**ECE 8527: Introduction to  
Machine Learning and Pattern Recognition**

# HW No. 12: Clustering

For the Yin-Yang data of HW #3, cluster the data into N clusters using K-Means and a top-down clustering approach (Linde-Buso-Gray algorithm). Compare the two clustering approaches by measuring (1) the overall distortion (compute the mean-square error for each cluster) on the training data and (2) the classification error rate using a simple Euclidean distance criterion from the cluster centers.

Comment on and justify your findings.

In this assignment, we are using two different methods to clustering the data generated via the YinYang script provided in Homework Assignment 3. With clustering, each class will be categorized into a single point to represent the whole class. Both algorithms are obtained via pre-built packages and then compared to each other with both the distortion of the true mean and the error rate determined by the Euclidean distance calculation.

Depending on how separate the data set is, the clustering performance for both algorithms performed relatively as expected. When the overlap value is less than -0.2, which means that they are sufficient distinguishable between classes, both algorithms shows a decent clustering result of zero distortion from the true mean and about 0.5 % error rate when classifying the data via the clustered center. As two classes getting closer to each, the overlap value is getting bigger, the performance of the clustering drops significantly. The distortion for both algorithms increases to about 0.2 while the classification error rate spikes up to the region of 49.68 %. Since with the increase of the overlap value, two classes are becoming less distinguishable and based on the PCA, which also utilized the Euclidean distance to determine the error rate of the classification, the performance in classification is as expected especially in cases of overlap value that mix two classes together, such as the value of 0.4, where K-Mean scored an error rate of 49.64% while the LBG scored an error rate of 49.79%.

Taking the K-Mean clustering as am example. When classes are very distinguishable, the determination result, clustering result, can be viewed as follow. Since the majority of the class is centered at the head of the graph (the top part of the “,” ), classes are determined mainly based on that. As two classes are getting closer to each other, the tail starts to get incorrectly classed and when they are mixed together, classes are determined and clustered based on the head portion of the shape, thus generating errors when classifying the evaluation data. The following diagrams shows the progression of the clustering via K-Mean.

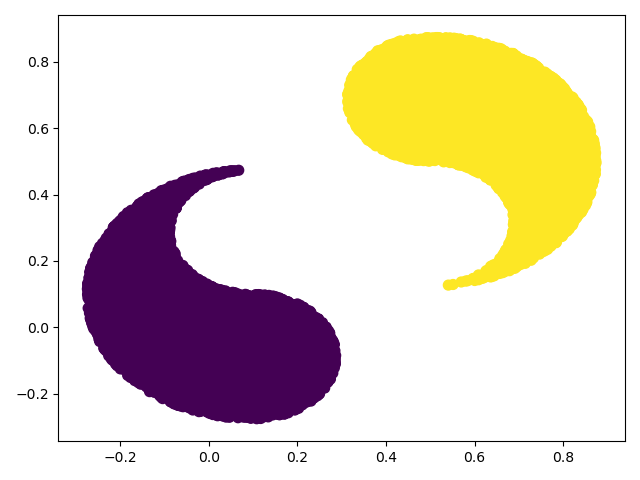


Figure K-Mean with Overlap of -0.8

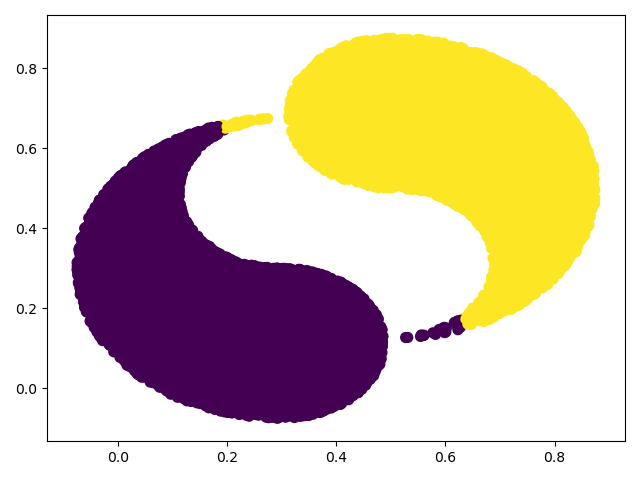


Figure K-Mean with Overlap of -0.4

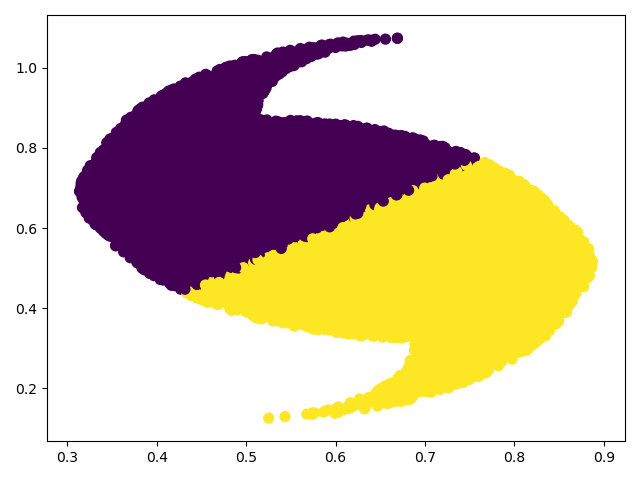


Figure K-Mean with Overlap of 0.6

In both algorithms, each class is clustered into one single cluster center in this assignment, while both algorithms are capable of clustering into multiple centers. Due to the special shape of the 2D YingYang data we are given, clustering them into one center per class is quite similar to obtaining the mean of the class and using circle to determine whether a data belongs to a class or the other. By increasing the number of clustering centers, it can help with determining the unique shape of the input data and increase the performance of the classification based on Euclidean distance while at the same time increase the classification complexity since multiple cluster center will be compute in order to generate an “educated” prediction. However, with multi-center clustering, it will perform more similar to KNN with the increase of cluster center. But with a slight increase of the number of clustering center, the algorithm will have a better performance as it shapes the data better and more efficient in comparison to KNN when the number of clustering center is relatively low.