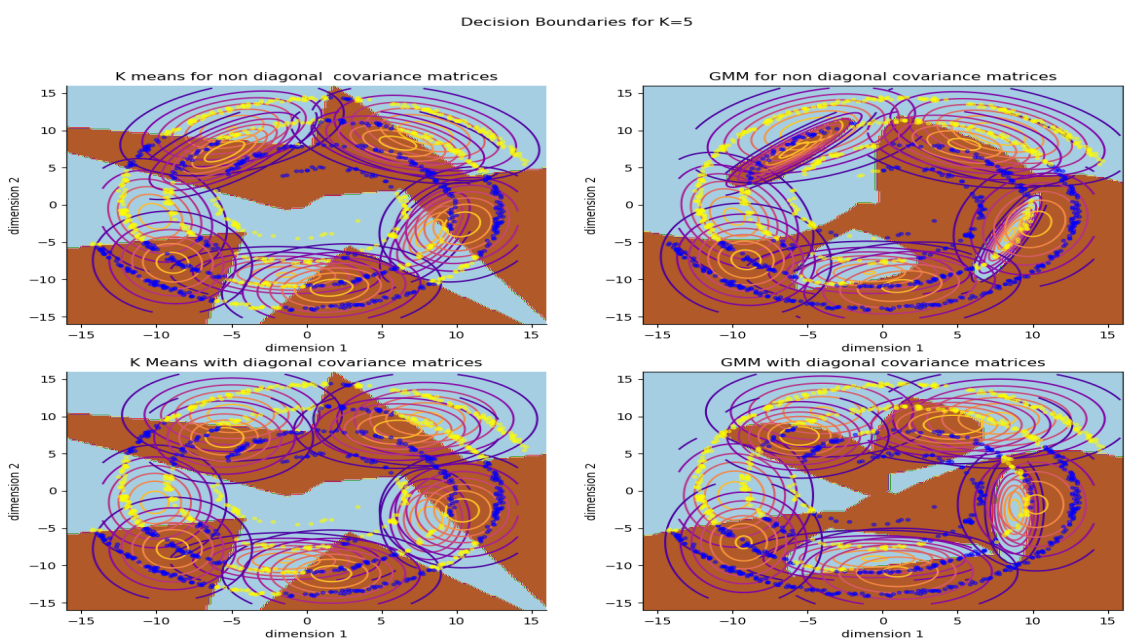
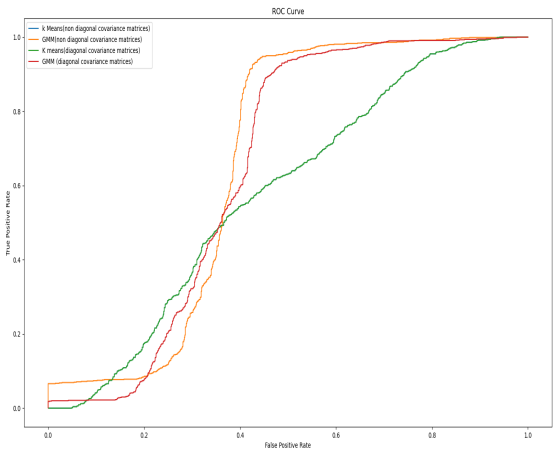
Synthetic Data Set

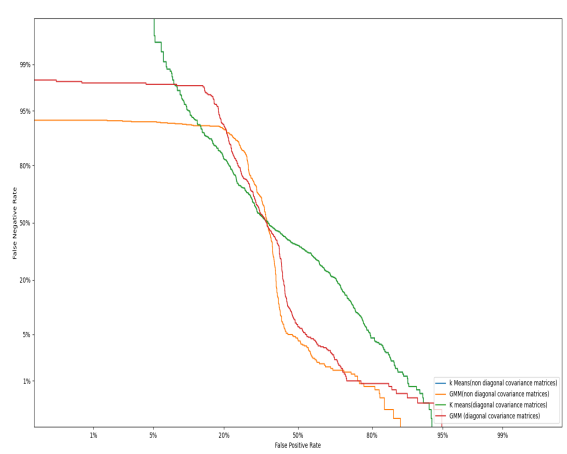
K Means And GMM:

we will see how accuracy of **K Means** and **GMM** varies as we wary the number of clusters in a given graph

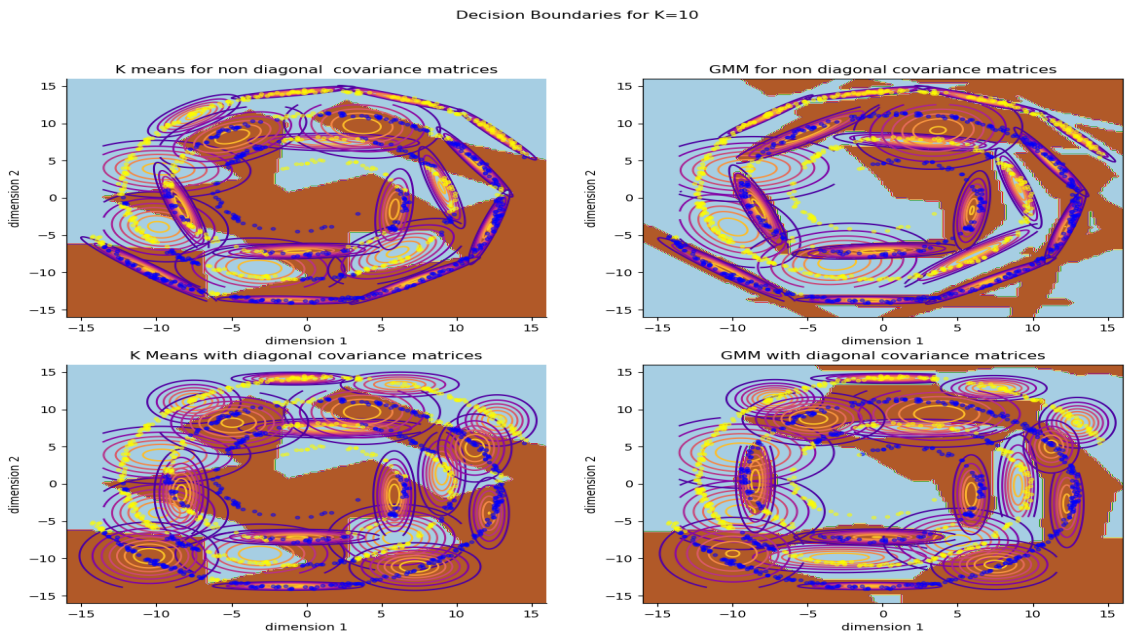
K=5:

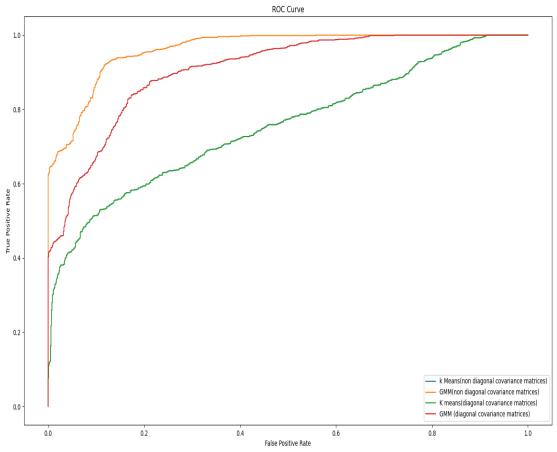


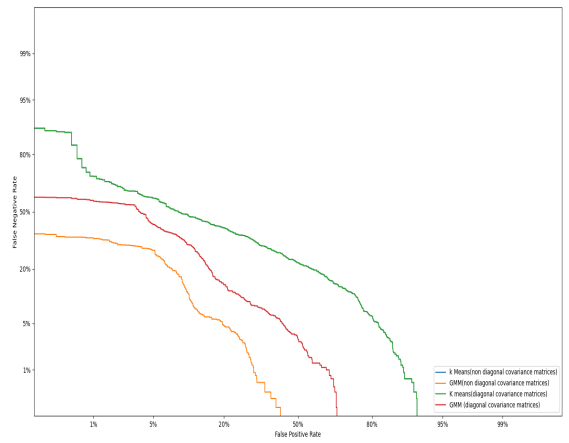




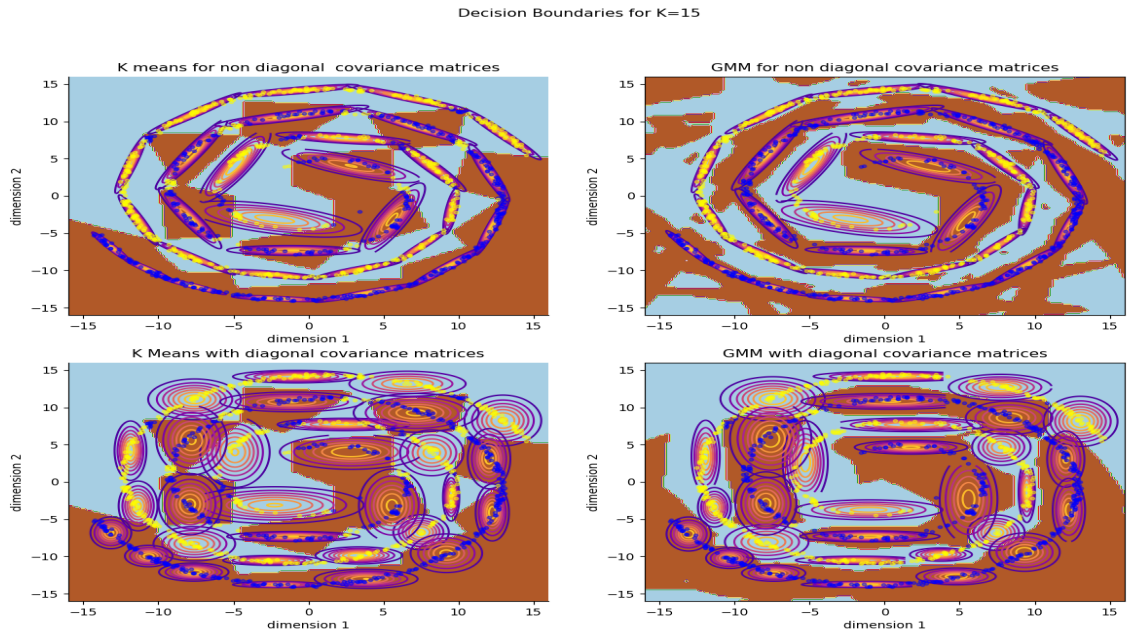
K=10:

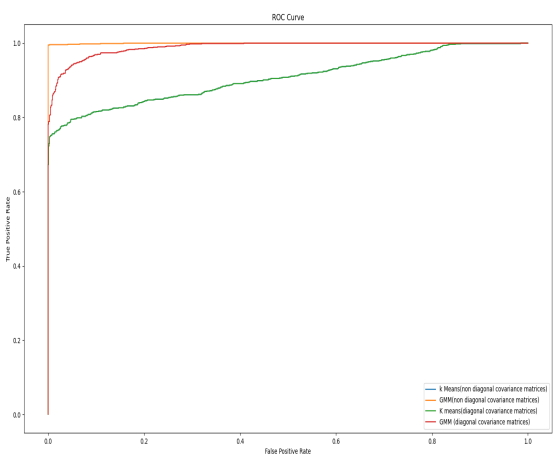


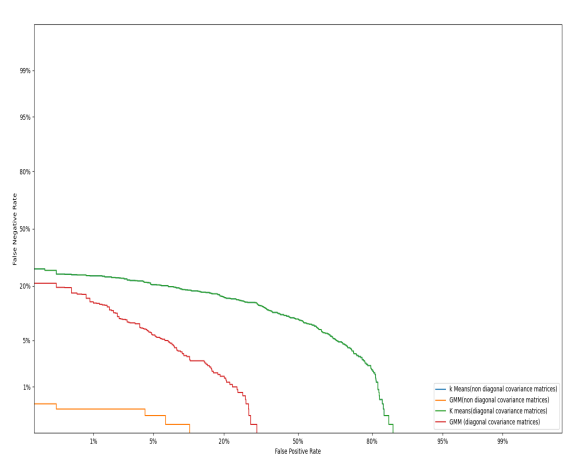

ROC Curve
True Positive Rate
False Positive Rate
k Means(non diagonal covariance matrices)
GMM(non diagonal covariance matrices)
K means(diagonal covariance matrices)
GMM (diagonal covariance matrices)


False Negative Rate
False Positive Rate
k Means(non diagonal covariance matrices)
GMM(non diagonal covariance matrices)
K means(diagonal covariance matrices)
GMM (diagonal covariance matrices)

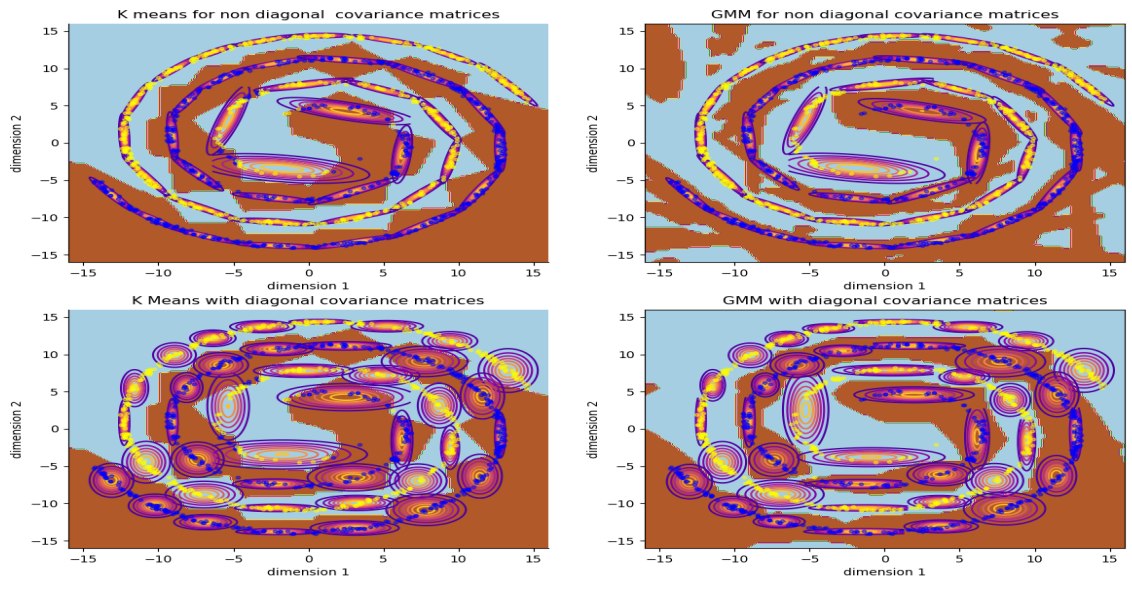
K=15:

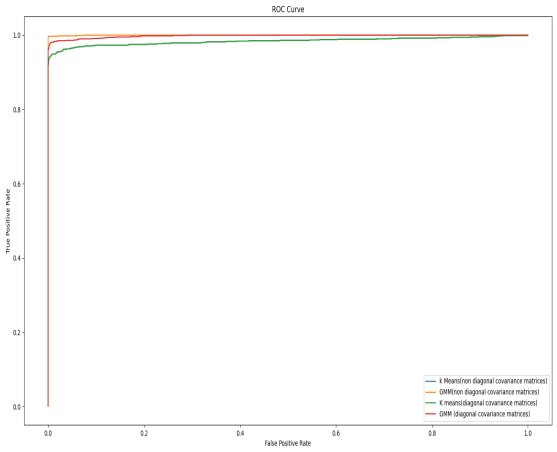

Decision Boundaries for K=15
K means for non diagonal covariance matrices
GMM for non diagonal covariance matrices
K Means with diagonal covariance matrices
GMM with diagonal covariance matrices
dimension 1
dimension 2

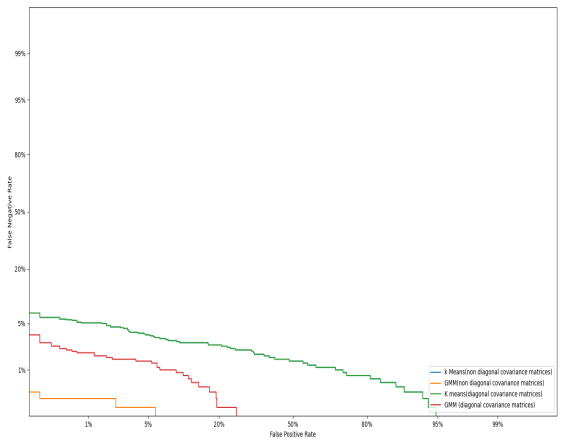

ROC Curve
True Positive Rate
False Positive Rate
k Means(non diagonal covariance matrices)
GMM(non diagonal covariance matrices)
K means(diagonal covariance matrices)
GMM (diagonal covariance matrices)


False Negative Rate
False Positive Rate
k Means(non diagonal covariance matrices)
GMM(non diagonal covariance matrices)
K means(diagonal covariance matrices)
GMM (diagonal covariance matrices)

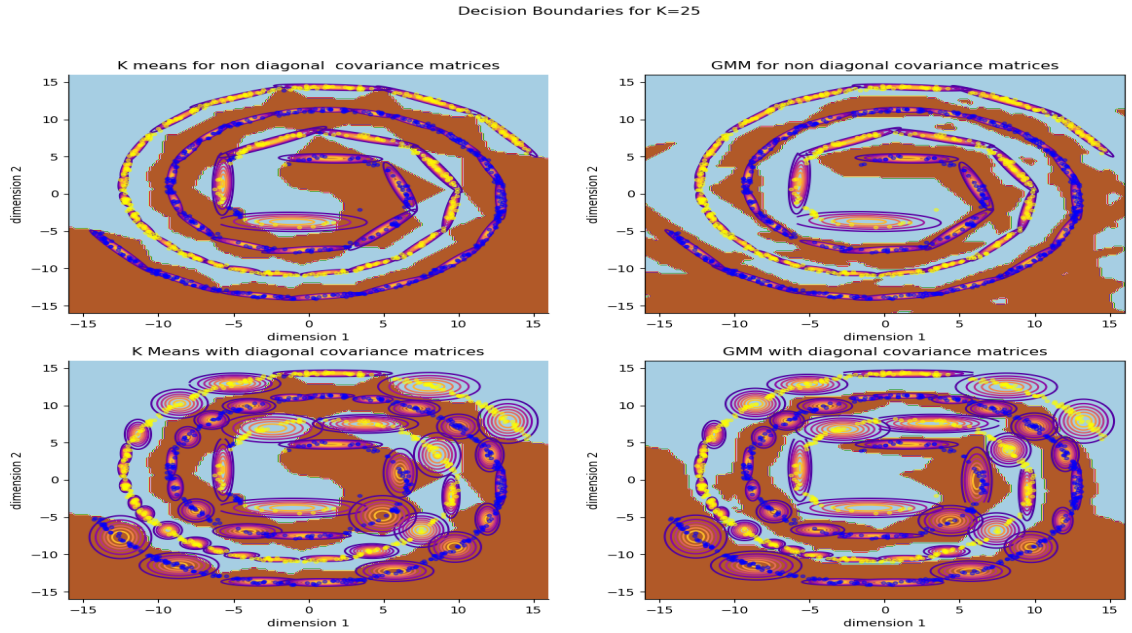
K=20:

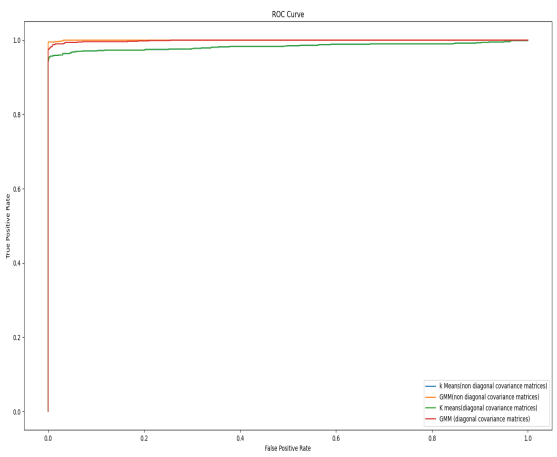

Decision Boundaries for K=20
K means for non diagonal covariance matrices
GMM for non diagonal covariance matrices
K Means with diagonal covariance matrices
GMM with diagonal covariance matrices
dimension 1
dimension 2

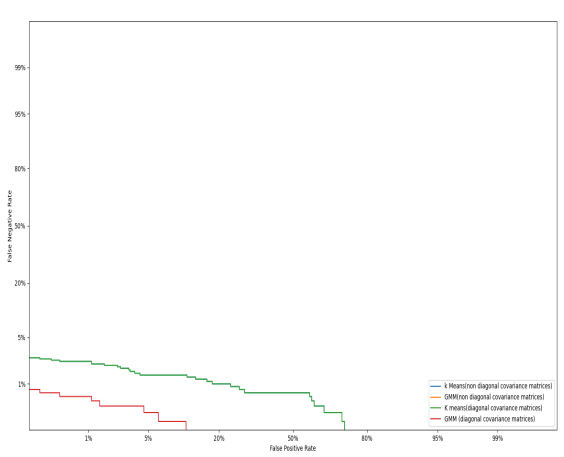




K=25:





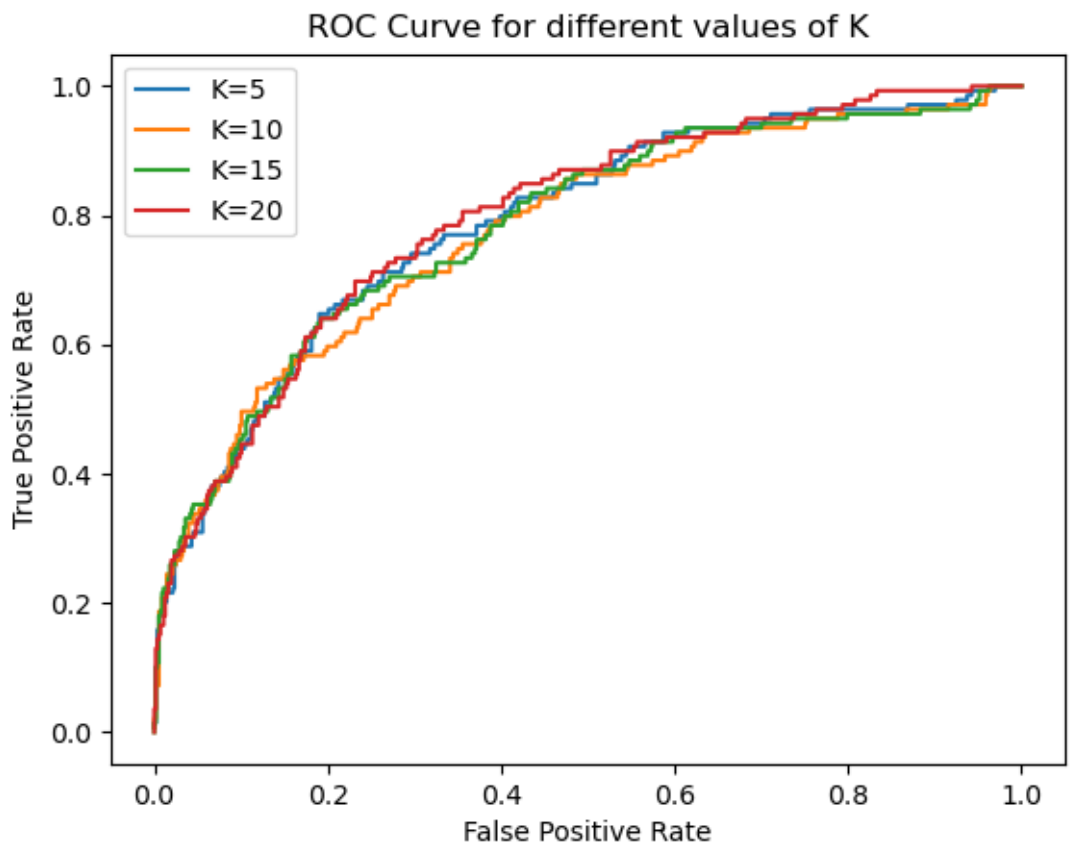


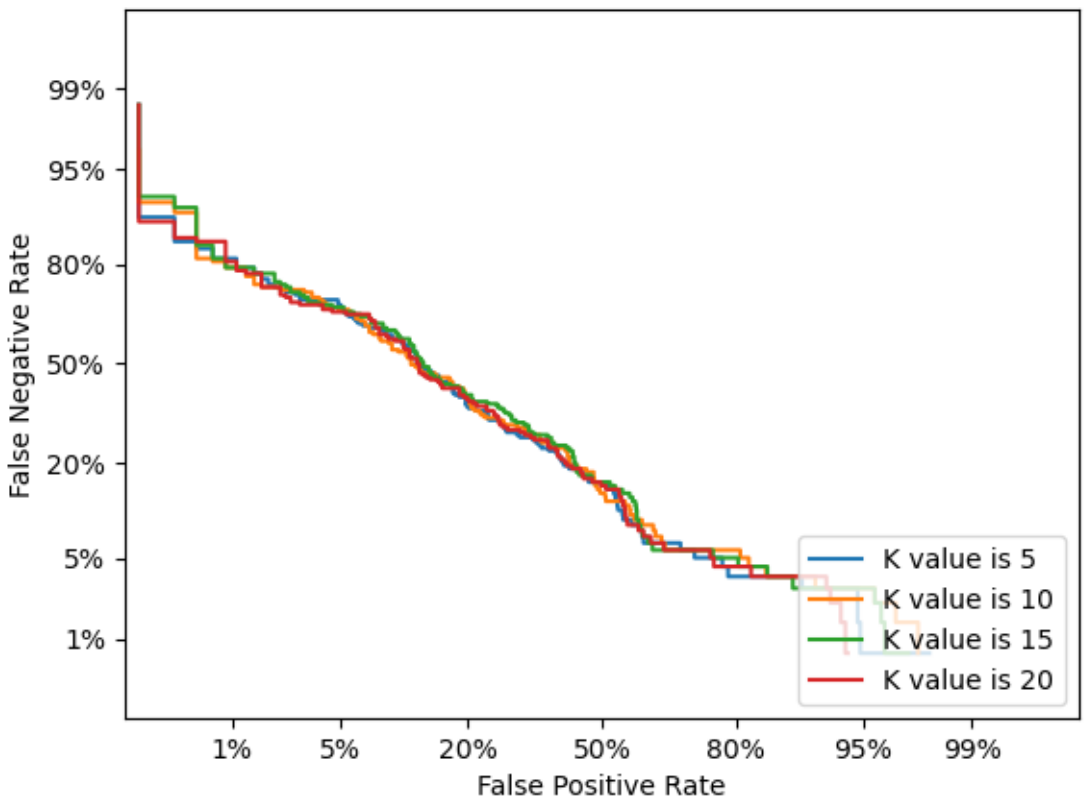
Inference:

1. As we can see that the accuracy of both K means and GMM increases as we increase the value of K as we can see in the graphs the ROC graphs and DET curves become more and more ideal.
2. The best classifier occurs when we have 25 clusters for each class
3. The number of iterations for K Means and GMM are around 10 and 3 respectively
4. The accuracy for GMM with non diagonal covariance matrices is more tham GMM with diagonal covarinace matrices.

Image Data Set

We first consider every vector of 23 length as a data point in a vector and taje all points form a data set as a class and do the K means and GMM for every dev data point we take every block of the 36 blocks and caclulate the probability of the each vector for every class and multiply the probablilities and classify the class with maxmimum product is classified as expected class. The accuracy of the algorithm is 64 percentage We will see how the ROC and DET curves come for different values of K for image set





DTW:

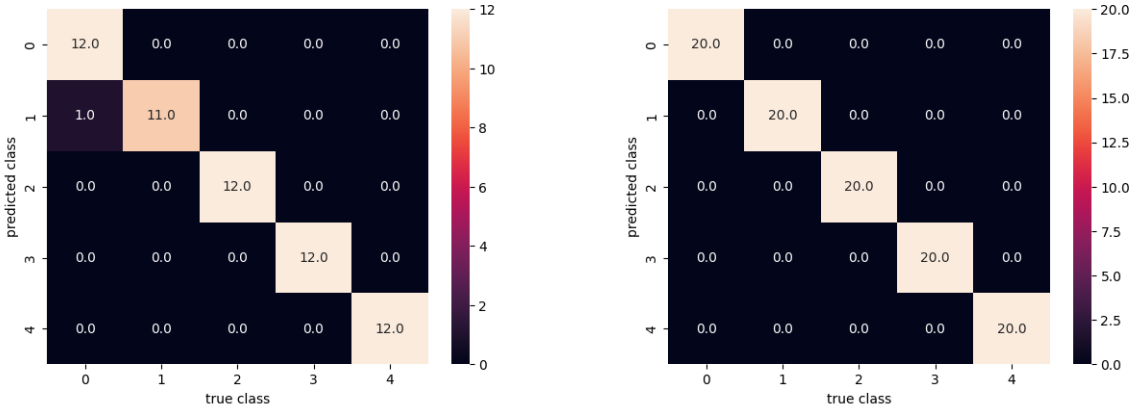
In DTW we are given two data sets one for spoken-digits and another for isolated hand-written data.

In this for both cases we will do as follows:

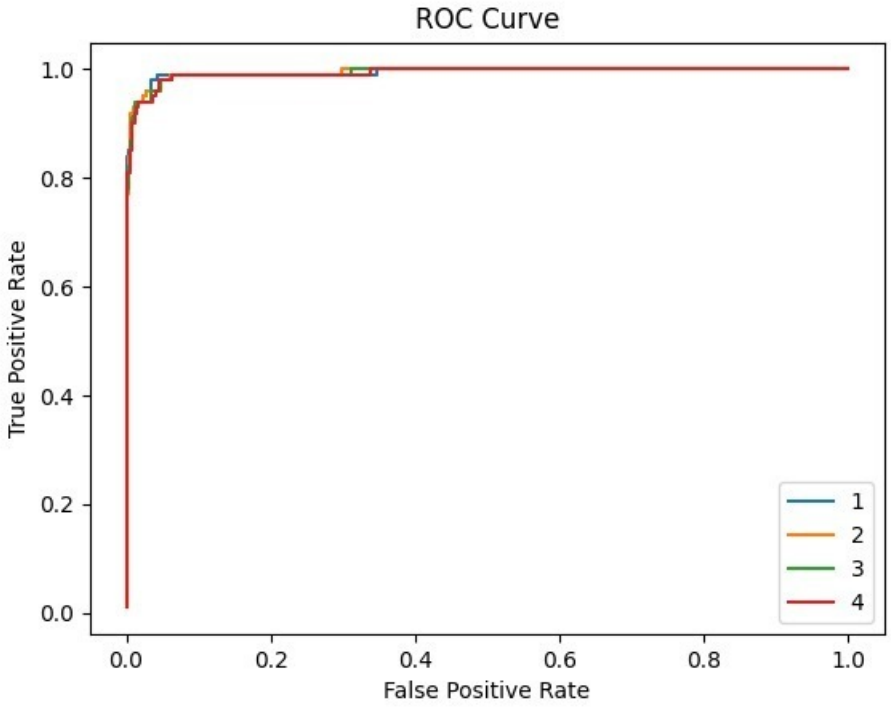
- For each development data point that is a feature vector in the given data we apply the DTW algorithm with each training data point's feature vector of a class, and pick average of the top most k(emperical) values from them and we do this for every class
- Now, we will assign the class which has the nearest distance from the above process.
- To plot the ROC and DET curves we will take inverse of these averages as the scores.

For the hand-written data-set we are given (x,y) co-ordinates and we need to normalize them by using the co-ordinates of the same file given.

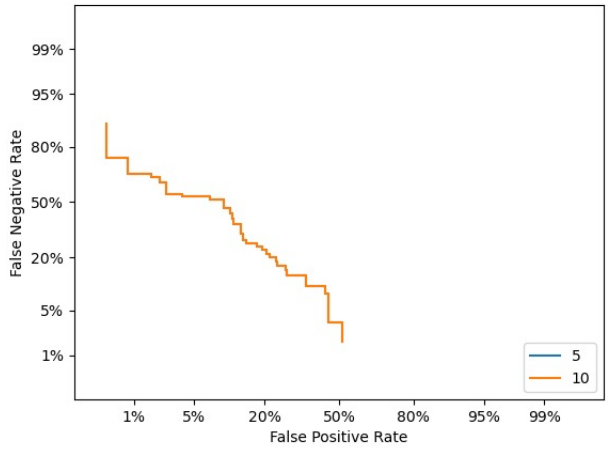
For both data-sets the confusion-matrices are as given below for k=5, that is picking top 5 values from each distance array obtained by implementing DTW algorithm on each class.

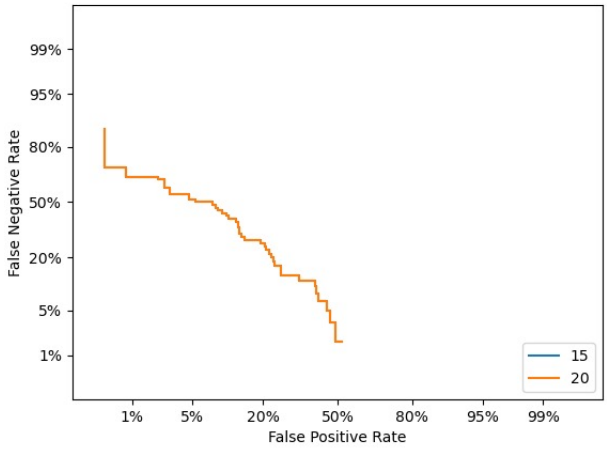


The ROC curve for Spoken-digit dataset for different values of K=[5,10,15,20] are as follows:

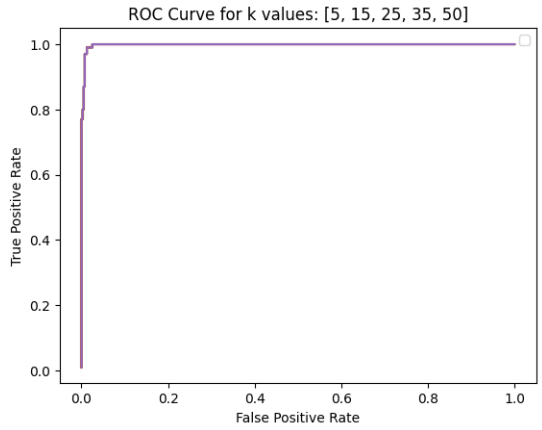


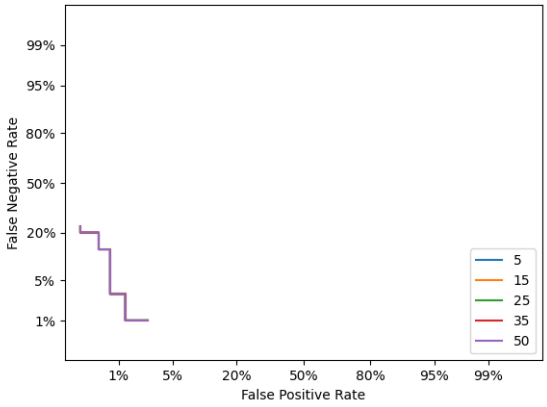
The DET curve for Spoken-digit dataset for different values of K=[5,10] and K=[15,20] are as follows:





The ROC and DET curves for Hand-written datset for different values of K-[5,15,25,35,50] are as follows:





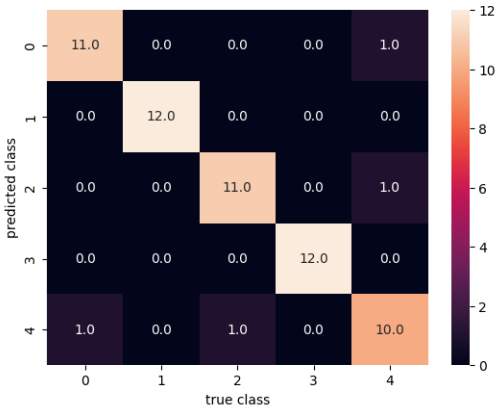
HMM:

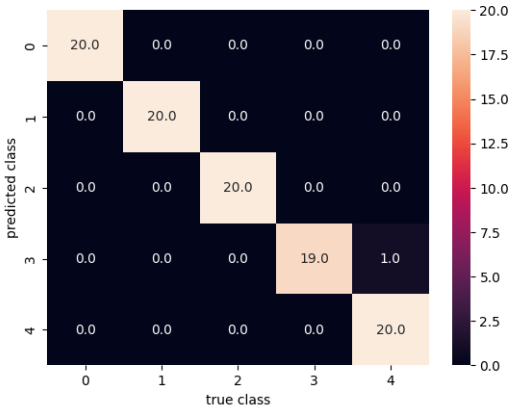
In HMM also we are given two data sets one for spoken-digits and another for isolated hand-written data.

In this for both cases we will do as follows:

- For each training data in the given data set we will assign a cluster which we find using K-means clustering for an emperical K.
- Now, we will choose an emperical number of states S and for each development data in the data set we will find the maximum probability of that data to be in the one of the given classes by dynamic left right probablity algorithm with the help of the state transitions which we find using the cluster preiction for the development data.

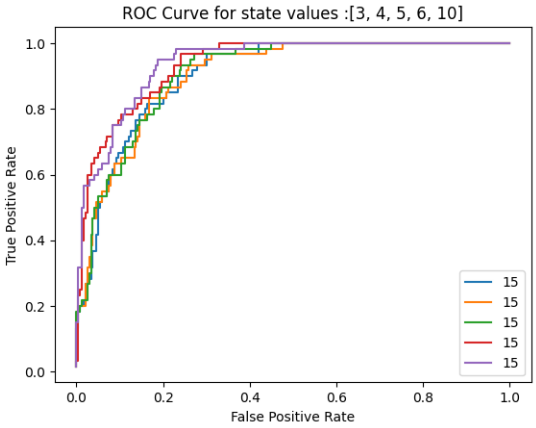
For, both data sets Spoken-digits and hand-written for no.of states=3 and no.of symbols/clusters=15 we will get the confusion matrices as below:

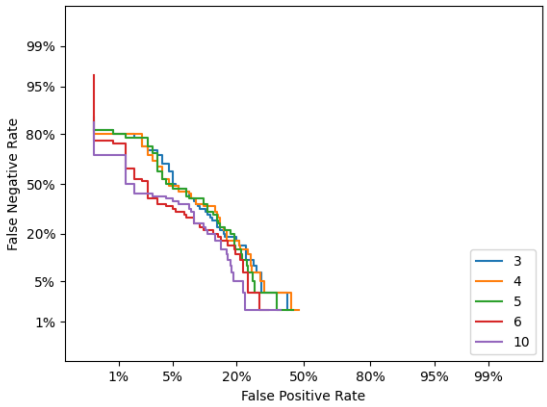




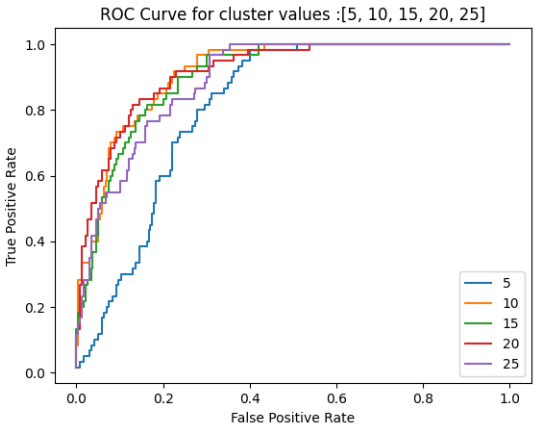
Spoken-digits :

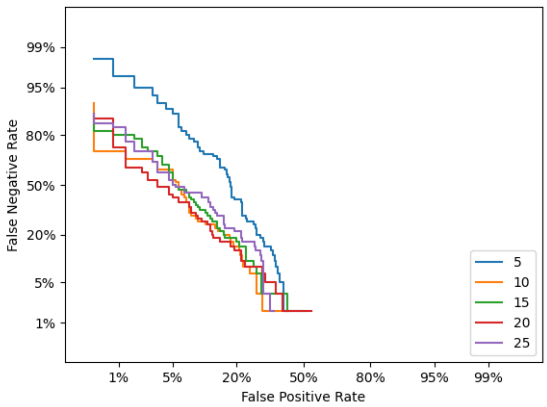
If we vary no.of states as [3,4,5,6,10] keeping no.of clusters K as constant 15, the ROC and DET curves are obtained as below:





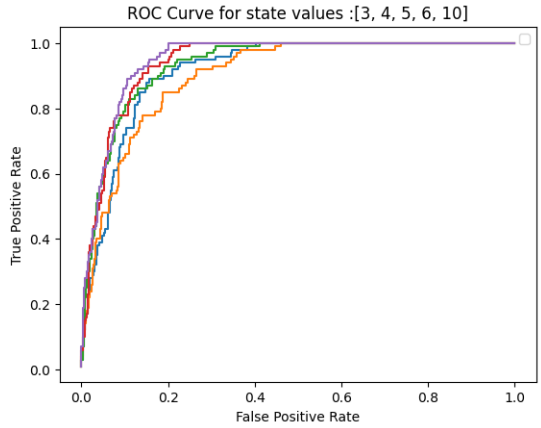
If we vary no.of clusters K as [5,10,15,20,25] keeping no.of states S as constant 3, the ROC and DET curves are obtained as below:

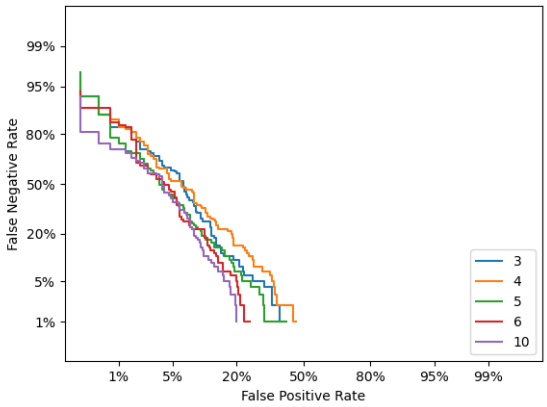




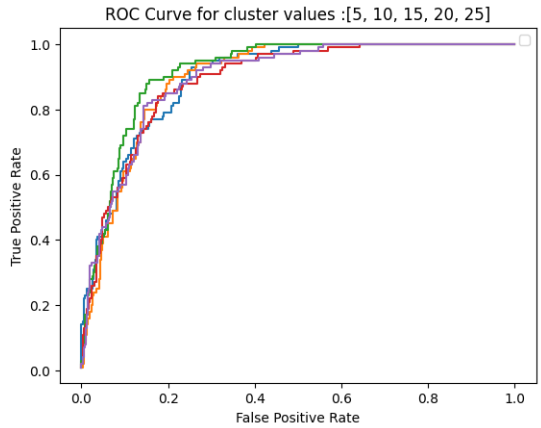
Hand-written :

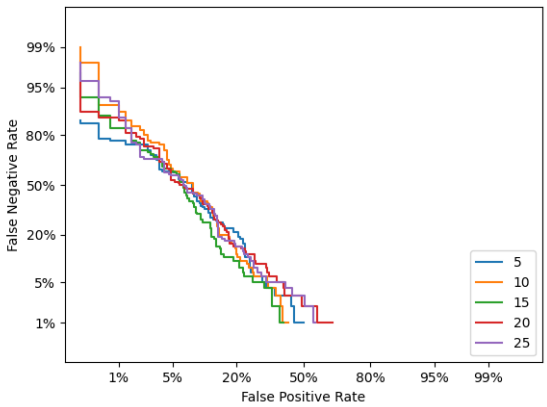
If we vary no.of states as [3,4,5,6,10] keeping no.of clusters K as constant 15, the ROC and DET curves are obtained as below:





If we vary no.of clusters K as [5,10,15,20,25] keeping states S as constant 3, the ROC and DET curves are obtained as below:





The best values of K,S are 15,3 for both data-sets.