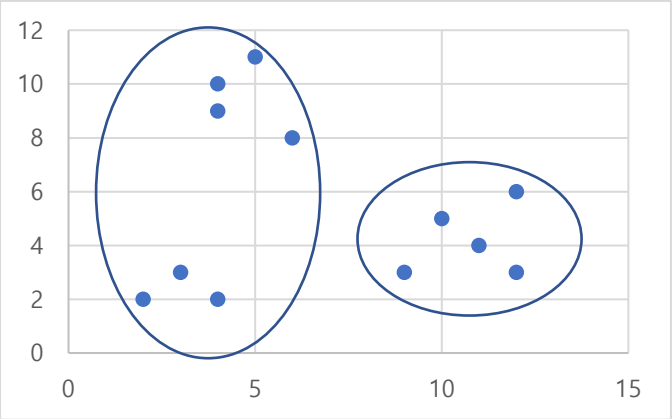
(2)

Method to select k is "Finding the Knee" method. Try different k, looking at the change in the average distance to centroid as k increases.(k is the number of clusters.) Average falls rapidly until right k(=best k), then changes little.

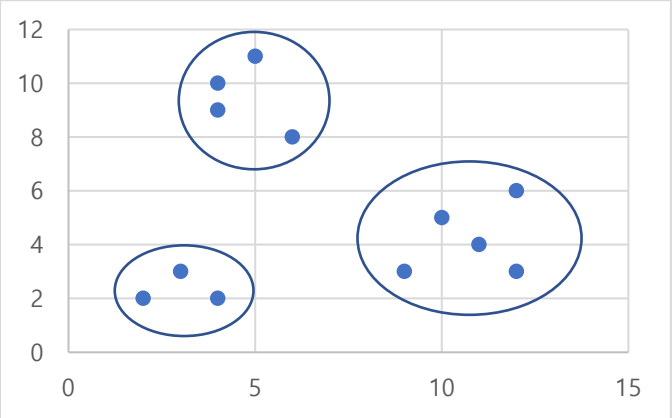
All result is rounded to the second digit after the decimal point.



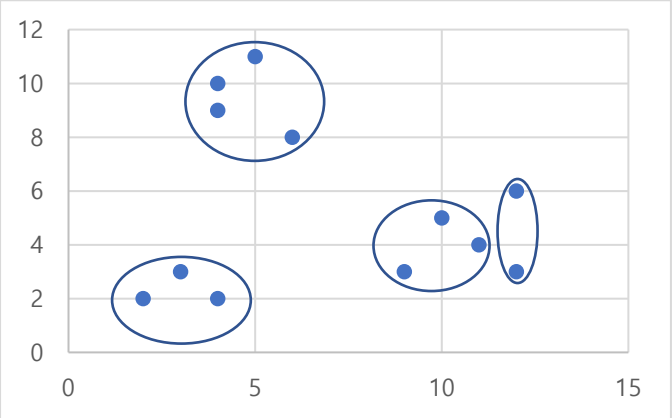
If k=1, centroid is (6.83, 5.5) and average distance to centroid of each point is 4.60



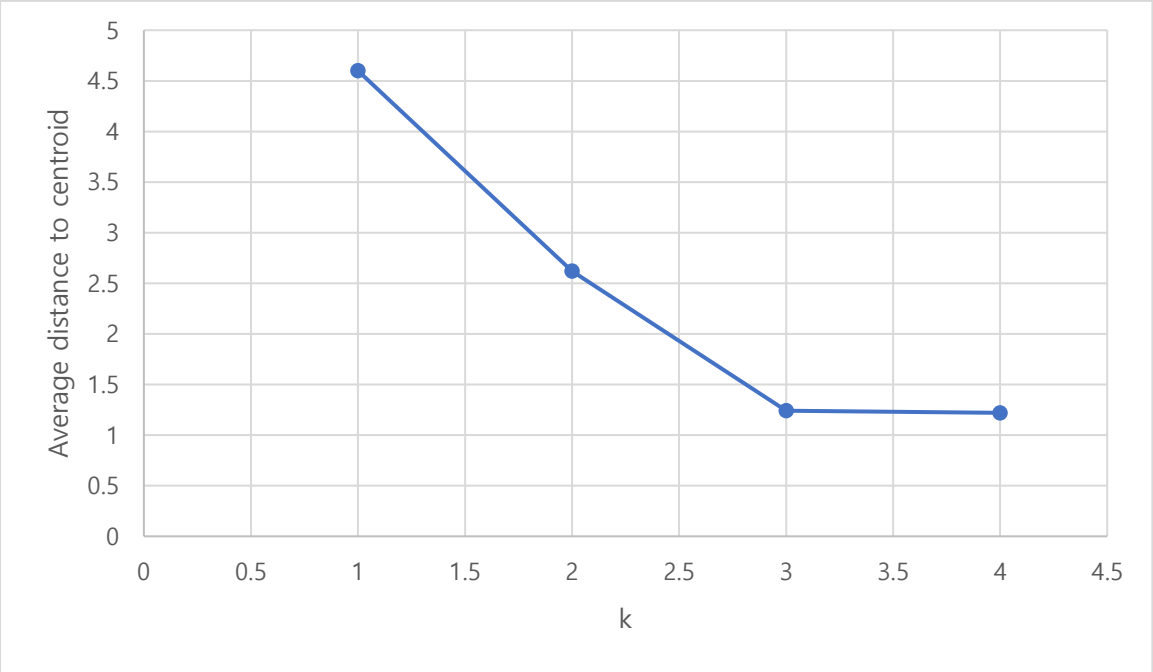
If k=2, centroids are (4, 6.43), (10.8, 4.2) and average distance to centroid of each point is 2.62



If k=3, centroids are (3, 2.33), (4.75, 9.5), (10.8, 4.2) and average distance to centroid of each point is 1.24



If k=4, centroids are (3, 2.33), (4.75, 9.5), (10, 4), (12, 4.5) and average distance to centroid of each point is 1.22



Because average falls rapidly until k=3 and after k=3 changes little, the best value of k is 3.

4.

(1) SUMSQ is sum of the squares of coordinates

cluster	points	N	SUM	SUMSQ
1	(4, 8) (4, 10) (6, 8) (7, 10)	4	(4+4+6+7, 8+10+8+10) =(21, 36)	(117, 328)
2	(2, 2) (3, 4) (5, 2)	3	(2+3+5, 2+4+2) =(10, 8)	(38, 24)
3	(9, 3) (10, 5) (11, 4) (12, 3) (12, 6)	5	(9+10+11+12+12, 3+5+4+3+6) =(54, 21)	(590, 95)

(2) the variance in the i-th dimension is $(SUMSQ_i/N)$ - $(SUM_i/N)^2$

Variance of cluster 1 is $((117/4) - (21/4)^2, (328/4) - (36/4)^2) = (27/16, 1)$

Variance of cluster 2 is ((38/3) - $(10/3)^2$, (24/3) - $(8/3)^2$) = (14/9, 8/9)

Variance of cluster 3 is ((590/5) - $(54/5)^2$, (95/5) - $(21/5)^2$) = (34/25, 34/25)

Standard deviation is the square root of that

Standard deviation of cluster 1 is $(\frac{3\sqrt{3}}{4}, 1)$

Standard deviation of cluster 2 is $(\frac{\sqrt{14}}{3}, \frac{2\sqrt{2}}{3})$

Standard deviation of cluster 3 is $(\frac{\sqrt{34}}{5}, \frac{\sqrt{34}}{5})$