Assignment 4

1. Consider the following data set shown in Table 1.

Table 1. Market basket Transaction.

|  |  |
| --- | --- |
| Transaction ID | Items Bought |
| 1 | Chips, Cookies, Regular Soda, Ham |
| 2 | Chips, Ham, Boneless Chicken, Diet Soda |
| 3 | Ham, Bacon, Whole Chicken, Regular Soda |
| 4 | Chips, Ham, Boneless Chicken, Diet Soda |
| 5 | Chips, Bacon, Boneless Chicken |
| 6 | Chips, Ham, Bacon, Whole Chicken, Regular Soda |
| 7 | Chips, Cookies, Boneless Chicken, Diet Soda |

1. Find the frequent itemsets (up to size 4) for the support threshold min\_support = 0.5 by using the Apriori approach.

1-itemset

|  |  |  |
| --- | --- | --- |
| **Item** | **Count** | **Support** |
| {Chips} | 6 | 0.8571 |
| {Cookies} | 2 | 0.2857 |
| {Bacon} | 3 | 0.4286 |
| {Boneless Chicken} | 4 | 0.5714 |
| {Diet Soda} | 3 | 0.4286 |
| {Ham} | 5 | 0.7143 |
| {Regular Soda} | 3 | 0.4286 |
| {Whole Chicken} | 2 | 0.2857 |

The highlighted cells have a min\_support of 0.5. With support-based pruning, cookies, bacon, diet soda, regular soda and whole chicken are removed.

2-itemset

|  |  |  |
| --- | --- | --- |
| **Item** | **Count** | **Support** |
| {Chips, Boneless Chicken} | 4 | 0.5714 |
| {Chips, Ham} | 4 | 0.5714 |
| {Boneless Chicken, Ham} | 2 | 0.2857 |

Here “Bonless Chicken and Ham” will be removed The highlighted because its support is less than the min\_support of 0.5.

3-itemset

|  |  |  |
| --- | --- | --- |
| **Item** | **Count** | **Support** |
| {Chips, Boneless Chicken, Ham} | 2 | 0.2857 |

There are no more frequent itemset.

1. Find the association rules for the support threshold **min\_support = 0.5** and confidence threshold **min\_confidence = 0.5** by using the Apriori and the rule generation approaches. Calculate the lift of each rule.

**Association rules**

1. Using Apriori principle done in 1(a), we know that **‘chips’, ‘boneless chicken’ and ‘ham’ are the items with** **min\_support = 0.5.**
2. **Generating rules** from the frequent itemsets {Chips, Boneless chicken}, {Chips, Ham}

Let Chips = C, Bonless chicken =B, Ham =H

Min\_confidence = 0.5

Three rules are constructed: **{C} ⇒{B}, {B}⇒{C} , {C}⇒{H} and {H}⇒{C}**

**Rule 1: {C} ⇒{B},**

Support ({C}⇒ {B})=0.5714

Confidence ({C}⇒ {B}) =

**Confidence({C}⇒ {B}) =**

**Rule 2: {B}⇒{C},**

Support ({B}⇒ {C}) = 4/7=0.5714

Confidence ({B}⇒ {C}) = (Support ({B}⇒{C}))/(support({B}))

**Confidence({B}⇒ {C}) = 0.5714/0.5714=1**

**Rule 3: {C} ⇒{H},**

Support ({C}⇒ {H})=0.5714

Confidence ({C}⇒ {H}) = (Support ({C}⇒{H}))/(support({C}))

**Confidence({C}⇒ {H}) = (0.5714)/(0.8571)=0.6667**

**Rule 4: For {H}⇒{C},**

Support ({H}⇒ {C}) = 4/7=0.5714

Confidence ({H}⇒ {C}) = (Support ({H}⇒{C}))/(support({H}))

**Confidence({H}⇒ {C}) = 0.5714/0.7143=0.8**

The confidence of the four rules {C}⇒{B} , {B}⇒{C} , {C} ⇒{H}, and {H}⇒{C} are 0.6667,1, and 0.8 respectively. **All the rules are returned because they have min\_confidence of 0.5**

**C1 = {{C},{B},{H}}**

K =2, V = {C,B,H}

Three rules are constructed: **{C,B} ⇒{H}, {C,H}⇒{B} and {B,H}⇒{C}**

Rule 5: {C,B} ⇒{H},

Support ({C,B}⇒ {H}) = support ({C,B} U {H}) = 0.2857

Rule 6: {C,H}⇒{B},

Support ({C,H}⇒ {B}) = support ({C,H} U {B}) = 0.2857

Rule 7: For {B,H}⇒{C},

Support ({B,H}⇒ {C}) = support ({B,H} U {C}) = 0.2857

These rules have less than the **min\_support=0.5** therefore no item is added to C2.

1. Lift for each rule; **{C}⇒{B} , {B}⇒{C} ,** **{C} ⇒{H} and{H}⇒{C}**

**Lift({C} ⇒{B})** =(Confidence ({C}⇒ {B}))/(support({B}))= 0.6667/0.5714=**1.1667**

Lift ({C} ⇒{B}) > 1 indicates that the occurrence of {C,H} has a positive effect on the occurrence of {B}.

**Lift({B} ⇒{C})** =(Confidence ({B}⇒ {C}))/(support({C}))= 1/0.8571=**1.1667**

Lift ({C} ⇒{B}) > 1 indicates that the occurrence of {B} has a positive effect on the occurrence of {C}.

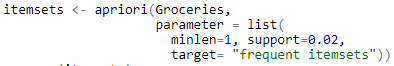
**Lift({C} ⇒{H})** =(Confidence ({C}⇒ {H}))/(support({H}))= 0.6667/0.714=**0.9338**

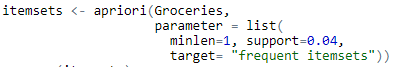
Lift ({C} ⇒{H}) < 1 indicates that the occurrence of {C} has a negative effect on the occurrence of {H}.

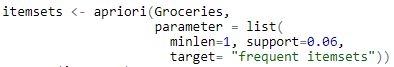
**Lift({H} ⇒{C})** =(Confidence ({H}⇒ {C}))/(support({C}))= 0.8/0.8571=**0.9338**

Lift({H} ⇒{C}) < 1 indicates that the occurrence of ({H} has a negative effect on the occurrence of({C}.

1. How many frequent itemsets will be generated if the support threshold min\_support = 0.02, 0.04, and 0.06, respectively?

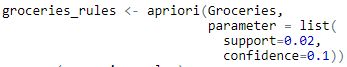
 

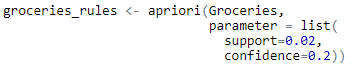
 

|  |  |
| --- | --- |
| Min\_support | Frequent itemsets |
| 0.02 | 122 |
| 0.04 | 41 |
| 0.06 | 21 |

Consider the frequent itemsets generated with the support thresholds min\_support = 0.02, 0.04, and 0.6, respectively. How many association rules will be generated from the frequent itemsets if the confidence threshold min\_confidence = 0.1, 0.2 and 0.3, respectively?

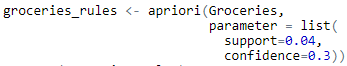
 

|  |  |  |
| --- | --- | --- |
| Min\_support | Min\_confidence | Association rules |
| 0.02 | 0.1 | 128 |
| 0.2 | 73 |
| 0.3 | 37 |
| 0.04 | 0.1 | 26 |
| 0.2 | 16 |
| 0.3 | 7 |
| 0.06 | 0.1 | 10 |
| 0.2 | 3 |
| 0.3 | 1 |

1. Consider the following data set shown in Table 2. Place the objects in the data set in two clusters using k-means. Andrew and Carolina are selected as the initial centroids of the two clusters. Find the cluster to which each object belongs to using k-means with one iteration and re-compute the centroids of the two clusters. Create a scatter plot for the data set and the re-computed centroids. The cluster to which each object belongs to is represented by color of the object in the plot.

Table 2. Social network data set

|  |  |  |
| --- | --- | --- |
| Name | Age | Educational level |
| Andrew (A) | 55 | 1 |
| Bernhard (B) | 43 | 2 |
| Carolina (C) | 37 | 5 |
| Dennis (D) | 82 | 3 |
| Eve (E) | 23 | 3.2 |
| Fred (F) | 46 | 5 |

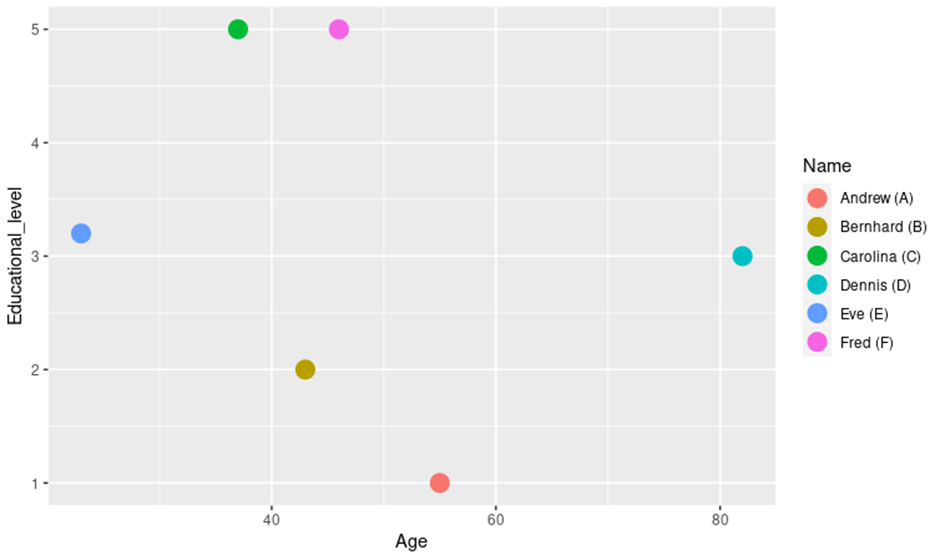


Fig. 1 Scatter plot showing the names in the social network data set.

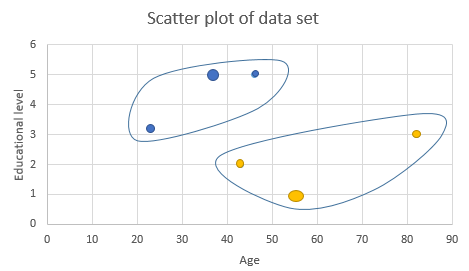


Fig. 2 Scatterplot of data with Andrew and Carolina the initial centroids of the two clusters and denoted with bigger circles.

Each object of the same cluster is represented with the same color.

Cluster 1 (Yellow) => Andrew, Bernhard and Dennis. Centriod 1 => Andrew

Cluster 2 (Blue) => Carolina, Eve and Fred. Centroid 2=> Carolina

First iteration: Recompute centroid

Cluster 1:

Age = (55+43+82)/3= 60

Educational level =(1+2+3)/3 =2

Cluster 2:

Age = (37+23+46)/3= 35.3

Educational level =(5+3.2+5)/3 =4.4

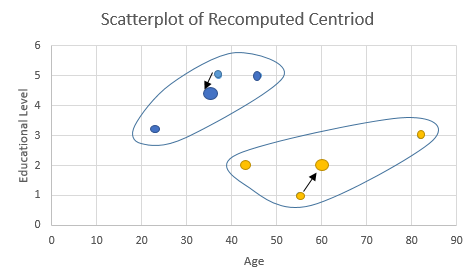


Fig. 3 Scatterplot of the recomputed centroids (bigger circles) of the two clusters.