

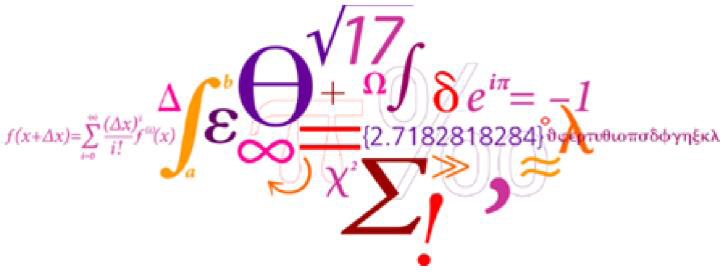
02450 Introduction to Machine Learning and Data Mining

*Unsupervised Machine Learning*

Report 3

Longfei Lin, s185882

Iraklis Chrysikopoulos, s182995



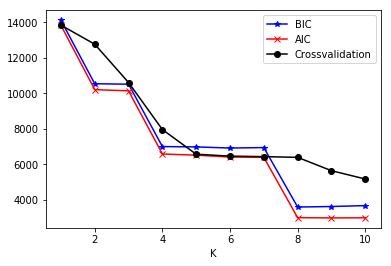
07/05/2019

1. **Clustering**

In unsupervised machine learning, a method named clustering of data is used and in this part it will be shown and discussed how our dataset about heart disease can be analyzed in this way. A known method in clustering is Gaussian Mixture Model (GMM) and with cross-validation we will find the optimum number of clusters for our dataset. After that another method of clustering called hierarchical clustering (HC) will be applied in the dataset using different linkage functions. Finally, a comparison will be made between GMM and HC methods in order to extract conclusions about the quality of the clustering that has been made.

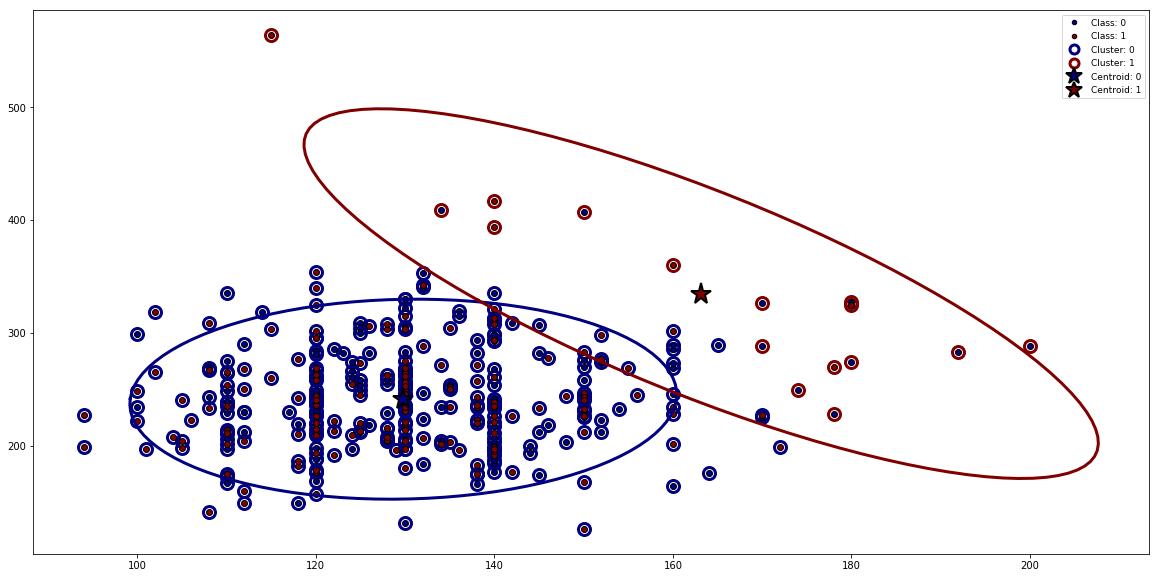
* 1. **Gaussian Mixture Model (GMM)**

Gaussian Mixture Model is a method of clustering and in other words is a set of multivariate normal distribution and is used to make a more flexible distribution to our data. In the first step we try to find a good number of cluster to our dataset. Thus we use the BIC (Bayesian information criterion), AIC (Akaike information criterion) and 10-fold cross validation.



*Figure 1. – BIC, AIC, 10-fold cross validation – score vs number of clusters*

From figure 1, we can extract information about the optimal number of clusters/components we should choose to proceed. As the score gets lower then that’s the optimal number of cluster we should choose. In our case we can see that the cross-validation curve stabilizes after K=4 so the best number of clusters is 4. Also, it’s not useful to choose a large number of cluster i.e. K=10. With the python scripts there are many options concerning covariance matrix constraints such as diagonal, full, tied etc.



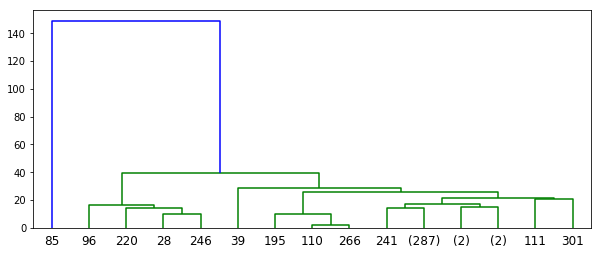
*Figure 2. – Cluster representation and centroids*

From figure 2 it is shown the 2 different clusters that occur and their centroids. The attributes that are used are TRESTBPS (resting blood pressure) and CHOL (serum cholesterol).

* 1. **Hierarchical Clustering**

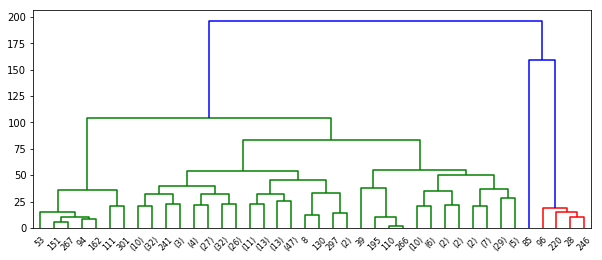
Hierarchical clustering arranges the data in a nested sequence of partitions organized as a hierarchy and as a result overcomes the limitation by instead of finding a single K. The low point of the hierarchy correspond to each observation in a unique cluster and the top point in the hierarchy is a unique cluster of all observations. This kind of clustering needs a specific method called linkage function. In this way, this function can find the closest neighbor and then merge the two clusters in a tree shape (dendrograms). The three function that we use in this project are:

* Minimum (single)
* Maximum (complete)
* Average



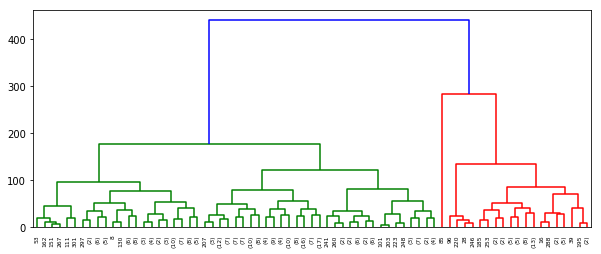
*Figure 3. –Dendrogram of Single Linkage Function*

In this case of the minimum linkage the distance between the groups is the distance between the closest pair of observations. We can see from the colors that our data are separated in 2 clusters.



*Figure 4. – Dendrogram of Average Linkage Function*

In this case of the average linkage, the distance between the groups is the average distance between all pairs in the groups. We can see a separation of our data which gives us 3 clusters.



*Figure 5. – Dendrogram of Complete Linkage Function*

In the last case of the maximum linkage the distance between the groups is the distance between the most distant pair of observations. We can see from the colors that our data are separated in 3 clusters. The complete linkage function seems to have better results.

* 1. **Hierarchical Clustering vs GMM – Quality Evaluation**

In order to compare and evaluate the two different methods of clustering we use 3 different indices for cluster validity: Rand statistic, Jaccard coefficient and normalized mutual information (NMI).

|  |  |
| --- | --- |
| **GMM** | |
| Rand | 0.12 |
| Jaccard | 0.12 |
| NMI | 0 |

Table A – Indices for GMM evaluation

|  |  |
| --- | --- |
| **Hierarchical Clustering** | |
| Rand | 0.50 |
| Jaccard | 0.50 |
| NMI | 0 |

Table B – Indices for hierarchical clustering evaluation

From table A and B is easy to compare and decide which method gives us better clusters. In our case Hierarchical clustering is better as we see from Rand and Jaccard indices. For some reason NMI in both cases is 0, so we don’t take it into account.