ML Assignment 1

Harshvardhan Kalra
2014043

Classification using K-means:

Dataset Name	K = 2				K = actual value				K = 12			
	MI	AMI	RI	ARI	MI	AMI	RI	ARI	MI	AMI	RI	ARI
Segmentation(7)	0.39	0.18	0.41	0.10	0.50	0.45	0.81	0.33	0.60	0.52	0.87	0.40
Iris(3)	0.67	0.51	0.76	0.53	0.70	0.67	0.82	0.63	0.62	0.41	0.75	0.34
Seeds(3)	0.55	0.42	0.72	0.46	0.69	0.69	0.87	0.71	0.54	0.35	0.73	0.28
Vertebral(3)	0.42	0.33	0.64	0.29	0.41	0.40	0.67	0.32	0.41	0.26	0.67	0.18

Quantitative results:
1) The segmentation dataset should give best performance with 7 clusters. Performance for 12 clusters is coming out to be slightly better, probably because of the erratic distribution of data in the ground truth.
2) The iris dataset gives best performance for 3 clusters. The performance for 2 clusters is better than 12 clusters, due to the absolute difference being less in the case of 2.
3) The seeds dataset gives best performance for 3 clusters. The performance for 2 clusters is slightly better than 12 clusters, due to the absolute difference being less in the case of 2.
4) The vertebral dataset gives best performance for 3 clusters. The performance for 2 clusters is similar to 12 clusters, probably due to the distribution of data in the ground truth.

Qualitative results:
1) In general, the seeds and iris datasets were well behaved, giving consistent results with the clustering observed from the ground truth and from the labelling obtained using K-means.
2) The segmentation dataset was slightly less so, giving poorer results though still indicating the clustering was behaving as expected in most iterations.
3) The vertebral dataset was segmented and hence did not give very good results on multiple iterations. This is evident given the plot of the ground truth, indicating that the dataset itself suffered from poor clustering to begin with.

Plots named *_pre.png are ground truth plots, while those named *_post.png are plots obtained from the clustered labels from K-means.

Classification using GMM:

Dataset Name	K = 2				K = actual value				K = 12			
	MI	AMI	RI	ARI	MI	AMI	RI	ARI	MI	AMI	RI	ARI
Segmentation(7)	0.04	0.00	0.41	0.00	0.65	0.59	0.85	0.48	0.61	0.54	0.86	0.39
Iris(3)	0.76	0.57	0.77	0.57	0.78	0.77	0.89	0.76	0.62	0.40	0.75	0.31
Seeds(3)	0.59	0.45	0.74	0.49	0.68	0.67	0.85	0.68	0.52	0.34	0.73	0.26
Vertebral(3)	0.37	0.30	0.70	0.41	0.36	0.29	0.70	0.41	0.30	0.19	0.64	0.11

Quantitative results:
Apart from a few extreme case, where GMM performs very badly, the overall performance is better than K-means, especially for the iris dataset.

Qualitative results:
The qualitative estimates for each dataset are as follows:
a) Seeds has slightly better clustering than K-means.
b) Iris has comparable clustering
c) Vertebral and segmentation result in poor clusterings with only a single point being assigned to a cluster.

Plots named *_gmm.png are GMM plots with actual values as K.