Introduction to Machine Learning and Data Mining: Project 3

Jens Leysen (s191908)
Maud Leclerc (s191975)

November 2019

1 Introduction

In our final report, we will apply Clustering, Anomaly Detection and Association mining to our Titanic data set.

2 Clustering

2.1 Hierarchical Clustering

For the distance measure between individual observations we used the Euclidean Distance measure. There's a variety of distance measures that we can choose from; Manhattan, Euclidean, Mahalanobis etc. The choice of distance measures is very important, as it has a strong influence on the clustering results. Euclidean distance will tend to cluster observations with high values of features together (same holds for low values). The Mahalanobis distance measure takes the covariance of observations into account. This means that observations with attributes that tend to increase/decrease together (according to the covariance matrix) will have a lower distance value than they would have using the euclidean distance measure. Distance measures are often chosen depending on the application and research question, euclidean distance is the default for many applications.

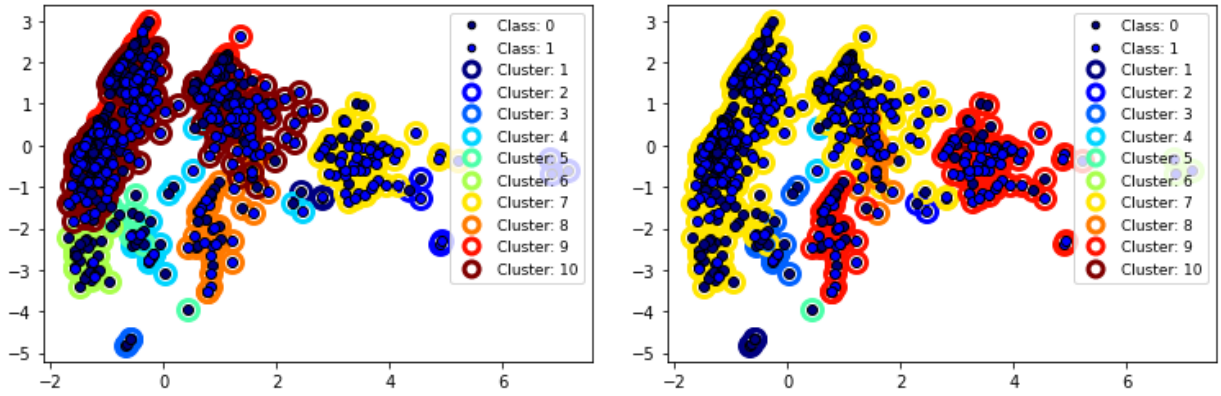
The linkage function defines the distance measure between groups of observations, we tried both the minimum and maximum linkage functions. A minimum linkage function tends to chain together connected components, while a maximum linkage function tends to favor round, equal sized clusters. A problem with minimum linkage is that it fails due to outliers or a very noisy dataset. When looking at the 3D plot of the 3 principal components (which capture more than 50% of the variation), we can immediately see some candidate outliers but also very unequal shapes and clusters of different sizes.

From the python documentation on "clusterplot": "If data is more than 2-dimensional it should be first projected onto the first two principal components." This is done to ensure we capture the most variance and can thus best distinguish between the different clusters that have been found by the hierarchical clustering.

We also standardized the data before applying the hierarchical clustering. For any of the other data operations (1-out-of-K encoding) we applied on our dataset we refer to our past projects.

Looking at the hierarchical clustering 2D plots (Figure 1.), we can immediately see that the clusters aren't identical because of the different linkage functions. The hierarchical clustering 3D plots confirm this (Figure 2.). Upon further inspection of our dataset, clusters seem to have more of a narrow elliptical shape, the single linkage function seems better at finding these clusters. The complete linkage function disregards these elliptical shapes and forms clusters of more equal size. We are however limited to this 3D plot of the 3 principal components.

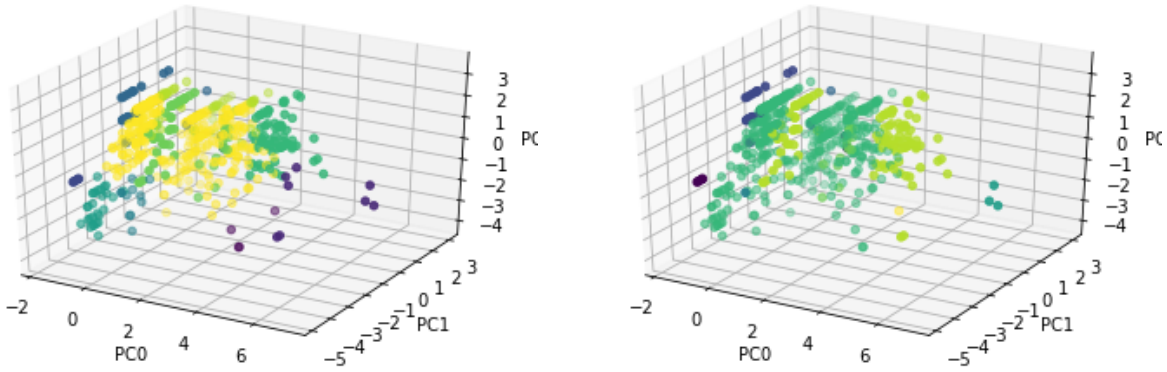
Figure 1: Hierarchical Clustering 2D Cluster Plots



(a) Clusterplot with complete linkage function

(b) Clusterplot with single linkage function

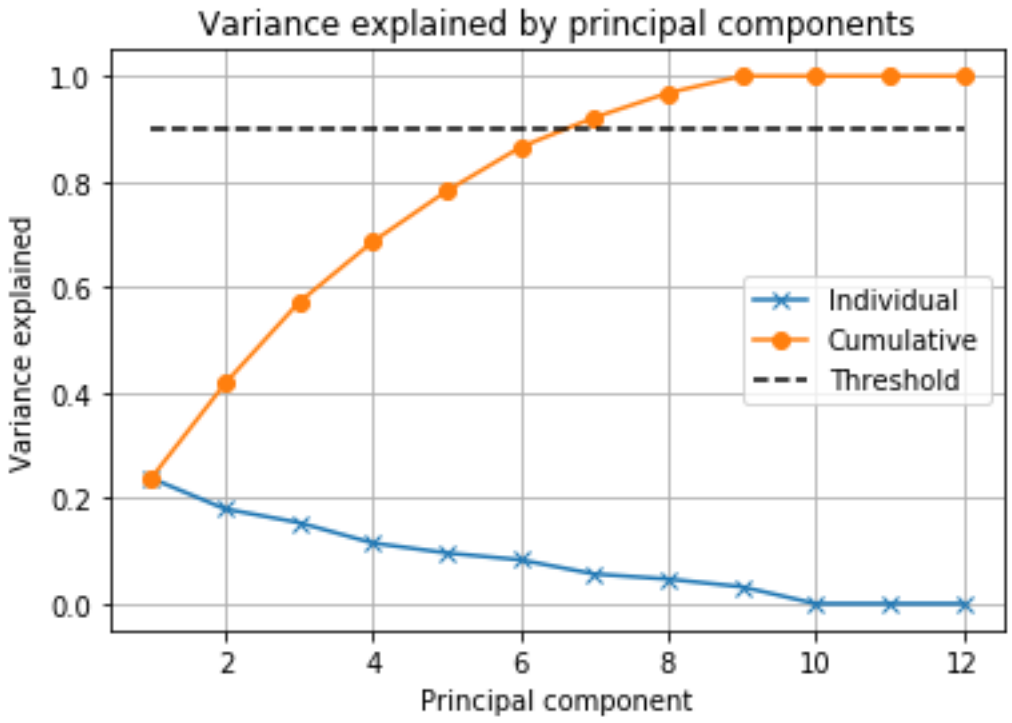
Figure 2: Hierarchical Clustering 2D Cluster Plots



(a) Clusterplot with complete linkage function (b) Clusterplot with single linkage function

In project 1, we applied a PCA analysis to our dataset: As can be seen from Figure 3, the first principal component accounts for nearly 25% of variation in the data. The first 3 principal components account for 60% of variation. The problem with this is that we lose nearly 40% of information in the 3D plots shown above, in other words, any interpretation we try to make is inherently based on only part of the information contained in the dataset.

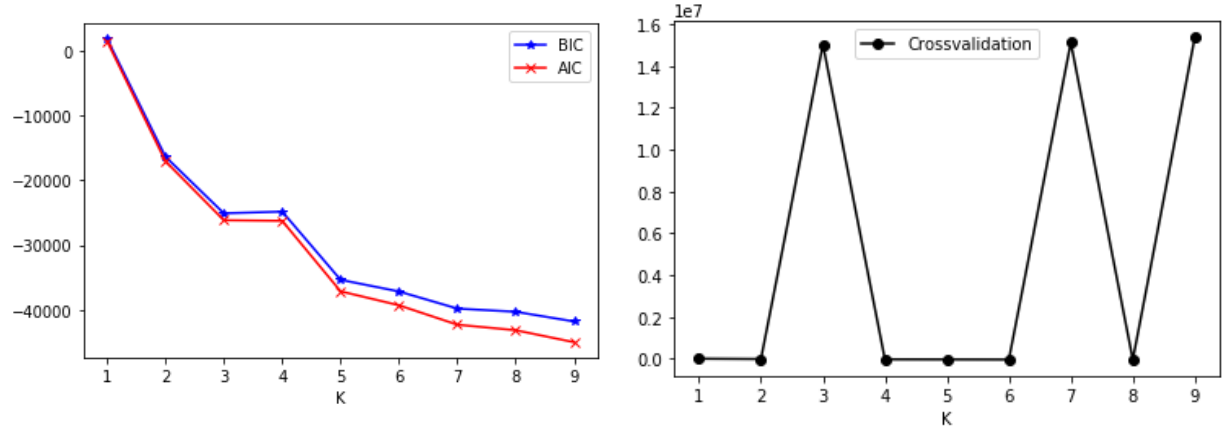
Figure 3: PCA Analysis



2.2 Gaussian Mixture Model

In the Gaussian mixture model (GMM) we use a mixture of K multivariate Gaussians to model the data. [1] We can easily determine K by using cross-validation. The goal when fitting the GMM using the EM algorithm is to maximise the log-likelihood. The performance of a GMM can then be measured on a test set. Apart from cross-validation the optimal number of clusters are sometimes derived by penalizing model complexity based on the Bayesian Information Criteria (BIC) or Akaike's Information Criteria (AIC). [1]

Figure 4: Finding K for the GMM



(a) BIC and AIC

(b) Negative Log-likelihood

We see that the AIC and BIC keep on decreasing and that the cross-validation is quite spurious for different cross-validation runs. (The minimum value is sometimes determined to be 5, sometimes 8.)

It's suggested to use the elbow method as a final resort, the elbow method is heuristic, there is no exact way to determine which value best describes this point. Looking at the AIC and BIC figure, we would argue that the optimal amount of clusters is around 5.

We applied a covariance matrix of type 'full'. Which means each component has its own general covariance matrix i.e. the components may independently adopt any position and shape. The 'Diag' option means each component has it's own diagonal covariance matrix i.e. the contour axes are oriented along the coordinate axes.

The covariance matrix of type 'full' is expected to perform best in general because of it's flexibility, it is prone to overfitting on small datasets however. Since our dataset is quite large, we used the 'full' covariance matrix type.

Discussion The goal of this part was to find any underlying patterns in our data we might have still been unaware of. The cluster centers are summarised in the "cds" variable, shown in the picture below.

Figure 5: Centroid Attribute Values

	0	1	2	3	4	5	6	7	8	9	10
0	-0.590326	0.296865	0.262215	-0.337238	-0.669716	0.906711	1.20415	-0.133329	0.23178	-0.202326	-0.147508
1	-0.590326	0.197865	0.347134	0.249909	-0.0055773	-0.373028	-0.506787	-0.442334	0.280164	-0.202326	-0.19955
2	1.69398	-0.566538	-0.997195	-0.180396	0.584082	-0.0741797	-0.0175752	1.01055	-0.444451	-0.202326	0.579846
3	-0.372773	-0.344476	0.621859	-0.233695	0.105268	-0.39911	-0.506787	-0.352043	-1.87252	4.94253	-0.472618
4	-0.590326	-0.566538	1.00281	0.162238	-0.744675	2.21215	1.3342	-0.176462	-1.87252	4.94253	-0.472618

We won't discuss all centroids separately. Interpreting Centroid 4; we see large values for attributes 5,6,8 and 9. Attribute 5 represent age, 6 the amount of siblings and parents, attribute 8 represents fare and 9 people that embarked in Queenstown. Meaning that cluster 4 represent a group of relatively older people, travelling with siblings/parents that paid a lot less than average and were very likely to embark in Queenstown.

Comparison of Clustering Methods Comparing partitions is done in the textbook by using a 'true clustering' as a reference, we do not have access to a true clustering for our data set. We could compare the different partitions found by the different clustering methods, this wouldn't make sense to us since there is no way of knowing if the partitions found in the data set make any sense. To say this in different words, different methods might find the same clusters and largely agree or disagree on clusters but this doesn't mean that the clusters are actually meaningful since you can't compare them to a 'true clustering'.

Note: We haven't included the 'survived' attribute in the data set we applied the clustering on.

3 Anomaly Detection

The outlier detection was computed on the full dataset; it includes the following attributes: Survived, Pclass, Sex, Age, SibSp, Parch, Fare and Embarked. As for the previous parts, Sex is encoded to be binary, and Pclass and Embarked are coded using one-of-K encoding. The dataset has been standardized.

3.1 Gaussian Kernel Density Estimation

The first method we used to detect outliers is by calculating the gaussian kernel density of the dataset, using leave-one-out cross validation. The goal of this part is to get an estimate of the probability density function of the attributes. To get the best possible estimate, we first calculate the most suitable bandwidth for the kernel probability density using leave-one-out cross-validation. A range of widths between 2^{-5} and 2^2 are tested to calculate a Gaussian kernel density, and the one giving the highest log of the likelihood is then used to compute the estimated probability densities for each observation. The optimal estimated bandwidth is 0,0625. We then displayed the observations having the 20 lowest densities, the first 10 can be see in table 1.

Two observations stand out because of their very low density. The first one is a 39 years old female, travelling in 3rd class with five children or parents. There is nothing odd about the values of her attributes, when compared to other similar passengers (of the same age and gender for example), apart maybe from the fact that she embarked in Queensborough, which is not that common. Maybe this is why this observation has a low probability density, but it is definitely not an outlier.

The second observation is a 64 years old male, travelling with 5 relatives. Again, when compared with similar passengers, there is nothing wrong with the data. A possible explanation for him being classified as an outlier is that he paid a very high fare, and other passengers that paid the same fare (or relatively close ones) are aged between 18 and 24. But this does not make this observation an outlier.

3.2 Nearest Neighbours

The second method used to find outliers is to estimate the probability density using the nearest neighbours method. Different amounts of neighbours have been used when running the algorithm, although they do not find the same densities, they do agree on the observations with the lowest densities. Since the value of the probability densities seem to stabilize at around 20 neighbours, this number is used to compare the possible outliers with the other techniques.

We notice that 3 estimations stand out. What these passengers have in common is the fare they paid: 512,239£. Since they are the only passengers paying this amount, and the highest fare below this is 263£, it seems logical that these passengers would be classified as outliers. However, these are real ticket fares and not erroneous. We have researched these passengers before: "It's stated in the description of Mr Thomas Drake Martinez Cardeza that he occupied one of the two most luxurious suites on board (B51/3/5, ticket 17755, £512, 6s). There are no strong reasons to think that the outliers represented in the boxplots below are erroneous values and we will thus not remove any of these outliers." [2]

3.3 Average Relative Density

Lastly, nearest neighbours average relative density (further referred to as ARD) was used to find outliers. The advantage of ARD is that is works on a dataset with uneven density among clusters. Instead of using probability densities, it enables us to consider that an observation is a potential outlier when its density is lower than what it is on average for its nearest neighbours.

We used the same amount of neighbours as for the previous method (20) to calculate the ARD. It can be seen in table 1, that there is one possible outlier. This possible outlier concerns a 28 year-old male passenger travelling in 3rd class, with no family. By comparing this passenger to passengers of roughly the same age and gender, there is nothing odd about this passenger (the paid fare is not very different, nor the fact that he is travelling alone).

We are not going to consider him as an outlier.

3.4 Discussion

The following table allows us to compare what observations have a low density according to the three methods. In the interest of simplicity, it displays only the 10 lowest density observations, although we made the comparison on the first 20. Out of these 20 observations, 9 are common to the first two methods (gaussian kernel density estimator and nearest neighbours), but none of these 9 has a low density according to the ARD.

	KDE		KNN		ARD	
Lowest density	Density	Observation	Density	Observation	Density	Observation
1	$1.692e^{-}82$	707	0.161	586	0.109	134
2	$7.9505e^{-}70$	350	0.162	536	0.12	0
3	$1.340e^{-}36$	196	0.164	207	0.135	531
4	$1.340e^{-}36$	330	0.167	707	0.143	295
5	$5.202e^{-}35$	494	0.197	350	0.165	675
6	$5.8083e^{-}35$	258	0.227	196	0.172	677
7	$1.306e^{-}30$	23	0.233	15	0.176	157
8	$2.903e^{-}20$	660	0.233	23	0.179	18
9	$2.467e^{-}17$	64	0.235	136	0.190	118
10	$3.058e^{-}17$	384	0.238	330	0.197	572

Table 1: Summary of the 10 lowest densities for the three methods

Our dataset has 8 attributes, we can hardly reduce its dimensions using PCA nor visualize it in a way that makes sense, it is not possible for us to determine visually if there would be any outliers. It is not really possible to estimate the density of any clusters either, since they are 8-dimensional. It is then quite hard to know if the ARD would be a better method to use than the two other ones because of the distribution of our data (which would explain why the low density observations are not the same), or if they are just no obvious outliers and that is why the methods do not agree.

From looking at the values of some attributes, however, and as discussed in the first report, some of the passengers do stand out, from the fare they paid. Some did not pay anything (because the ticket was purchased by their master, or by the H&W Guarantee Group), and others paid an extremely high amount. Although passengers paying no ticket fee are not detected by any of the three algorithms used, a very high fare does have an impact. More precisely, what we considered to be potential outliers (even if it turned out they were not) are the observations found by the nearest neighbours method. We can then consider that it was the most relevant method for our dataset, since it found relevant possible candidates for outliers, and conclude that our dataset does not contain any outliers that should be removed.

4 Association mining

The relevant attributes to research associations are the following : Survived, Pclass, Sex, Age, SibSp, Parch, Fare and Embarked.

In order for the algorithm to work, the data had to be binarized first. The sex attribute has been binarized, and the Embarked status has been one-of-K encoded. Age, SibSp, Parch and Fare have been changed to 1 when the value was strictly above the median, or 0 if it was lower than or equal to the median of the attribute. We should precise here that the median for both SibSp and Parch is 0, meaning that a value lower than the median means no relatives, and a value above the median means the passenger is not travelling alone. We then have the following categories, for which each observations holds a 0 or 1 : died, survived, 1st class, 2nd class, 3rd class, female, male, age<median, age>median, no siblings/spouse, siblings/spouse, no parents/children, parents/children, fare<median, fare>median, S, Q, C.

From what we knew about our dataset, we expected to find strong relations between the gender of a passenger and their chances of survival, and more precisely, we expected to find association rules stating that females were likely to survive. This would be even more obvious for 1st class female passengers. Since there were not a lot of female passengers compared to males, and even less 1st class female passengers, the support of these associations is going to be very low. We decided to set the support threshold for the Apriori algorithm accordingly, to 0.09. To avoid having an extremely huge amount of associations, and keep only the most certain ones, we then set the confidence threshold to 0.8.

With these threshold, the algorithm finds 1042 rules, which is more rules than what we can exploit. This is due to the fact that the support threshold is very low, but we need it to be this low in order to have associations including females. Moreover, the rules involving 3 and more attributes are a combination of the simpler rules, so looking at the first few rules only will be enough to understand the dataset.

Here are a few of the main associations found by the algorithm. First, we will start with general rules :

1. A passenger travelling 1st class is likely to have paid more than the median fare (supp: 0.253, conf: 0.978)

2. A passenger travelling 2nd class has probably embarked in Southampton (supp: 0.219, conf: 0.902)

3. A male passenger was probably travelling without parents or children (supp: 0.517, conf: 0.812)

4. If a passenger has paid less than the median fare, was probably travelling alone (no parents/children: supp: 0.459, conf: 0.919, no siblings/spouse: supp: 0.442, conf: 0.885)

5. A passenger older than the median, and accompanied by siblings or a spouse is likely to have paid more than the median fare (supp: 0.139, conf: 0.908)

These rules are fairly logical : a 1st class ticket is certainly more expensive than a 3rd class ticket. Most passengers were from Southampton, so this is why the algorithm finds this second rule. The third rule could maybe due to the fact that a lot of single, and not very rich, male passengers were travelling to America in hope for a better life. On the opposite, if females were travelling, it was because they were following their husband or some family members. The last two rules are somewhat complementary, in the sense that people who were travelling accompanied paid more than people travelling alone.

Since we were interested from the beginning in understanding which passengers were more likely to survive, the following rules concern specifically associations wich implied the death of the survival of a passenger.

1. A female is travelling 1st class was extremely likely to survive (supp: 0.112, conf: 0.964)
2. So was a female travelling in 2nd class (supp: 0.096, conf: 0.919)
3. But a male in 2nd class would most likely die (supp: 0.118, conf: 0.848)
4. And a male in 3rd class as well (supp: 0.302, conf: 0.850)
5. A passenger older than the median, travelling 3rd class, would probably also die (supp: 0.147, conf: 0.827)
6. And for the same passenger, travelling alone would not increase his chances of survival (supp: 0.105, conf: 0.824)
7. And if a passenger is dead, he was probably male (supp: 0.506, conf: 0.849)

And more generally, if a set of attributes induces "died", one of the attributes is almost always "male". This confirms what we first thought from the data : females, especially rich ones, were more likely to survive than males. This phenomenon has an underlying social explanation. There is a rule about "women and children first" [3], that would have been used when boarding people on the life boats when the Titanic was sinking. Although not all females survived, and not all men died, this explains why there is such a difference in the ratio of survivors from each gender. If we were to divide the "age" attribute between children and adults (and not with the median age of passengers), we would also probably find some similar rules. This breakdown of the passengers: [4] finds that 86% of children travelling 1st class survived, which confirms this hypothesis. This being said, there were very few children matching this criterion, so we would need an extremely low support threshold to come up with this rule. There also seems to be an advantage in being from a high social status, and being able to travel in 1st class, since in general the rules (not disclosed here for a matter of simplicity) find associations between "1st class" or "fare>median" and survived, as well as "fare<median" and "died".

5 Attribution Table

Section	Responsibility
Introduction	Jens (100%)
Clustering	Jens (80%) / Maud (20%)
Anomaly Detection	Maud (80%) / Jens (20%)
Association Mining	Maud (80%) / Jens (20%)
Discussion	Maud (100%)

References

[1] EXERCISE 11, (November, 2019), Mixture models and density estimation with PYTHON

[2] Encylopedia Titanica: Thomas Drake Martinez Cardeza,
`https://www.encyclopedia-titanica.org/titanic-survivor/thomas-cardeza.html`

[3] Women and children first, section on the Titanic,
`https://en.wikipedia.org/wiki/Women_and_children_first#21st_century`

[4] Demographics of the Titanic passengers: Deaths, Survivals, Nationality, and Lifeboat Occupancy,
`http://www.icyousee.org/titanic.html`