## Due Date: Friday, March 11th, by midnight , Total number of points: 50 points

**Problem 1 (25 points):**

Download the seeds dataset from [http://archive.ics.uci.edu/ml/datasets/seeds#](http://archive.ics.uci.edu/ml/datasets/seeds)

1. (15 points) Perform k-means clustering using all the attributes with the except of the class label, vary the number of clusters from 3 to 4 to 5 to 6 and report:
   1. How the cluster centers were calculated: **The clusters were calculated using the squared Euclidean distance from the center of each cluster.**
   2. What similarity measure was used: **high intra-class similarity and low inter-class similarity.**
   3. For each k, report the following:
      1. Final cluster centers



* + 1. Number of elements in each cluster

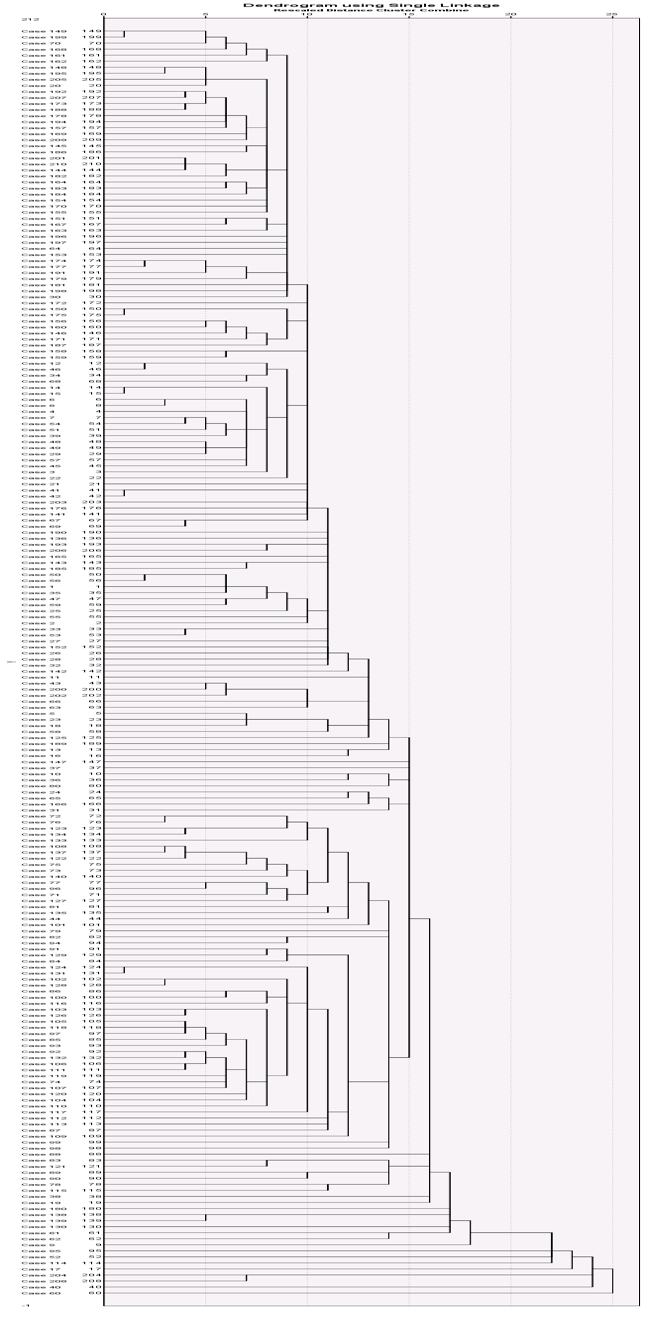


* + 1. The class distribution within each cluster



* + 1. In your opinion, which k should be selected? Explain your selection. **In my opinion I’m looking for the k number which has the purest within-class distribution from the above cross tabulation matrix. I believe that k=3 should be selected.**
    2. For the selected k in iv, analyze and report if the normalization of the attributes will influence the clustering results. **The normalization of the attributes will influence the clustering result because of the range in values for the variables (Area versus Compactness). The final cluster centers (described above) show the value ranges.**

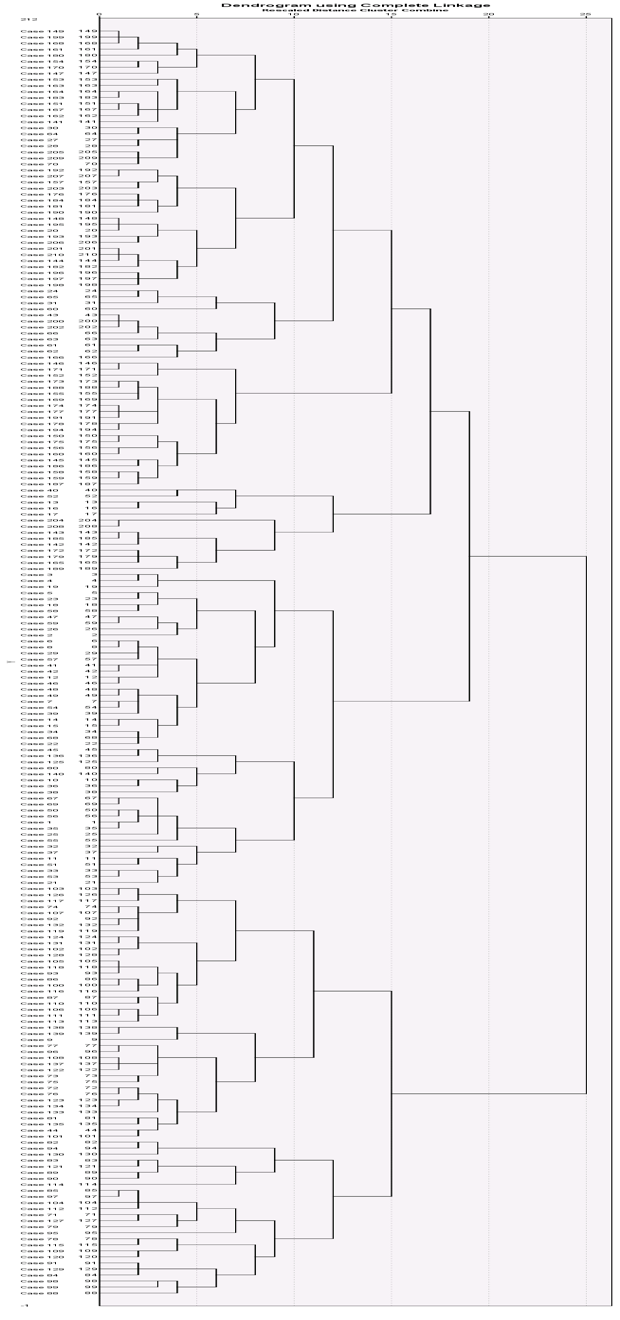
1. (10 points) Perform hierarchical clustering using all attributes except the class label as follows:
   * 1. Apply single linkage algorithm and report (**Nearest Neighbor**)
        1. The dendogram



* + - 1. The class distribution at the level of the dendogram where there are only three clusters.



* + 1. Apply complete linkage and report (**Farthest Neighbor**)
       1. The dendogram



* + - 1. The class distribution at level of the dendogram where there are only three clusters.



1. (2.5 points) Compare the results with hierarchical clustering and k-means algorithm. **The k-means clustering method produced more within-class similarity as shown by comparing the two class distribution tables above and therefore the best method to utilize would be the k=3 means method.**
2. (2.5 points) Create an executive summary (~half a page) that outlines the problem, summarizes the data, describes the methodology, summarizes the results, and makes recommendations. When creating it, imagine that you will give this summary to someone who is not an expert in data mining.

# Problem Statement: This is a clustering analysis problem to determine and assign a label to the seeds dataset which will indicate its wheat group based on 7 continuous values.

# Summary of Data: The seeds data set comes from the UCI Machine Learning Repository. The research examined a group comprised kernels belonging to three different varieties of wheat: *Kama, Rosa and Canadian*, 70 elements each, randomly selected for the experiment. Measurements of geometrical properties of kernels belonging to three different varieties of wheat. A soft X-ray technique and GRAINS package were used to construct all seven, real-valued attributes.

Attribute information:

To construct the data, seven geometric parameters of wheat kernels were measured:   
1. area A,   
2. perimeter P,   
3. compactness C = 4\*pi\*A/P^2,   
4. length of kernel,   
5. width of kernel,   
6. asymmetry coefficient   
7. length of kernel groove.   
All of these parameters were real-valued continuous.

The last attribute in the data file represents the class label.

# Methods:

1. Download the data from <http://archive.ics.uci.edu/ml/datasets/seeds#>
2. Clean up data in excel, that has been shifted and add corresponding column headers (not necessary but it will make it easier)
3. Open up IBM SPSS 23 and load the dataset and ensure everything transferred nicely
4. For k-means clustering:
   1. Analyze > Classify > k-means
   2. Select all of the independent variables (except the class label) in put in the ‘variables’
   3. Put in the amount of clusters on the main window
   4. In the save settings: ☑ Cluster Membership, Set max refractions to 100
   5. Press OK
5. For Hierarchical clustering:
   1. Analyze > Classify > k-means
   2. Select all of the independent variables (except the class label) in put in the ‘variables’
   3. In the statistics settings: ☑ Agglomerative, ☑ Dendrogram
   4. Select in the Methods settings: Nearest Neighbor (Single linkage), Farthest Neighbor (Complete linkage), Euclidian Distance measure, implement z-score normalization.

# Results:

The results can be assessed in a variety of methods from the output tables but the cross tabulation tables from the actual and predicted from each analysis methods can lead to recommendations:



# Recommendations:

According to the results my recommendation would be to proceed with (k=3) k-means clustering as it produces the purest within-class similarity for classification.

**Problem 2 (25 points):**

On the same data used in Problem 1, create a decision tree classification model for the three different varieties of wheat: Kama, Rosa and Canadian.

* 1. Use 10-fold cross validation and at least five different configurations to produce a decision tree classifier. Report the results obtained for the different configurations and chose one as being the best among the configurations you tried. Explain your answer. **The tables below represent the 6 configurations I tried (with the criteria that was changed in green). The last configuration for with maximum tree depth = 25, Parent node minimum cases = 10, and Child node minimum cases = 5 is the best model compared to the other configurations. At that configuration is the point of improvement from the other previous configurations while also the percent correct does not improve as the parameters are decreased. [The appendix includes the tree diagrams and table calculations for all tested configurations]**

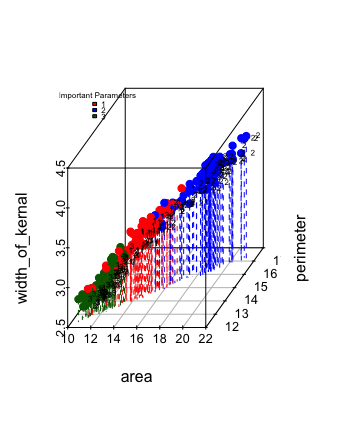


* 1. For the best tree configuration, report the misclassification matrix and interpret it. In your opinion, is accuracy a good way to interpret the performance of the model? If not, suggest other measures. **For the misclassification table below the model configuration produces very pure within-class similarity – meaning there are very few misclassifications in the predicted space. Accuracy is a good way to interpret the model as the sample size is large, the class are evenly distributed and the cross-validation method is robust.**



* 1. What are the most important three attributes for classifying the wheat data? **The three most important variables for classifying the wheat data are: area, perimeter, and width of kernel as listed below in the table output.**



* 1. Create a graph that will allow you to visualize the data in the 3-dimensional space of the most important attributes. Interpret the graph. **A 3D scatterplot can expose underlying relationships in the data of 3 continuous variables and in this case the most important variables. The Figure 2 created in RStudio, shows that between the three variables are moderately define classes indicated by the colors which indicate the distinct classification which is ideal in determining the most important variables.**

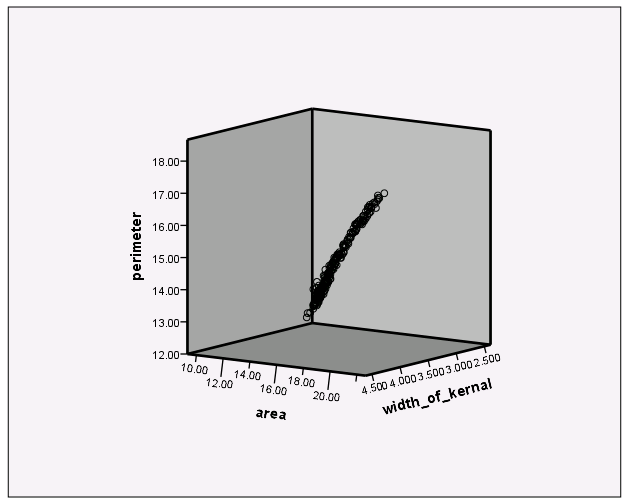


Figure : SPSS 3D Scatterplot

Figure 2: R 3D Scatterplot

* 1. Are there any other techniques that can help identify variables for data visualization? Explain your answer and include any analysis you will perform to answer this question. **Other Data visualization techniques include scatter matrices, boxplots, and violin plots. You can use an automated feature selection algorithm for variable selection such as forward, backward, stepwise selection, which will identify the most important variables to be included in the model. The following R script shows the algorithm for forward selection which will identify the variables that reduce the AIC value. It is very interesting to note that the algorithm did not include the variables (area, perimeter, width\_of\_kernal) highlighted in yellow in the code as the Decision tree picked as most important.**

# linear model  
fit <- lm(class ~ area + perimeter + compactness +   
 length\_of\_kernal + width\_of\_kernal +   
 asymmetry\_coefficient + length\_of\_kernal\_grove, data = seeds)   
summary(fit)

##   
## Call:  
## lm(formula = class ~ area + perimeter + compactness + length\_of\_kernal +   
## width\_of\_kernal + asymmetry\_coefficient + length\_of\_kernal\_grove,   
## data = seeds)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.30568 -0.24785 -0.01632 0.24198 1.22362   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 53.44356 7.44511 7.178 1.32e-11 \*\*\*  
## area 1.48907 0.26133 5.698 4.25e-08 \*\*\*  
## perimeter -3.22038 0.53815 -5.984 9.77e-09 \*\*\*  
## compactness -30.67744 5.24108 -5.853 1.92e-08 \*\*\*  
## length\_of\_kernal -2.31510 0.45444 -5.094 8.01e-07 \*\*\*  
## width\_of\_kernal 0.24598 0.78571 0.313 0.755   
## asymmetry\_coefficient 0.11489 0.02257 5.089 8.19e-07 \*\*\*  
## length\_of\_kernal\_grove 2.19260 0.20358 10.770 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.415 on 202 degrees of freedom  
## Multiple R-squared: 0.7515, Adjusted R-squared: 0.7428   
## F-statistic: 87.25 on 7 and 202 DF, p-value: < 2.2e-16

# forward selection  
min.model = lm(class ~ 1, data=seeds) # aka intercept only model  
biggest = formula(lm(class ~ area + perimeter + compactness +   
 length\_of\_kernal + width\_of\_kernal +   
 asymmetry\_coefficient + length\_of\_kernal\_grove, data = seeds) ) # the full model  
model = step(min.model, direction='forward', scope=biggest)

## Start: AIC=-83.15  
## class ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + asymmetry\_coefficient 1 46.654 93.346 -166.267  
## + compactness 1 39.476 100.524 -150.708  
## + width\_of\_kernal 1 25.105 114.895 -122.649  
## + area 1 16.766 123.234 -107.934  
## + perimeter 1 15.053 124.947 -105.035  
## + length\_of\_kernal 1 9.266 130.734 -95.528  
## <none> 140.000 -83.148  
## + length\_of\_kernal\_grove 1 0.083 139.917 -81.272  
##   
## Step: AIC=-166.27  
## class ~ asymmetry\_coefficient  
##   
## Df Sum of Sq RSS AIC  
## + compactness 1 18.1451 75.201 -209.66  
## + width\_of\_kernal 1 11.3020 82.044 -191.37  
## + area 1 6.7386 86.607 -180.00  
## + perimeter 1 6.0216 87.324 -178.27  
## + length\_of\_kernal 1 3.6115 89.734 -172.55  
## <none> 93.346 -166.27  
## + length\_of\_kernal\_grove 1 0.1319 93.214 -164.56  
##   
## Step: AIC=-209.66  
## class ~ asymmetry\_coefficient + compactness  
##   
## Df Sum of Sq RSS AIC  
## + length\_of\_kernal\_grove 1 1.99051 73.210 -213.29  
## <none> 75.201 -209.66  
## + length\_of\_kernal 1 0.25395 74.947 -208.37  
## + perimeter 1 0.15260 75.048 -208.09  
## + width\_of\_kernal 1 0.09095 75.110 -207.91  
## + area 1 0.02433 75.176 -207.73  
##   
## Step: AIC=-213.29  
## class ~ asymmetry\_coefficient + compactness + length\_of\_kernal\_grove  
##   
## Df Sum of Sq RSS AIC  
## + length\_of\_kernal 1 31.935 41.275 -331.64  
## + perimeter 1 25.289 47.921 -300.29  
## + area 1 20.475 52.735 -280.18  
## + width\_of\_kernal 1 15.820 57.390 -262.42  
## <none> 73.210 -213.29  
##   
## Step: AIC=-331.64  
## class ~ asymmetry\_coefficient + compactness + length\_of\_kernal\_grove +   
## length\_of\_kernal  
##   
## Df Sum of Sq RSS AIC  
## <none> 41.275 -331.64  
## + area 1 0.296664 40.978 -331.15  
## + width\_of\_kernal 1 0.146829 41.128 -330.39  
## + perimeter 1 0.040096 41.235 -329.84

summary(model)

##   
## Call:  
## lm(formula = class ~ asymmetry\_coefficient + compactness + length\_of\_kernal\_grove +   
## length\_of\_kernal, data = seeds)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.11443 -0.25620 0.00504 0.29770 1.22321   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.53844 1.25242 8.414 6.73e-15 \*\*\*  
## asymmetry\_coefficient 0.13656 0.02344 5.827 2.17e-08 \*\*\*  
## compactness -7.54868 1.52941 -4.936 1.65e-06 \*\*\*  
## length\_of\_kernal\_grove 2.58439 0.19984 12.932 < 2e-16 \*\*\*  
## length\_of\_kernal -2.92179 0.23199 -12.594 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4487 on 205 degrees of freedom  
## Multiple R-squared: 0.7052, Adjusted R-squared: 0.6994   
## F-statistic: 122.6 on 4 and 205 DF, p-value: < 2.2e-16

**Submission Instructions**

1. Solve the problems and write your answers in a Word document.
2. Submit your file online at the website at <http://d2l.depaul.edu> in the Submission page
3. Keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.