Assignment 5

Lavinia Wang #1474703

Due Date: Wednesday, March 14th, 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#**

The examined data group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin.

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.

i. (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:
   a. How the cluster centers were calculated

Ans:
Given k, we pick k seed points as initial centers for k clusters. Then calculate the distance to each seed point and assign point to the nearest clusters. After assignment, recalculate new cluster center and iterate calculation and assignment until no more assignment is needed.

   b. What similarity measure was used

Ans：

Case1	15.26	14.84	0.87	5.76	3.31	2.22	5.22
Cluster Center3	14.65	14.46	0.88	5.56	3.28	2.65	5.19

Table 1: Example of SPSS output calculation

Calculate the distance between case 1 and cluster center 3:

$$d = \sqrt{(15.26-14.65)^2+(14.84-14.46)^2+(0.87-0.88)^2+(5.76-5.56)^2+(3.31-3.28)^2+(2.22-2.65)^2+(5.22-5.19)^2} = 0.86189$$

This number matches what we got from SPSS. So the similarity measure was Euclidean distance.

   c. For each k, report the following:

i. Final cluster centers

Final Cluster Centers			
	Cluster		
	1	2	3
area A	18.72	11.96	14.65
perimeter P	16.30	13.27	14.46
compactness C = 4*pi*A/P^2	0.8851	0.8522	0.8792
length of kernel	6.2089	5.2293	5.5638
width of kernel	3.723	2.873	3.278
asymmetry coefficient	3.6036	4.7597	2.6489
length of kernel groove	6.066	5.089	5.192

Final Cluster Centers				
	Cluster			
	1	2	3	4
area A	11.94	14.42	17.75	19.52
perimeter P	13.27	14.35	15.88	16.65
compactness C = 4*pi*A/P^2	0.8515	0.8795	0.8840	0.8844
length of kernel	5.2292	5.5239	6.0476	6.3501
width of kernel	2.867	3.253	3.614	3.812
asymmetry coefficient	4.8040	2.5904	3.1649	4.1641
length of kernel groove	5.095	5.127	5.921	6.184

Final Cluster Centers					
	Cluster				
	1	2	3	4	5
area A	16.56	14.69	19.15	12.09	11.98
perimeter P	15.39	14.47	16.47	13.31	13.29
compactness C = 4*pi*A/P^2	0.8782	0.8809	0.8871	0.8571	0.8508
length of kernel	5.8882	5.5721	6.2689	5.2174	5.2414
width of kernel	3.481	3.286	3.773	2.901	2.880
asymmetry coefficient	4.1095	2.4079	3.4604	3.3438	5.6733
length of kernel groove	5.725	5.159	6.127	5.005	5.122

Final Cluster Centers						
	Cluster					
	1	2	3	4	5	6
area A	11.83	14.24	16.41	18.95	12.32	19.58
perimeter P	13.22	14.26	15.32	16.39	13.42	16.65
compactness C = 4*pi*A/P^2	0.8500	0.8793	0.8783	0.8868	0.8580	0.8877
length of kernel	5.2156	5.4935	5.8640	6.2475	5.2659	6.3159
width of kernel	2.844	3.234	3.463	3.745	2.951	3.835
asymmetry coefficient	4.1684	2.3165	3.8501	2.7235	6.3367	5.0815
length of kernel groove	5.076	5.062	5.690	6.119	5.122	6.144

Table 2: Final Cluster Centers for k = 3 - 6

ii. Number of elements in each cluster

Number of Cases in each Cluster		
Cluster	1	61
	2	77
	3	72
Valid		210
Missing		0

Number of Cases in each Cluster		
Cluster	1	75
	2	67
	3	40
	4	28
Valid		210
Missing		0

Number of Cases in each Cluster		
Cluster	1	25
	2	51
	3	48
	4	44
	5	42
Valid		210
Missing		0

Number of Cases in each Cluster		
Cluster	1	56
	2	54
	3	31
	4	33
	5	21
	6	15
Valid		210
Missing		0

Table 3: Number of cases in each cluster for k = 3 - 6

iii. The class distribution within each cluster

Class distribution with each cluster						
k=3	1	2	3			
Actual class = 1	1	9	60			
Actual class = 2	60	0	10			
Actual class = 3	0	68	2			
k=4	1	2	3	4		
Actual class = 1	8	58	4	0		
Actual class = 2	0	6	36	28		
Actual class = 3	67	3	0	0		
k=5	1	2	3	4	5	
Actual class = 1	6	48	0	14	2	
Actual class = 2	19	3	48	0	0	
Actual class = 3	0	0	0	30	40	
k=6	1	2	3	4	5	6
Actual class = 1	7	52	9	0	2	0
Actual class = 2	0	0	22	33	0	15
Actual class = 3	49	2	0	0	19	0

Table 4: Class distribution within each cluster

iv. In your opinion, which k should be selected? Explain your selection.

	S_W	S_B	S_W/S_B
k=3	587.3186	10.24373	57.33445
k=4	516.3976	11.91074	43.35563
k=5	385.5073	11.12936	34.63877
k=6	336.626	12.12873	27.75443

Table 5: Scatter matrix

From the calculated scatter matrix, we can see as we increase k, ratio of within-cluster and between-cluster decreases. But this also increases the percentage of impurity within-clusters. 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.

v. 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). After normalization, number of cases in each cluster is closer to the actual class.

Number of Cases in each Cluster		
Cluster	1	76

	2	69
	3	65
Valid		210
Missing		0

Table 6: Number of cases in each cluster for k = 3 on normalized data

ii. (10 points) Perform hierarchical clustering using all attributes except the class label as follows:

i. Apply single linkage algorithm and report

1. The dendrogram

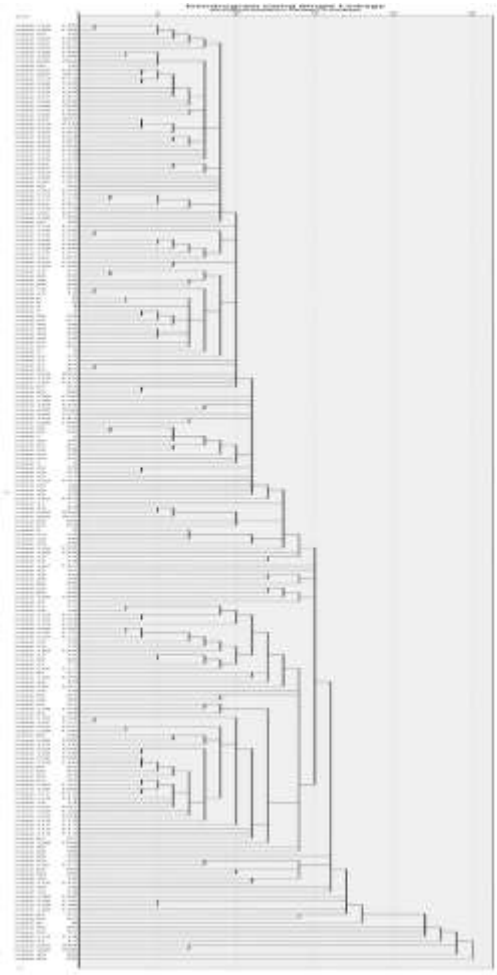

Figure 1: Dendrogram using single linkage

2. The class distribution at the level of the dendrogram where there are only three clusters.

Class distribution with each cluster(Single linkage)				
Count	Class			Total
	1	2	3	
Actual class = 1	69	1	0	70
Actual class = 2	69	1	0	70
Actual class = 3	69	1	0	70
Total	208	5	3	210

Table 7: Class distribution within each cluster using single linkage algorithm

ii. Apply complete linkage and report

1. The dendrogram

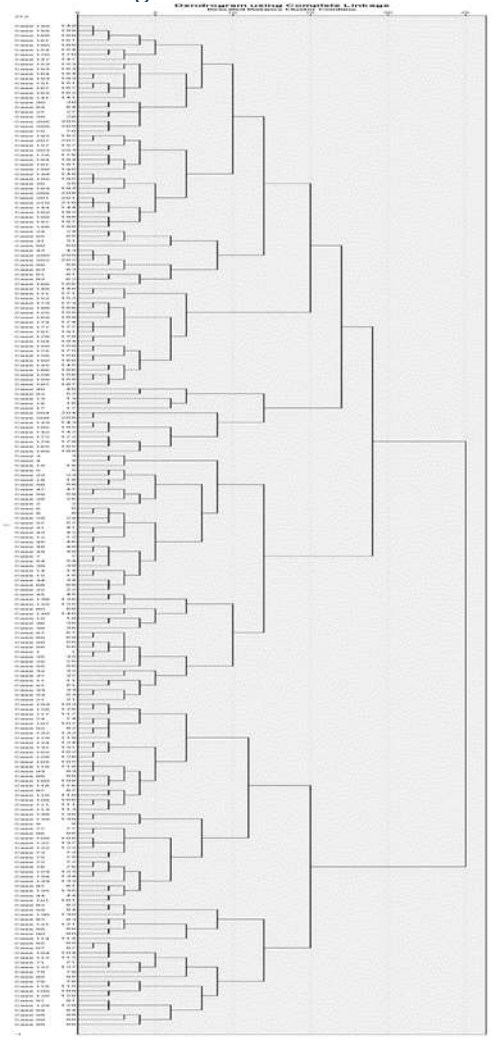

Figure 2: Dendrogram using complete linkage

2. The class distribution at level of the dendrogram where there are only three clusters.

Class distribution with each cluster(Complete linkage)				
Count	Class			Total
	1	2	3	
Actual class = 1	49	4	17	70
Actual class = 2	13	57	0	70
Actual class = 3	0	0	70	70
Total	63	63	90	210

Table 8: Class distribution within each cluster using complete linkage algorithm

iii. (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.

iv. (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 labels to the seeds dataset which will indicate wheat group based on 7 continuous attributes.

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. Mismatched data were shifted and add corresponding column.
3. Run IBM SPSS 24 and load the dataset and ensure everything transferred nicely
4. For k-means clustering:
    a. Analyze > Classify > k-means cluster analysis
    b. Select all of the independent variables (except the class label) in put in the 'variables'
    c. Put in the number of clusters on the main window (from 3 - 6)
    d. In the save settings: ☑ Cluster Membership ☑ Distance from Cluster Center
    e. Press OK

5. For Hierarchical clustering:
    a. Analyze > Classify > hierarchical cluster analysis
    b. Select all of the independent variables (except the class label) in put in the 'variables'
    c. In the statistics settings: ☑ Agglomerative, ☑ Dendrogram
    d. Select in the Methods settings: Nearest Neighbor (Single linkage), Furthest 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 crosstabs of actual and predicted from each analysis methods can lead to recommendations:

Class distribution with each cluster(k-means where k=3)				
Count	Class			Total
	1	2	3	
Actual class = 1	1	9	60	70
Actual class = 2	60	0	10	70
Actual class = 3	0	68	2	70
Total	62	79	75	210

Class distribution with each cluster(Single linkage)				
Count	Class			Total
	1	2	3	
Actual class = 1	69	1	0	70
Actual class = 2	69	1	0	70
Actual class = 3	69	1	0	70
Total	208	5	3	210

Class distribution with each cluster(Complete linkage)				
Count	Class			Total
	1	2	3	
Actual class = 1	49	4	17	70
Actual class = 2	13	57	0	70
Actual class = 3	0	0	70	70
Total	63	63	90	210

Table 9: Class distribution within each cluster comparison

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.

a. 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 9 configurations I tested. The 6th configuration for with tree depth = 5, Parent node minimum cases = 15, Child node minimum cases = 7 and complexity = 6 is the best model compared to the other configurations. At that configuration is the point of improvement of other previous configurations while accuracy does not improve as the parameters are decreased. [The appendix includes the tree diagrams and table calculations for all tested configurations]

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	55	1	14	78.6%
2	2	68	0	97.1%
3	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%
Growing Method: CRT Dependent Variable: wheat class				
Np=**35**, Nc=**17**, complexity = **3**, depth=**2**				

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	55	1	14	78.6%
2	2	68	0	97.1%
3	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%
Growing Method: CRT Dependent Variable: wheat class				
Np=**25**, Nc=**12**, complexity = **3**, depth=**2**				

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	55	1	14	78.6%
2	2	68	0	97.1%
3	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%
Growing Method: CRT Dependent Variable: wheat class				
Np=**20**, Nc=**10**, complexity = **3**, depth=**2**				

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	55	1	14	78.6%
2	2	68	0	97.1%
3	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%
Growing Method: CRT Dependent Variable: wheat class				
Np=**18**, Nc=**9**, complexity = **3**, depth=**2**				

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	62	1	7	88.6%
2	2	68	0	97.1%
3	1	0	69	98.6%
Overall Percentage	31.0%	32.9%	36.2%	94.8%
Growing Method: CRT Dependent Variable: wheat class				
Np=**17**, Nc=**8**, complexity = **5**, depth=**4**				

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	68	1	1	97.1%
2	2	68	0	97.1%
3	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%
Growing Method: CRT Dependent Variable: wheat class				
Np=**15**, Nc=**7**, complexity = **6**, depth=**5**				

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	68	1	1	97.1%
2	2	68	0	97.1%
3	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%
Growing Method: CRT Dependent Variable: wheat class				
Np=**14**, Nc=7, complexity = **6**, depth=**5**				

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	68	1	1	97.1%
2	2	68	0	97.1%
3	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%
Growing Method: CRT Dependent Variable: wheat class				
Np=**12**, Nc=**6**, complexity = **7**, depth=**5**				

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	68	1	1	97.1%
2	2	68	0	97.1%
3	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%
Growing Method: CRT Dependent Variable: wheat class				
Np=**10**, Nc=**5**, complexity = **7**, depth=**5**				

Table 10: Class distribution using decision tree

b. 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.

Classification				
Observed	Predicted			
	1	2	3	Percent Correct
1	68	1	1	97.1%
2	2	68	0	97.1%
3	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%
Growing Method: CRT Dependent Variable: wheat class				
Np=**15**, Nc=**7**, complexity = **6**, depth=**5**				

Table 11: Best configuration of decision tree misclassification matrix

For the misclassification table, the model configuration produces very pure within-class similarity – meaning there are very few misclassifications in the predicted value. Accuracy is a good way to interpret the model as the class are evenly distributed, false positive and false negative are very small and the cross-validation method is robust.

c. What are the most important three attributes for classifying the wheat data?

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
area A	.499	100.0%
perimeter P	.484	97.0%
width of kernel	.426	85.4%
length of kernel	.395	79.1%
length of kernel groove	.381	76.3%
asymmetry coefficient	.200	40.0%
compactness C = 4*pi*A/P^2	.175	35.0%

Growing Method: CRT

Dependent Variable: wheat class

Table 11: Independence variable importance

The three most important variables for classifying the wheat data are: area A, perimeter P, and width of kernel.

d. 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. From figure 3, it shows that between the three variables moderately define classes denoted by different colors which indicate the distinct classification which is ideal in determining the most important variables.

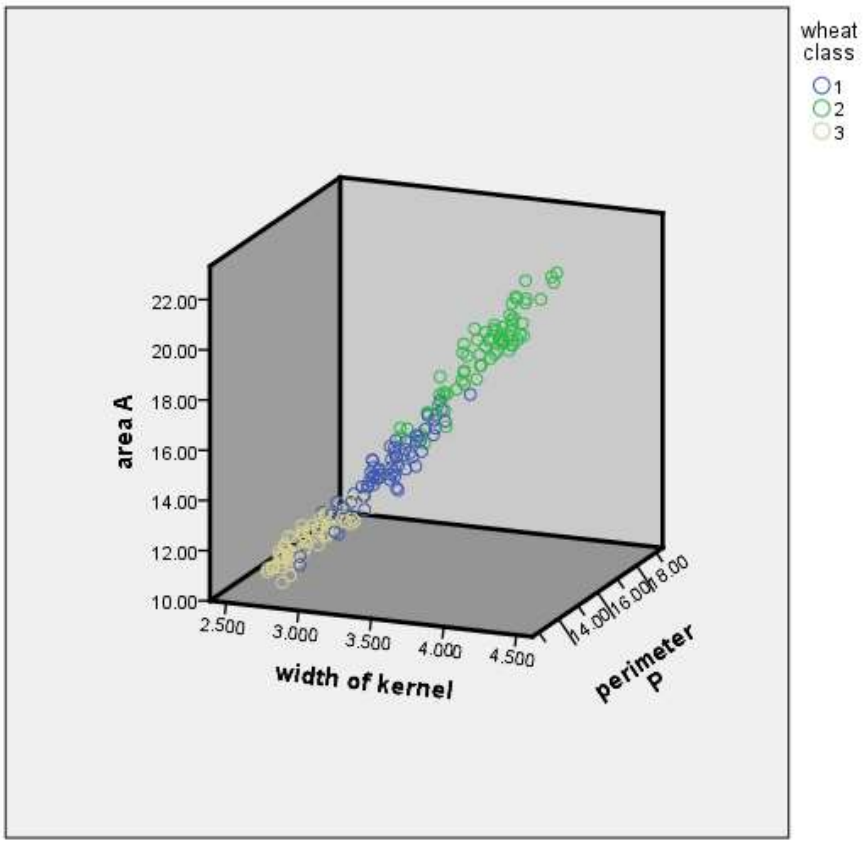

Figure 3: 3D scatterplot of wheat data

e. 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 matrix, histogram, and network. 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. Because for this data set, our dependent variable is discrete, instead of using linear regression which is predicting continuous output, I would perform multinomial logistic regression analysis and apply stepwise selection for independent variables.