**COMP5121 Data Mining & Data Warehousing Applications**

**Assignment #2**

Instructions: - Answer all three questions.

* Interpret the questions logically, show your steps and write down your assumption(s) when necessary.
* You may use this word file to prepare your answers and submit it to L@PU before the due date.

1. (35 marks) Suppose you are asked to provide data mining consulting services to an Internet DVD shop. After interviewing the shop’s manager and the database administrator, the following movie database is collected.

**Movie Database**

|  |  |  |
| --- | --- | --- |
| **Movie ID** | **Movie Name** | **Types** |
| 0001 | Maze Runner | Romance, Sci-Fiction, Drama, Mystery |
| 0150 | Dear John | Drama, Romance, Fiction |
| 0553 | Avengers | Action, Sci-Fiction, Thriller, Horror |
| 1011 | Poltergeist | Horror, Thriller |
| 3997 | Amazing Spiderman | Action, Crime, Sci-Fiction |

1. If you are asked to cluster the movies, propose an appropriate dissimilarity measure for it and prepare the following dissimilarity matrix for the five movies in the database above.
2. Based on your dissimilarity matrix obtained in part (a), apply the COMPLETE linkage agglomerative clustering algorithm to cluster the five movies. Draw the dendrogram obtained.
3. The COMPLETE linkage agglomerative clustering has been suffering from the weakness of low scalability (high time complexity). Other than the sampling approach, propose a way to speed up its computation.

2. (30 marks) You are asked to use the k-means algorithm to cluster the following 8 examples into 3 clusters:

P1=(2,10), P2=(2,6), P3=(8,4), P4=(5,8), P5=(7,4), P6=(6,4), P7=(1,2), P8=(4,9).

1. Compute the distance matrix based on the Euclidean distance.
2. Suppose that the initial seeds (centroids of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 iteration ONLY and then write down:

i) The new clusters (i.e. the examples belonging to each cluster)

ii) The centroids of the new clusters

3. (35 marks) In some social network services, the generated social network data should be modeled by directed graphs, i.e., the edges have directional indications as shown in Fig.1. Based upon such kind of data, a user classification problem can be formed as listed in Table I. Here, Table I records the number of in-links (A1) and the number of out-links (A2) as “many” or “few”. In addition, it introduces one more attribute, i.e., A3, to denote the privacy level as “high” or “low”. The class attribute is user attention, i.e., the social network user is classified as “Superstar”, “Activist” or “Follower”.

**/**

*Fig.1 Directed social network graph*

*Table I. Social network user classification data*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| User ID | A1: Number of in-links | A2: Number of out-links | A3: Privacy level | User Attention |
| A | Many | Few | Low | Activist |
| B | Few | Many | High | Activist |
| C | Few | Many | High | Follower |
| D | Many | Many | Low | Superstar |
| E | Many | Few | Low | Superstar |
| F | Few | Many | High | Follower |

Suppose you make use of the naive Bayesian classification method to solve the problem and all the six records above are used for training.

a) Compute the classification rate of the training data obtained by the naive Bayesian classifier.

b) WITHOUT guessing/inducing the missing values, show how the following data records with missing value “---“ should be classified by the naive Bayesian classifier constructed in part (a).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| User ID | A1: Number of in-links | A2: Number of out-links | A3: Privacy level | User Attention |
| G | Many | Few | --- | ? |
| H | --- | Many | High | ? |

**E N D**