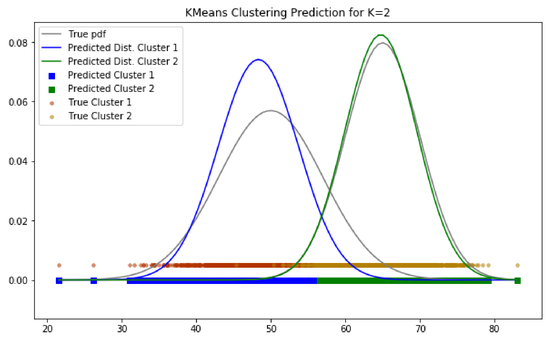
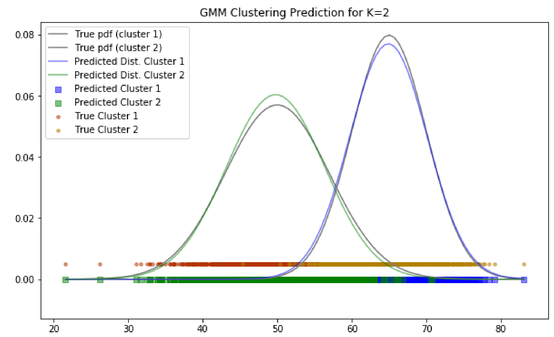
Comparative Study of Gaussian Mixtures and Clustering on Health Data

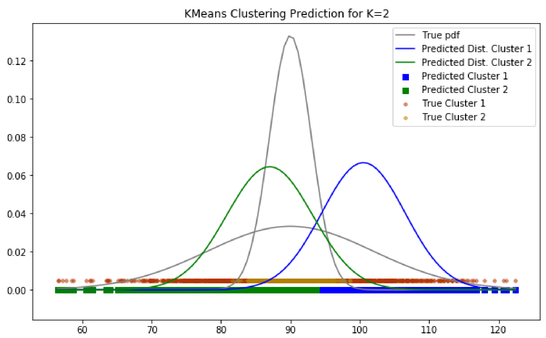
Researcher: Sarah Harkins Advisor: Dr. Ivan Dungan
Sarah.Harkins@g.fmarion.edu

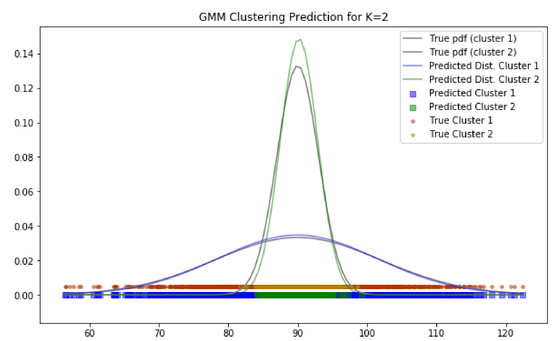
Abstract:

In this project, K-Means clustering and Gaussian Mixture Models (GMM) were compared on their abilities to decipher clusters in 1-dimensional data sets. The K-Means is a popular clustering algorithm, but there are aspects that it fails to capture that possibly leave the GMM to be the superior algorithm. The abilities of the algorithms were compared with a simulated data set. The algorithms were then further compared on a data set from the National Health and Nutrition Examination Survey. This data set includes 10,004 people ages 8 to 19 years old and their respective BMI. The goal of this application was to see if the algorithms would accurately discern between the group of males versus the group of females. GMM is typically used with higher-dimensional datasets. However, only one is used for this project, for example, features like BMI. The GMM delivers a more accurate distribution than K-Means due to the consideration of the standard deviation.









Bayes' Theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

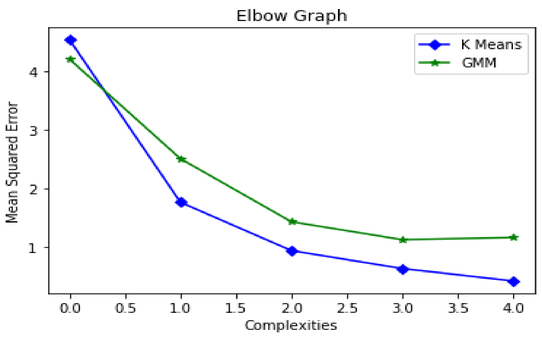
Clustering Methodology:

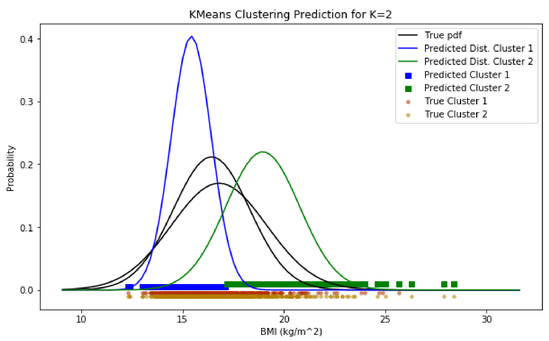
K Means: Hard clustering method considers the mean of the predicted cluster and stabilizes on a centroid
Gaussian Mixture Model: Soft clustering based on the normal distribution considers the means and standard deviation of the predicted cluster and stabilizes on a means and standard deviation using Bayes Theorem

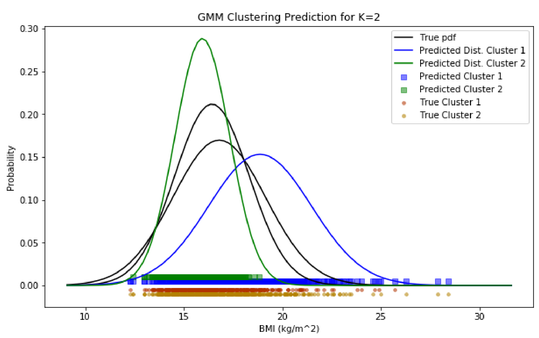
Background on Data Set:

This data set is from the National Health and Nutrition Examination Survey (NHANES) form the year 2015. here were 10,004 total individuals in the data set of ages ranging from 8 to 19 years old. The population in this study is 18-year-old males and females. This sample includes individuals and the data that describes their body mass index (BMI), height, and weight.

The elbow graph shows that two clusters is the optimal amount for this data set.







Simulation 1: Distinct Means

- Data set one → Mean 65, Std Dev 5
- Data set two →Mean 50, Std Dev 7

Simulation 2: Overlapping Means

- Data set one → Mean 90, Std Dev 3
- Data set two → Mean 90, Std Dev 12

Observations of Simulation 1:

- K means merely split the data set in half. This is almost as good as a 50-50 guess.
- GMM was very successful at determining the means and standard deviations of the data sets

Observations of Simulation 1:

- K means did a better job predicting the means and standard deviations of the given data sets compared to the previous simulation
- GMM was still able to make a stronger prediction at the means and standard deviations of the data sets

Conclusion:

The Gaussian Mixture Model more accurately depicts the true cluster curves compared the K Means clustering based on the resulting predicted distribution.

Future Work:

In a future project, I would like to reflect the work of the clustering done in the article *A review of machine learning in obesity* (by K.W. DeGregory, et al) but with the GMM method. This would involve scaling up to six dimension instead of one. The article originally uses k means clustering method so using the GMM method may glean different results. Additionally, comparing the computation time between the two algorithms and comparing the percent error of the predicted clusters.

Citations:

Sharma, P. (2020, October 18). *K means Clustering: K means clustering algorithm in Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/.

Singh, A. (2020, April 22). *Gaussian mixture models: Clustering algorithm python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/.