M11109124 梁[illegible]

Data mining makeup (2022/12/26)

1. (20%) (Frequent Itemsets)

Given the following transactional dataset, answer the following questions:

Table 1. Transactional dataset

Transaction id.	Items included
1	{A, B, D}
2	{A, C, D}
3	{A, B, C, D, E}
4	{B, C, D}
5	{A, B, D}
6	{A, B}
7	{A, B, D}
8	{A, C,E}

(a) Find the closed itemsets with minimum support equal to 0.5. (10%)

(b) Find the confidence and lift of rule {B} -> {D}. (5%)

(c) Find any two rules that satisfy the minimum confidence of 0.5. (5%)

2. (20%) (Linear Regression) The following dataset contains two columns X and Y. Please build a regression equation with X as the input variable and Y as the output variable using simple linear regression (10%). Then, find the R-square for this linear regression. (10%)

Subject	X	Y	**Predicted Value**	Error in Prediction	(Error in Prediction)^2
1	2	10			
2	3	12			
3	3	13			
4	4	13			
					$SSE = \sum(y - \hat{y})^2$ =

Mean(Y) =12, Mean(X)=3

b1=S_{xy}/S_{xx} ; b0=mean(y)-b1*mean(x) ; $R^2=S^2_{xy}/(S_{xx}*S_{yy})$=SSR/SST=(SST-SSE)/SST

S_{xy}=cov(X,Y); S_{xx}=Varance(X) (find b1, then find b0 , use the equation to find the predicted values)

3. (20%) Given a dataset with values of 3, 5, 7, 11,25, 23, 26, 18, 19, answer the following questions:
   (a) Cluster the dataset into THREE clusters using the hierarchical clustering algorithm with "COMPLETE LNK" for calculating the distance between two clusters. Draw the dendrogram for the clustering result. (10%)
   (b) Cluster the dataset into TWO clusters using the K-Means algorithm with 3 and 11 as the initial two means of the K-Means algorithm. (10%)

4. (20%) (Text Mining) Listed below is a corpus with three documents.

   D1: A dog barks at a cat and it fell from a Tree.
   D2: A dog watches ants on the bark of a Tree.
   D3: The bark falls from the tree as a cat Watches.

   (a) Please perform text preprocessing on each document of the corpus. (5%)
   (b) Please find the DTM matrix for this corpus using "tf-idf" as the measure of importance for a term in the DTM matrix. (10%)
   (c) Based on the DTM matrix, please use hierarchical clustering to cluster these five documents into two clusters using Manhattan distance as the distance measure and "Single link" as the grouping criterion in the hierarchical clustering. (5%)

Note that the stop words in this corpus include { a, at, and, it, from on, of, the, from}

The TF-IDF for term t_k in document d_i denoted by w_{ik}, *is equal to* $tf_{i,k} * idf(t_k)$, where
$tf_{i,k}$ = *frequency of term* t_k *in document* d_i;

$idf(t_k)$

$= log_{10}\left(\frac{N}{n_k}\right)$, *where* N *is the number of documents in the corpus, and* n_k *is the number of documents which contain term* t_k.

註: (N 是文章總篇數，n_k 是包含t_k 的文章總篇數)

5. (20%) Answer the following questions:

(a) List the major steps in data preprocessing. (10%)

(b) How is a boxplot look like (Please draw a boxplot (盒鬚圖.)? (10%)