COMP7703 – Homework Task 4
Joel Thomas 44793203

1.

The formula used by Bishop for the image compression example in Fig. 9.3 is given by:

$24K + N\log_2(K)$

$K = \text{number of classes}$

$N = \text{number of pixels (samples in dataset)}$

The first expression $24K$ is the total number of bits required to transmit the K code book vectors $\boldsymbol{\mu_k}$ $(8 \times 3 \times K)$. The second expression $N\log_2(K)$ is the total number of bits to transmit the identity of the nearest vector μ_k for K vectors per pixel ($\log_2(K)$), over the total number of pixels ($N\log_2(K)$).

Thus, for $K = 6$ classes, assuming the original images in Fig. 9.3 have $240 \times 180 = N = 43200$ pixels each:

$24 \times 6 + 43200\log_2(6) \approx 111815$ bits are required to transmit one of his images. The associated compression ratio compared to the original image is:

$$\frac{24K + N\log_2(K)}{24N} = \frac{24 \times 6 + 43200\log_2(6)}{24 \times 43200} \approx 0.1078 = 10.78\%$$

2.

If we keep the total number of clusters K the same and simply repeat the experiment with different initial cluster centres, we expect the algorithm to still converge and hence follow the same shape of the error curve (regardless if it gets stuck in a local minimum). This is a reasonable guess since we know that the algorithm converged successfully in the previous experiment (same number of clusters).

$$E\left(\{\boldsymbol{m}\}_{i=1}^{k}\middle|\mathbb{X}\right) = \sum_t \sum_i b_i^t \left|\left|\boldsymbol{x}^t - \boldsymbol{m}_i\right|\right|$$

$$b_i^t = \begin{cases} 1, & \text{if } \left|\left|\boldsymbol{x}^t - \boldsymbol{m}_i\right|\right| = \min_j \left|\left|\boldsymbol{x}^t - \boldsymbol{m}_j\right|\right| \\ 0, & \text{otherwise} \end{cases}$$

There is no reason to believe the new curve would start/end at the same points because as evident from the above formulas, the new cluster centres $\{\boldsymbol{m}\}_{i=1}^{k}$ appear in both the indicator function b_i^t as well as the objective function $E \rightarrow$ different $\{\boldsymbol{m}\}_{i=1}^{k}$ guarantee that the starting value for E will be different. Since there is a chance that starting with different initial cluster centres can yield a different local optimum after convergence → the ending value for E is also highly likely to be different.

3.

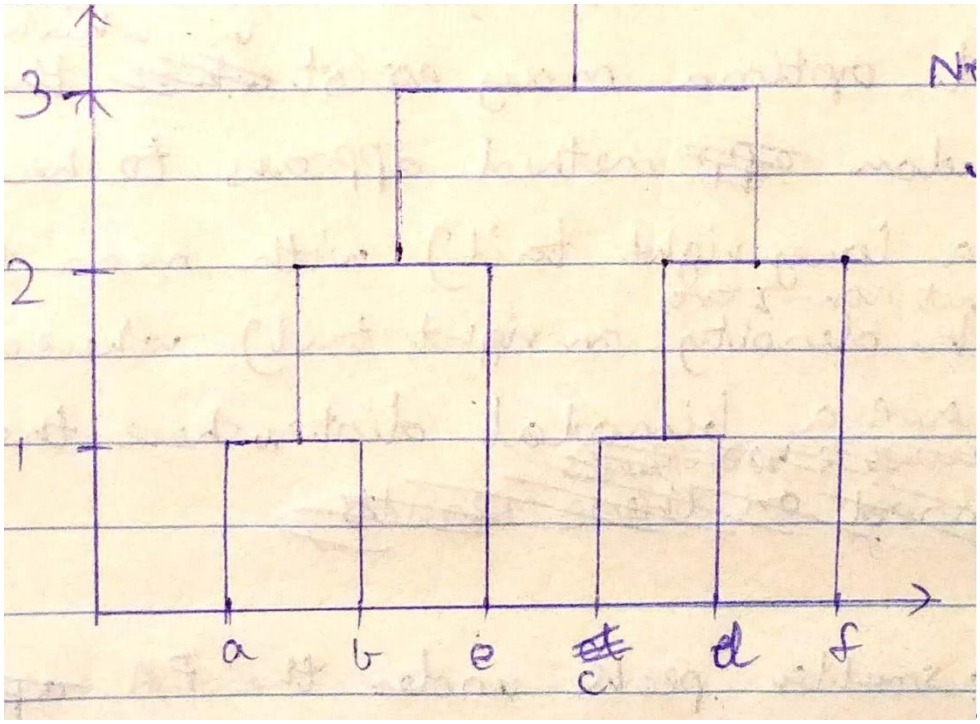

Figure 1: New dendrogram using city-block (Manhattan) distance metric

As evident in the figure above, the key difference is that the height at which groups $\{a, b\}$ and $\{e\}$ merge is at $1 + 1 = 2$ (horizontal + vertical distance from a to e) according to the city-block (Manhattan) distance metric instead of $\sqrt{1^2 + 1^2} = \sqrt{2} \approx 1.414$ (diagonal distance from a to e) which is seen in the original dendrogram using the Euclidean distance metric.

4.

1) Row 1 – this dataset is an example of non-spherical true clusters. Algorithms that make assumptions about clusters being spherical (points concentrated inside a circular shape) e.g. k-means will be unable to accurately identify the true clusters.
2) Row 3 – this dataset has the true clusters located very close to each other and there is also varying variance/dispersion around each of the clusters e.g. bottom left cluster has much smaller variance compared to middle cluster which is clearly more dispersed in terms of the density of the points. This makes it more challenging in general for algorithms to identify and decide which data points to include as a cluster.
3) Row 5 – this dataset is an example of classic spherical data where each cluster is spherical and is located far away from the other clusters (i.e. easily identifiable) together with small variance/points densely packed together. It is the easiest (least difficult) type of dataset to use clustering algorithms on as evident by every featured algorithm correctly identifying the three separate clusters.

5.

First note that depending on the initialisation method used in the experiment performed on the Iris dataset, a different number of local optima exist in which the algorithm converges into. The RANDOM method appears to have a right skewed distribution with mean error ≈ 57.69 (note mean $>$ median $>$ mode for such distributions) whereas the Forgy Approach method appears to have a bimodal distribution where the second peak sits at an error of 72.5 (considerably higher than 57.69). Since the second smaller peak under the FA method is a local optimum that occurs roughly 400 times in the experiment, this is evidence that this technique gets stuck in the local optimum very commonly and is hence undesirable. Comparing this to the RANDOM method, although this technique has a very low but non-zero probability of getting caught in local optima (evident by the far right tail), it is able to reach an optima reasonably close to the global optimum most of the time (i.e. with high probability) so the RANDOM method appears more favourable than the FA method based solely on Figure 5 of the paper.