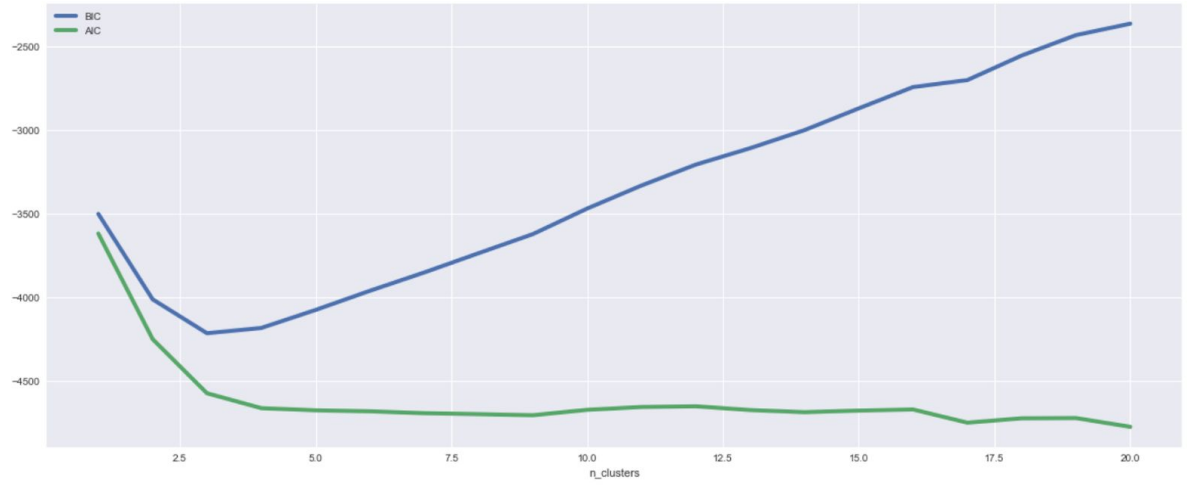
Team members: Olav Gjerde, Malin Wist Jakobsen (ogj005, xun007)
| INFO284 | Date: 30.04.18 |

INFO 284 - Spring 2018

Second Obligatory Group Assignment

Between 1000 - 5000 characters of report describing how the clustering algorithms comparatively performed on the task. State your opinions as to why the results were as such. The report should also include a visualisation of the clustered data in a 2D scatter plot. You may want to reduce the dimensionality of your data before you visualise it.

Fine tuning of clusters



Before we can fit our dataset to the KMM clustering algorithm, we had to fine tune the amount of clusters needed. We used two different methods for this, the Akaike information criterion and Bayesian information criterion. We see from the figure above that the results from AIC and BIC hit a low point around 3-4, this would be a reasonable starting point for how many clusters/categories we are looking for. But we had the categories in our dataset already present, so we know that 3 clusters are the amount we were looking for. This is usually not the case with datasets.

Team members: Olav Gjerde, Malin Wist Jakobsen (ogj005, xun007)
| INFO284 | Date: 30.04.18 |

K-means clustering

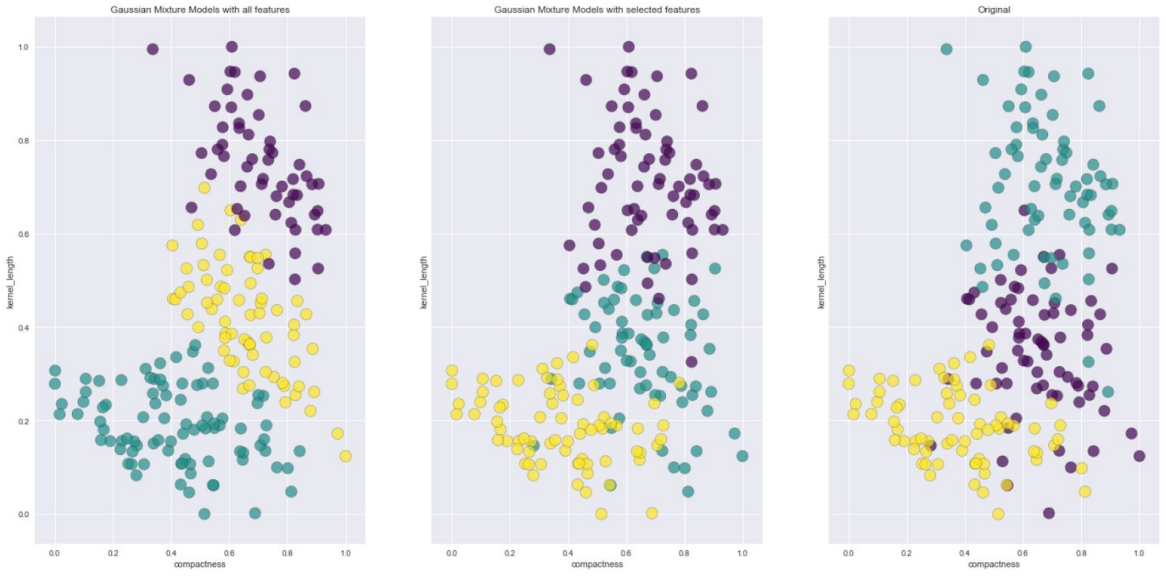
As is known with the K Means clustering algorithm, it does not lend itself well to data that is spread out in a non-spherical fashion. This implies that if the clusters have complicated geometries, the cluster centers may not end up where you would expect them to. This is further confirmed by looking at the plot where K Means have considered all the features in the data set. In the second plot where we have removed some of the features that had a diagonal spread (area, perimeter), we see that the centers align more closely to what you might have chosen by "eyeballing" where to put the center.

Using all features: Precision: 0.8809523809523809

Using selected features: Precision: 0.9190476190476190

Team members: Olav Gjerde, Malin Wist Jakobsen (ogj005, xun007)
| INFO284 | Date: 30.04.18 |

Gaussian Mixture Models for Clustering



As known with gaussian mixture models for clustering, it is better with non-spherical shapes than k-means clustering. This is because it is not technically a clustering model, but a generative probabilistic model describing the distribution of the data. This means that points that land in between the "circular bounds" of the K Means clusters may get mixed, and there is no definition of how probable it is that one point belongs to a cluster versus another. Gaussian Mixture Models does not suffer from the aforementioned problems. It will be able to fit data sets that have more "problematic" shapes, e.g elliptical shapes. As we can see in the second plot, dropping the features *area* and *perimeter* increases the performance significantly as it did for the K Means clustering.

Using all features: Precision: 0.8333333333333334

Using selected features: Precision: 0.9571428571428572

Conclusion

When we look at the precision, we see that GMM has a higher percentage than the K-means, this is explained by that GMM is not technically a clustering model, but a generative probabilistic model describing the distribution of the data. So if the data was more spherical, the k-means would have had a higher percentage. Another thing that affected the result was the scaling of dataset. In a setting where we do not know exactly which types of measurements that have been used for each feature, it is good precaution to scale these measurements so that one feature does not take precedence over another. We used the min-max scaler since it gave a general higher precision. The biggest difference in precision was when we dropped the features: perimeter and area. The percentage in K-means went from 0.88% to 0.919%, but GMM showed the most difference from 0.83% to 0.95%. The reason the precision increased is because perimeter and area was the two features that was the most diagonal and non-spherical compared to the other features.

We used *t-distributed Stochastic Neighbor Embedding* to reduce the dimensionality of our dataset so that we can have a 2 feature representation of the whole dataset and map our result to the data points from the reduction.

Below you will see the best results compared to the original in scatter plots based on the data set with reduced dimensionality