**BUDT 758T: Data Mining & Predictive Analytics: “Assignment 7”**

This assignment illustrates several of the unsupervised data techniques we used in class. You’ll need to use subjective judgement and domain knowledge to evaluate the quality of the results of these data mining methods. You may find these techniques useful for exploring your project data!

The data for this assignment can be found in *RecipeData.csv*. This data set contains information on a set of recipes and their features, including what type of cuisine the recipe is (American, Asian, European, or Latin) and whether given ingredients are used (1) or not (0).

The goal of this analysis is to identify clusters and ingredient groups in the data rather than to predict a given outcome variable. For each problem, use the code from the slides provided in class or otherwise specified. You should not need any additional code that is not explicitly included in the slides or in the problem itself (or is a function you have already seen, such as sample() or table())!

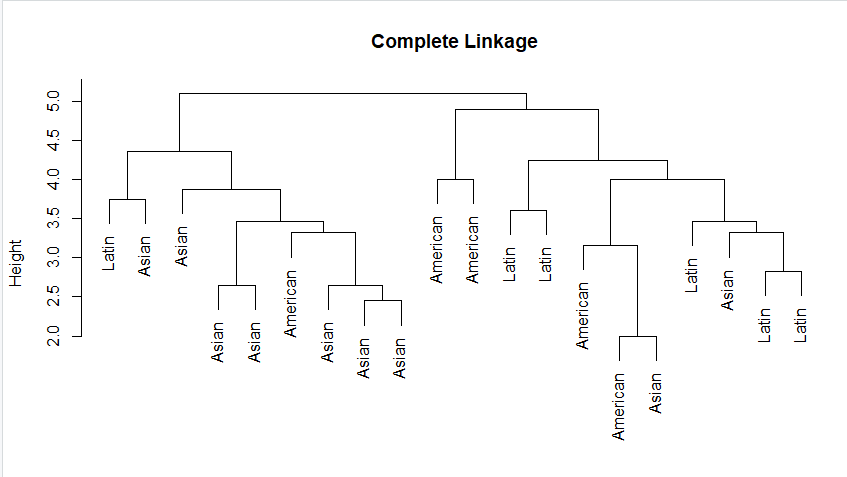
**Part 1: Hierarchical Cluster Analysis**

1. Set the seed to 1001. Take a sample of 20 instances/observations from your data. Perform a hierarchical clustering analysis on all variables except *cuisine* using the complete linkage approach (and Euclidean distance). Plot the dendrogram, using the cuisine variable as the label.

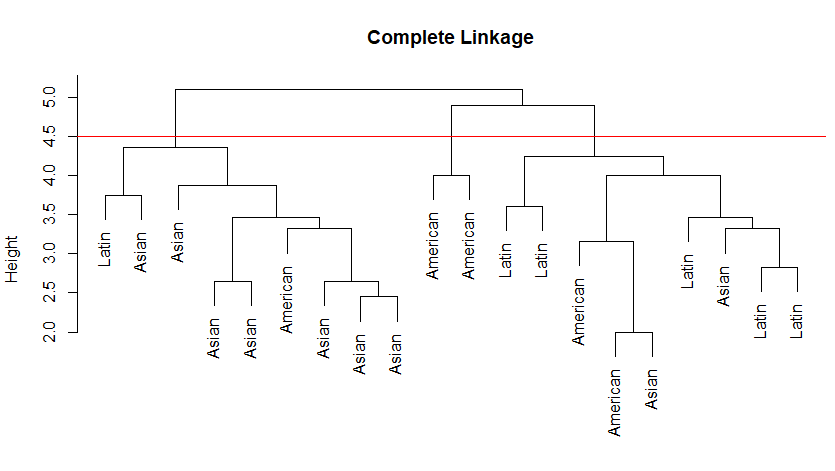
Note: to put labels on instances on a dendrogram, use the labels= option, for example:

plot(your.clustering, main=“Your Title”, **labels=recipe\_subset[,1]**, xlab=“”,sub=“”)

* 1. Paste the dendrogram for this clustering here:



* 1. If you decide that there are three main clusters in this data, indicate on your plot how you would determine the three clusters.



* 1. Describe the three clusters in terms of the main cuisines that they contain.

Cluster 1 – Asian

Cluster 2 – American

Cluster 3 – Latin

* 1. Should we have scaled (or normalized) our variables here? Explain your answer.

Scaling is only required when the units of our variables differ by large amounts (for example, a categorical variable of 0 and 1 compared to a numerical variable from 0 to 1000). Here all the variables have the same values of 0 and 1, so scaling is not required.

**Part 2: K-Means Cluster Analysis**

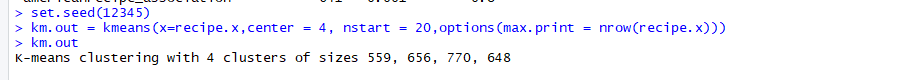
1. Set the seed to 12345. Perform a K-Means clustering of the data with *K* = 4 using all of the data except *cuisine*. Use 20 random starts to ensure that you have good clusters.
   1. How many recipes are included in each cluster?

Cluster 1 🡪 559 recipes

Cluster 2 🡪 656 recipes

Cluster 3 🡪 770 recipes

Cluster 4 🡪 648 recipes



* 1. Report the cluster mean results for five ingredients: *butter, cream, cocoa, lime\_juice,* and *cheese.* For each ingredient, how would you describe the differences among the clusters?

**Cluster mean results for butter:**

1🡪 0.79964222

2🡪 0.02743902

3🡪 0.36753247

4🡪 0.11419753



Butter was present in almost 80% of the 559 observations in cluster 1. Butter is an important ingredient of the recipes in cluster 1. In contrast, butter was rarely present in cluster 2 and cluster 4 (2% and 11% respectively), so it is not an important ingredient for recipes in cluster 2 and 4.

**Cluster mean results for cream:**

1🡪0.55456172

2🡪0.01676829

3🡪 0.14415584

4🡪 0.16512346



More than half of the recipes (55% of 559 recipes) in cluster 1 included cream. It is used much more often for recipes in cluster 1 than in the other clusters, where cream is not an important ingredient.

**Cluster mean results for cocoa:**

1🡪 0.196779964

2🡪 0.000000000

3🡪 0.001298701

4🡪 0.020061728



Again, cocoa appears to be an important ingredient for Cluster 1 compared to the other clusters. 19.68% of the recipes in Cluster 1 include cocoa, but it is used much more rarely in the other clusters.

**Cluster mean results for lime\_juice:**

1🡪 0.01967800

2🡪 0.17225610

3🡪 0.01948052

4🡪 0.29938272



Lime juice appears to be an important ingredient for Clusters 2 and 4 compared to the other clusters. 29.94% of the recipes in Cluster 4 and 17.23% of the recipes in Cluster 2 include lime juice, but it is used much more rarely in Clusters 1 and 3.

**Cluster mean results for cheese:**

1🡪 0.07871199

2🡪 0.00304878

3🡪 0.04155844

4🡪 0.22839506



Cheese appears to be an important ingredient for Cluster 4 compared to the other clusters. 22.84% of the recipes in Cluster 4 include cheese, but it is used much more rarely in the other clusters.

* 1. For each cluster, identify five “characteristic” ingredients. That is, which five ingredients are most representative of each cluster?

Note: you could have looked at this in two different ways. The easiest way is simply to sort the cluster centers and find the top ingredients by how often they appear in recipes for each cluster (i.e., the largest cluster means):

Cluster 1: egg, butter, wheat, vanilla, cream

Cluster 2: soy sauce, ginger, garlic, rice, scallion

Cluster 3: olive oil, garlic, onion, black pepper, butter

Cluster 4: cayenne, onion, garlic, tomato, cilantro

However, a better way would be to look at how often certain ingredients appear in recipes in a cluster relative to the other clusters. For example, let’s say salt appeared in 95% of all recipes. Then salt is likely to have a high cluster mean for all clusters, but that wouldn’t mean salt is a good way to describe the recipes of all four clusters. It only means salt is used heavily in all recipes. In contrast, an ingredient like cocoa is not used in most of the recipes, so even though it only has a cluster mean of 0.1968 for Cluster 1, it’s very characteristic of recipes for Cluster 1—it very rarely appears in the other clusters.

You could have answered this question either using the cluster means or by identifying the ingredients that clearly differentiate recipes, which is a better way to use clustering in practice. It frequently leads to a better overall picture of your data, though it’s also much more subjective.

* 1. Do the ingredients you identified in part (c) give you any insight into what type of recipes might be included in each cluster? Explain.

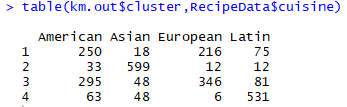
The ingredients in cluster 1 are general ingredients, such as butter, milk, egg, cane molasses, cocoa, lard, almond, nutmeg, yeast, etc. These frequently appear in baking recipes for American/European cuisine, so it is likely Cluster 1 represents baking recipes since it does not include spices at all. (Note: if you did not connect this to baking, that’s fine. It should be noted these ingredients are not associated with typically Latin or Asian cuisine.)

Cluster 2 includes ingredients like rice, soy sauce, ginger, garlic, vinegar, vegetable oil, sesame oil, fish, chicken, soybean, chicken broth, etc. These ingredients are frequently found in Asian food.

Cluster 3 includes ingredients like olive oil, garlic, onion, black pepper, butter, parsley, thyme, chicken broth, tomato, and vinegar. These ingredients are typical of savory dishes in American and European cuisine.

Cluster 4 include ingredients like cayenne, corn, onion, garlic, tomato, cilantro, lime juice, cheese, cumin, bell pepper, etc. which suggest spicy Latin food given these ingredients are frequently used in Latin recipes.

* 1. The output allows you to identify, for each recipe, the cluster they belong to using (for example) *km.out$cluster*. Create a table showing how many recipes from each cuisine belong to each cluster by combining the cluster assignments with the *cuisine* column from the original data and paste it here.



* + 1. Does this table support your conclusions in (d)? Explain.

Yes, the table output matches with the conclusions in d. Cluster 2 is primarily Asian recipes, while Cluster 4 is primarily Latin recipes. Clusters 1 and 3 are much harder to distinguish, because they contain both American and European recipes.

* + 1. Which pair of cuisines are the most similar?

American and European cuisines are most similar. They are the most represented (and approximately equal) cuisines in Clusters 1 and 3, and they both are the lowest in Clusters 2 and 4.

* + 1. Which cuisine is most associated with cluster 2? Which cuisine is most associated with cluster 3? Do the ingredients you listed in part (c) correctly describe these cuisines (according to your domain knowledge)?

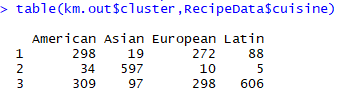
As discussed in answer d, each cuisine has unique combination of ingredients and taste. Cluster 2 completely associates with Asian cuisine. Cluster 3 is mostly associated with European and American cuisine. The ingredients in part (c) do agree with these cuisines, since as mentioned above, the characteristic ingredients appear in these cuisines often.

* + 1. If I have a new recipe that uses cane molasses, nutmeg, cinnamon, and vanilla, which cluster would you assign it to?

These ingredients are all well-represented in Cluster 1, so I should assign it to Cluster 1.

* 1. Run kmeans with *K=3* and decide which set of clusters (K=3 or K=4) is more meaningful to you.

Although the sum of squares decreases for K=4 compared to K=3 (which is usually an indication that the clustering is improving), we noted above that Clusters 1 and 3 seem to have similar cuisines. Therefore K=3 might be better. Here is the table for three clusters:



This actually separates worse than doing K=4 clusters, because now only the Asian recipes are clearly identified. The other two clusters include American, European, and Latin cuisines without drawing clear differences among them. Therefore, it appears to be better to use K-means clustering with K=4 here.