CS910 Exercise Sheet 5: Clustering

Please present your results as a single printed document. Limit your answers to 4 sides of paper – this should be more than enough to express your findings; any more and you are including a lot of unnecessary detail.

1. (a) Consider the following two data points: [12]

   $$x : (5, 3, 7, 9) \text{ and } y : (7, 3, 6, 7)$$

   Calculate the following distances. In each case, show your working.

   i. The Hamming distance between x and y

   ii. The Manhattan (L_1) distance between x and y

   iii. The Euclidean (L_2) distance between x and y

   (b) Suppose you have data on patients that records their height (in millimetres) and the length of their eyelashes (in millimetres). The aim is to perform clustering on this data. Explain why clustering using a distance measure such as Euclidean distance would be problematic on this data, and outline how you could modify the data to address this. [8]

   (c) Suppose you are given the following points in two-dimensional Euclidean space, with x and y values representing location. [10]

Point id	A	B	C	D	E	F	G	H
x	2	2	8	5	7	6	1	4
y	10	5	4	8	5	4	2	9

   The task is to form these points into three clusters. Suppose initially we assign A, D, and G as the centre of each cluster, respectively.

   i. How will the k-means algorithm assign the remaining points to these three cluster centres (A, D or G)? Show your calculations clearly.

   ii. What are the new cluster centres after this step? Show your working.

2. We again use the breast cancer data provided with Weka or available from `http://tunedit.org/repo/UCI/breast-cancer.arff`. We will use the "class" attribute to see if we can find clusters that align with this class (cluster mode is "classes to clusters evaluation").

   (a) Use Hierarchical Agglomerative Clustering to cluster the data into two clusters. Try all combinations of distance function (Euclidean or Manhattan distance) and linkage type (single, complete, average) seen in lectures. Which methods perform best for accuracy of predicting the class value (whether or not there is recurrence)? [10]

   (b) Use Furthest Point Clustering (FarthestFirst) on the data with 2 clusters. The "seed value" initializes the random number generator, which affects which point is chosen to start the clustering. Try different values of the seed value, in the range 1 to 10. What is the range of accuracy results you observe, and what do you conclude about this method? [10]

(c) Use DBSCAN to cluster the data. Search for different settings of the parameters ϵ and MinPts that can find two similarly sized clusters for this data. What do you conclude about the suitability of this clustering method for this data? [10]

Note: Weka may give an error message if trying to evaluate the accuracy of a clustering but all data points are classed as "noise".

(d) Compare the results of the clustering to the accuracy of classifiers on this dataset. What do you conclude about the ability of clustering to predict the class value in this data? [10]

3. For this question, the "cluster mode" in Weka should be "use training set". Download the bmw-browsers data set, from `http://www2.warwick.ac.uk/fac/sci/dcs/teaching/material/cs910/exercises/bmw-browsers.arff`. This contains information on visitors to a (fictional) car dealers. There are multiple (binary) attributes in the data:

dealership	whether the visitor went into the main dealership
showroom	whether they went into the car showroom
computer search	whether they used a terminal to do a search
M5	whether they looked at the (most expensive) M5 model
3series	whether they looked at the (entry level) 3series model
Z4	whether they looked at the (mid-range) Z4 model
financing	whether they enquired about financing a car
purchase	whether they completed a purchase

(a) Apply DBSCAN using parameters $\epsilon = 1.1$ and MinPts = 10. This should find two clusters. Visualize the clustering to find which attributes are separated between the two clusters. Use the "Jitter" slider to add some small random perturbation to each point, so it is easier see the different data points. Which attributes are mostly divided by the clustering, so examples with a 1 are in a different cluster to those with a 0? [6]

(b) Apply the EM algorithm to find 5 clusters in the data. For each cluster that is found, Weka reports its mean and standard deviation in each dimension. We are interested in dimensions where there is a significant concentration – say, a mean of more than 0.9 or less than 0.1. Run EM in Weka with seed values 1, 2, 3 (please mention which version of Weka you are using, as theer is some variation across versions). For each clustering that is found, describe what are the significant dimensions for each cluster. Is there consistency across the different clusterings? [12]

(c) Find a k-means clustering for 5 clusters. The clustering is described in terms of the centroid locations (there is no standard deviation given). Again, report the dimensions in each cluster where there is a strong preference for a value (above 0.9 or below 0.1) for seeds = 1, 2, 3, and describe how consistent the clusterings are with each other and with those found by EM. [12]