COMP4702/COMP7703/DATA7703 - Machine Learning
Homework 4 - Clustering
Solutions

Marcus Gallagher

Core Questions

1. k-means clustering can be viewed as a method for compression via a set of codebook vectors. In the book [Bishop 2006], he discusses the compression achieved for the images shown in his Fig. 9.3 (see also this week's lecture). Using the same calculation method, how many bits would be required to transmit one of his images for ($K = 6$) and what is the corresponding compression ratio?

   Answer: from [Bishop 2006], $N = 43200$ and the total number of bits required is given by $24K + N \log_2 K$ (rounding up to the nearest integer). $\log_2 6 \approx 2.58$, but we must round up because we need a integer number of bits per pixel: $\lceil \log_2 6 \rceil = 3$.

   $$24K + N \log_2 K = 24 \times 6 + 43200 * 3 = 129744$$

   Compression ratio (given the original image required 1036800 bits):

   $$\frac{129744}{1036800} = 12.5\%$$

2. The k-means algorithm performs optimisation on an objective function which Alpaydin calls the "reconstruction error". [Bishop 2006] (p.424) calls it a "distortion measure". In Fig.9.2, Bishop shows the reduction of error as the k-means algorithm runs on the Old Faithful dataset. After one iteration (M-step), the error is around 300. Using this value and assuming that each cluster center is the closest for exactly half of the points in the dataset, calculate the average squared Euclidean distance between data points and cluster centres at this moment.

   Answer: this question was unclear initially and turned out a bit redundant. There are 272 points in the Old Faithful dataset and the error (sum of squared Euclidean distance between points and closest centres) is 300. So on average, error per point is $300/272 = 1.1029$. The bit about half the points doesn't matter: $150/136 = 1.1029$.

3. In Fig.7.5 of Alpaydin, the Euclidean distance metric has been used (single-link hierarchical clustering). How would the dendrogram change if the city-block (aka Manhattan) distance metric was used instead?

   Answer: the height of the branch connecting e to (a, b) will increase to 2 on the y-axis, equal to the branch connecting (c, d) and f.

Advanced Questions

4. The python `scikit-learn` toolbox contains implementations of many different clustering algorithms. A nice illustration comparing some of the algorithms on some 2-D datasets is shown here: https://scikit-learn.org/stable/modules/clustering.html

   Think about *why* these particular example datasets (i.e. each row of subplots) might have been chosen for this illustration. Describe the most important properties of the datasets in rows 1, 3 and 5 (at most two sentences for each dataset).

   Answer: note here is a possible answer here but plenty of scope for other thoughtful, well-written answers!

   Row 1: A ring of data inside another ring of data, with points densely distributed around the rings. Euclidean (and other) distance metrics struggle to separate the rings because many points are close to points located on the other ring.

   Interestly, note that hierarchical (agglomorative) gets it. And what is a cluster anyway? :)

   Row 3: Three spherical clusters but of different sizes, close together (relative to within-cluster variance) with seemingly more noise for the larger clusters. The clusters are not easy to separate and with noise create a challenging dataset, leading to variable results across algorithms.

   Row 5: Perhaps the simplest dataset for clustering - each cluster has the same spherical shape and small within-cluster variance compared to the distance between the clusters. Note clustering is unsupervised so the ordering of the colours (labels) is arbitrary.

5. The method for initializing the k-means algorithm discussed in lectures (and in Alpaydin) is known as the *Forgy* approach. [1] compares this with three other techniques. Refer to the top two plots in Figure 5 of this paper (available on the course website). Based on these results only, would you use the RANDOM or the Forgy (FA) technique? Explain why in no more than 4 sentences.

   Answer: It's a pretty hard choice and could probably be argued either way. There is a lot more variability in RANDOM but the FA finds a solution (approx. 40% of the time) that is worse than any of the RANDOM solutions. Sensible answers that demonstrate understanding will get full marks.

[Pena et al. 1999] Pena, J. M., Lozano, J. A., & Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. Pattern recognition letters, 20(10), 1027-1040.