ENGG5103 Techniques for Data Mining

Assignment1

Q1. (Preprocessing) Given the following table:

Student id	**Name**	**Gender**	**Current GPA**	**CS Student**	**Course Enrolled**
12345	Jerry	Male	3.2	Yes	Data Mining
12345	Jerry	Male	3.2	Yes	Physics
23456	Tom	Male	3.8	No	Data Mining
23456	Tom	Male	3.8	No	Linear Algebra
34567	Spike	Male	2.0	No	Linear Algebra
45678	Lily	Female	4.0	No	Physics
56789	Lucy	Female	3.7	Yes	Data Mining

(1) Do the coding process with feature selection and data transformation. (15')
(2) Calculate the **mean**, **median** and **standard deviation** of current GPA **after Preprocessing**. (5')

Q2. (K-means) Given the following 2-D points

P1	P2	P3	P4	P5	P6	P7
(1.0, 1.0)	(1.5, 2.0)	(3.0, 4.0)	(5.0, 7.0)	(3.5, 5.0)	(4.5, 5.0)	(3.5, 4.5)

If the k-means clustering (c = 2) is initialized by **(1.8, 2.3)** and **(4.1, 5.4)**. Show the iteration of k-means algorithm until the algorithm converges. (You should indicate the centroid and member of each cluster by table or graph) (20')

Q3. (Hierarchical Clustering)

Given following points:

P1: (3, 4) **P2**: (3, 3) **P3**: (6, 2) **P4**: (10, 12) **P5**: (11, 11) **P6**: (12, 10)

(1) Calculate the Euclidean distance between every two points with the following table (Just fill the colored cells): (5')

Euclidean Distance	P1	P2	P3	P4	P5	P6
P1						
P2						
P3						
P4						
P5						
P6						

(2) With Euclidean distance and agglomerative algorithm, perform hierarchical clustering. Show the resulting matrices with intermediate steps after performing **complete linkage** clustering.(15')

(3) With Euclidean distance and agglomerative algorithm, perform hierarchical clustering. Show the resulting matrices with intermediate steps after performing **single linkage** clustering.(15')

(4) Assume that the points have been divided into two sub-clusters as following:

C1	C2
P1, P2, P3	P4, P5, P6

Try to calculate the distance between two clusters by **group average**, **centroid** and **median** clustering rules. (5')

Q4. The lecture notes of this course have shown the Gaussian mixture methods with EM algorithm. Actually, the algorithm of k-means could also be explained by EM-algorithm. Try to briefly explain the relationship between EM-algorithm and k-means algorithm. (20')