1 Team Details

Group 19 - Pavan Ravishankar(**CS17S026**), Gouthaman KV(**CS16D018**)

2 K-Means vs Gaussian Mixture Models

K-Means cannot generate complicated decision boundaries. GMM can generate complicated decision boundaries as it assumes various templates within a class,thereby coming up with soft clustering technique.

3 Normalizing Data

As a rule of thumb, since K-Means involves distance calculation, **Z-norm** feature normalization is done per feature per class to avoid 1 feature dominating another feature in distance calculation.

4 Diagonal vs Full Covariance

Diagonal matrix cannot capture tilt in data unlike Full Covariance matrix. This leads to higher accuracy of a full covariance classifier compared to diagonal matrix which can be confirmed by ROC plot [1]. But classifying output using diagonal covariance is faster compared to full covariance.

5 Number of Mixtures

For **sythetic data**, approximate number of mixtures is first observed by looking at the data(which is atleast 5) as 5 different curvatures [1] involved for every class. One can confirm by measuring validation accuracy for different mixtures. In full covariance matrix, 8 mixtures gives accuracy > 0.9 and 30 mixtures gives 0.96 accuracy. In diagonal matrix, 25 mixtures is required to achieve 0.92 accuracy.

For **real data** since data cannot be visualized, choose a rough estimate of mixtures which gives close to minimum distortion which turns out to be 10 [3]. Then decide number of mixtures by varying mixtures around 10 and choose one based on validation accuracy. With 16 mixtures, diagonal matrix, validation accuracy on real data is close to 0.92. [4]

6 Importance of number of mixtures

More mixture means more intricate patterns are being captured by the model. In most situations, increasing the number of mixtures should make decision

making more accurate. Look at the performance of 10 mixture diagonal covariance classifier and 5 mixture diagonal covariance classifier on real data in ROC plot [3]. There might be few situations where accuracy might dip too[1]. Synthetic data set has 0.82 accuracy on 4 mixtures and 0.72 on 5 mixtures which could have been due to boundary points misclassification when mixtures were increased.

7 Number of EM iterations

EM converges really fast when K-Means is performed initially to determine mixtures. One can observe from plot that accuracy on real data is stabilizing after 4 iterations which indirectly indicates that log likelihood is stabilizing with just 5 iterations[3].

8 Performance of classifier

Confusion matrices, ROC and DET plots for various classifiers are given below. Every vector is classified into a particular class which gives maximum likelihood for that vector. While dealing with real data, result of all 36 vectors that comprise the image is taken into consideration and the class that occurs most of the times is class of the image. Probit scaling is used for plotting DET curves and hence negative values are present on x-axis and y-axis.

9 Code Optimizations

Working with real data is time consuming as data size increases. Even though numpy gives matrix multiplcation refrain from using it unless required. For example in finding PDF of normal distribution, exponent calculation is $e^{\frac{Data \cdot \Sigma^{-1} \cdot Data^T}{2}}$ which takes $O(m \cdot n \cdot n + m \cdot n \cdot m)$. If $e^{\frac{Data[i,:] \cdot \Sigma^{-1} \cdot Data^T[:,i]}{2}}$ is used time taken is $O(m \cdot n \cdot n + m \cdot n)$ where Data is (m, n) matrix. Calculate determinant and inverse of covariance matrix beforehand than computing it every time in equation. There were overflow issues when greater than 20 mixtures were used for real data but since desired accuracy was already acheived we have not looked into solving this issue.

10 Real Data Result

We are unsure about the performance of full covariance matrix on real data. For any given mixture, EM iterations and on any data, accuracy is 100%

11 Diagrams and Tables

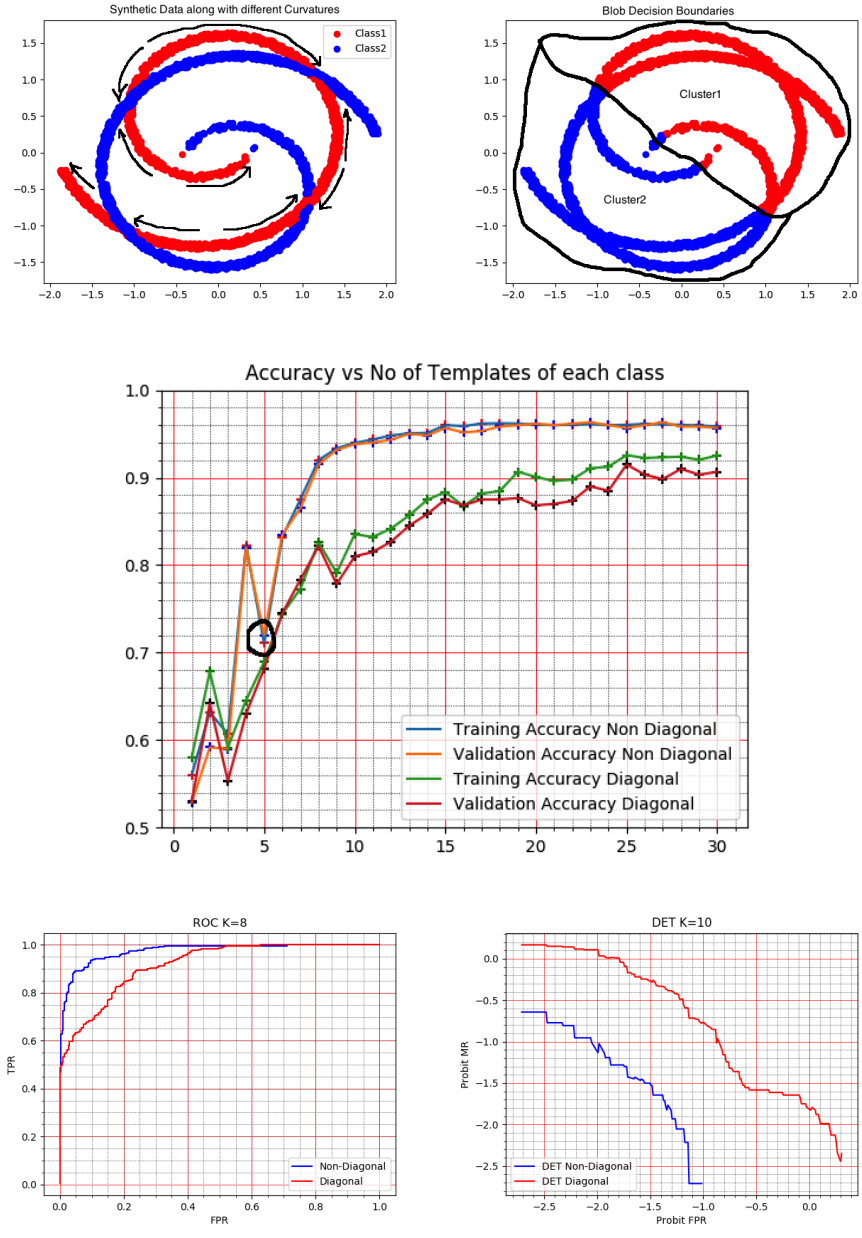


Figure 1: Synthetic Data Plots

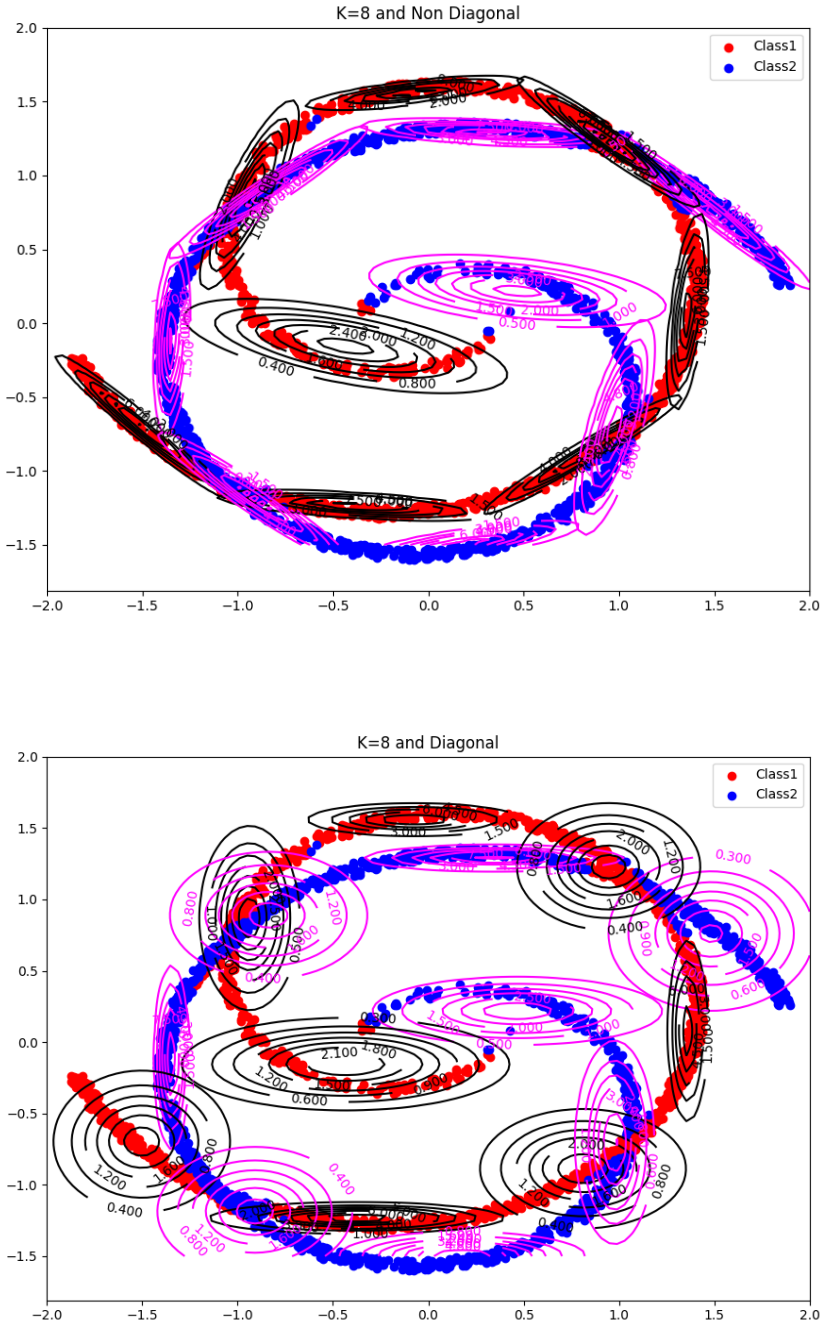


Figure 2: Contour plots of Gaussian

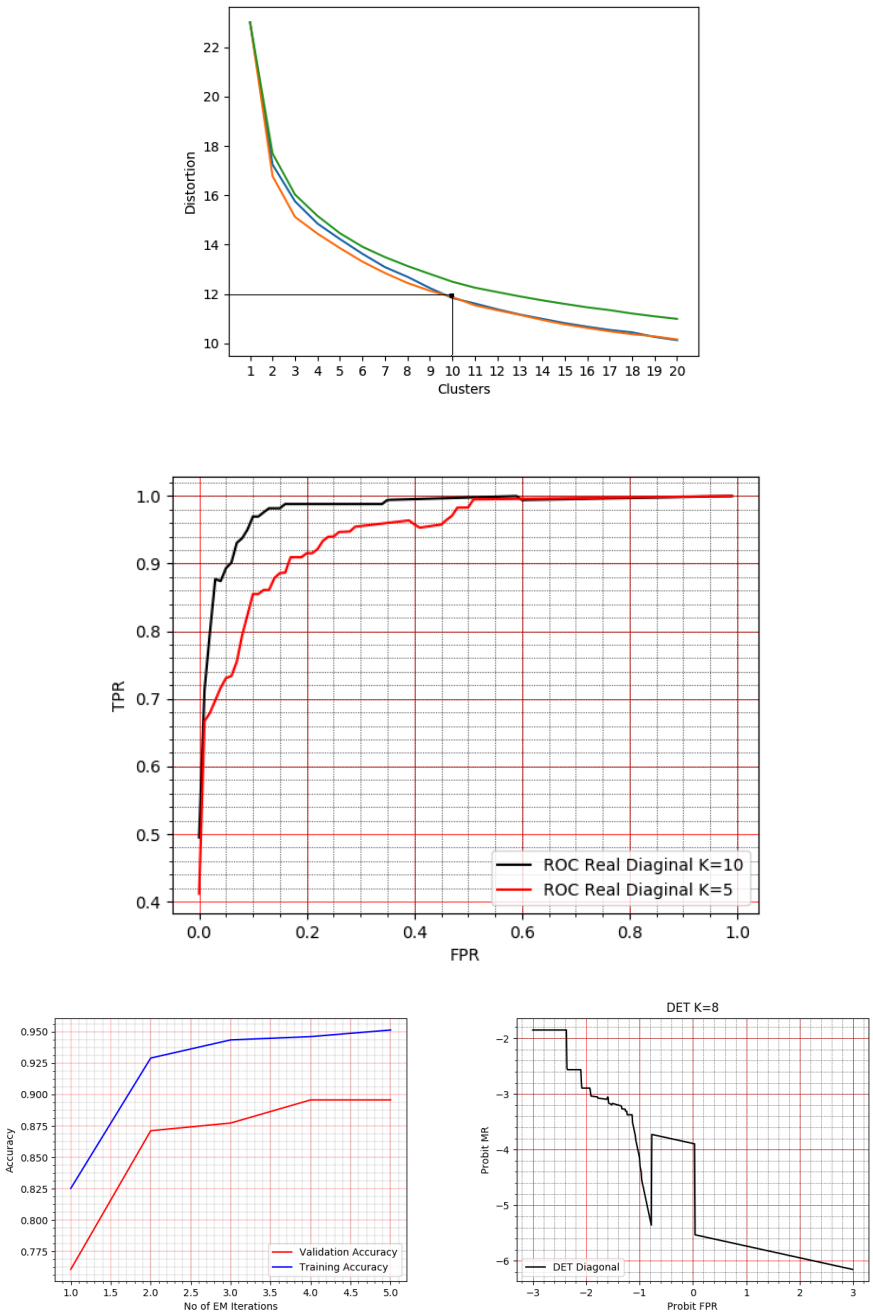
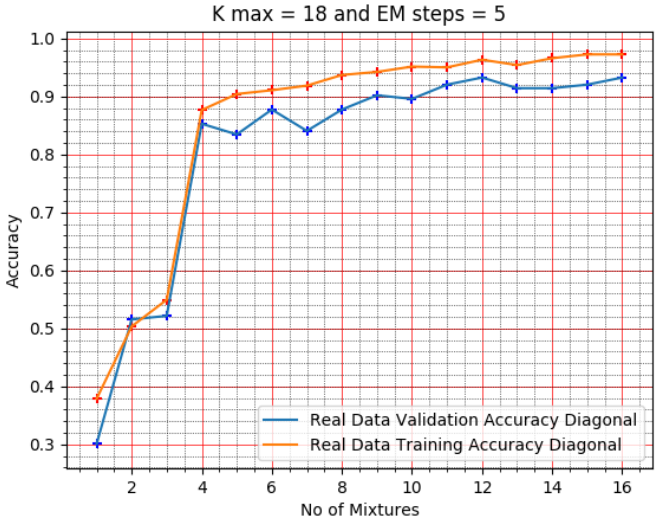
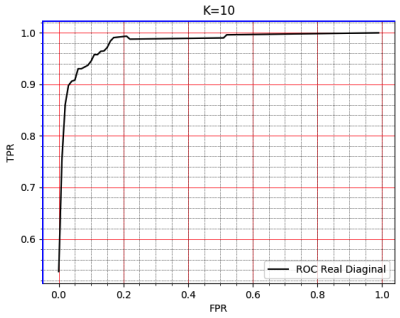


Figure 3: Real Data plots





Input	Prediction	
	C1	C2
C1	243	57
C2	57	243

Synthetic Diagonal
(K=8)

Input	Prediction	
	C1	C2
C1	286	14
C2	23	277

Synthetic Non Diagonal
(K=8)

Input	Prediction		
	C1	C2	C3
C1	53	1	0
C2	0	43	13
C3	0	3	50

Real Diagonal
(K=10)

Figure 4: Confusion Matrices