Predictive Modelling Tutorial 10 & 11: Clustering

Dr Liew How Hui

Jan 2021

Preparing for Final

- Week 13: This is the last Tutorial class. No more tutorial class starting next week.
- Week 14: Assignment Report, Programming Code, Oral Presentation marks to be announced. Apologies if not all of them is rushed out.

Preparing for Final

- 26 April 2021, 2pm–5pm. 4 compulsary questions, each 10 marks.
- Similar to year 2020. Show the steps!
- Programming code cannot be regarded as answer (I am not going to run your code for you). Computer output without explanation will not be accepted.
- If your writing/scan is blur, marks will be deducted.
- Leave margin of 2.5cm for marking.

Tutorial 2, Q1

The first step of k-means clustering is to decide the number of clusters, k. After a series of iterations, can k-means ever give results which contain

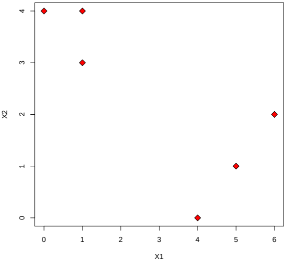
(a) More than k clusters?

(b) Less than k clusters?

Tutorial 2, Q2

You are given a small example with $n = 6$ observations and $p = 2$ variables. The observations are as follows:

Obs	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0



Tutorial 2, Q2 (cont)

(a) Plot the observations.

(b) Rescale the observations to [0,1].

(c) Perform k-means clustering to the observations with $k = 2$. The initial centroids are 2, 5.

(d) In the plot from (a), colour the observations according to the cluster labels obtained.

Tutorial 2, Q3

The table below shows the marks for 2 assessments (Test and Assignment) of 6 students in a class. Group the students into 3 clusters based on the mark for these 6 students using Manhattan distance (and Euclidean distance if time permits).

Student	Test	Assignment	Initial Cluster
1	89	34	A
2	45	27	A
3	56	30	B
4	89	44	B
5	81	46	C
6	83	36	C

Tutorial 2, Q4

Suppose that we have four observations, for which we compute a distance matrix:

$$\begin{bmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{bmatrix}$$

Tutorial 2, Q4

(a) Sketch the dendrogram that results from hierarchically clustering these four observations using **complete linkage**. Plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram. Suppose that we cut the dendrogram such that two clusters result. What are the observations in each cluster?

(b) Repeat (a) using single linkage clustering.

FA May 2020 Q3 (a)

Given the unlabelled data in Table 3.1.

Table 3.1: Unlabelled data.

	V1	V2	V3
1	7.5205	4.6564	-0.1947
2	-1.1824	-1.1174	1.8383
3	-0.3576	-0.4739	-1.1603
4	-1.422	-0.5891	-0.8287
5	3.2287	0.7141	0.6208
6	3.2926	3.1609	2.7553
7	8.2304	3.8832	-1.7378
8	4.2079	0.4964	4.361
9	3.8443	5.7565	1.0293
10	1.493	3.525	-2.9904

Use the k-means algorithm with $k = 2$ (unsupervised learning) to find the final cluster centres if the **first** and **sixth** rows are chosen as the **initial cluster centres**.

(4 marks)

FA May 2020 Q3 (b)

Given an appropriate example to explain why the Minkowski distance

$$M(\mathrm{x},\mathrm{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^r\right)^{\frac{1}{r}}, \quad \mathrm{x},\ \mathrm{y} \in \mathbb{R}^p$$

will no longer be a distance function when $r = \frac{1}{2}$.
(2 marks)

FA May 2020 Q3 (c)

Group the observations in Table 3.1 using hierarchical clustering and the **Minkowski distance** with $r = 3$ (refer to part (b) for the definition of Minkowski distance) and **complete linkage** and draw the dendrogram formed by the hierarchical clustering.

Table 3.1: Unlabelled data.

Obs	x_1	x_2	x_3
A	1	3	2
B	5	7	9
C	6	9	8
D	7	8	9
E	2	3	5
F	1	4	3

(4 marks)