Technical University of Denmark

02450
Introduction to Machine Learning and Data Mining

Report 3

Data
Breast Cancer Wisconsin (Diagnostic) (BCWD)

Work Distribution
Celia Sullca Cailloux (s132534) **50%**
Oseze Esther Iyore (s121922) **50%**

Technical University of Denmark
May 1st, 2018

Contents

1 Introduction

The focus of this report is unsupervised learning. The first part will go through clustering by use of Gaussian Mixture Model and Hierarchical Clustering. The second part will be about outlier and anomaly detection. The third part will go through association mining.

2 Clustering

In clustering the idea is to cluster or label the data into groups with high similarity within each group and large dissimilarity between the groups. There are two main types of clustering, partitional clustering and hierarchical clustering. In partitional clustering the data is partitioned into non-overlapping clusters in such a way that each observation is in one cluster. In hierarchical clustering a single number of clusters is not defined instead the data is arranged in a nested sequence of subsets organized as a hierarchy where the bottom has the finest partition and the top has the most coarse partition (one cluster). A cutoff can later be decided. In the coming subsection a partitional clustering method called the Gaussian mixture model is used to cluster the Breast Cancer Wisconsin (Diagnostic) (BCWD). Following, a hierarchical clustering is performed. Finally the two methods are compared.

2.1 Clustering with the Gaussian Mixture Model

To find the best suited number of clusters K, a 10 fold cross-validation is performed on the data. Bayesian Information Criteria (BIC) and Akaike's Information Criteria (AIC) are also used to determine optimal number of clusters.

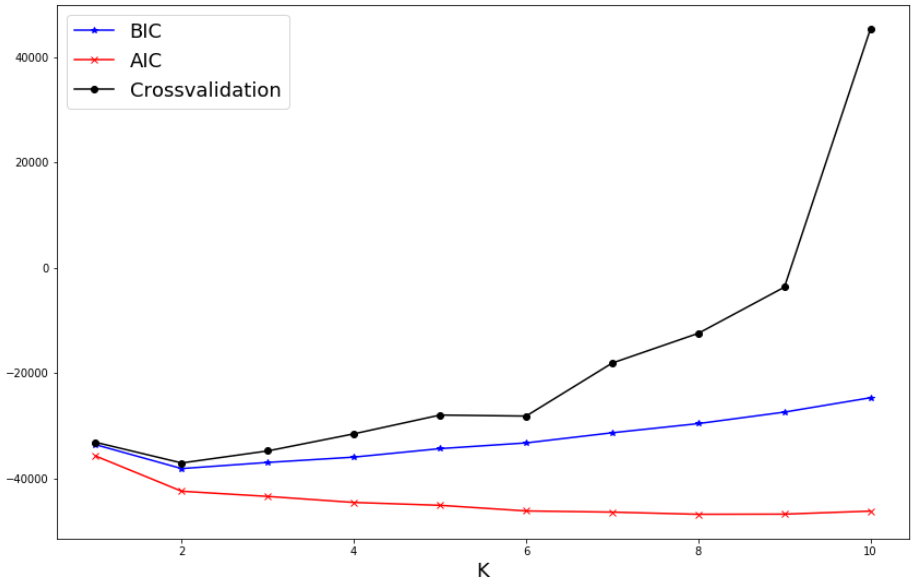


Figure 1: Bayesian information criterion, Akaike information criterion and cross-validation error as a function of clusters for a 10 fold cross-validation.

Figure 1 shows the cross-validation error, BIC and the AIC as a function of number of clusters. The two information criteria are given as:

$$\text{BIC} = -2\log(L) + p\log(N), \qquad \text{AIC} = -2\log(L) + 2p$$

where L is the log likelihood of observing the data, p is the number of attributes and N is the number of observations. Regarding figure 1 it is seen that the cross-validation error and the Bayesian information criterion are lowest for 2 clusters but for the Akaike criterion the value is the lowest for 8 clusters. The formulas for the two information criteria define a trade-off between a good model and a complex one where model complexity is penalized. The AIC penalizes model complexity less than

BIC which explains the difference in result of these two criteria. For this analysis the cross-validation error has the highest weight and since the result is consisten with also the result of the BIC this is the value of optimal clusters chosen to continue the analysis. This also corresponds to the two class labels: benign and malignant for the breast lump diagnosis.

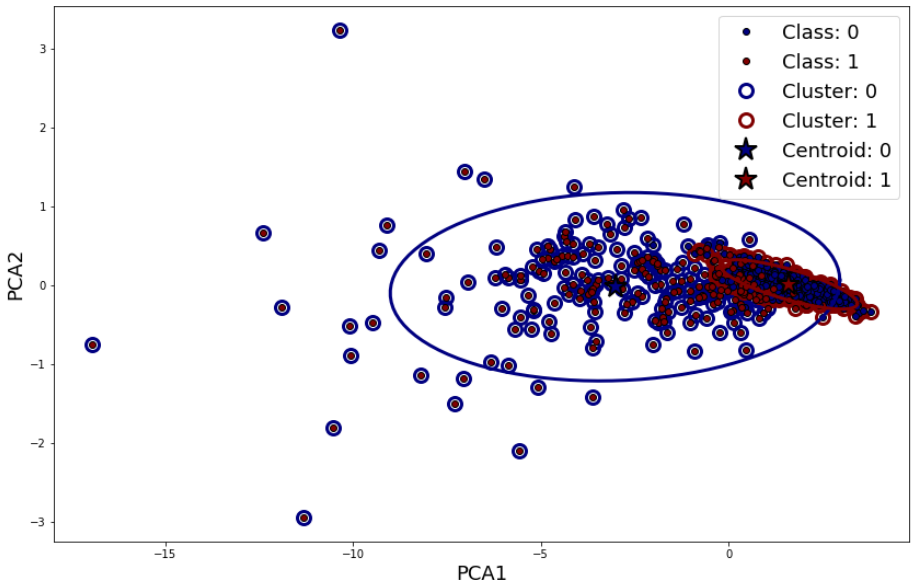


Figure 2: BCWD data in two Clusters by Gaussian Mixture Model plotted for the two principal components.

Figure 2 shows the attributes plotted in the first two principal component axis, in two clusters. The two first principal components account for 98.8 % of the variance of the WBDC data. The outer color represents the cluster which the observation is placed in and the inner the true class. The figure shows that almost all the observations are classified correctly compared to their true class. The stars show the cluster centroids and it is clear that the data points in class 0 are much more closely packed which is shown also in the centroid position. The distribution of the observations in class 1 are much more spread out and some observations are very far from the centroid of the data.

2.2 Hierarchical Clustering

When performing Hierarchical clustering we need to decide which linkage function and dissimilarity measure to use. Regarding our data it seems that the data is clustered in what is most similar to center based cluster, for that reason the group average could be used as a linkage function. The dissimilarity measure usually does not have a very large impact in the result and both the euclidean distance and city-block distance could be used. To ensure the assumptions made are correct the Rand index, Jaccard similarity and NMI are computed for different linkage functions and dissimilarity measures to compare the errors and decide based on this.

	Rand		Jaccard		NMI	
	City-block	Euclidean	City-block	Euclidean	City-block	Euclidean
Average	0.5498	0.5521	0.5312	0.5314	0.1050	0.1130
Single	0.5326	0.5326	0.5315	0.5315	0.0188	0.0188
Complete	0.5521	0.5521	0.5314	0.5314	0.1129	0.1129

Table 1: Rand Index, Jaccard similarity and Normalized Mutual Information for the three linkage functions: average, single and complete and for the two dissimilarity measures city block and euclidean distance.

Table 1 shows the values for the different linkage functions and dissimilarity measures. The table confirms that the difference when using the two different dissimilarity measures is very small or non-existent. The assumption that group average should be a good linkage for our data is disproved

and the table shows that the complete function would be preferred. Finally the linkage function complete and the dissimilarity measure euclidean is chosen for the Hierarchical clustering.

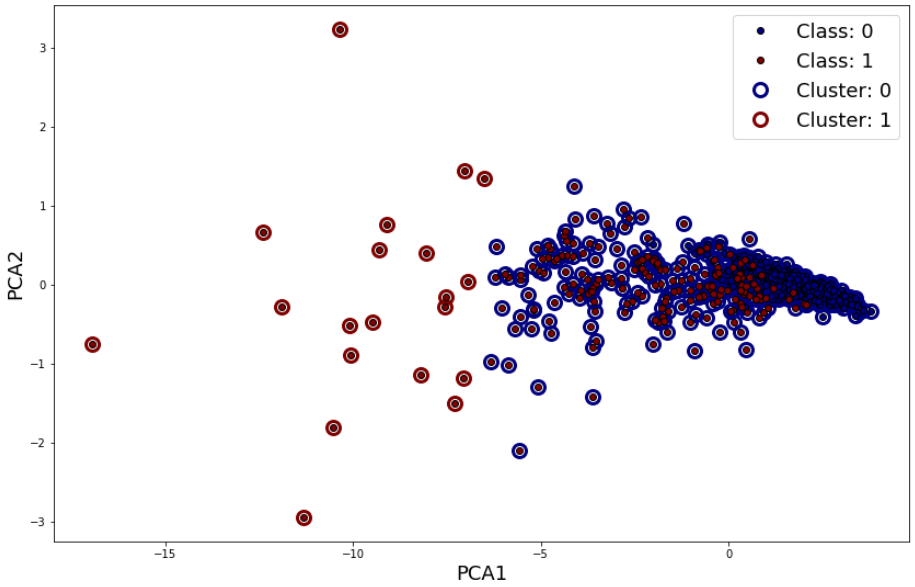


Figure 3: BCWD data in two Clusters by Hierarchical Clustering plotted for the two principal components

Figure 3 shows the data attributes in two clusters by Hierarchical clustering plotted for the two principal components. The figures shows that a lot of the data that was previously put into the second cluster is now put into the first and all the data points in the middle of the plot are misclassified compared to the true values.

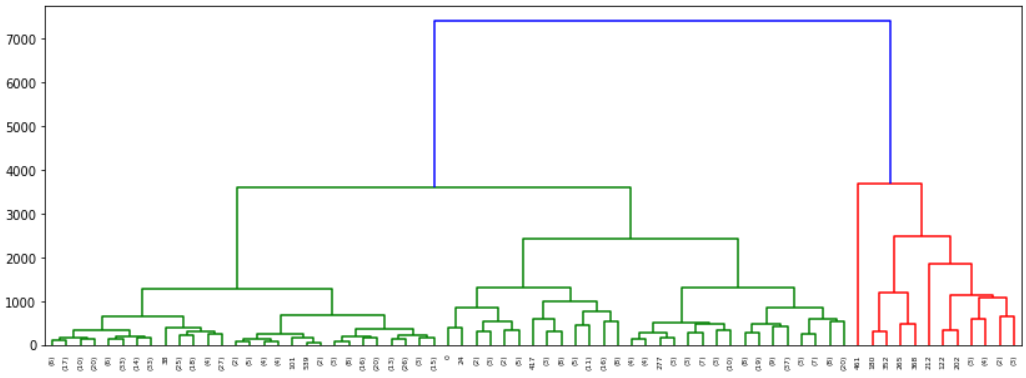


Figure 4: BCWD data Dendrogram

Figure 4 shows the dendrogram for the data points. It is seen that the data point in the middle are merged to the cluster on the left where they should have been merged with the cluster on the right to get the correct classification.

2.3 Evaluation of Clusters

It has been decided to evaluate the clusters based on the rand index, Jaccard and NMI. Table 2 shows these result. Here it is seen that the Gaussian mixture model classified much more precise compared to hierarchical in this case. This confirms what the PCA-plots already showed.

Cluster Method	Rand	Jaccard	NMI
Gaussian Mixture Model	0.8751	0.7942	0.6472
Hierarchical	0.5521	0.5314	0.1129

Table 2: Evaluation of Cluster Method

3 Outlier/Anomaly Detection

Anomaly detection tries to find observations that may be considered different from the other observations. In this report anomaly detection will be conducted using the probabilistic definition of an outlier. Hence, an observation is regarded as an outlier if its in the low density regime of the data density function. In this section three different methods will be presented: Gaussian Kernel Density, K-nearest Neighbour Density and K-nearest Neighbour Average Relative Density. Prior to all methods, the data has been normalized.

3.1 Gaussian Kernel Density Estimator

The gaussian kernel density estimator (GKDE) uses multivariate gaussians around each data point and estimates the density by summing the found gaussians. The density is thus described by a mean and the covariance $\lambda^2 \boldsymbol{I}$, where the kernel width λ is the parameter to be optimized. Using KDE gives the advantage of an efficient leave-one-out to estimate the kernel width λ. Figure 5 shows the evaluation of each kernel width λ and the found optimum for $\lambda = 0.21$. If λ is too low, the found Gaussian distribution only includes the point itself. At too high λ the probability is likewise low due to dilution.

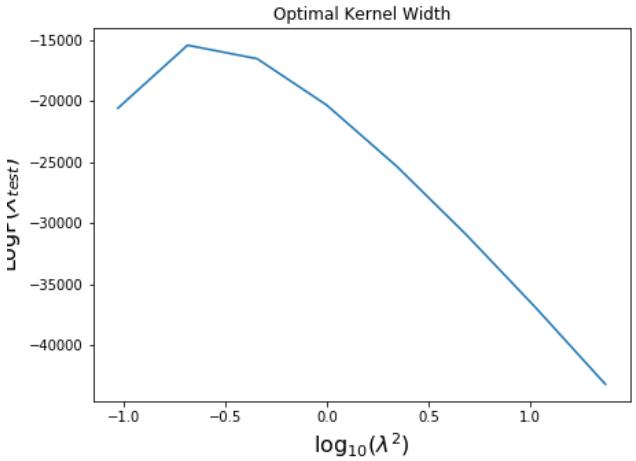
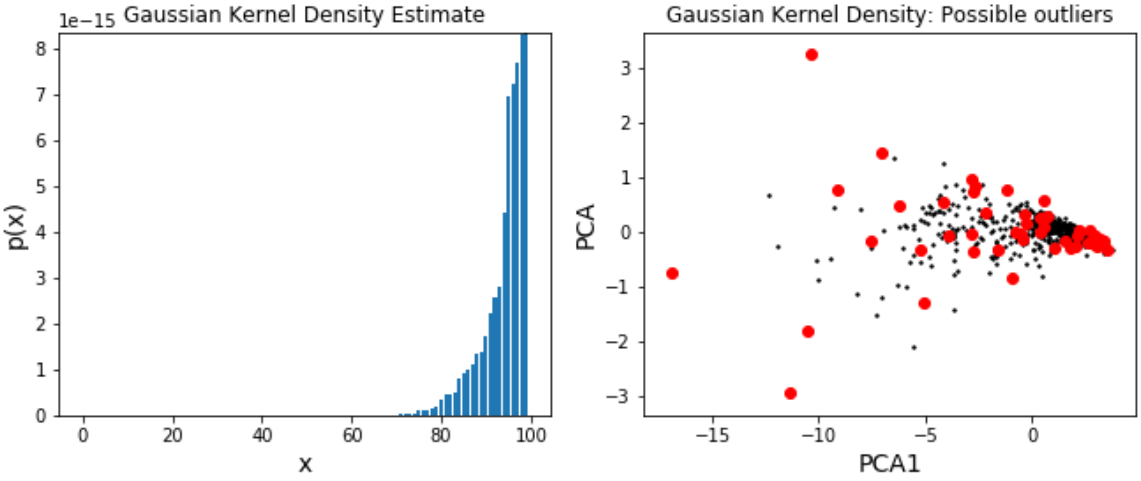


Figure 5: Leave-one-out cross-validation performed on a range of kernel widths λ. Optimum is found be $\lambda = 0.21$

Figure 6a show GKDE run on the BCWD data for $\lambda = 0.21$. The attributes with the 100 lowest probability are plotted in the bar plot (Figure 6a). The number of observations are 569, thus approximately 1/6 of the observations have probabilities of the order $\sim 1 \times 10^{-15}$ or orders of magnitudes lower making them highly unlikely. This suggest that the GKDE method may not be the best choice to describe the density function of BCWD data. The 50 observations with the lowest probabilities ($p_x \sim 1 \times 10^{-20}$) are plotted in a plot with the two principal components, PCA1 and PCA2, in figure Figure 6. GKDE is not suitable for a dataset where the density varies, which seems to be the case for BCWD data.

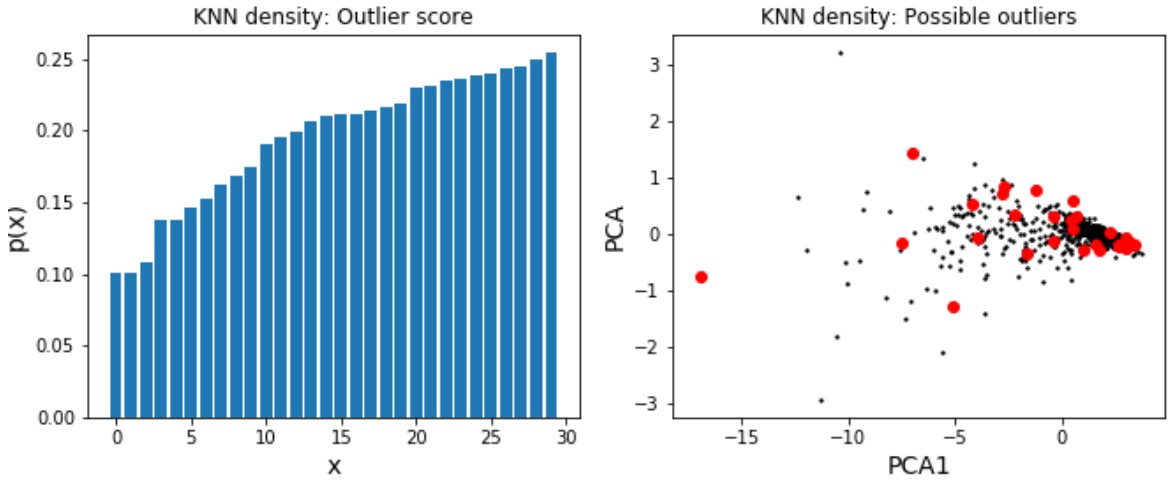


(a) (b)

Figure 6: (a) The 100 observations with lowest probabilities in the GKDE denstiy. This is approximately 1/6 of the observations that have probabilities of 1×10^{-20} or much lower. (b) First two principal components PCA1 and PCA2 plotted, where the possible outliers (the 50 observations with lowest probability) are marked red.

3.2 K-nearest Neighbour

K-nearest neigbour (KNN) estimates the data density as the inverse of the average distance to the k nearest neighbour. Hence, a choice of distance and number of nearest neigbour is required. In this report, the euclidean distance has been chosen together with K = 5. K=3,4,5,6,7 was tried and showed no significant difference. Figure 7 show the results of KNN KDE.



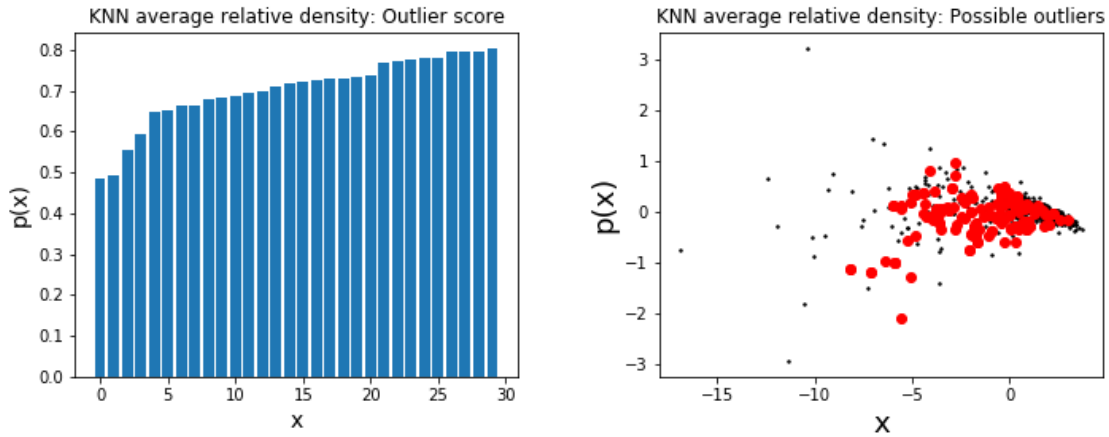
(a) (b)

Figure 7: KNN KDE applied on the BCWD data. Both figures show the 30 observations that encountered lowest probabilities. (a) The probabilities of the 30 possible outliers. (b) PCA plot of the 30 possible outliers.

Only the 30 observations with lowest probabilities are included in both Figure 7a and Figure 7b. All 30 observations have probabilities of the same magnitude ($\sim 1 \times 10^{-1}$), thus KNN KDE doess not seem to detect any outliers. The possible outliers show now obvous pattern in the PCA plot. Comparing GKDE and KNN KDE (Figure 6 and Figure 7b) the estimated outliers from both methods do not seem to overlap.

3.3 K-nearest Neighbour Average Relative Density

A disadvantage GKDE and KNN KDE is the cluster densities are not taken into account. However, the K-nearest neighbor average relative density (KNN ARD) accounts for possible varying densities. In the section about *Clustering*, the two found clusters clearly showed different densities, why KNN ARD could happen to be more suitable that the two previous methods applied.



(a) (b)

Figure 8: (a) Outlier score for KNN ARD of the 30 least probable observations. (b) PCA plot of the 30 leat probable observations.

The results of KNN ARD applied on the BCWD data are shown in Figure 8. In Figure 8a all 30 observations with the lowest probabilities have probabilities $\sim 1 \times 10^{-1}$ thus there seem to be no outliers. The same thirty observations are plotted in a PCA plot in Figure 8b. These possible outliers do not resemble either of the possible outliers found using the two previous methods GKDE and KNN KDE.

3.4 Conclusion

As mentioned in the first report, the BCWD data has been tested for outliers and was reported not to have any outliers. This agrees with the results in this section, where the GKDE computed a density function where most of the observations were highly unlikely, whereas both KNN KDE and KNN ARD showed no apparent outliers.

4 Association Mining

Association mining is an area of machine learning where the interest lies in finding relations between different attributes. The market-basket analogy is often used to describe association rules: if a person buys a beer then how likely is that person also to buy milk or another item in the store? For the BCWD we wish to find relations between the 10 different attributes and how frequently these relations appear. The attribute relations are called association rules describes by two factors: (1) the support and (2) the confidence. For a given association rule $X \rightarrow Y$ where both X and Y are attributes in a given observation $\boldsymbol{x}_i$ it means that if X appears then also Y will appear. The support is then the probability of the joint distribution of X and Y given the entire data set. The confidence is the probability of the joint distribution given the attributes that contain X. Thus if all X only occurs when Y occurs then the confidence interval will be 100%. To apply the apriori algorithm used for association mining, the BCWD must be binarized. All the attributes in BCWD are continuous, why binarization is enabled by introducing a threshold that assigns to 0 when not meeting the threshold and 1 when meeting the threshold. The threshold defined for the binarization is the median Q_2 of the N observations of each attribute. Hence, each attribute is split into two attributes.

After a series of testing, the minimum support was set to 45% for association between only two attributes. The minimum confidence was set to 50% to find frequent association rules. Tab. 3 shows the found associations rules and their respective confidence and support. *(0th−Q_2)* means that is the given attribute is below the median Q_2 whereas *(Q_2-100th)* means the given attribute is above the median Q_2.

Association Rule			Confidence	Support
Area (0th−Q_2)	←	Radius (0th−Q_2)	99	49
Radius (0th−Q_2)	←	Area (0th−Q_2)	99	49
Radius (Q_2-100th)	←	Area (Q_2-100th)	99	49
Area (Q_2-100th)	←	Radius (Q_2-100th)	99	49
Perimeter (0th−Q_2)	←	Radius (0th−Q_2)	97	49
Radius (0th−Q_2)	←	Perimeter (0th−Q_2)	97	49
Perimeter (Q_2-100th)	←	Radius (Q_2-100th)	97	48
Radius (Q_2-100th)	←	Perimeter (Q_2-100th)	97	48
Area (0th−Q_2)	←	Perimeter (0th−Q_2)	97	48
Perimeter (0th−Q_2)	←	Area (0th−Q_2)	97	48
Area (Q_2-100th)	←	Perimeter (Q_2-100th)	97	48
Perimeter (Q_2-100th)	←	Area (Q_2-100th)	97	48
Concavity (0th−Q_2)	←	Concave Points (0th−Q_2)	90	45
Concave Points (0th−Q_2)	←	Concavity (0th−Q_2)	90	45

Table 3: Associations rules for BCWD.

Due to the defined threshold, the support can not be higher than 50% for each attribute ((0th−Q_2) and (Q_2-100th)). The BCWD are measurement of cell nuclei, thus it is expected that the attributes that are related to size (Area, Radius, Perimeter etc.) show association rules due to their mathematical correlation. Thus if an observation has area above the median Q_2 then the radius, perimeter etc. are also expected to be above median, simply due to the mathematical relation between the attributes. This is the case for the shown association rules in Tab. 3. The size-related attributes are completely dominant. This is also the reason why such high confidence are found. The two last association rules Concavity (0th−Q_2) ← Concave Points (0th−Q_2) and Concave Points (0th−Q_2) ← Concavity (0th−Q_2) show slightly less confidence which is possibly due to a higher uncertainty when determining concavity and concave points. It would have been more interesting to find association rules between the non-size-related attributes (such as Texture, Smoothness, Compactness etc.). These occured for supports lower that 40%. However, due to the highly dominant size-related attributes, filtering the association rules would be a greater task.

Appendix

Work distribution

We worked in collaboration for all three parts: clustering, anomaly detection and association mining. However, Oseze (s121922) had the main responsibility the for clustering whereas Celia (s132534) had the main responsibility for anomaly detection and association mining.