Data mining (2022/12/19)

1. (20%) (Association rules) Table 1 shows a transaction database. Answer the following questions with minimum support of 40%.

Table 1

Transaction ID	Items Bought
T100	{b,d,e,f}
T200	{ b,c,d}
T300	{ b,d,e,f}
T400	{b,d }
T500	{b,c,d,e}
T600	{b,d,e}

(a) Use the Apriori algorithm to find all the frequent itemsets. (10%)

6 * 0.4 =2.4, min_sup_count=3;

b:6, ~~e:2~~, d:6, e:4, ~~f:2~~; bd:6, be:4, de:4; bde:4

(b) Compute the confidence and lift of rule "bd→e." (5%)

(#{bde}/6)/(#{bd}/6 *#{e}/6)=4/6 * (6/6) * (6/4) =1

(c) What is the anti-monotone principle in the Apriori algorithm? How is it used in the Apriori algorithm? (5%)

Support count of an itemset is smaller than that of any of its sub-itemsets.

Used to reduce the number of candidate itemsets in Apriori algorithm.

2. (20%) (Regression) Table 2 shows a dataset with the values of X and Y attributes. Build a regression equation for the dataset with Y as the response variable and X as the predictor (15%). Please find the $\mathbf{R^2}$ value for this regression. (5%)

Table 2. Dataset for regression

X	Y
2	5
2	6
3	5
3	8
4	7
4	10

5	10
5	12

Mean(Y) =7.875, Mean(X)=3.5

b1=Sxy/Sxx；b0=mean(y)-b1*mean(x)；

$R^2=S^2xy/(Sxx*Syy)=SSR/SST=(SST-SSE)/SST$; S_{xy}=cov(X,Y); S_{xx}=Varance(X)

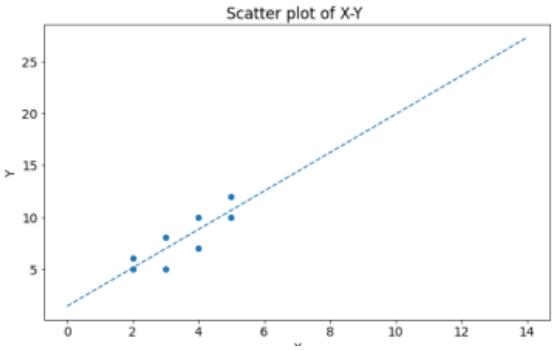
b0=1.4, b1=1.85, R-square=0.73

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.730
Model:	OLS	Adj. R-squared:	0.685
Method:	Least Squares	F-statistic:	16.23
Date:	Tue, 10 Jan 2023	Prob (F-statistic):	0.00689
Time:	15:26:38	Log-Likelihood:	-13.184
No. Observations:	8	AIC:	30.37
Df Residuals:	6	BIC:	30.53
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.4000	1.687	0.830	0.438	-2.728	5.528
X	1.8500	0.459	4.029	0.007	0.726	2.974

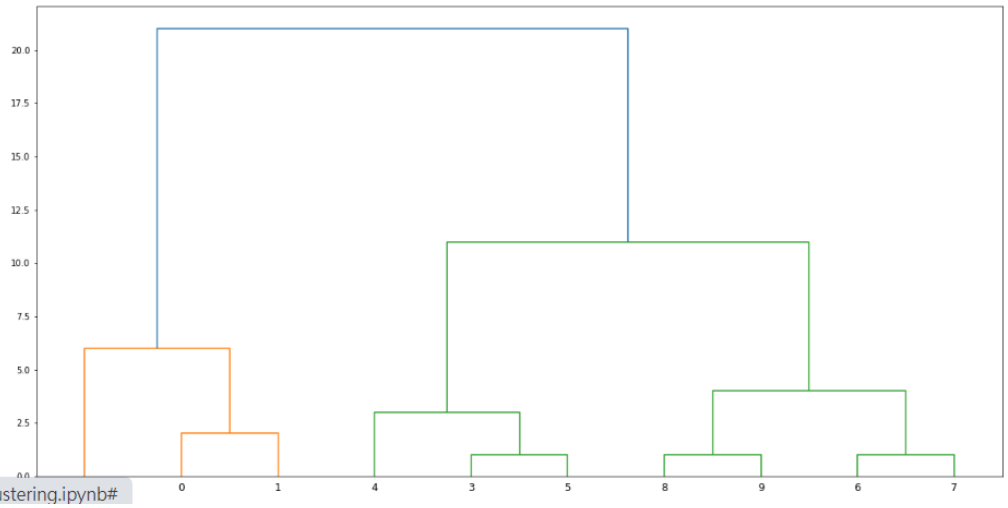
Omnibus:	1.896	Durbin-Watson:	3.373
Prob(Omnibus):	0.387	Jarque-Bera (JB):	0.927
Skew:	-0.468	Prob(JB):	0.629
Kurtosis:	1.620	Cond. No.	12.9



3. (20%)A dataset contains ten integers of 5, 7, 11,25, 23, 26, 18, 19, and 15,16. Answer the following questions:

(a) Use the hierarchical clustering method with "**Complete link**" as the grouping criterion to cluster these integers into three clusters. Please show the members of each cluster and draw the resulting dendrogram.

In [6]:
```
X = [[i] for i in [5,7,11,25,23,26, 18, 19, 15, 16]]
#X = [[i] for i in [5,7,11,25,23,26, 18, 19]]
#X = [2, 4, 10, 12,3,20, 30, 11, 25]
#list = [2,4,10,11,3,20,30,12,25]
#X= np.array(list)
linked = linkage(X, 'complete')
plt.figure(figsize=(20, 10))
dn=dendrogram(linked)
plt.show()
```

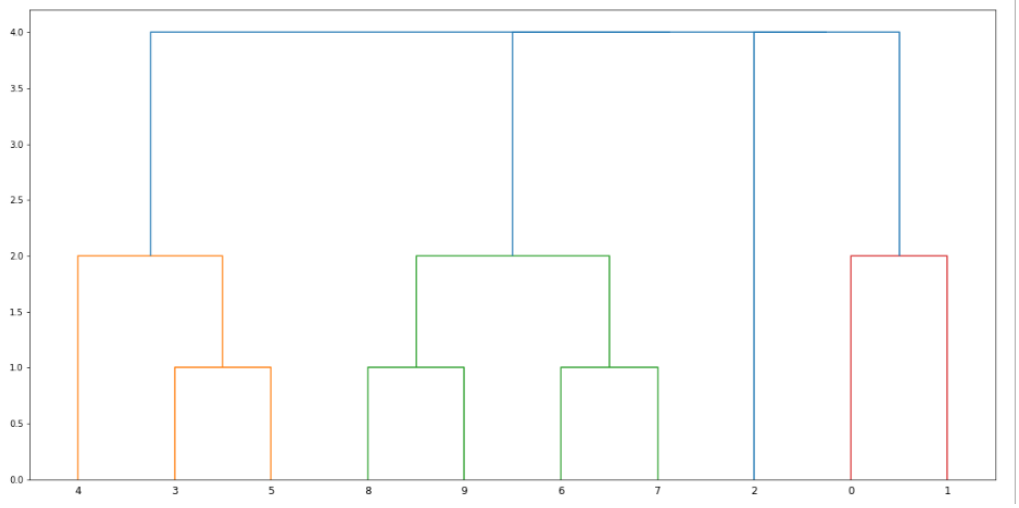


ebooks/clustering.ipynb#

(b) Use the hierarchical clustering method with "**Single link**" as the grouping criterion to cluster these integers into three clusters. Please show the members of each cluster and draw the resulting dendrogram.

In [7]:

```
X = [[i] for i in [5,7,11,25,23,26, 18, 19, 15, 16]]
#X = [[i] for i in [5,7,11,25,23,26, 18, 19]]
#X = [2, 4, 10, 12,3,20, 30, 11, 25]
#list = [2,4,10,11,3,20,30,12,25]
#X= np.array(list)
linked = linkage(X, 'single')
plt.figure(figsize=(20, 10))
dn=dendrogram(linked)
plt.show()
```



4. (20%) （Data preprocessing）

(a) What are the general steps in data preprocessing？ (10%)

data collection, data cleaning, data integration, data transformation, dimension reduction

(b) What are the data preprocessing steps for text mining? (5%)

noise reduction, stop_word removal, stemming, attribute reduction, DTM construction

(c) According to the contingency table, determine whether smoking and lung cancer are correlated based on the significant level of 0.01. Note that "L =1" stands for having lung cancer while "L=0" stands for not having lung cancer. Similarly, "S=1" stands for smoking and "S=0" stands for not smoking. (5%) Independent

	L=1	L=0	Total
S=1	8 (6.39)	19 (20.6)	27
S=0	1 (2.6)	10 (8.39)	11
Total	9	29	38

Chi-Square distribution table for Probability level (alpha)

Degrees of Freedom	0.1	0.05	0.02	0.01	0.001
1	2.706	3.841	5.412	6.635	10.827

2	4.605	5.991	7.824	9.210	13.815

Note that $\sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} = \frac{(8-6.39)^2}{6.39} + \frac{(19-20.6)^2}{20.6} + \frac{(1-2.6)^2}{2.6} + \frac{(10-8.39)^2}{8.39} = 1.81 << 6.635$

follows a chi-square distribution with a degree of freedom 1; o_i and e_i are the observation value and the expectation value of cell i, respectively.

Execution result of chisq.test
chisq.test(x)

Pearson's Chi-squared test with Yates' continuity correction

data: x
X-squared = 0.86474, df = 1, p-value = 0.3524;

Cannot reject H0, which is the independent assumption.

5. (20%) (Text Mining) Listed below is a corpus with five 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: A dog watches another dog which watches a Cat.
D4: A dog barks at a cat that watches another Cat.
D5: The bark falls from the tree as a cat Watches.

(a) Please perform text preprocessing on each document of the corpus. (5%)

D1: dog bark cat fall tree
D2: dog watch ant bark tree
D3: dog watch dog watch cat
D4: dog bark cat watch cat
D5: bark fall tree cat watch

	dog	bark	cat	fall	tree	watch	ant
D1	1 (0.2)	1 (0.2)	1 (0.2)	1 (0.2)	1 (0.2)	0 (0)	0(0)
D2	1 (0.2)	1 (0.2)	0	0	1 (0.2)	1 (0.2)	1 (0.2)
D3	2 (0.4)	0	1 (0.2)	0	0	2 (0.4)	0
D4	1 (0.2)	1 (0.2)	2 (0.4)	0	0	1 (0.2)	0
D5	0	1 (0.2)	1 (0.2)	1 (0.2)	1 (0.2)	1 (0.2)	0

IDF	0.097	0.097	0.097	0.398	0.222	0.097	0.699

TF-IDF

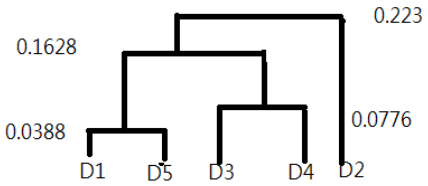
	dog	bark	cat	fall	tree	watch	ant
D1	0.0194	0.0194	0.0194	0.0796	0.0444	0 (0)	0(0)
D2	0.0194	0.0194	0	0	0.0444	0.0194	0.1398
D3	0.0388	0	0.0194	0	0	0.0388	0
D4	0.0194	0.0194	0.0388	0	0	0.0194	0
D5	0	0.0194	0.0194	0.0796	0.0444	0.0194	0
IDF	0.097	0.097	0.097	0.398	0.222	0.097	0.699

(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%)

Distance matrix

	D1	D2	D3	D4	D5
D1					
D2	0.2582				
D3	0.2016	0.2618			
D4	0.1628	0.223	0.0776		
D5	0.0388	0.2582	0.2016	0.1628	



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

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)