Group Assignment

INFO284 / Machine Learning

Assignment 2

Candidate numbers:

90 | 36 | 18 | 107 | 64

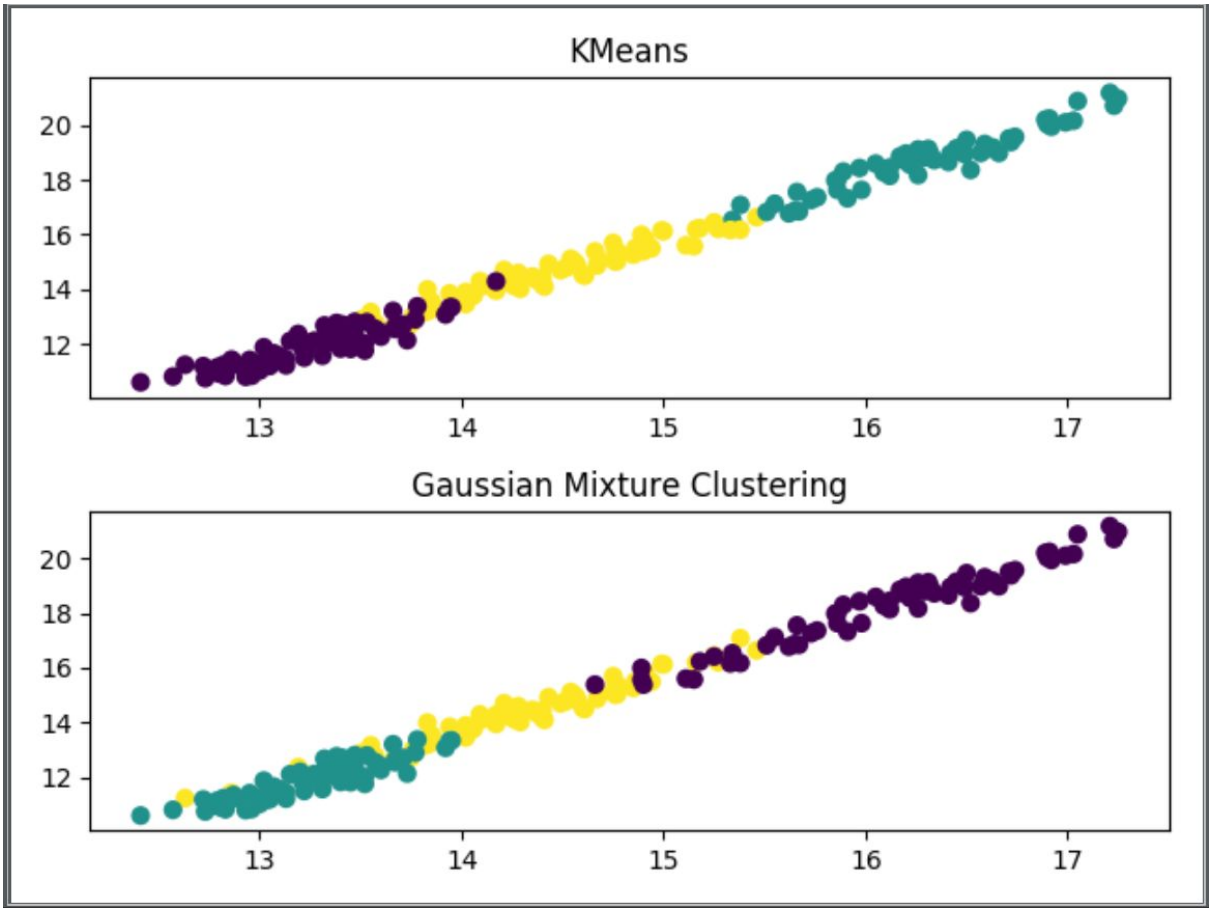
DEPARTMENTS OF INFORMATION SCIENCE AND MEDIA STUDIES

UNIVERSITY OF BERGEN

Spring 2018

Here's the overview of this task, and how the two different data models managed to handle the dataset given. This provides a great information when it comes to comparing the two different models, and how they stack up against each other when it comes to decide which one is the best one for this type of data and which one has the best visualization model of the dataset.

Below is the visualization of the dataset using Gaussian Mixture Model and using K-Means.



This shows that Gaussian Mixture Model is the best one to use for this type of data, because of how it handles the dataset and the clusters in the way that represent the data properly.

When it comes to differencing the two models for displaying clustered data, we have to take a look at what the two different methods do to the data. When it comes to Kmeans, this method assumes that each point belongs to a specific cluster, and that it can't belong to two different clusters. Here is where Gaussian Mixture Model has an advantage. It can tell which cluster the point belongs to, but also how much it can belong to other different clusters. The one with the highest probability is the cluster the point belongs to. The speed is also an advantage for K-Means over Gaussian Mixture Model, since K-Means is faster to cluster data. But when it comes to this complex dataset , Gaussian Mixture Model might be the best choice when it comes to displaying the results. But then the results has to be reduced to fully show the visualization of the plot.

This shows again when it comes to how it handles our data, with Gaussian Mixture Model you have a greater sense of how the data overlaps and interacts. Showing us a greater view of which point belongs to which cluster, while also being a significant point to other clusters. What's surprising is that how well K-Means adapts to this kind of dataset. Managing to difference the points in each clusters, while also maintaining some overlaps of clusters.

When it comes to using the Gaussian Mixture Model we had to stick with using fewer components to get a satisfying results, without limiting the amount of components, the data clusters were to clustered to see the results properly. Because the Gaussian Mixture Model will always use all of the components that are available in the set, we have to find out how many different components we want to show in our cluster diagram. We went with a maximum amount of clusters to three, giving us the best view for our diagram. This can be distinguished in the diagram above, with the three colors; green, lilac and yellow.

What's surprising about this data is how well K-Means handle it, as you can see how close the two different clustering methods are on our data. K-Means has it difficult when it comes to complex shapes, which this dataset is somewhat. But it still manages to provide an excellent overview of the different clusters. It's also surprising how well it manages to handle non spherical clusters. To prevent this loss when it comes to using K-Means for datasets that it usually struggles with, it's possible to use more clusters, managing to cover these complex structures with several small spherical clusters. But this lead on to the difficulty of having overlapping data points. This provides it possible to apply a linear model on complex structured data.