CIDM 6355 Data Mining Methods Exam 2 Part 2 Instruction

(70 points in total; due 11:59 pm CST, November 12th, 2023)

Requirements: Please read, understand, and comply with the following requirements in this exam.

- This exam is open book, open slides, and open notes, but you are not allowed to collaborate nor discuss with anyone else during the exam period. Any question about the exam should be addressed to the instructor.
- This part of the exam is not timed, but you have to submit all the required responses by the deadline to be accepted; a late submission will not be acceptable and a zero point will be assigned.
- This is an individual exam, so sharing your RM processes, R script, screenshots, or answers with other students or parties is considered as cheating, which will be reported to the university authority.
- It is your responsibility to make your responses and deliverables meet the required format; otherwise, you would lose points because of wrong format.

Background

A wine company attempted to discover distinct groups in the current wines and examine whether they differ in wine quality. You, a newly hired data scientist, were asked to accomplish this job. In order to achieve your company's goals, you collected 300 wines, which was stored in the dataset titled "winequality.csv". The dataset includes chemical characteristics (such as fixed acidity, alcohol level, pH, density, etc.) of 300 wines as well as the quality ratings of these wines graded by experts. Note that attribute values in the dataset are all numerical values (real), but on different scales. Based on your previous discussion and research with customers, industry experts, and distributors, you decide to cluster those wines based on their chemical characteristics to three groups as a starting point. **Every attribute except "Quality" is considered a chemical characteristic.**

You plan to use two analytics tools (RM and R) and two clustering methods (K-means and Agglomerative Clustering) to generate four models as below.

		Clustering Methods	
		K-Means Clustering	Agglomerative Clustering
Clustering Tools	RapidMiner	Model 1	Model 3
	R or R Studio	Model 2	Model 4

Instructions

1. Data Import and Preparation
   1.1. Import the dataset (Store your dataset as Exam2Data1 in R).
   1.2. Check whether or not there are highly-correlated attributes (absolute value of correlation coefficients greater than 0.85 in this case).
   1.3. After examining the dataset, you decide to normalize chemical attributes that are used for clustering.
   1.4. You plan to use range transformation (min=0, and max=1) for all the chemical attributes.
   1.5. For the range transformation in R, you can either use the method provided in Week 9 R Lab Instruction, or use the data.Normlization function, which is under the library clusterSim (see the code as below).

```
#install and use the library clusterSim for data normalization
install.packages("clusterSim")
library(clusterSim)
#different from Lab 9, here we normalize the first six columns using data.Normalization function with type=n4, meaning unitization with zero minimum ((x-min)/range))
Exam2Data1[1:6]<-data.Normalization(Exam2Data1[1:6],type="n4",normalization="column")
#check if the data is normalized
summary(Exam2Data1)
```

Attention: Step 1 applies to each of the following four steps (models), which means that all the attributes except Quality must be normalized before clustering.

2. K-Means Clustering and Post-Clustering Analysis in RM (Model 1)
   2.1. Use the same parameters that we used in Week 8 RM Lab to develop a k-means clustering model in RM.
   2.2. Record the size for each cluster.
   2.3. Take a screenshot of the centroid table with date and time (Screenshot 1).
   2.4. Generate a bar chart to show the average quality ratings of the three clusters.
   2.5. Perform a Grouped ANOVA analysis to assess whether and why there are differences in the average quality ratings of the three clusters of wines at the 0.05 significance level (Alpha = 0.05).
   2.6. Take a screenshot of the ANOVA table with date and time (Screenshot 2) and briefly describe your conclusion. Your conclusion must be based on both Steps 2.4 and 2.5.

3. K-Means Clustering and Post-Clustering Analysis in R (Model 2)
   3.1. Set 123 as the random seed and use Week 8 R Lab as a reference to build a k-means clustering model in R and store it as kcluster.
   3.2. Generate a data frame to show the size and centroid for each cluster (i.e., the number of observations in each cluster and the mean of each attribute in each cluster).
   3.3. Take a screenshot of your output (cluster size and centroids) with date and time (Screenshot 3).
   3.4. Generate a new column called klabel to save the cluster in the dataset. You may refer to the following code.

```
#3.4. Generate a new column called klabel to save the cluster in the dataset
Exam2Data1$klabel<-kcluster$cluster
```

   3.5. Perform an ANOVA analysis to assess whether and why there are differences in the average quality ratings of the three clusters of wines at the 0.05 level (Alpha = 0.05). Please note that you need to convert klabel as a factor in the ANOVA analysis.

```
#3.5 Perform ANOVA analysis
summary(aov(quality ~ factor(klabel), data=Exam2Data1))
```

   3.6. Take a screenshot of the ANOVA table with date and time (Screenshot 4) and briefly describe your conclusion based on the ANOVA table.
4. Agglomerative Clustering and Post-Clustering Analysis in RM (Model 3)
   4.1. Use CompleteLink, NumericalMeasures, and EuclideanDistance to cluster 300 wines into three clusters.
   4.2. Record size of each cluster.
   4.3. Generate a bar chart to show the average quality ratings of the three clusters. Take a screenshot of the bar chart with date and time (Screenshot 5) and briefly describe your conclusion. Your conclusion must include each cluster's size and their average quality ratings.
   4.4. Perform a Grouped ANOVA analysis to assess whether and why there are differences in the average quality ratings of the three clusters of wines at the 0.05 significance level (Alpha = 0.05). No deliverable is required for this step.

5. Hierarchical Clustering and Post-Clustering Analysis in R (Model 4)
   5.1. Set 123 as the random seed and use Week 9 R Lab as a reference to build a hierarchical clustering model in R.
   5.2. Cut the dendrogram to three cluster and store your clustering result as hcluster in R.
   5.3. Generate a table to show the size (the number of items) of each cluster.
   5.4. Generate a new column called hlabel to save the cluster in the dataset.

5.5. Create a bar chart to show the average quality ratings of the three clusters. Take a screenshot of the bar chart with date and time (Screenshot 6) and briefly describe your conclusion. Your conclusion must include each cluster's size (Step 5.3) and their average quality ratings.

5.6. Perform an ANOVA analysis to assess whether and why there are differences in the average quality ratings of the three clusters of wines at the 0.05 level (Alpha = 0.05). Please note that you need to convert klabel as a factor in the ANOVA analysis.

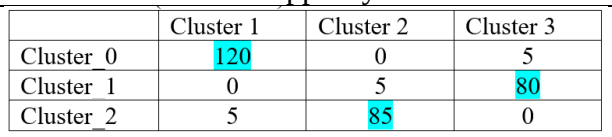
6. Summary and Comparison

6.1. Compare the clustering results of the four models (You may refer to Deliverable R4 in Week 8 R Lab) and compute the match rate for each pair of models. For example, you can compute the match rate of the following two models as below: (120+85+80)/300=95%.

	Cluster 1	Cluster 2	Cluster 3
Cluster_0	120	0	5
Cluster_1	0	5	80
Cluster_2	5	85	0

6.2. Calculate the match rate for each pair and include the corresponding screenshot below. You must demonstrate how the match rate is computed, and your screenshot (e.g., a PivotTable) should illustrate how clusters from each model correspond to one another. Your screenshots are not required to display date and time.

Attention: To calculate the match rate, please illustrate how it is computed, as demonstrated in the following example; otherwise, 1 point will be deducted. You should provide a screenshot that clearly displays how each pair of clusters is matched, similar to the example shown; otherwise, 1 point will be deducted.

Model Pair	Match Rate	A screenshot to support your match rate
Models 1 & 2	(120+85+80)/300 =95%	
Models 1 & 3		
Models 1 & 4		
Models 2 & 3		
Models 2 & 4		
Models 3 & 4		

If you have a hard time to determine the matching pattern, please check the appendix at Page 4.

You do not need to include date and time in your screenshots here. You can use snip part of the screen to include your evidence such as a PivotTable in the table above. Instruction for Mac users; Instruction for Windows users; Instruction for Linux users.

----------The end of Exam 2 Part 2 ----------

Appendix: Finding the Highest Match Rate

For each pair of models, there are six possible matching scenarios. You can calculate the match rate for each situation and identify which one results in the highest match rate. However, the following method may help you quickly locate the best matching scenario.

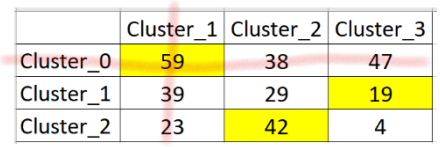
In some cases, the matching pattern is quite evident, making it straightforward to determine by selecting the highest number in each row or column, as shown below.

	Cluster 1	Cluster 2	Cluster 3
Cluster_0	120	0	5
Cluster_1	0	5	80
Cluster_2	5	85	0

However, in some cases, the matching is not easily determined when the matching results are scattered or not clearly defined. In the following case, all the numbers at the first row are greater than any numbers in the other two rows, making the matching more challenging.

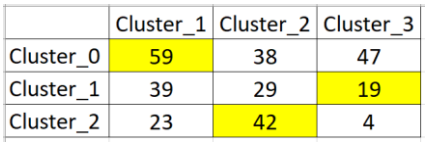
	Cluster_1	Cluster_2	Cluster_3
Cluster_0	59	38	47
Cluster_1	39	29	19
Cluster_2	23	42	4

You can begin by selecting the highest number and subsequently crossing out both the corresponding row and column associated with that number.



	Cluster_1	Cluster_2	Cluster_3
Cluster_0	59	38	47
Cluster_1	39	29	19
Cluster_2	23	42	4

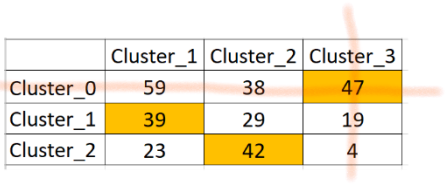
Next, select the highest number among the remaining options and then proceed to cross out the corresponding row and column of that number.



	Cluster_1	Cluster_2	Cluster_3
Cluster_0	59	38	47
Cluster_1	39	29	19
Cluster_2	23	42	4

Finally, you will find the matching result. In this case, the match rate is (59+42+19)/300=40%.

This method is effective in most cases, but when the other numbers at the same row or column is very close in value to the highest number, this method may not consistently yield the highest match rate. You may try to start with the second highest one, that is 47 in this case. Then, you will find the match rate is (47+42+39)/300=42.7%.



	Cluster_1	Cluster_2	Cluster_3
Cluster_0	59	38	47
Cluster_1	39	29	19
Cluster_2	23	42	4

In such cases, if you encounter a situation where the other numbers at the same row or column is very close in value to the highest number, you might need to make multiple attempts to find the highest match rate, trying different approaches or methods two or three times.