Homework 4: Saurav Maheshwary (samahesh)

**9 April 2016**

Answer 1:

From reference [6] I figured that if we select a high threshold we increase the specificity of the test but then lose sensitivity. Also if we select a low threshold we increase the test’s sensitivity but lose specificity. So we need to find a balance and calculate the Optimum Threshold. To do that we need to pick a threshold for each tuple. We will do that by picking the one with maximal sum of true-positive and false-negative values.

Now, to find the minimum and maximum number of thresholds we have to do work on the probability. We have to sort the data points on the basis of the probabilities and then we choose each distinct posterior probability as one of the thresholds. So now it can be that we get the maximum threshold that is equal to the total number of data points that is

n = n+ + n-

Further, from the paper I realized that for the minimum thresholds we need to find at least 2 thresholds so that we can plot them on the axes.

In this way we can find the maximum and the minimum of the available probabilities as our threshold. Also interestingly we can join these two points to create a line which will be considered as the ROC.

Now, if we are given the points (0, 0) and (1, 1) from beforehand then we need to only find a point to form the ROC.

An algorithm for the ROC can be found here: Algorithm 1 on Page 866 of reference [6].

This algorithm generates a ROC curve and displays the maximum number of thresholds taken.

Calculating maximum number of thresholds:

If there are n number of tuples (data points) which we arrange in decreasing order of their probabilities. In the worst case if all these values turn out to be the same then we can have either of the probabilities of X as a threshold, so the maximum number of thresholds is n.

Calculating minimum number of thresholds:

We only need two points to create a line segment. So if the threshold values are arranged in the sorted order we can choose the highest and the lowest values. So minimum number of thresholds required = 2.

-------------------------------------------------------------------------------------------------------------------------------

Answer 2­­­:

I have tried to solve it using the following method:

For each point in a single cluster A, I will calculate the average of its distance to all the other points in another cluster B that I am assuming I’ll merge with A.

I’ll repeat the above step for all points in A and then add all the calculated distances D for all points in A to get Sum.

Then over all the clusters I’ll pick the lowest two Sum values that I get and merge those two clusters. This will give me the two clusters with lowest sum of squared errors.

-------------------------------------------------------------------------------------------------------------------------------

Answer 3:

I have used the following data sets from the UCI repository: [Car Evaluation](http://archive.ics.uci.edu/ml/datasets/Car+Evaluation) (1728 examples), [Contraceptive Method Choice](http://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice) (1473 examples) and [Nursery](http://archive.ics.uci.edu/ml/datasets/Nursery) (12960 examples).

Sample Output for this Question on Car Evaluation dataset:

Enter Support Threshold Value (double):

0.1

Please wait.

----------Itemsets Produced Using k-1 X 1 Method------------------

------------------------------------------------------------------

1-Itemset:[[0], [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]]

2-Itemset:[[0, 21], [1, 21], [2, 21], [3, 21], [4, 21], [5, 21], [6, 21], [7, 21], [8, 21], [9, 21], [10, 21], [11, 21], [12, 15], [12, 16], [12, 17], [12, 18], [12, 19], [12, 20], [12, 21], [13, 15], [13, 16], [13, 17], [13, 18], [13, 19], [13, 20], [13, 21], [13, 22], [14, 15], [14, 16], [14, 17], [14, 18], [14, 19], [14, 20], [14, 21], [14, 22], [15, 18], [15, 19], [15, 20], [15, 21], [16, 18], [16, 19], [16, 20], [16, 21], [17, 18], [17, 19], [17, 20], [17, 21], [18, 21], [19, 21], [19, 22], [20, 21], [20, 22]]

3-Itemset:[[12, 15, 21], [12, 16, 21], [12, 17, 21], [12, 18, 21], [12, 19, 21], [12, 20, 21], [13, 18, 21], [14, 18, 21], [15, 18, 21], [16, 18, 21], [17, 18, 21]]

Number of Itemsets considered: 418

Number of Frequent Itemsets Generated: 86

------------------------------------------------------------------

----------Itemsets Produced Using k-1 X k-1 Method----------------

1-Itemset: [[0], [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]]

2-Itemset: [[0, 21], [1, 21], [2, 21], [3, 21], [4, 21], [5, 21], [6, 21], [7, 21], [8, 21], [9, 21], [10, 21], [11, 21], [12, 15], [12, 16], [12, 17], [12, 18], [12, 19], [12, 20], [12, 21], [13, 15], [13, 16], [13, 17], [13, 18], [13, 19], [13, 20], [13, 21], [13, 22], [14, 15], [14, 16], [14, 17], [14, 18], [14, 19], [14, 20], [14, 21], [14, 22], [15, 18], [15, 19], [15, 20], [15, 21], [16, 18], [16, 19], [16, 20], [16, 21], [17, 18], [17, 19], [17, 20], [17, 21], [18, 21], [19, 21], [19, 22], [20, 21], [20, 22]]

3-Itemset: [[12, 15, 21], [12, 16, 21], [12, 17, 21], [12, 18, 21], [12, 19, 21], [12, 20, 21], [13, 18, 21], [14, 18, 21], [15, 18, 21], [16, 18, 21], [17, 18, 21]]

Number of Itemsets considered: 375

Number of Frequent Itemsets Generated: 86

------------------------------------------------------------------

Maximal Frequent ItemSet Count: 43

Closed Frequent ItemSet Count: 73

------------------------------------------------------------------

Total number of rules generated by bruteforce method: 170

------------------------------------------------------------------

Enter Confidence Threshold Value (double):

0.5

Please wait.

Total number of rules generated after confidence-based pruning: 39

Top 10 are:

Rule 1:

[2 person car, small luggage boot] -> [class: unacceptable]

Rule 2:

[2 person car, medium luggage boot] -> [class: unacceptable]

Rule 3:

[2 person car, big luggage boot] -> [class: unacceptable]

Rule 4:

[2 person car, low safety] -> [class: unacceptable]

Rule 5:

[2 person car, medium safety] -> [class: unacceptable]

Rule 6:

[2 person car, high safety] -> [class: unacceptable]

Rule 7:

[4 person car, low safety] -> [class: unacceptable]

Rule 8:

[more than 4 person car, low safety] -> [class: unacceptable]

Rule 9:

[small luggage boot, low safety] -> [class: unacceptable]

Rule 10:

[medium luggage boot, low safety] -> [class: unacceptable]

------------------------------------------------------------------

Top 10 rules generated using lift are:

Rule 1:

[2 person car] -> [high safety, class: unacceptable]

Rule 2:

[high safety, class: unacceptable] -> [2 person car]

Rule 3:

[low safety] -> [4 person car, class: unacceptable]

Rule 4:

[4 person car, class: unacceptable] -> [low safety]

Rule 5:

[low safety] -> [more than 4 person car, class: unacceptable]

Rule 6:

[more than 4 person car, class: unacceptable] -> [low safety]

Rule 7:

[2 person car] -> [medium safety, class: unacceptable]

Rule 8:

[medium safety, class: unacceptable] -> [2 person car]

Rule 9:

[high safety] -> [class: acceptable]

Rule 10:

[class: acceptable] -> [high safety]

3a)

As asked both the Fk-1 X F1 and Fk-1 X Fk-1 methods have been implemented. All the resulting itemsets have been displayed. Also the total number of frequent itemsets generated has been displayed.

-------------------------------------------------------------------------------------------------------------------------------

3b)

* Car Evaluation:

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Minimum Support Threshold | Number of itemsets considered | Number of frequent itemsets generated |
|  | 0.1 | 418 | 86 |
| Fk-1 X F1 | 0.05 | 1821 | 349 |
|  | 0.01 | 11829 | 2291 |
|  | 0.1 | 375 | 86 |
| Fk-1 X Fk-1 | 0.05 | 1669 | 349 |
|  | 0.01 | 6981 | 2291 |

* Contraceptive Method Choice:

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Minimum Support Threshold | Number of itemsets considered | Number of frequent itemsets generated |
|  | 0.1 | 2432 | 464 |
| Fk-1 X F1 | 0.05 | 9126 | 1650 |
|  | 0.01 | 75546 | 16852 |
|  | 0.1 | 1237 | 464 |
| Fk-1 X Fk-1 | 0.05 | 4014 | 1650 |
|  | 0.01 | 36035 | 16852 |

* Nursery:

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Minimum Support Threshold | Number of itemsets considered | Number of frequent itemsets generated |
|  | 0.1 | 1283 | 179 |
| Fk-1 X F1 | 0.05 | 4707 | 664 |
|  | 0.01 | 43928 | 8094 |
|  | 0.1 | 969 | 179 |
| Fk-1 X Fk-1 | 0.05 | 4315 | 664 |
|  | 0.01 | 29816 | 8094 |

Observed Saving:

It can be seen from the values above that the Fk-1 X Fk-1 method is more efficient as the number of itemsets that are generated using it are much less than the number of itemsets that are generated using Fk-1 X F1 method.

-------------------------------------------------------------------------------------------------------------------------------

3c)

* Car Evaluation:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Number of frequent closed itemsets | Number of maximal frequent itemsets | Number of frequent itemsets |
| 0.1 | 73 | 43 | 86 |
| 0.05 | 309 | 195 | 349 |
| 0.01 | 1950 | 1000 | 2291 |

* Contraceptive Method Choice:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Number of frequent closed itemsets | Number of maximal frequent itemsets | Number of frequent itemsets |
| 0.1 | 456 | 88 | 1237 |
| 0.05 | 1559 | 280 | 1650 |
| 0.01 | 11689 | 2007 | 16852 |

* Nursery:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Number of frequent closed itemsets | Number of maximal frequent itemsets | Number of frequent itemsets |
| 0.1 | 155 | 123 | 179 |
| 0.05 | 578 | 416 | 664 |
| 0.01 | 6800 | 3781 | 8094 |

Observation:

The number of frequent closed itemsets is always greater than the maximal frequent itemsets. Also not all frequent itemsets are closed; and not all frequent itemsets are maximal frequent itemsets.

-------------------------------------------------------------------------------------------------------------------------------

3d)

* Car Evaluation:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Minimum Confidence Threshold | Number of generated confidence rules (confidence based pruning) | Number of generated confidence rules (brute force) |
| 0.1 | 0.6 | 30 | 170 |
| 0.05 | 0.5 | 140 | 1076 |
| 0.01 | 0.3 | 4557 | 19164 |

* Contraceptive Method Choice:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Minimum Confidence Threshold | Number of generated confidence rules (confidence based pruning) | Number of generated confidence rules (brute force) |
| 0.1 | 0.6 | 1254 | 3800 |
| 0.05 | 0.5 | 5808 | 19864 |
| 0.01 | 0.3 | 141735 | 485808 |

* Nursery:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Minimum Confidence Threshold | Number of generated confidence rules (confidence based pruning) | Number of generated confidence rules (brute force) |
| 0.1 | 0.6 | 26 | 346 |
| 0.05 | 0.5 | 447 | 2400 |
| 0.01 | 0.3 | 23581 | 93334 |

-------------------------------------------------------------------------------------------------------------------------------

Answers 3e and 3f:

For the following answers 3e and 3f the rules have been mapped back from sparse matrix column indexes to real attribute names using the mapping given in README file.

Note: While generating rules I am getting rules for example as follows:  
If A -> B is generated as a rule then B -> A is also generated as a rule:

* always in case of Lift (because lift values will always be the same.)
* and sometimes in case of Confidence (values might turn out to be the same.)

This is clearly possible because that is how Lift and confidence work.

Comment: In general the rules generated using Confidence are not as interesting as the rules generated using Lift. This is because confidence does not capture the correlation between the left hand side and the right hand side of the rule, whereas Lift does in the sense that it tells us whether the left hand side of the rule influences the right hand side of the rule positively or negatively.

3e)

Using Confidence as a measure of interestingness:

As the rules have been sorted based on the decreasing order of confidence values the Rules generated are the same (even though confidence threshold values are changed) after a certain lower bound of confidence threshold.

* Car Evaluation:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Minimum Confidence Threshold | Top 10 Association Rules generated (Sorted on decreasing order of confidence values) | Comments |
| 0.1 | 0.6 | Rule 1:  [2 person car, small luggage boot] -> [class: unacceptable]  Rule 2:  [2 person car, medium luggage boot] -> [class: unacceptable]  Rule 3:  [2 person car, big luggage boot] -> [class: unacceptable]  Rule 4:  [2 person car, low safety] -> [class: unacceptable]  Rule 5:  [2 person car, medium safety] -> [class: unacceptable]  Rule 6:  [2 person car, high safety] -> [class: unacceptable]  Rule 7:  [4 person car, low safety] -> [class: unacceptable]  Rule 8:  [more than 4 person car, low safety] -> [class: unacceptable]  Rule 9:  [small luggage boot, low safety] -> [class: unacceptable]  Rule 10:  [medium luggage boot, low safety] -> [class: unacceptable] | From the rules that were generated I was able to safely infer that if the car is a 2 person car then the class of the car is unacceptable.  So in general people from the region where the data was collected people do not prefer cars which fit only 2 people. |
| 0.5 |
| 0.3 |
| 0.05 | 0.6 | Rule 1:  [v-high buying price, v-high maintenance cost] -> [class: unacceptable]  Rule 2:  [v-high buying price, high maintenance cost] -> [class: unacceptable]  Rule 3:  [v-high buying price, 2 person car] -> [class: unacceptable]  Rule 4:  [v-high buying price, low safety] -> [class: unacceptable]  Rule 5:  [high buying price, v-high maintenance cost] -> [class: unacceptable]  Rule 6:  [high buying price, 2 person car] -> [class: unacceptable]  Rule 7:  [high buying price, low safety] -> [class: unacceptable]  Rule 8:  [medium buying price, 2 person car] -> [class: unacceptable]  Rule 9:  [medium buying price, low safety] -> [class: unacceptable]  Rule 10:  [low buying price, 2 person car] -> [class: unacceptable] | From the rules that were generated it can be inferred that if the buying price of the car or the maintenance price of the car is very high then the car is classified as unacceptable.  So in general people from the region where the data was collected from are thrifty. |
| 0.5 |
| 0.3 |
| 0.15 | 0.6 | Rule 1:  [2 person car] -> [class: unacceptable]  Rule 2:  [low safety] -> [class: unacceptable]  Rule 3:  [v-high buying price] -> [class: unacceptable]  Rule 4:  [v-high maintenance cost] -> [class: unacceptable]  Rule 5:  [small luggage boot] -> [class: unacceptable]  Rule 6:  [2 doors car] -> [class: unacceptable]  Rule 7:  [high buying price] -> [class: unacceptable]  Rule 8:  [high maintenance cost] -> [class: unacceptable]  Rule 9:  [3 doors car] -> [class: unacceptable]  Rule 10:  [medium luggage boot] -> [class: unacceptable] | From the results it can be safely inferred that  the car is unacceptable if:   1. the safety of the car is low 2. the buying price is very high 3. the luggage boot is small or medium in size 4. it’s a 2 or a 3 door car |
| 0.5 |
| 0.3 |

* Contraceptive Method Choice:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Minimum Confidence Threshold | Top 10 Association Rules generated (Sorted on decreasing order of confidence values) | Comments |
| 0.1 | 0.9 | Rule 1:  [Wife's Education: 4=high, Husband's Education: 4=high, Wife Not working now, Husband's occupation: 1, Standard of living index: 4=high] -> [Media Exposure: Good]  Rule 2:  [Wife's Education: 4=high, Wife Not working now, Husband's occupation: 1, Standard of living index: 4=high] -> [Media Exposure: Good]  Rule 3:  [Husband's Education: 4=high, Wife's Religion: Islam, Wife Not working now, Standard of living index: 3] -> [Media Exposure: Good]  Rule 4:  [Wife's Education: 4=high, Husband's Education: 4=high, Contraceptive method used: 2=Long-term] -> [Media Exposure: Good]  Rule 5:  [Wife's Education: 4=high, Wife's Religion: Islam, Contraceptive method used: 2=Long-term] -> [Media Exposure: Good]  Rule 6:  [Wife's Education: 4=high, Standard of living index: 4=high, Contraceptive method used: 2=Long-term] -> [Media Exposure: Good]  Rule 7:  [Husband's Education: 4=high, Wife Not working now, Standard of living index: 3] -> [Media Exposure: Good]  Rule 8:  [Wife's Education: 4=high, Contraceptive method used: 2=Long-term] -> [Media Exposure: Good]  Rule 9:  [Wife's Education: 4=high, Wife Not working now, Standard of living index: 4=high] -> [Media Exposure: Good]  Rule 10:  [Wife's Education: 4=high, Husband's Education: 4=high, Wife Not working now, Standard of living index: 4=high] -> [Media Exposure: Good] | Inferences:  If the wife and the husband both have high education and the contraceptive method used is long term then it implies that the media exposure is good. |
| 0.6 |
| 0.3 |
| 0.2 | 0.5 | Rule 1:  [Wife's Education: 4=high, Standard of living index: 4=high] -> [Media Exposure: Good]  Rule 2:  [Wife's Education: 4=high, Husband's Education: 4=high, Standard of living index: 4=high] -> [Media Exposure: Good]  Rule 3:  [Wife's Education: 4=high, Wife's Religion: Islam] -> [Media Exposure: Good]  Rule 4:  [Wife's Education: 4=high, Wife's Religion: Islam, Wife Not working now] -> [Media Exposure: Good]  Rule 5:  [Wife's Education: 4=high, Husband's occupation: 1] -> [Media Exposure: Good]  Rule 6:  [Wife's Education: 4=high, Wife Not working now] -> [Media Exposure: Good]  Rule 7:  [Wife's Education: 4=high, Husband's Education: 4=high, Husband's occupation: 1] -> [Media Exposure: Good]  Rule 8:  [Wife's Education: 4=high, Husband's Education: 4=high, Wife's Religion: Islam] -> [Media Exposure: Good]  Rule 9:  [Wife's Education: 4=high] -> [Media Exposure: Good]  Rule 10:  [Wife's Education: 4=high, Husband's Education: 4=high, Wife Not working now] -> [Media Exposure: Good] | Inferences:  If the wife’s education is high then it implies that the media exposure is good. |
| 0.4 |
| 0.3 |
| 0.3 | 0.6 | Rule 1:  [Wife's Education: 4=high] -> [Media Exposure: Good]  Rule 2:  [Wife's Education: 4=high, Husband's Education: 4=high] -> [Media Exposure: Good]  Rule 3:  [Husband's Education: 4=high, Standard of living index: 4=high] -> [Media Exposure: Good]  Rule 4:  [Wife Not working now, Standard of living index: 4=high] -> [Media Exposure: Good]  Rule 5:  [Standard of living index: 4=high] -> [Media Exposure: Good]  Rule 6:  [Husband's Education: 4=high, Wife Not working now] -> [Media Exposure: Good]  Rule 7:  [Husband's Education: 4=high] -> [Media Exposure: Good]  Rule 8:  [Husband's Education: 4=high, Wife's Religion: Islam, Wife Not working now] -> [Media Exposure: Good]  Rule 9:  [Husband's Education: 4=high, Wife's Religion: Islam] -> [Media Exposure: Good]  Rule 10:  [Wife's Religion: Islam, Standard of living index: 4=high] -> [Media Exposure: Good] | Some inferences from the generated rules:  If the standard of living is high then the media exposure is good. Also if the husband’s education is high then the media exposure is good. |
| 0.5 |
| 0.4 |

* Nursery:

|  |  |  |  |
| --- | --- | --- | --- |
| Minimum Support Threshold | Minimum Confidence Threshold | Top 10 Association Rules generated (Sorted on decreasing order of confidence values) | Comments |
| 0.1 | 0.6 | Rule 1:  [Parent's Occupation: usual, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 2:  [Parent's Occupation: usual, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 3:  [Parent's Occupation: pretentious, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 4:  [Parent's Occupation: pretentious, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 5:  [Parent's Occupation: great\_pret, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 6:  [Parent's Occupation: great\_pret, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 7:  [Housing conditions: convenient, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 8:  [Housing conditions: convenient, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 9:  [Housing conditions: less\_conv, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 10:  [Housing conditions: less\_conv, Class: not\_recom] -> [Health Conditions of family: not recommended] | From the rules it can be inferred that the children in Ljubljana, Slovenia will not get admission to schools there if the health conditions of their family is not recommended. |
| 0.5 |
| 0.3 |
| 0.02 | 0.6 | Rule 1:  [Parent's Occupation: usual, Child's Nursery: proper, Health Conditions of family: priority] -> [Class: priority]  Rule 2:  [Parent's Occupation: usual, Child's Nursery: proper, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 3:  [Parent's Occupation: usual, Child's Nursery: proper, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 4:  [Parent's Occupation: usual, Child's Nursery: less\_proper, Health Conditions of family: priority] -> [Class: priority]  Rule 5:  [Parent's Occupation: usual, Child's Nursery: less\_proper, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 6:  [Parent's Occupation: usual, Child's Nursery: less\_proper, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 7:  [Parent's Occupation: usual, Child's Nursery: improper, Health Conditions of family: priority] -> [Class: priority]  Rule 8:  [Parent's Occupation: usual, Child's Nursery: improper, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 9:  [Parent's Occupation: usual, Child's Nursery: improper, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 10:  [Parent's Occupation: usual, Child's Nursery: critical, Health Conditions of family: not recommended] -> [Class: not\_recom] | From these rules it can be inferred that if the health conditions of a family is priority, the parent has got a usual occupation and the child’s nursery is proper, improper or less proper then the child will be admitted to the school on priority. |
| 0.3 |
| 0.1 |
| 0.01 | 1 | Rule 1:  [Parent's Occupation: usual, Child's Nursery: proper, Financial Standing of family: convenient, Health Conditions of family: priority] -> [Class: priority]  Rule 2:  [Parent's Occupation: usual, Child's Nursery: proper, Financial Standing of family: convenient, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 3:  [Parent's Occupation: usual, Child's Nursery: proper, Financial Standing of family: convenient, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 4:  [Parent's Occupation: usual, Child's Nursery: proper, Financial Standing of family: inconvenient, Health Conditions of family: priority] -> [Class: priority]  Rule 5:  [Parent's Occupation: usual, Child's Nursery: proper, Financial Standing of family: inconvenient, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 6:  [Parent's Occupation: usual, Child's Nursery: proper, Financial Standing of family: inconvenient, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 7:  [Parent's Occupation: usual, Child's Nursery: less\_proper, Financial Standing of family: convenient, Health Conditions of family: priority] -> [Class: priority]  Rule 8:  [Parent's Occupation: usual, Child's Nursery: less\_proper, Financial Standing of family: convenient, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 9:  [Parent's Occupation: usual, Child's Nursery: less\_proper, Financial Standing of family: convenient, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 10:  [Parent's Occupation: usual, Child's Nursery: less\_proper, Financial Standing of family: inconvenient, Health Conditions of family: priority] -> [Class: priority] | From the rules that have been generated I inferred that if the financial standing of the family is inconvenient and the parent’s occupation is usual and the health of the family is priority then the child would be considered for admission in priority. |
| 0.6 |
| 0.3 |

-------------------------------------------------------------------------------------------------------------------------------

3f) Using Lift as a measure of interestingness:

* Car Evaluation:

|  |  |  |
| --- | --- | --- |
| Minimum Support Threshold | Top 10 Association Rules generated (Sorted on decreasing order of lift values) | Comments |
| 0.1 | Rule 1:  [2 person car] -> [high safety, class: unacceptable]  Rule 2:  [high safety, class: unacceptable] -> [2 person car]  Rule 3:  [low safety] -> [4 person car, class: unacceptable]  Rule 4:  [4 person car, class: unacceptable] -> [low safety]  Rule 5:  [low safety] -> [more than 4 person car, class: unacceptable]  Rule 6:  [more than 4 person car, class: unacceptable] -> [low safety]  Rule 7:  [2 person car] -> [medium safety, class: unacceptable]  Rule 8:  [medium safety, class: unacceptable] -> [2 person car]  Rule 9:  [high safety] -> [class: acceptable]  Rule 10:  [class: acceptable] -> [high safety] | I was able to infer from these rules that if a car is very safe and is not a 2 person car the class of the car is acceptable. |
| 0.05 | Rule 1:  [4 person car, high safety] -> [class: acceptable]  Rule 2:  [class: acceptable] -> [4 person car, high safety]  Rule 3:  [more than 4 person car, high safety] -> [class: acceptable]  Rule 4:  [class: acceptable] -> [more than 4 person car, high safety]  Rule 5:  [4 person car, medium safety] -> [class: acceptable]  Rule 6:  [class: acceptable] -> [4 person car, medium safety]  Rule 7:  [more than 4 person car, medium safety] -> [class: acceptable]  Rule 8:  [class: acceptable] -> [more than 4 person car, medium safety]  Rule 9:  [2 person car] -> [high safety, class: unacceptable]  Rule 10:  [high safety, class: unacceptable] -> [2 person car] | These rules confirm that if the safety is high or if the safety is medium but the car is 4 or more people car then that car is acceptable. |
| 0.005 | Rule 1:  [low buying price, big luggage boot, high safety] -> [class: v-good]  Rule 2:  [class: v-good] -> [low buying price, big luggage boot, high safety]  Rule 3:  [low buying price, more than 4 person car, high safety] -> [class: v-good]  Rule 4:  [class: v-good] -> [low buying price, more than 4 person car, high safety]  Rule 5:  [low buying price, 4 person car, high safety] -> [class: v-good]  Rule 6:  [class: v-good] -> [low buying price, 4 person car, high safety]  Rule 7:  [4 person car, big luggage boot, high safety] -> [class: v-good]  Rule 8:  [class: v-good] -> [4 person car, big luggage boot, high safety]  Rule 9:  [more than 4 person car, big luggage boot, high safety] -> [class: v-good]  Rule 10:  [class: v-good] -> [more than 4 person car, big luggage boot, high safety] | As expected: if all the features of the car are exceptional the class of the car is v-good. That is buying price is low, luggage boot is big and the safety of the car is high then the class is acceptable. |

* Contraceptive Method Choice:

|  |  |  |
| --- | --- | --- |
| Minimum Support Threshold | Top 10 Association Rules generated (Sorted on decreasing order of lift values) | Comments |
| 0.25 | Rule 1:  [Wife's Education: 4=high] -> [Husband's Education: 4=high, Standard of living index: 4=high, Media Exposure: Good]  Rule 2:  [Husband's Education: 4=high, Standard of living index: 4=high, Media Exposure: Good] -> [Wife's Education: 4=high]  Rule 3:  [Husband's Education: 4=high, Standard of living index: 4=high] -> [Wife's Education: 4=high, Media Exposure: Good]  Rule 4:  [Wife's Education: 4=high, Media Exposure: Good] -> [Husband's Education: 4=high, Standard of living index: 4=high]  Rule 5:  [Wife's Education: 4=high] -> [Husband's Education: 4=high, Standard of living index: 4=high]  Rule 6:  [Husband's Education: 4=high, Standard of living index: 4=high] -> [Wife's Education: 4=high]  Rule 7:  [Wife's Education: 4=high, Standard of living index: 4=high] -> [Husband's Education: 4=high, Media Exposure: Good]  Rule 8:  [Husband's Education: 4=high, Media Exposure: Good] -> [Wife's Education: 4=high, Standard of living index: 4=high]  Rule 9:  [Husband's Education: 4=high] -> [Wife's Education: 4=high, Standard of living index: 4=high]  Rule 10:  [Wife's Education: 4=high, Standard of living index: 4=high] -> [Husband's Education: 4=high] | Inferences form the rules:  If the wife’s education is high then it implies that the husband’s education is also high and their standard of living is high with good media exposure. |
| 0.20 | Rule 1:  [Wife's Education: 4=high] -> [Husband's Education: 4=high, Husband's occupation: 1, Media Exposure: Good]  Rule 2:  [Husband's Education: 4=high, Husband's occupation: 1, Media Exposure: Good] -> [Wife's Education: 4=high]  Rule 3:  [Husband's Education: 4=high, Husband's occupation: 1] -> [Wife's Education: 4=high, Media Exposure: Good]  Rule 4:  [Wife's Education: 4=high, Media Exposure: Good] -> [Husband's Education: 4=high, Husband's occupation: 1]  Rule 5:  [Wife's Education: 4=high] -> [Husband's Education: 4=high, Husband's occupation: 1]  Rule 6:  [Husband's Education: 4=high, Husband's occupation: 1] -> [Wife's Education: 4=high]  Rule 7:  [Wife's Education: 4=high, Husband's Education: 4=high] -> [Husband's occupation: 1, Media Exposure: Good]  Rule 8:  [Husband's occupation: 1, Media Exposure: Good] -> [Wife's Education: 4=high, Husband's Education: 4=high]  Rule 9:  [Husband's occupation: 1] -> [Wife's Education: 4=high, Husband's Education: 4=high, Media Exposure: Good]  Rule 10:  [Wife's Education: 4=high, Husband's Education: 4=high, Media Exposure: Good] -> [Husband's occupation: 1] | Inferences from the rules:  If the husband has occupation = 1 then it implies that the wife’s education is high the husband’s education is high and they have good media exposure. |
| 0.01 | Rule 1:  [Husband's Education: 4=high, Number of children ever born: 0, Standard of living index: 4=high] -> [Wife's Education: 4=high, Wife working now, Media Exposure: Good, Contraceptive method used: 1=No-use]  Rule 2:  [Wife's Education: 4=high, Wife working now, Media Exposure: Good, Contraceptive method used: 1=No-use] -> [Husband's Education: 4=high, Number of children ever born: 0, Standard of living index: 4=high]  Rule 3:  [Wife's Education: 4=high, Number of children ever born: 0] -> [Husband's Education: 4=high, Wife working now, Standard of living index: 4=high, Media Exposure: Good, Contraceptive method used: 1=No-use]  Rule 4:  [Husband's Education: 4=high, Wife working now, Standard of living index: 4=high, Media Exposure: Good, Contraceptive method used: 1=No-use] -> [Wife's Education: 4=high, Number of children ever born: 0]  Rule 5:  [Husband's Education: 4=high, Number of children ever born: 0, Standard of living index: 4=high, Media Exposure: Good] -> [Wife's Education: 4=high, Wife working now, Contraceptive method used: 1=No-use]  Rule 6:  [Wife's Education: 4=high, Wife working now, Contraceptive method used: 1=No-use] -> [Husband's Education: 4=high, Number of children ever born: 0, Standard of living index: 4=high, Media Exposure: Good]  Rule 7:  [Husband's Education: 4=high, Number of children ever born: 0, Standard of living index: 4=high] -> [Wife's Education: 4=high, Wife working now, Contraceptive method used: 1=No-use]  Rule 8:  [Wife's Education: 4=high, Wife working now, Contraceptive method used: 1=No-use] -> [Husband's Education: 4=high, Number of children ever born: 0, Standard of living index: 4=high]  Rule 9:  [Wife's Education: 4=high, Number of children ever born: 0, Media Exposure: Good] -> [Husband's Education: 4=high, Wife working now, Standard of living index: 4=high, Contraceptive method used: 1=No-use]  Rule 10:  [Husband's Education: 4=high, Wife working now, Standard of living index: 4=high, Contraceptive method used: 1=No-use] -> [Wife's Education: 4=high, Number of children ever born: 0, Media Exposure: Good] | Results found via these rules include: |

* Nursery:

|  |  |  |
| --- | --- | --- |
| Minimum Support Threshold | Top 10 Association Rules generated (Sorted on decreasing order of lift values) | Comments |
| 0.1 | Rule 1:  [Health Conditions of family: not recommended] -> [Parent's Occupation: usual, Class: not\_recom]  Rule 2:  [Parent's Occupation: usual, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 3:  [Class: not\_recom] -> [Parent's Occupation: usual, Health Conditions of family: not recommended]  Rule 4:  [Parent's Occupation: usual, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 5:  [Health Conditions of family: not recommended] -> [Parent's Occupation: pretentious, Class: not\_recom]  Rule 6:  [Parent's Occupation: pretentious, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 7:  [Class: not\_recom] -> [Parent's Occupation: pretentious, Health Conditions of family: not recommended]  Rule 8:  [Parent's Occupation: pretentious, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 9:  [Health Conditions of family: not recommended] -> [Parent's Occupation: great\_pret, Class: not\_recom]  Rule 10:  [Parent's Occupation: great\_pret, Health Conditions of family: not recommended] -> [Class: not\_recom] | From the rules the following can be inferred:  Even if the parent’s occupation is pretentious is the family health is not recommended then the child will not be recommended for admission. |
| 0.04 | Rule 1:  [Child's Nursery: very\_crit, Health Conditions of family: priority] -> [Class: spec\_prior]  Rule 2:  [Class: spec\_prior] -> [Child's Nursery: very\_crit, Health Conditions of family: priority]  Rule 3:  [Health Conditions of family: not recommended] -> [Child's Nursery: proper, Class: not\_recom]  Rule 4:  [Child's Nursery: proper, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 5:  [Class: not\_recom] -> [Child's Nursery: proper, Health Conditions of family: not recommended]  Rule 6:  [Child's Nursery: proper, Class: not\_recom] -> [Health Conditions of family: not recommended]  Rule 7:  [Health Conditions of family: not recommended] -> [Child's Nursery: less\_proper, Class: not\_recom]  Rule 8:  [Child's Nursery: less\_proper, Health Conditions of family: not recommended] -> [Class: not\_recom]  Rule 9:  [Class: not\_recom] -> [Child's Nursery: less\_proper, Health Conditions of family: not recommended]  Rule 10:  [Child's Nursery: less\_proper, Class: not\_recom] -> [Health Conditions of family: not recommended] | From the rules the following inferences can be made:  If a child gets special priority for admission then it means that the child’s nursery is very critical and the health conditions of the family is a priority. |
| 0.01 | Rule 1:  [Housing conditions: convenient, Financial Standing of family: convenient, Health Conditions of family: recommended] -> [Class: very\_recom]  Rule 2:  [Class: very\_recom] -> [Housing conditions: convenient, Financial Standing of family: convenient, Health Conditions of family: recommended]  Rule 3:  [Parent's Occupation: usual, Financial Standing of family: convenient, Health Conditions of family: recommended] -> [Class: very\_recom]  Rule 4:  [Class: very\_recom] -> [Parent's Occupation: usual, Financial Standing of family: convenient, Health Conditions of family: recommended]  Rule 5:  [Parent's Occupation: great\_pret, Health Conditions of family: priority] -> [Child's Nursery: proper, Class: spec\_prior]  Rule 6:  [Child's Nursery: proper, Class: spec\_prior] -> [Parent's Occupation: great\_pret, Health Conditions of family: priority]  Rule 7:  [Parent's Occupation: great\_pret, Health Conditions of family: priority] -> [Child's Nursery: less\_proper, Class: spec\_prior]  Rule 8:  [Child's Nursery: less\_proper, Class: spec\_prior] -> [Parent's Occupation: great\_pret, Health Conditions of family: priority]  Rule 9:  [Housing conditions: convenient, Health Conditions of family: recommended] -> [Financial Standing of family: convenient, Class: very\_recom]  Rule 10:  [Financial Standing of family: convenient, Class: very\_recom] -> [Housing conditions: convenient, Health Conditions of family: recommended] | From the rules generated the following can be inferred:  A child will be highly recommended for admission if the financial standing of the family is convenient and the health conditions of the family is recommended.  Also, if the Parent’s occupation is great pretentious and the health of the family is a priority then the child will get special priority for admission. |

-------------------------------------------------------------------------------------------------------------------------------

Answer 4:

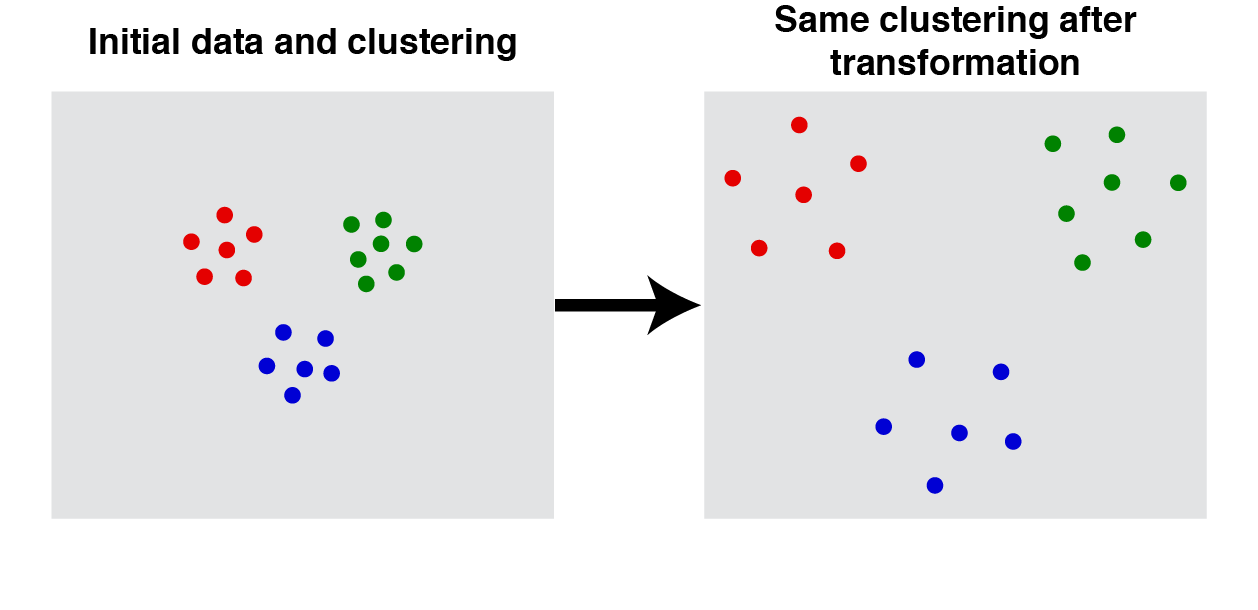
4a) An Impossibility Theorem for Clustering – Jon Kleinberg

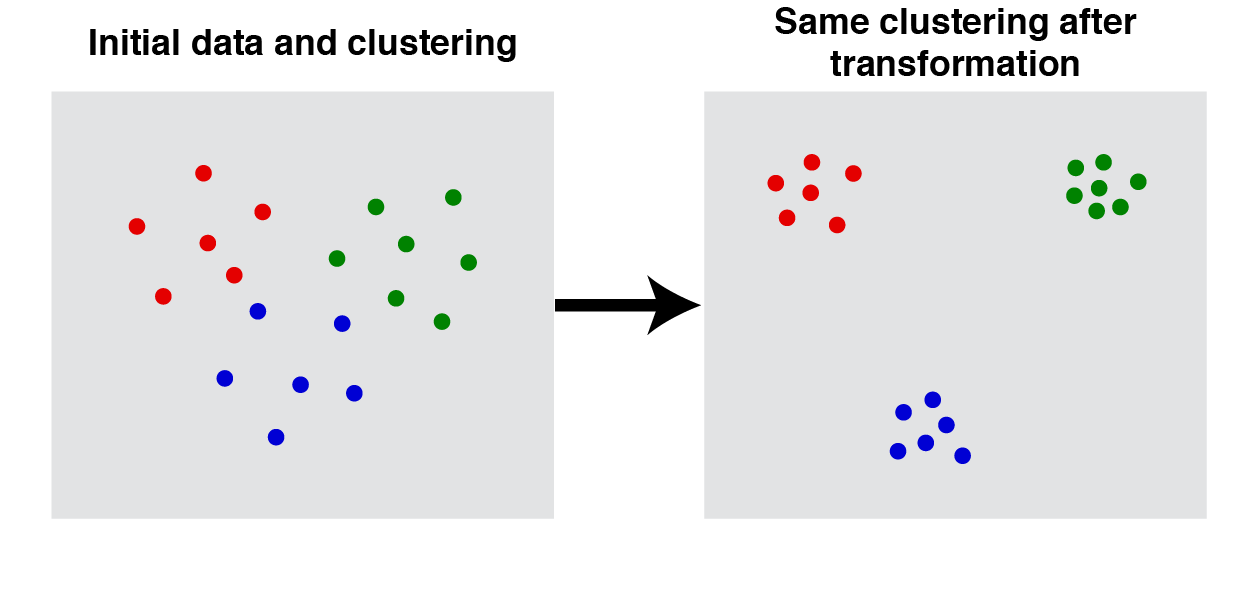
In the paper the author has talked about difficulties pertaining to clustering.

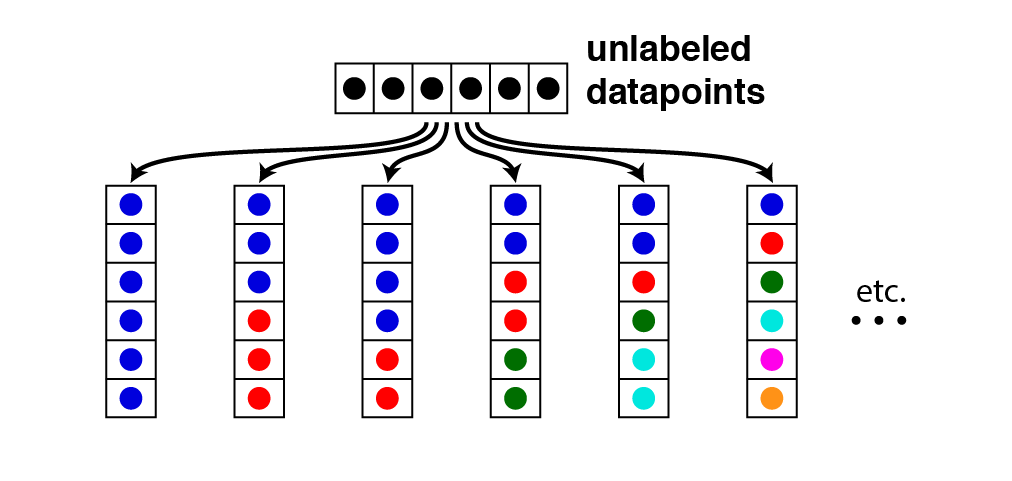
Clustering is grouping of points (real world objects’ elements) to form a meaningful set called a partition. The author says that: Clustering of data points can be defined formally through a function which satisfies 3 properties – scale-invariance, richness and consistency.

*Scale-invariance* enforces a strict rule that the elements in the clusters must remain intact within the cluster even after multiplying the distances between them with a constant factor say alpha.

f(d)=f(α⋅d)



*Consistency* enforces this intactness even after increasing the inter-cluster distances and/or decreasing the intra-cluster distances. *Richness* expects the function to determine the number of clusters in the dataset by itself without having a knowledge of distances between any pair of points.



The author proves through his paper [Theorem 2.1] that a function that can satisfy all the three above mentioned properties does not exist.

Hierarchical clustering is a method of generating clusters in either top-down (Divisive) or bottom-up (Agglomerative) fashion. Complexity-wise, agglomerative is cubic (can be improved and optimized into quadratic) while divisive is exponential, being exhaustive in nature. Single-Linkage clustering is based on agglomerative clustering. It begins from considering a point to be a cluster and then merging points into clusters repeatedly, thereby employing a bottom-up style of cluster generation.

It creates a Graph of nodes where nodes are the data points and edges represent distances. Merging of nodes is made by taking edges in a non-increasing order. The merge process has to have a condition which when reached stops further merging of data points. This condition is called stopping-condition and once reached, we can have a function f, called clustering function which maps input data to clustered output.

Following three stopping-conditions are discussed in the paper-

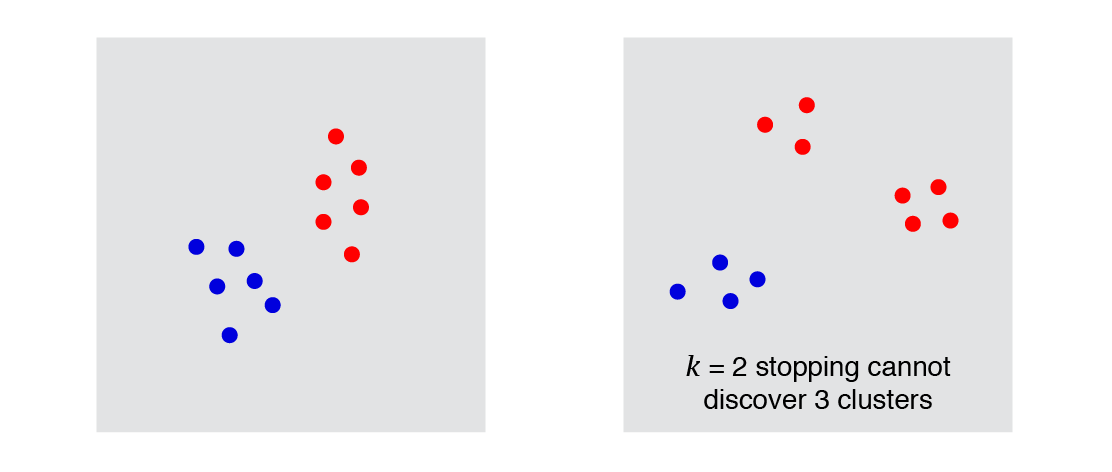
1. K-Cluster stopping condition: Merge till we get ‘k’ number of clusters.

(Min. number points = k)

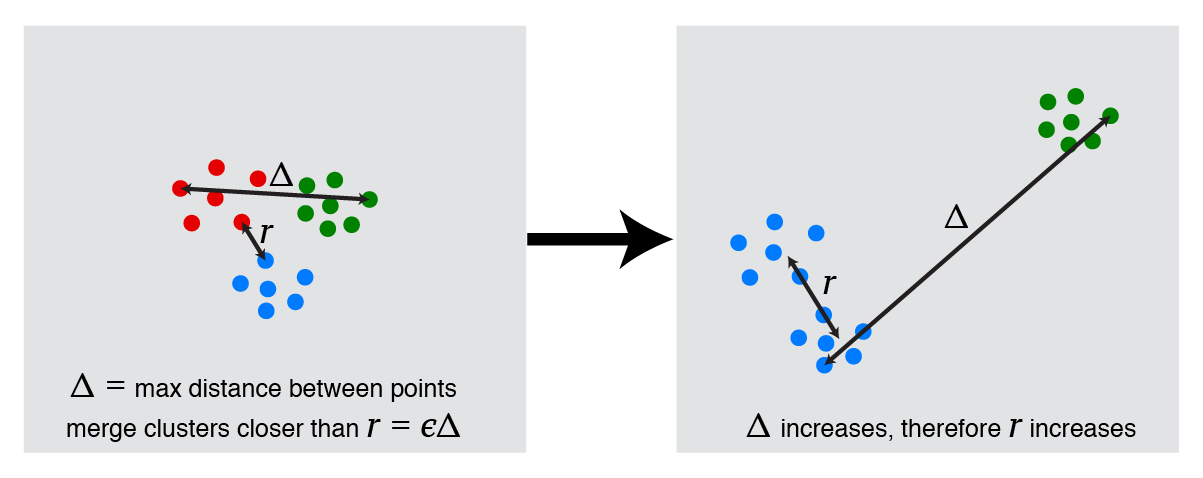
1. Distance-r stopping condition: Merge two nodes only if the edge’s weight is less than or equal to ‘r’.
2. Scale-α stopping condition: Merge two nodes only if their weight is less than or equal to αρ∗ where α is a scale-factor and ρ∗ is maximum distance between those nodes.

The paper further illustrates how each of these three stopping conditions violate one of the three essential properties required for function [1]

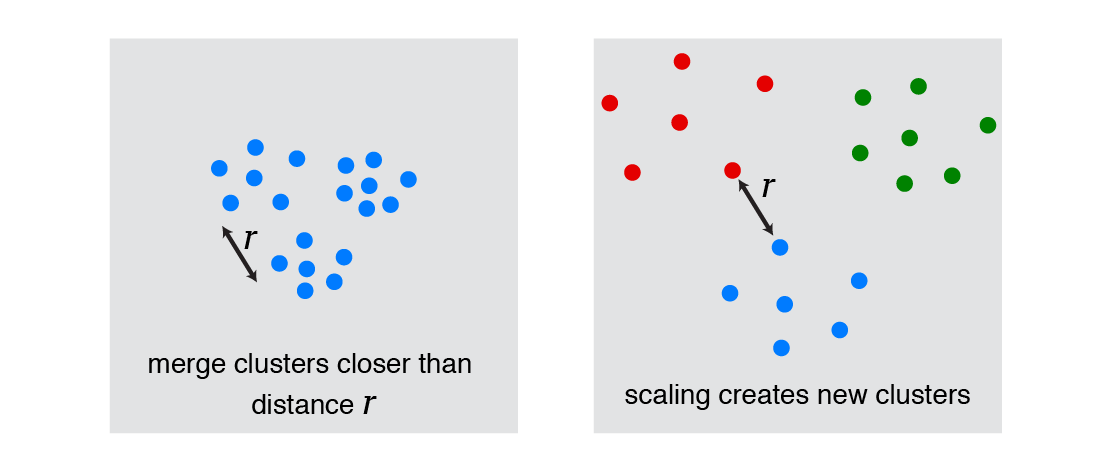
1. Richness is not satisfied by k-cluster stopping condition: As the stopping condition itself pre-defines numerical constraint on clusters, richness does not gets satisfied.



1. Scale- ϵ doesn’t satisfy Consistency: As we are restricting a merge up to ϵ⋅ρ∗ weight, consistency fails because we aren’t allowing intra-distance compressions.



1. Distance-r fails for Scale-invariance because we are not allowed data-point stretches.



The impossibility- No clustering function practically exists that can satisfy all the three conditions. However, many functions are practically used which work for 2 of the 3 conditions. Therefore, a relaxation in non-satisfying condition comes into picture in which a trade-off between properties or assumptions is required so as to formulate a unified framework satisfying all the three properties (the property which was previously not getting satisfied is now getting weakly-satisfied)

Arrow’s impossibility theorem [5] - The theorem states that no rank-order voting system can be designed that always satisfies these three "fairness" criteria:

1. If every voter prefers alternative X over alternative Y, then the group prefers X over Y.
2. If every voter's preference between X and Y remains unchanged, then the group's preference between X and Y will also remain unchanged (even if voters' preferences between other pairs like X and Z, Y and Z, or Z and W change).
3. There is no "dictator": no single voter possesses the power to always determine the group's preference.

Kleinberg’s impossibility theorem is similar to Arrow’s impossibility wherein the fairness is defined by the 3 conditions which cannot be satisfied together at the same time with a unified framework/clustering function.

The formal definitions on conditions imposed is the major reason for impossibility of a unified framework. This analysis is useful but only a good starting point for characterization.

Margareta Ackerman and Shai Ben-David’s “Measures of Clustering Quality: A Working Set of Axioms for Clustering” clearly argues that impossibility is a result of specific formalism for example consider ‘consistency’ property. If we consider 3 clusters such that 2 of those are closer while third is not to both. Consistency if applied in this example will move the clusters in such a way that third cluster goes so farther that we intuitively feel the need to merge other 2 clusters. This puts a genuine question over consistency as formally specified in the paper.

-------------------------------------------------------------------------------------------------------------------------------

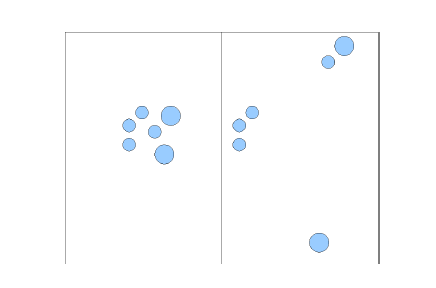
4b) Measures of Clustering Quality: A Working Set of Axioms for Clustering

Via this paper the authors have tried to oppose what Jon Klienberg’s paper proposes. They argue that the impossibility theorem can only be valid because of the way in which it is formalized in Klienberg’s paper. The authors via the paper have tried not to define what the clustering function should obey. They have tried to measure the quality of a given clustering function. Also they point out that this method includes all the axioms expressed by the previous paper.

Clustering Quality Measure is a formulation which given an input of data points and the output of the clustering function outputs an objective score which says how good the quality of the cluster is. The final clustering quality measure is also affected by the total number of clusters. Basically, the authors say that the Clustering quality measure measures if the clusters that have been formed are worth pursuing for further data mining.

This quality measure is founded on the center based and linkage based clustering. The authors have pointed out in the paper that the Sum of Squared Error loss function is not scale invariant hence cannot be used to measure the cluster quality. From the paper I understand that the authors agree with many points that Klienberg had proposed but do not agree with his formalization.

The following diagram from the paper clearly shows that the consistency axiom proposed by the Klienberg paper is not accurate fully:



The clusters do not have to remain the same as said by Klienberg because there may be new groupings which might even be better.

The authors in their paper have defined the following CQM axioms analogous to Klienberg’s axioms and they are applied to quality functions instead of to the clustering functions:

*Scale Invariance*: It says that even if there is a change in the distance function, the quality of existing groupings is not troubled even though after that change there might be better candidates.

*Consistency*: If we make a consistent change in the distance function it would still yield the same clustering.

*Richness*: This would ensure that the quality measure will take the best clustering into account via the CQM function.

After this the authors have covered several new measures for cluster quality. They are:

*Relative Margin*: What I understand via the paper is that smaller this value the better is the clustering quality. This is used for centroid based clustering.

*Weakest Link*: The quantification that this measure does is done by measuring the longest chain distance amongst points in a cluster divided by the smallest distance between two clusters. This is used for linkage based clustering.

*Additive Margin*: The average additive margin of a clustering is the average additive point margin divided by the average of the intra cluster distance. This normalization as pointed out in the paper is necessary for scale invariance. This is also like relative margin applicable to centroid based clustering.

Further in the paper the authors have introduced the axiom of isomorphic invariance via which the CQM measures are made independent of clustered element’s domain identity. Finally the authors have concluded their paper by discussing about clustering quality evaluation being dependent on the number of clusters. They introduce the concept of L-normalization which they define as a loss function measure which unlike SSE is scale invariant. L-normalization is introduced as a loss function measure which is scale invariant. They also talk about intuitive notions of refinement and coarsening of clusters.

Via this paper the authors are hopeful that by discussing about the sample measures more people would research about these measures which should be incorporated in the CQM framework. They have laid down a good starting point having defined the measures and also proposed rules to be followed by new measures in CQM. I think if these measures are put to further test on real data to check we would be able to find if the clustering quality gain is actually being supported by CQM. In conclusion this paper convinces me that the impossibility theorem is not very accurate.

-------------------------------------------------------------------------------------------------------------------------------

References:

[1] <http://alexhwilliams.info/itsneuronalblog/2015/10/01/clustering2/>

[2] <http://stats.stackexchange.com/questions/173313/clustering-intuition-behind-kleinbergs-impossibility-theorem>

[3] <https://en.wikipedia.org/wiki/Single-linkage_clustering>

[4] <https://en.wikipedia.org/wiki/Hierarchical_clustering>

[5] <https://en.wikipedia.org/wiki/Arrow%27s_impossibility_theorem>

[6] <https://ccrma.stanford.edu/workshops/mir2009/references/ROCintro.pdf>