04250 - Introduction to Machine Learning and Data Mining

Project 3 - Supervised Learning, E17

April 30, 2017

This report is written in collaboration by the following students:

Christian Kirstein Thygesen, s123871

Henriette Steenhoff, s134869

DTU Compute
Institut for Matematik og Computer Science

Contents

1 Clustering

1.1 Gaussian Mixture Model

The Gaussian Mixture Model was used to cluster our data, and along with cross-validation the optimal number of clusters was estimated.

A way of evaluating what the optimal number og clusters is, clustering was performed for each K in the range 1 to 20. For each K 10-fold cross validation was performed to estimate the best possible clustering. In each iteration three parameters were calculated; Akaikes Information Criteria (AIC), Bayesian Information Criteria (BIC) and the log-likelihood of the clustering performed in the cross validation.

BIC and AIC are defined by

$$BIC = -2 * logL + p * log(N) \tag{1}$$

and

$$AIC = -2 * logL + 2p \tag{2}$$

where

$$logL = \sum_{i=1}^{N} log[p(x_i|w, \{\mu_{(1)}, \Sigma_{(1)}, ..., \mu_{(K)}, \Sigma_{(K)}\})] \tag{3}$$

Thus BIC and AIC are information criteria that define a trade-off between modeling the data well and penalizing complexity of the model. The best model, will therefore have the smallest BIC and AIC, as well as the log-likelihood computed in the cross-validation folds. By computing these for each K, we find that the optimal number of clusters is 5 (Figure 1), as this is where all three parameters have leveled out. The optimal number of clusters is thus not equal to the number of class labels (2).

This is not ideal, as what we are trying to predict from the data set is whether a person suffers from coronary heart disease or not, which means that in order for the clustering to work perfectly in this case, the ideal number (in a perfect world) would be two. However, it could be that several clusters predict the same class, so this is not necessarily dooming to this approach.

From each of these 5 clusters, the centroids have been identified, and the coordinates of these are illustrated in figure 2.

It is difficult to interpret the exact qualitative meaning of the coordinates of the centroids, since we are operating in a 9-dimensional space. However, it can be seen that each centroid seems to put different emphasis on each of the 9 attributes. Some of the centroids looks like they are very located very close to each other, as some of the cordinates are more or less equal. This does not necessarily mean that they are overlapping, but simply that they are placed on the same position along one of the planes. This is emphasized by the fact, that even the clusters that seem to agree on certain coordinates differ markedly on others.

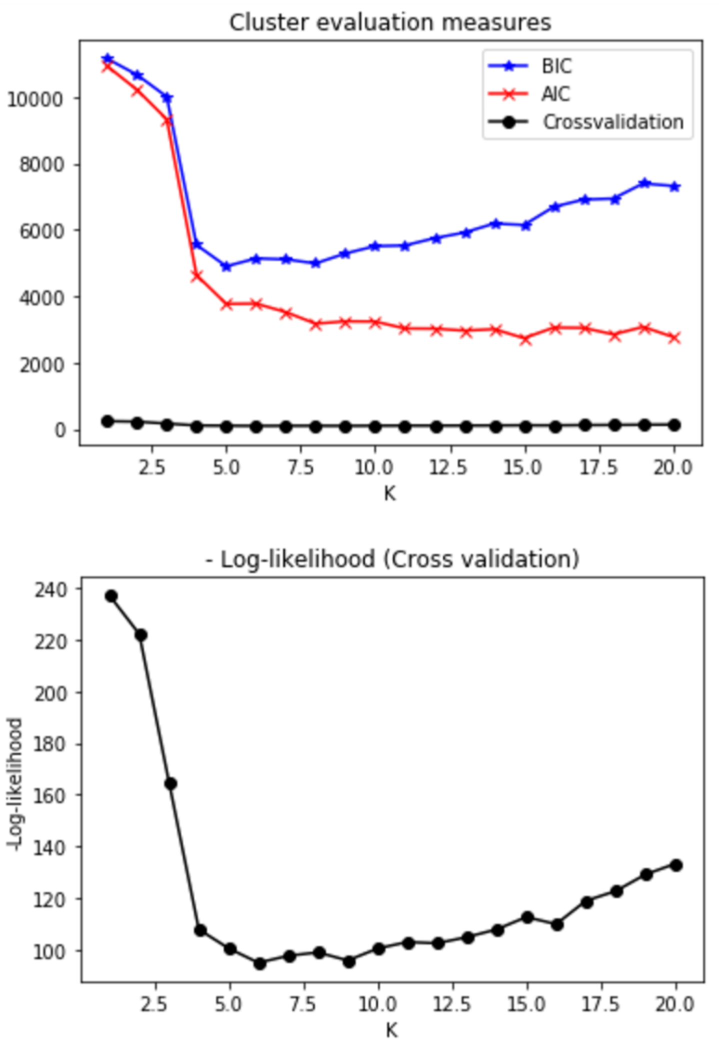
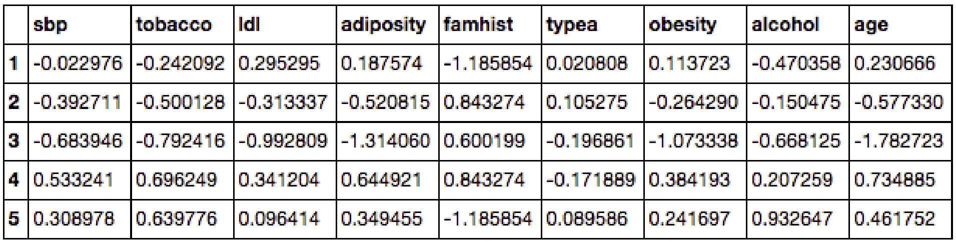


Figure 1: Optimal number of clusters visualized by the three parameters AIC, BIC, and cross validation log-likelihood. Given the scale differences, the log-likelihood has been visualized by itself.



	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age
1	-0.022976	-0.242092	0.295295	0.187574	-1.185854	0.020808	0.113723	-0.470358	0.230666
2	-0.392711	-0.500128	-0.313337	-0.520815	0.843274	0.105275	-0.264290	-0.150475	-0.577330
3	-0.683946	-0.792416	-0.992809	-1.314060	0.600199	-0.196861	-1.073338	-0.668125	-1.782723
4	0.533241	0.696249	0.341204	0.644921	0.843274	-0.171889	0.384193	0.207259	0.734885
5	0.308978	0.639776	0.096414	0.349455	-1.185854	0.089586	0.241697	0.932647	0.461752

Figure 2: Coordinates of the centroids of the 5 clusters previously identified.

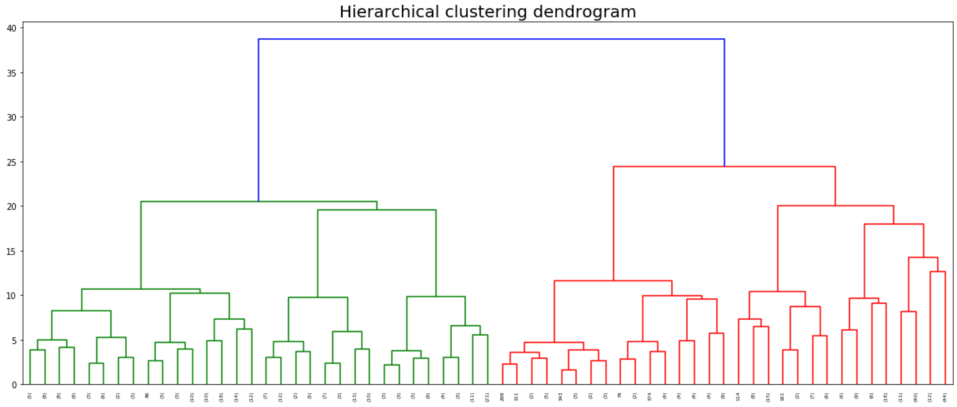


Figure 3: Hierarchical clustering using Ward's method. Clear separation between two clusters were achieved.

1.2 Hierarchical clustering

Hierarchical clustering was performed using the ward's method. This method was chosen, since our clusters are tightly connected, and distinguishing between simply using K-means have proven unfruitful. However, by using Wards method two clear clusters emerged when looking at the dendrogam for the hierarchical clustering approach (figure 3). This seems promising, as a separation into two clusters could potentially mean, that we would be able to distinguish between people based on whether they suffer from coronary heart disease or not.

1.3 Quality evaluation

The performance of the clustering techniques performed previously are evaluated based on their labelling information. Both in GMM and in hierarchical clustering, the 5 clsuters were obtained for comparability of the two methods. Each cluster will be classified based on the majority of the class labels in the cluster, and the frequency of the "wrong" class label in the cluster will be used as a misclassification rate. It should be noted, that even though we have more clusters (5) than classes (2), this does not mean, that the methods did not perform well.

In table 1, the clusters obtained using Gaussian Mixture Model clustering have been labelled given most frequent observation class in the cluster. Cluster 2 does not carry a lot of information, as only 1 observation has been placed in this cluster, which could indicate an outlier/anomaly. Clusters 1, 3, and 4 carries more information, as relatively many observation are placed into these, and with a large majority in each exhibiting the same class. Cluster 5 is inconclusive as equally many observations from each class has been placed here. Notably, this model is not suitable for our aim, as none of the clusters exhibit a clear indication of CHD classification.

In table 2 the class distribution for the hierarchical clusters has been elucidated. It seems, that this hierarchical approach may be inadequate as well for our particular need. This is exemplified by the

Gaussian Mixed Model Clusters			
Cluster	CHD	No CHD	Cluster class
1	34	53	No CHD
2	1	0	CHD
3	28	112	No CHD
4	1	41	No CHD
5	96	96	N/A

Table 1: GMM clusters investigated and classified using the number of observations in each cluster.

fact, that only one relatively small cluster (4) is classified as a CHD cluster with a high misclassification rate.

Hierarchical Clusters			
Cluster	CHD	No CHD	Cluster class
1	26	89	No CHD
2	31	95	No CHD
3	18	129	No CHD
4	25	15	CHD
5	70	74	No CHD

Table 2: Hierarchical clusters investigated and classified using the number of observations in each cluster.

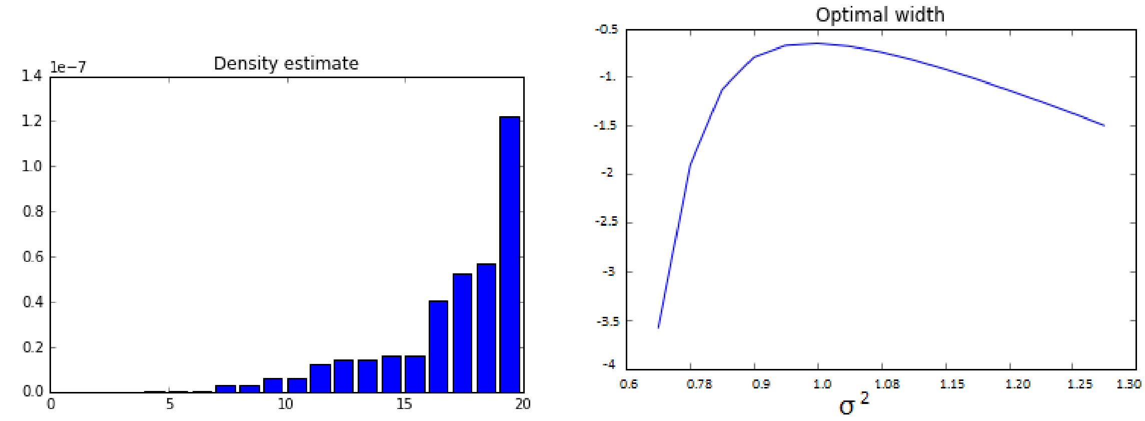
2 Outlier detection

2.1 Observation ranking

Kernel density estimation is a non-parametric method of estimating the probability density function of a random variable. The estimated kernel density for a variable x is given by:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i),$$

where K is the kernel function, N is the sample size, $f(x)$ is the density function. For this assignment, the density will be estimated with leave-one-out cross-validation, such that each observation is estimated from all other observations, not including itself in the estimate. The optimal kernel width is the width with the highest logP value. Prior to performing these analyses, the data was normalized as described in report 2.



(a) Gaussian Kernel Density with use of leave-one-out cross-validation.

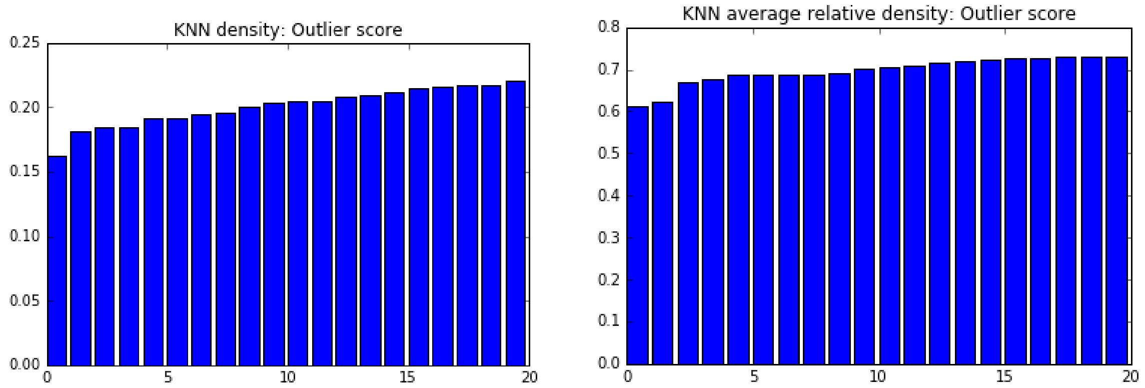
(b) Optimal width, $\sigma^2 = 1.0$ for the Gaussian Kernel Density.

Figure 4: Density information for Gaussian Kernel Density.

From figure 4a we see the 20 observations with the lowest density. The density was evaluated at each point, but only plotted for some, as there are more observations that what would be approporiate to plot. The 30 observations with the lowest densities (by observation number) were saved to match against the outliers found in KNN and ARD below. The density, $\sigma^2 = 1.0$, based on leave-one-out cross-validation.

Working with KNN, the number of neighbours was set to 30, as in report 2 for consistency. When this was done, the KNN density and the KNN average relative density was calculated for all observations and plotted for the 20 observations with the lowest density value.

In the figure below is shown the result from the K Nearest Neighbour Average Relative Density (ARD). Outliers are data objects in low density areas. ARD tries to find anomalies/outliers in low density regions like GMM and KDE by combining the principles from K nearest Neighbours while looking for points where the density is lower than what it typically is for the surrounding points. This is done by considering the density of a given point relative to the average density of the K nearest neighbours.



(a) KNN Density

(b) KNN Average Relative Density

Figure 5: Density information using KNN Density 5a) and ARD 5b).

As for Gaussian Kernel Density the KNN and ARD density for the 30 lowest density observations were saved. Looking at the three sets generated by the different methods all three sets have no outliers in common. Testing the different sets against each other two and two however gives a different result. For the Gaussian Kernel Density and ARD, there are 35 outliers in common, observations with unmbers: $334, 407, 442, 452, 457$. These however, are the only outliers in common using the different methods. This might be because more outliers need to be included in order to find outliers in common or because the number of neighbors should have been chosen differently. Based on these results we can therefore not make any conclusions on outliers that are in common for all methods, however, there is no doubt in the fact that there are outliers.

3 Association mining

3.1 Apriori

To apply the Apriori algorithm on the data, we need to binarize it. After normalization, the data was binarized using `sklearn's preprocessing.binarize`. As association mining aims at finding associations between the different features of the data, the matrix X now consists of the feature `chd`, which through the other reports have been separated as the vector y, the predictors. This way we will try to find associations between the features that males with coronary heaarth disease have in common.

After formatting the binarized data as a string on the proper form (as can be seen in `writeaprioritest` example), the `WriteAprioriToFile()` function from the toolbox was applied to translate the data into correct format for the executable. An excerpt of the result of running the `WriteAprioriToFile()` can be seen below:

```
sbp,tobacco,ldl,alcohol,age
sbp,adiposity,famhist,typea,age
adiposity,age
sbp,tobacco,ldl,adiposity,alcohol,age
tobacco,adiposity,typea,alcohol,age
tobacco,ldl,adiposity,typea,age
sbp,tobacco,famhist,typea
tobacco,typea,age
ldl,adiposity,typea,age
...
```

Here, each line correspond to an observation for a male in the original dataset. The features (separated by commas) on each line, indicates that the feature has been binarized to a one, meaning that it is present for the current male.

The set of all items I is the set of all features from the dataset, therefore we can write I as:

$$I = \{spb, tobacco, ldl, adiposity, famhist, typea, alcohol, age\}$$

The observations for all males is denoted by $M = \{m_1, m_2, ..., m_N\}$, where N is the number of observations. Each male is then an itemset, which is a subset of I, such that $m_i \subseteq I$. Using this notation and the results from the binarized data, the itemset for the first observation can be written as $m_1 = \{spb, tobacco, ldl, alcohol, age\}$. Which means that the levels of the different features of m_1 is significant enough to define that male as having a high systolic blood pressure, tobacco intake etc..

Using the Apriori algorithm it is possible to look into whether there can be found any associations between the different features of each observations by discovering itemsets with high support.

Here, one tries to find sets of features that are frequent. This is done by finding the items from each itemset that are more frequent than others. Using the market basket example from the course book[1], one tries to discover patterns, i.e. ”poeple that buy milk also tend to buy bread”. This can be formalized as

$$X \to Y,$$

where $X, Y \subseteq I$. Using the South African Heart disease data, one could for example have $X = \{bloodpressure, alcohol\}$ and $Y = \{adiposity\}$, which could be interpreted as: males with high blood pressure and alcohol intake tend to have adiposity etc. This will only be the case if many of the observations show the same tendency, i.e. the itemset should have a *high support*, ϵ. If this rule is triggered, if a person does have high blood pressure and a high alcohol intake, it should be very likely that they also have adiposity which means that there should be a *high confidence*, ρ, of the probability of the itemset.

When running the Apriori algorithm, one need to set the support and confidence levels of the rules, but what will be good values for the two? Since we are trying to find rules for heart disease cases, we want to make as correct associations as possible, but we would also like to find the (possibly) less frequent itemsets that might be missed if the support level is set too high, which is one of the limitations of the Apriori algorithm. Because what if there is a high confidence that high blood pressure and alcohol also tend to mean adipostiy, but this itemset is left out because the support of the itemset lies under the threshold of the support. Because of this, it was chosen to use a lower value for the support and a higher value for the confidence to include more of the rare itemsets. One should however, be careful with concluding something based on itemsets with low support as this might not be representative.

There is however also other data problematiques. For the features alcohol and tobacco, the value is the sum of your consumption historically, which means that the older you are, the higher the consumption (assuming that everyone drinks/smokes). As these are binarized, you can either be in group 1 or group 2, for age this means either you are ”younger” or you are ”older”. There are no information about what separates these two. The same with tobacco, where you are either smoking or not, this should somehow be more correct as there will be people that do not smoke, which makes the range for tobacco larger than the range for age, as the age range is smaller, from $15 - 64$. Also, making sure that adiposity is grouped correctly such that the healthy/overweight threshold matches a certain index would have been preferable. Here one could have chosen to create different groups based on an age-range and within this age group define the threshold for healthy/unhealthy to observe whether the itemsets would change for different age-groups. This has not been done but could help in improving the analysis.

After trying different values for support and confidence, support was set to 25% and confidence to 70% meaning that the itemset should be discovered in at least 25% of the observations, roughly 53 of the males. The association rule should have a confidence of 70%, which means that among the 53 males, the association rule should be found within at least 40 of the males. By using these values there will be many different itemsets and association rules which gives many different possibilities to interpret the rules.

Frequent itemsets

Below is listed all of the itemsets with a support above 25%. Among the itemsets with the highest support we find: $\{adiposity\}$, $\{age\}$, $\{typea\}$, $\{famhist\}$ with a support level of 55%, 55%, 51% and 60%, respectively.

```
famhist           [Sup. 60]
age               [Sup. 55]
adiposity         [Sup. 55]
typea             [Sup. 51]
ldl               [Sup. 44]
age adiposity     [Sup. 41]
tobacco           [Sup. 39]
sbp               [Sup. 38]
chd               [Sup. 35]
ldl adiposity     [Sup. 34]
alcohol           [Sup. 33]
tobacco age       [Sup. 31]
ldl age           [Sup. 30]
typea famhist     [Sup. 30]
famhist adiposity[Sup. 28]
tobacco adiposity[Sup. 27]
famhist age       [Sup. 27]
sbp age           [Sup. 27]
typea adiposity   [Sup. 26]
chd age           [Sup. 26]
sbp adiposity     [Sup. 26]
ldl age adiposity[Sup. 25]
```

Figure 6: The frequent itemsets found after running the Apriori algorithm on the binarized data with $\epsilon = 0.25$ and $\delta = 0.7$. For convenience, the itemsets are sorted descendingly.

By means of this we can tell that 60% of the males have a family history of coronary heart disease, more than 50% of the males have type a behaviour, are categorized as older and are adipose. In general adiposity can be linked to many of the other features – this could also have been concluded using some basic summary statistics on the different features as this is only describing the amount of people with the given feature.

From looking at some of the itemsets with less support, we find that high cholesterol, being older and overweight is the case for 25% of the males. Adiposity is found for men with either a family history of coronary heart disease, men that are smoking or men with type a behaviour with a support of 27%. In all 35% of the males have coronary heart disease. There is a tendency of smoking a lot and being older, but since the value for tobacco is historic, this is hard to use as basis for any conclusion as this does not differentiate between someone smoking a lot at an early age and then quitting, someone starting at a later age or someone having smoked all his life. If one would like to use the tobacco consumption properly it would be better to divide your total tobacco amount with the number of years you have been smoking, but that also depend on how you want to link the consumption to your remaining data. Again this would have been easier to work with, had the data been divided into grouped age-ranges. Trying to say something about coronary heart disease and it's association with any of the other features is hard. From the itemset, the only group found with a support $\geq 25\%$ is coronary heart disease and age with a support of 26%.

For all the itemsets with support $\geq 25\%$ the highest level of support is 60% found for a single item. A maximmum support level for itemsets of two items is 41%. If we want groupings of more than two items the support gets no higher than 25%. As we are working with data describing a persons health it might be hard to find high support for any items. The features depend highly on your age which will increase the complexity of the way the Apriori algorithm should be executed, should we also handle this. Probably also the certainty that the results for the different groups are correct increase with this change, however, we do not have much data to work on which could make the results after further splitting of the data less representative.

Association rules

The rules associated with the different itemsets can be seen in the below figure. Not all rules will be interpreted, some of the interesting ones will be discussed and a conclusion as to what this can be used for will be described at the end.

It is clear to see that many of the associations all can be associated with age. The highest support found for any of the rules is `age <- adiposity`, which means that adipose men tend to be older, there is also a rule stating that `adiposity <- age`, stating the opposite, that older men tend to be adipose. While these rules somewhat seem to be an expression for the same conclusion, it is important to remember that the set of older men that also tend to be adipose is not the same as the set of adipose men that tend to be old and these two rules could very well be rules for different parts of the data.

```
age       <- adiposity         [Conf. 75, Sup. 41]
adiposity <- age               [Conf. 75, Sup. 41]
adiposity <- ldl               [Conf. 77, Sup. 34]
age       <- tobacco           [Conf. 80, Sup. 31]
age       <- sbp               [Conf. 72, Sup. 27]
age       <- chd               [Conf. 75, Sup. 26]
age       <- ldl adiposity     [Conf. 75, Sup. 25]
adiposity <- ldl age           [Conf. 85, Sup. 25]
adiposity <- tobacco age       [Conf. 75, Sup. 24]
age       <- tobacco adiposity[Conf. 87, Sup. 24]
age       <- sbp adiposity     [Conf. 83, Sup. 22]
adiposity <- sbp age           [Conf. 80, Sup. 22]
adiposity <- chd age           [Conf. 75, Sup. 20]
adiposity <- famhist age       [Conf. 73, Sup. 19]
```

Figure 7: The association rules found after running the Apriori algorithm on the binarized data with $\epsilon = 0.25$ and $\rho = 0.7$. For convenience, the association rules have been sorted descendingly.

Other than this we can also see that people with a high cholesterol tend to be adipose, this is the case for more than $\frac{1}{3}$ of the observations with a confidence of 77%. Following this with a suppport of 25% we have that high cholesterol and being adipose tend to happen for older men with a confidence of 75%. This fits very well with the known fact that one's cholesterol increases with age and is a larger problem for people who is adipose. Looking at a closely related rule, older people with high cholesterol tend to be adipose with the same support as the previous example, but a confidence level of 85%. For almost $\frac{1}{4}$ of the men, smoking and being adipose tend to happen when you are old with a confidence of 87%, this is what we find the highest confidence for. Looking at one of the things that is more interesting when trying to discover some of the leaing causes to coronary heart disease could be that the rule that people with coronary heart disease that are old tend to be adipose. This is the closest we get to any prediction on who is in danger getting coronary heart disease.

Overall age and adiposity is a tendency that can be extracted from the itemsets. If you have a family history of coronary heart disease and are older (but no more than 64 years old) you tend to be adipose, a high systolic blood pressure and age also tend to mean that you are adipose but here systolic blood pressure probably is common for older people, especially if they do not keep in shape. For all of the itemsets, the general confidence level is high ($\geq 70\%$, this is by the grouped deemed as an acceptable level. However one always have to look at both the support and the confidence level in order to get a good idea of whether the rule can be generalized or only holds for too small a sample size. A support of $\frac{1}{4}$ seems plausible dejectedly this only leaves 8 of the generated rules. As mentioned many times over, the age-range might have been able to increase the support for the different rules which could have given a better group-wise result.

What we hoped to see was that specific itemsets would lead to coronary heart disease, sadly, this is not something that we have found. If i.e. systolic blood pressure and high cholesterol lead to coronary heart disease, as a doctor, one could have used this information to make sure that the male's cholesterol decreased by the right exercise and diet which could by extension lead to a lower systolic blood pressure and minimize the risks of getting coronary heart disease.

References

[1] T. H. et Al., *Introduction to Machine Learning and Data Mining, chapter 19*. Technical University of Denmark, course notes, 2017.