ML : Assignment 1

Report (with Bonus)

Anshuman Suri (2014021)

Classification using K-Means

Dataset Name	k=2				k=real value				k=12			
	MI	AMI	RI	ARI	MI	AMI	RI	ARI	MI	AMI	RI	ARI
Iris (3)	0.68	0.52	0.76	0.54	0.72	0.70	0.83	0.66	0.63	0.41	0.76	0.33
Segmentation (7)	0.35	0.18	0.49	0.11	0.52	0.47	0.84	0.34	0.58	0.50	0.87	0.37
Seeds (3)	0.55	0.43	0.73	0.47	0.70	0.69	0.87	0.71	0.54	0.35	0.74	0.29
Vertebral (3)	0.42	0.33	0.64	0.29	0.42	0.40	0.68	0.32	0.40	0.26	0.67	0.19

Based on **quantitative** results:

- The *iris* data-set has best performance on *k* = *real value* (3). As 2 is closer to 3, the performance on *k=2* is better than *k=12*.
- The *segmentation* data-set has best performance on *k* = *real value* (7). 12 and 2 are equi-different fro 7, so performance should be comparable. However, 12 performs better than 2. This may be because the way is distributed.
- The *seeds* data-set has best performance on *k* = *real value* (3). As 2 is closer to 3, the performance on *k=2* is better than *k=12*.
- The *iris* data-set has best performance on *k* = *real value* (3). As 2 is closer to 3, the performance on *k=2* is better than *k=12* (in fact, MI is the same for *k=2* and *k=3*).

The performance for clustering is best when the chosen value of *k* is closest to the actual number of clusters in all of the cases. Every metric for each of the 4 data-sets is closer to 1, as compared to the cases of *k=2* and *k=12*.

Based on **qualitative** plots and graphs:

- The *iris* data-set has been grouped very close to the actual labels. The error function has also stagnated after 5 runs. This can be considered as a successful realization for k-means for this set of data.
- The *seeds* data-set seems to be well clustered. Based on the *t-SNE* visualization, the clustering seems to be in accordance with the actual data-labels.

- The *segmentation* data-set doesn't seem to be as well clustered as the others. Based on the *t-SNE* visualization, the original clusters do not segregate the data well. The k-means based clustering seems to do a better job, but has a few outliers here and there. In all, it produces better clusters than the actual labels.
- The *vertebral* data-set seems to be well clustered. Based on the *t-SNE* visualization, the original clusters do not segregate the data well, with mixing in two categories. The k-means based clustering seems to do a great job if we look at it without the original labels. However, the clusters produced do not seem to entirely match with the actual data.

Classification using Mixture of Gaussians

Dataset Name	k=2				k=real value				k=12			
	MI	AMI	RI	ARI	MI	AMI	RI	ARI	MI	AMI	RI	ARI
Iris (3)	0.76	0.57	0.77	0.57	0.79	0.78	0.89	0.76	0.62	0.40	0.75	0.31
Segmentation (7)	0.04	0.0	0.15	0.0	0.4	0.23	0.57	0.18	0.61	0.54	0.86	0.40
Seeds (3)	0.59	0.45	0.74	0.50	0.68	0.67	0.86	0.68	0.53	0.34	0.74	0.26
Vertebral (3)	0.37	0.30	0.71	0.41	0.36	0.30	0.71	0.41	0.31	0.20	0.65	0.12

Based on **quantitative** results:

- The performance for clustering using mixture of Gaussians seems to be better overall. For some cases, the metrics are considerably better than the case of k-means (for example, the *iris* and *vertebral* data-sets) . However , in the other 2 cases, it seems to be far worse, with scores as low as 0 for some of them. In all, the performance is slghtly better, if we look at the best metrics per class. In all, it has comparable performance.

Based on **qualitative** plots and graphs:

- For *iris*, clustering is comparable to k-means.
- For *seeds*, clustering is comparable, with the k-means clustering seem to perform better.
- For *segmentation*, k-means seems to perform better.
- For *vertebral*, performance is horrible for GMM, with the third cluster almost being non-existent (only 1-2 points).

All plots are contained in *./plots* folder (they weren't fitting nicely here)