Data Mining in Agriculture

Shweta Garg
Student of MS in Data Science program
The Graduate Center of the City University of New York
May 8, 2018

Introduction

Data mining is a tool to discover patterns by extracting information from the datasets and transforming it into a meaningful structure by techniques used in machine learning and statistics. It involves different methods such as association rule, clustering, classification, regression and dimension reduction. One or more of these methods is selected based on requirement of the user. Now a day's big-data is coming from different fields such as healthcare, education, product analysis, agriculture and may more. By applying data mining techniques, raw or unordered data can be converted to useful information. In this report, we have worked on data from agricultural field and applied some of the above-mentioned methods on it to get structure and meaningful information from the data.

Dataset Description:

In this report, the data used is available on UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/seeds). The dataset consists of geometrical properties of different variety of wheat kernels. It was first compiled by M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. The seeds are explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. Aim of this research was to differentiate between three types of wheat kernels (Kama, Rosa and Canadian) based on their geometrical properties. Studies were conducted using combine harvested wheat grain originating from experimental fields. These wheat kernels are required to be separate, since all have different financial revenues.

Soft X-ray technique was used for high quality visualization; they scanned X-ray photograms using the Epson Perfection V700 table photo-scanner with a built-in transparency adapter, 600 dpi resolution and 8bit gray scale levels. GRAINS package was used for analysis procedures of obtained bitmap graphics files, specially developed for X-ray diagnostic of wheat kernels. These experiments were conducted to observe the internal kernel structure of randomly selected 70 samples of each type of wheat (210 in total). The images were recorded on 13x18 cm X-ray KODAK plates

The dataset consists of 7 physical features of each variety of wheat. All these parameters are real-valued continuous. The goal of current study is to differentiate between three types of wheat kernels based on their geometrical properties.

The features are as follows:

1.Area (A)	5.Perimeter(P)
2.Compactness ($C= 4 * \pi * A/P^2$)	6.Length of kernel
3.Width of kernel	7.Asymmetry Coefficient
4.Length of groove	

Moreover, class of each datapoint was also given. All features were quantitative type.

Methods

Clustering: Clustering is an unsupervised learning technique and a descriptive analysis method. It gives a way of grouping objects in a subgroup by keeping the partition in a way that objects within groups are more similar (in some sense) than objects between different groups. Here, object means data point. The notion of similarity varies in different algorithms. Clustering is done using distance between data points or the distance from centroid.

Few methods of clustering are as follows:

K-means Clustering: It targets to partition the n datapoints into k subgroups or clusters in which each datapoint belongs to the cluster with the nearest mean.

Requirements: All the attributes of dataset should be quantitative and the measure of distance is squared Euclidean distance. Moreover, predefined number of clusters and initial center of cluster is a basic need of this algorithm.

The algorithm consists of following steps:

1. Choose the number of clusters k and choose k initial center points C_1, C_2 , ….C_k.
2. Calculate the Euclidean distance from each point to each center and assign it to the cluster with nearest center.
3. Calculate the mean of the subgroups generated by step 2 and consider these means as new center points C_1, C_2 , ….C_k.

   Repeat step 2 and 3 until algorithm converges.

K-means clustering algorithm works well under the following assumptions:

a. No outlier in the dataset.
b. Data is spherically distributed within clusters.
c. All clusters have almost same size.

d. Data has equal variance of each attribute.
e. Numbers of cluster are known.

It can give us disastrous result if any of the above conditions are not met. Moreover, choosing a wrong k and choosing a wrong initial center can also affect our result terribly. For resolving this issue, we can choose the value of k by using elbow curve (percentage of variance explained vs number of clusters) and we can also wisely pick k datapoints as our initial cluster centers.

Computational complexity:

If k (number of clusters and d (the dimension) are fixed, the problem can be exactly solved in time $O(n^{(dk+1)})$, here n is the number of datapoints to be clustered

Hierarchical Clustering:

Hierarchical clustering is a method which tries to find a hierarchy of clusters. It falls into two types:

Agglomerative: Agglomerative clustering is also known as bottom up approach. We consider each datapoint as its own cluster and merge the clusters based on linkage methods as we go up on hierarchy until one cluster is left. Compute the distance matrix and update it after each step of hierarchy.

Divisive: It is also known as top down approach. All datapoints start in one cluster, and splits are performed recursively as one moves down the hierarchy based on some linkage method until we get cluster consisting one single point. Compute the distance matrix and update it after each step of hierarchy.

The results of hierarchical clustering are can be shown as a dendrogram. The greater the difference in height in dendrogram (Tree like structure), more is dissimilarity.

Linkage methods and their condition of applicability:

Let C_i and C_j are two clusters and for a given ε

Single Linkage: Single-link distance between clusters C_i and C_j is the minimum distance between any datapoint in C_i and C_j. It is defined as:

$$\min \{ d(x,y) \mid \exists\, x \in C_i, y \in C_j \} < \boldsymbol{\varepsilon}$$

It is sensitive to noise and outliers and it may produce long, elongated clusters, but it can handle non-elliptical shapes.

Complete Linkage: Complete-link distance between clusters C_i and C_j is the maximum distance between any datapoint in C_i and C_j. It is defined as:

$\max \{ d(x,y) | \exists\ x \in C_i, y \in C_j \} < \varepsilon$

It gives more balanced clusters (with equal diameter) and less susceptible to noise or chaining, but all clusters tend to have the same diameter. As a result, small clusters can merge with larger ones.

Average Linkage: Average distance between clusters C_i and C_j is the average distance between any datapoint in C_i and C_j..

It is less susceptible to noise and outliers but biased towards globular clusters

Ward's method: It merges the two cluster when the pair minimize the increase in total within cluster variance.

It is less susceptible to noise and outliers but biased towards globular clusters.

COMPLEXITY:

For a dataset consisting of n points, Hierarchical agglomerative clustering has a time complexity of O^{n^3} and requires O^{n^2} memory. Because of its complexity we usually apply hierarchical clustering on small datasets.

Dimension reduction: Dimension reduction provides a way to visualize high dimensional data and summarizes the most important information. Principal component analysis is one of the methods of dimensionality reduction.

Principal component analysis (PCA): It is a statistical procedure which transforms a list of possibly correlated variables to linearly independent variables (called principal components). These new set of independent variables are orthogonal to each other. This transformation is defined in such a way that each preceding component captures as much of the original variance in the data as possible.

Mathematically, it can be viewed as a rotation of the existing axes to new positions in the space defined by original variables. New axes are orthogonal and represent the directions with maximum variability. Steps of PCA are as follows:

1. Subtract mean from datapoint for all attributes.
2. Calculate the co-variance matrix, its eigen values and orthonormal eigen vectors.

3. Choose first eigen vector as a direction of first principal component whose eigen value is maximum among the rest. Follow the process, this gives you the components in order of significance
4. Ignore the components of less significance which will be measured by variance or diagonal entries of co-variance matrix. Suppose we have 7 attributes and we are choosing 2 PCA components then dataset will reduce from n dimension to 2 dimensions.

Classification: It is a technique of supervised learning. It is a predictive analysis method, where we classify the new datapoint based on classifier generated on training dataset. Training dataset is a set of labeled data. We will discuss few algorithms of classification:

Logistic regression: A linear regression would be unsuitable for estimating probability, as it gives the output in the range of (-∞, ∞). Logistic regression overcomes the limitation of linear regression by providing class probability estimates. We achieve this target by using the logit function which gives us the output in a range [0,1]. Suppose (x_1 , y_1), (x_2 , y_2)….., (x_n , y_n) is a list of labeled training data where x_i is an observation and y_i is a class. $y_i \in \{ 1,2\}$classes and x is a n dimensional vector. We define the logistic function as:

$$p(y_i /x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}$$

Here β is log odds ratio associated with predictors. We can also write it as:

$$\ln\left(\frac{p(y_i/x)}{1-p(y_i/x)}\right) = \beta_0 + \beta_1 x$$

In case of multiple class, we calculate the probability class y_i with respect to every other class and finally assign x to the class which has maximum probability.

$$\ln\frac{Pr(Y_i=1)}{Pr(Y_i=K)} = \beta_1 \cdot X_i$$

Here k= 1, 2, 3….m are different class. In other words we compute the softmax function:

$$\text{softmax}(k, x_1, \dots, x_n) = \frac{e^{x_k}}{\sum_{i=1}^{n} e^{x_i}}$$

This function behaves as an equivalence of application of 2 class versus k class logistic regression.

Assumptions: Logistic regression assumes the datapoints to be independent of each other. It also requires no co-linearity in the variables and independent variables are linearly related to the log adds.

Linear Discriminant analysis (LDA): It is method of classification which assumes that class conditional probability $P(x/y_i)$ follows Gaussian distribution for each class and they all shares a common covariance matrix. The main difference between logistic regression and LDA is that here we estimate P(x/y) instead of estimating P(y/x). For this we calculate the discriminant δ as follows.

$$\delta_k = x^T \Sigma_1^{-1} \overrightarrow{\mu_1} - \frac{1}{2} \overrightarrow{\mu_j^T} \Sigma_0^{-1} \overrightarrow{\mu_j} + \ln(p_k)$$

This expression is known as discriminant and we will assign x to the class which has highest discriminant value. Here p_k is the prior probability of class k, $\overrightarrow{\mu_j}$is the mean of x in class j and $\sum$ is a covariance matrix.

Finally, we calculate decision boundary between class j and k is defined as:

$$x^T \Sigma_1^{-1} \overrightarrow{\mu_1} - \frac{1}{2} \overrightarrow{\mu_j^T} \Sigma_0^{-1} \overrightarrow{\mu_j} + ln(p_i) = x^T \Sigma_1^{-1} \overrightarrow{\mu_1} - \frac{1}{2} \overrightarrow{\mu_k^T} \Sigma_0^{-1} \overrightarrow{\mu_k} + ln(p_k)$$

Assumptions: Each feature is normally distributed. Variance among features should be same and high co-relation in features will decrease the power of prediction.

Quadratic Discriminant Analysis (QDA): The assumption that the inputs of every class have the same covariance Σ can be quite restrictive in LDA. In quadratic discriminant analysis we estimate a mean $\overrightarrow{\mu_k}$

and a covariance matrix Σ_k for each class separately under the assumption that p(x/y) follows gausian distribution. Given a input x we compute δ as follows:

$$\delta_j(x) = -\frac{1}{2} \log|\Sigma_j| - \frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) + \log \pi_j$$

This objective is now quadratic in x and so are the decision boundaries.

Decision Tree: It is a tree like structure, where we split the data space by choosing one attribute at a time. We choose the attribute in a way which minimize impurity or maximize the probability of classes within nodes and leaves of tree. We make regions as homogeneous as possible. Here we calculate the impurity or misclassification error by Gini index or cross-entropy

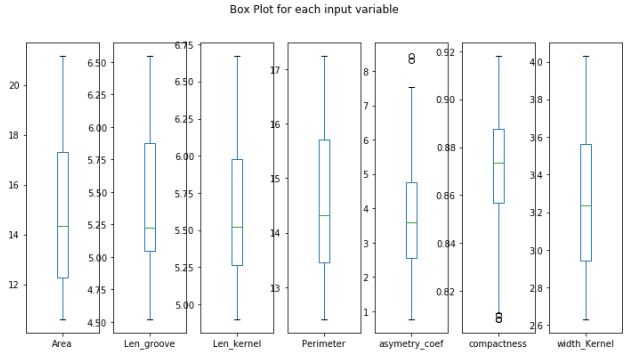
K-cross validation: We split our data into k sets. Take (k-1) together as a training set and remaining k^{th} as a test set. We keep on doing this k times by exchanging test set with a part of training set. We develop

our model/classifier on each training set and validate our results on test set. We average the rate of accuracy given by k test set and consider it as a final accuracy of classifier.

Exploratory Data Analysis and Data Visualization:

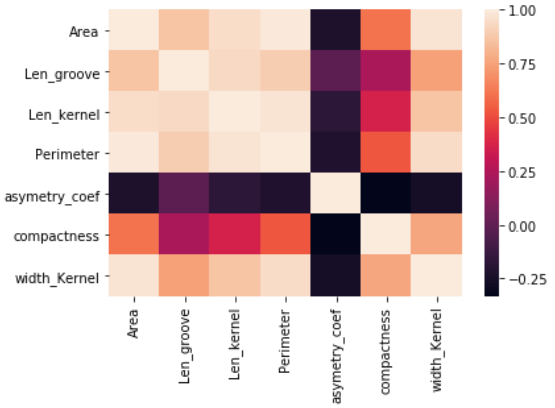
We performed the descriptive data analysis to know the structure of our data. We also created Box plot, co-relation coefficient diagram and scatter plot & density curve of features to know the relation between the features and structure of each feature as showed in Fig 1, Fig 2 and Fig. 3.

Figure 1.



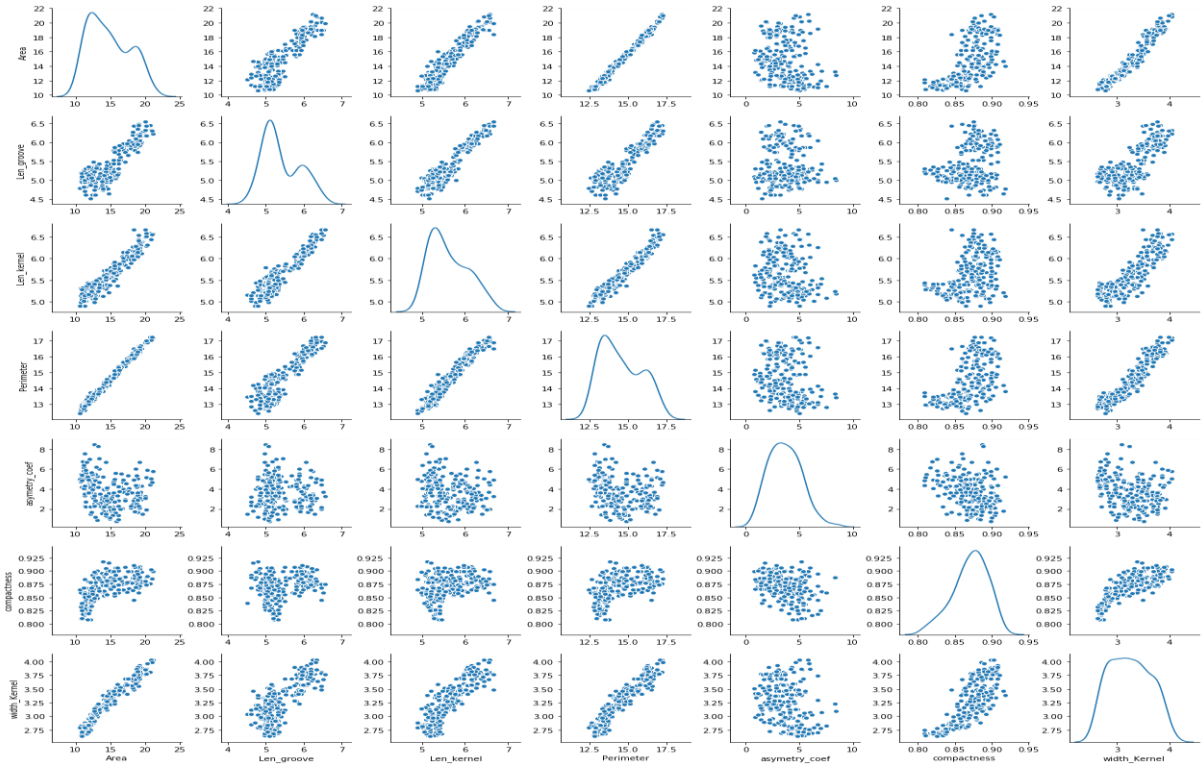
Box Plot

Figure 2.



Correlation Coefficients

Figure 3



Scatter plot and Density plot

EDA	Area	Perimeter	Compactness	Length of Kernel	Width of kernel	Asymmetry Coefficient	Length of Groove
Count	210	210	210	210	210	210	210
Mean	14.84752	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
Std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.49148
Min	10.59	12.41	0.8081	4.899	2.63	0.7651	4.519
25%	12.27	13.45	0.8569	5.26225	2.944	2.5615	5.045
50%	14.355	14.32	0.87345	5.5235	3.237	3.599	5.223
75%	17.305	15.715	0.887775	5.97975	3.56175	4.76875	5.877
Max	21.18	17.25	0.9183	6.675	4.033	8.456	6.55

Summary of EDA:

Attribute	Area	Perimeter	Compactness	Length of Kernel	Width of Kernel	Asymmetry Coefficient	Length of groove
Area	Bimodal Gaussian mixture model	.99	.608	.94	.97	-.22	.863
Perimeter	Linearly	Bimodal Gaussian mixture model	.52	.97	.94	-.21	.89
Compactness	Quadratically	Quadratically	Negatively skewed	.52	.76	-.33	.22
Length of Kernel	positively	positively	Uncorrelated	Bimodal Gaussian mixture model	.86	-.17	.93
Width of Kernel	positively	positively	Quadratically	Linearly	Nearly Normal	-.25	.74
Asymmetry Coefficient	uncorrelated	uncorrelated	Uncorrelated	uncorrelated	uncorrelated	Nearly normal	-.01
Length of groove	positively	positively	Uncorrelated	Linearly	low positive	uncorrelated	Bimodal Gaussian mixture model

1. Diagonal entries are distribution of the features and off-diagonal entries are correlation between the features.
2. By observing the box plot, we concluded that attributes asymmetry coefficient and compactness have two outliers each.
3. The variance of attributes is also different.

Data visualization helps prepare the data for further analysis and decide the best modeling method for the data.

The original creator did clustering on this sample dataset. However, aim of observing the sample datapoint is to predict some outcome for population. Therefore, we used this sample data to generate a classifier which can predict the class (Rosa, Kama and Canadian) of any new wheat kernel seed. We applied clustering and classification techniques on this dataset.

Data Preparation for K-means clustering:

We applied k-means on our dataset, but before applying the k-means algorithm we verified the assumptions of k-means. To do that, we did the following:

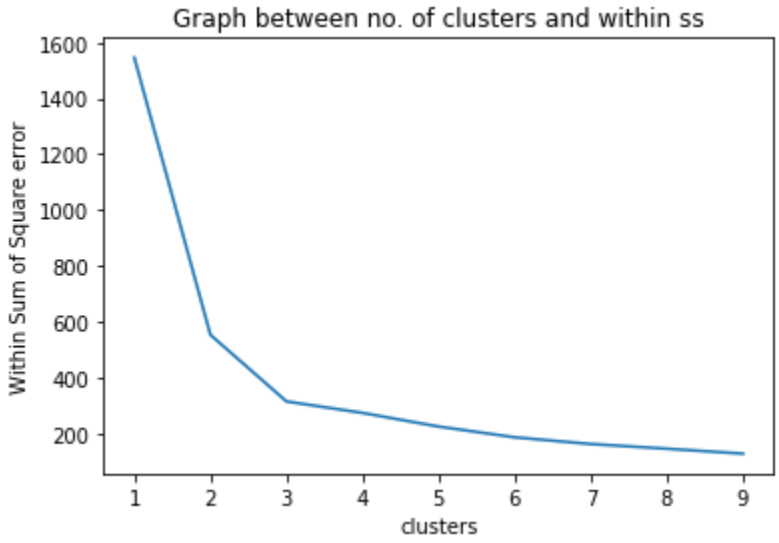
normalize each feature of dataset to make the variance of each cluster equally.

K-means is also sensitive to outliers therefore, we did the analysis with outliers and without outliers and compared the results.

Used within sum of square error versus number of clusters curve to predict number of clusters for K-means. (choose a point as k where declination in within some of square error relatively lesser than previous values of clusters.)

Results:

Figure 4.



Clearly **3** clusters are the best choice to pick.

K-means Clustering with outliers: We applied k-means algorithm on our dataset. Results are mentioned in table 1.

Table1:

Class	True Classification	False Classification	Total	Percentage of True Classification
1	62	8	70	88.57
2	65	5	70	92.85
3	66	4	70	94.28

K-means clustering without outliers: We have 4 outliers in our dataset. We deleted them and applied k-means algorithm on 206 remaining datapoints. Results are mentioned in below table 2.

Table 2:

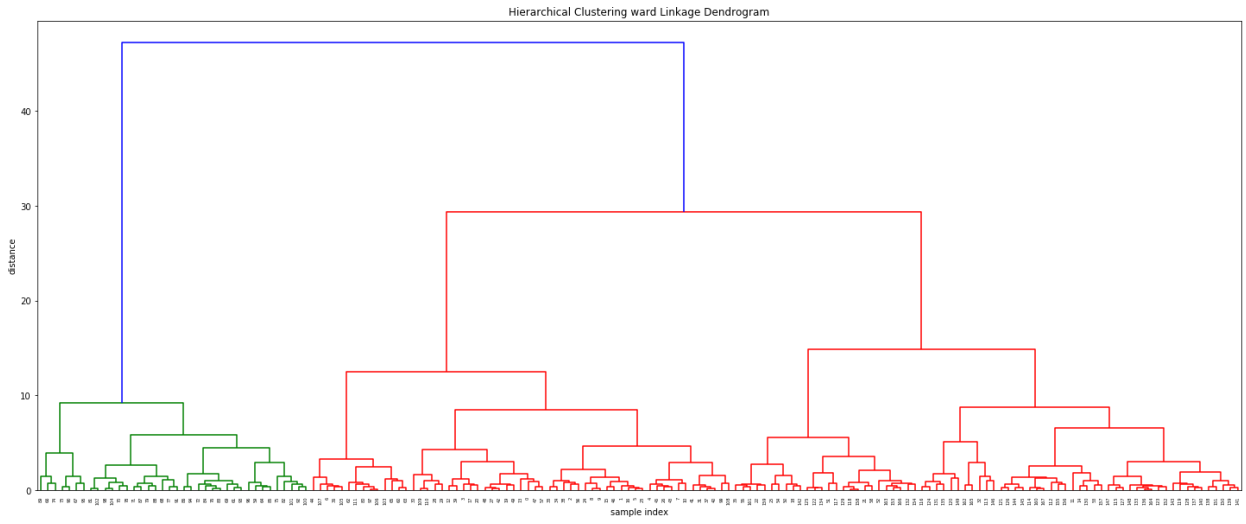
Class	True Classification	False Classification	Total	Percent of true classification
1	63	7	70	90
2	66	4	70	94.28
3	62	4	66	93.93

By comparing table 1 and table 2, we concluded that the accuracy of true classification improved further in table 2. This was because the outliers were shifting the cluster center.

We applied Hierarchal algorithm on our dataset. It is computationally costly but we have only small dataset.

Agglomerative clustering: By observing scatter plot we could clearly say that there is no chaining in between the datapoints so we avoided single linkage method. We decided to apply agglomerative clustering with ward linkage and complete linkage. Firstly, we created a dendrogram (using ward Linkage) mentioned below:

Figure 5. (ward linkage)



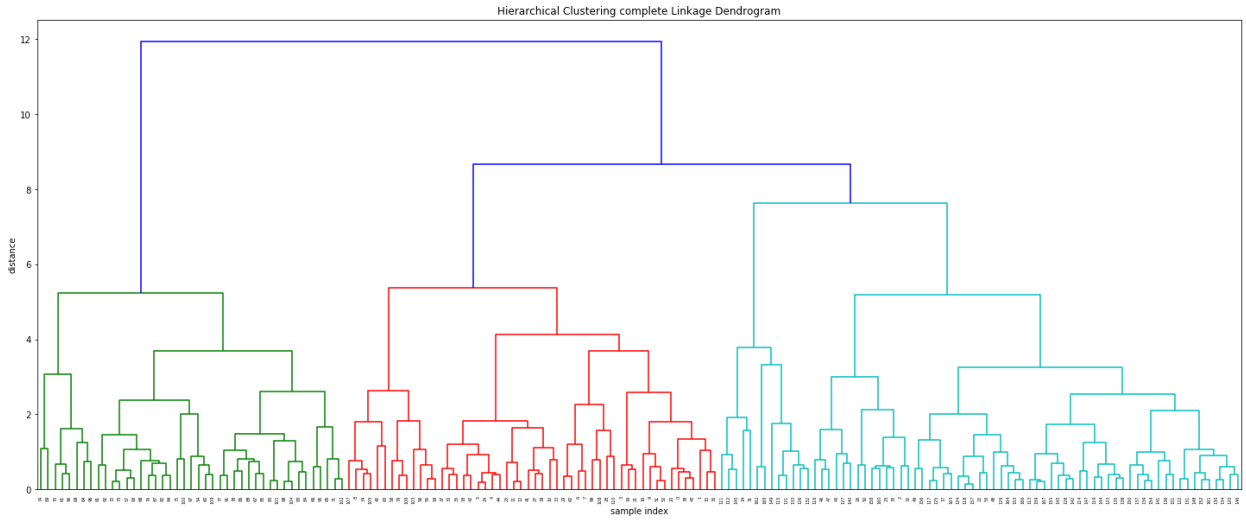
The vertical lines in the dendrogram is the measure of dissimilarity. As a result, we can clearly say that there are 3 significantly different clusters.

Distribution of the data points among clusters are as mentioned in Table 3:

Class	True Classification	False Classification	Total	Percent of true classification
1	54	16	70	80
2	63	7	70	90
3	70	0	70	100

Next, we created the dendrogram with complete linkage mentioned below:

Figure 6 (complete linkage)



The vertical lines in the dendrogram is the measure of dissimilarity. As a result, we could cut the dendrogram at the vertical distance of 7. This will give us 3 clusters. We could also cut it at vertical distance of 4, this will give us 7 clusters.

Firstly, we cut it at vertical height 7 and got **3** clusters. We compared our results to the true label. Results are mentioned in below table 2.

Class	True Classification	False Classification	Total	Percent of true classification
1	52	18	70	74.28
2	47	23	70	67.14
3	70	0	70	100

As we can see in above table, complete linkage is not giving better in classifying class 1,2, but 100% correctly classifying class 3.

We have further cut the dendrogram at height 4. This time our results were incomparable to true class. It was completely mixing all class together.

Summary of clustering results:

K-means: Between k-means with outliers and without outliers, the latter gave better accuracy.

Agglomerative clustering: Both linkage methods (Ward linkage and Complete linkage) were 100% correctly classifying class 3 but ward linkage is doing better job for classifying class 2 and 1.

In summary, we chose k-means without outlier in comparison to hierarchical clustering with ward linkage as it was giving us accuracy of more than 90% for all the classes. Although, the method selection would also depend on end user requirements. What I mean is that if class 3 gives us more financial revenue than other class than picking ward linkage would be better than any other method. Similarly, if class 3 revenue is not more than other class, than K-means clustering without outliers would be more beneficial.

Note: Clustering is an unsupervised learning. We are only comparing our results because we know the true label, otherwise, one can choose any of the methods among k-means without outlier, hierarchical clustering with ward linkage or complete linkage and will get the accuracy as explained above. Moreover, we can also apply the soft k-means clustering to get better accuracy.

We can further visualize the result of clustering by dimension reduction method Principal analysis component (PCA).

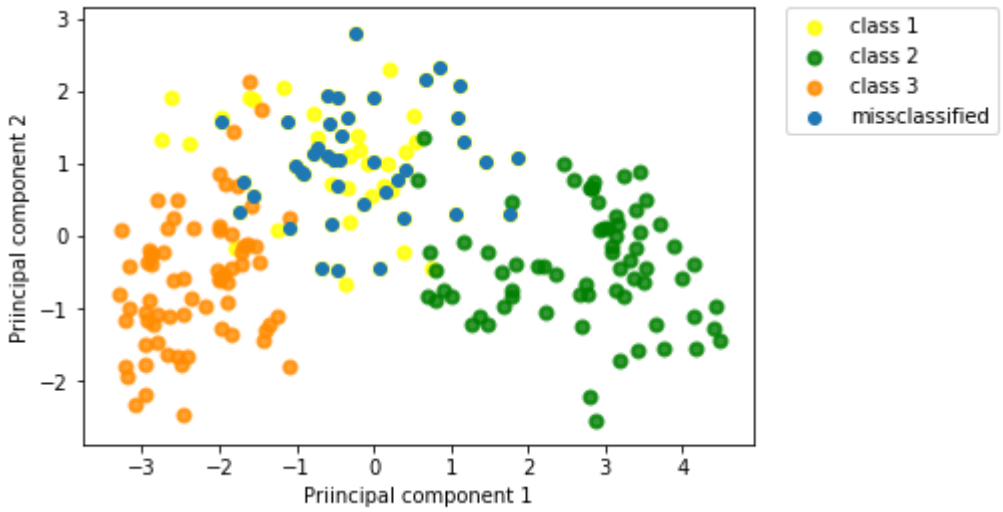
Variance of 1st 2 principal components are mentioned in table 4:

Principal Component	Variance
Principal Component 1	0.72
Principal Component 2	0.171
	Total=.89

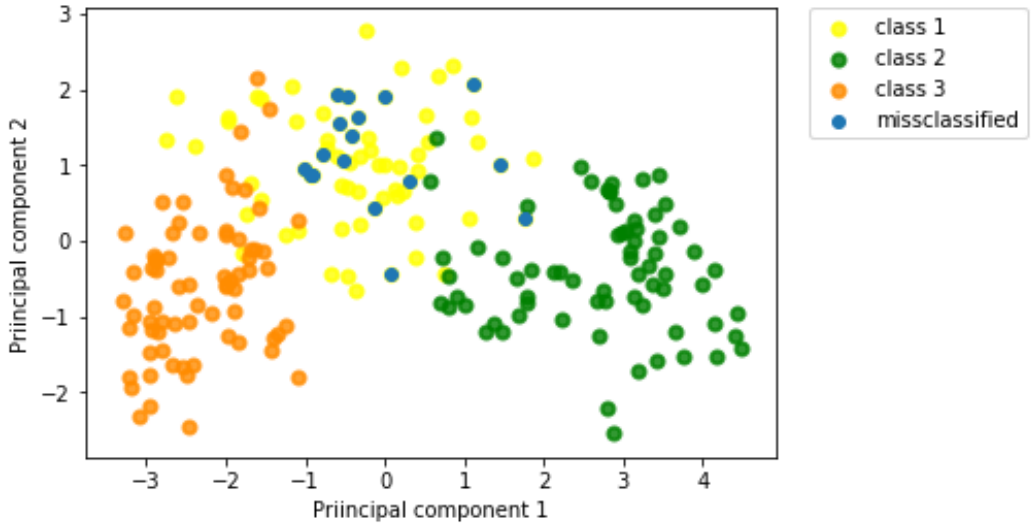
Table 4

First two principal components are covering 89% (variance) information of our data.

We can visualize the ward linkage and K-means in figure7 and figure8, respectively.



Ward Linkage (figure7.)



K-means (figure7)

Classification: We applied some techniques of classification such as multinomial logistic regression, linear discriminant analysis, quadratic discriminant analysis and decision tree.

Moreover, we applied 5-cross-validation and the results are mentioned below:

Logistic regression:

By observing the co-relation coefficient of features, we found that Area, Perimeter, Length of kernel, width of kernel is highly co-related. To overcome the multi-co-linearity, we retained the area and dropped the rest of the variable specified above. Now we have 4 remaining variables (Area, compactness, length of groove and asymmetric coefficient. We ran our algorithm of these variables after applying 0-1 rescaling it and our results are mentioned below:

```
cross validation set 1
Accuracy of Logistic regression classifier on training set: 0.93
Accuracy of Logistic regression classifier on test set: 0.88
cross validation set 2
Accuracy of Logistic regression classifier on training set: 0.90
Accuracy of Logistic regression classifier on test set: 0.88
cross validation set 3
Accuracy of Logistic regression classifier on training set: 0.88
Accuracy of Logistic regression classifier on test set: 0.95
cross validation set 4
Accuracy of Logistic regression classifier on training set: 0.92
Accuracy of Logistic regression classifier on test set: 0.86
cross validation set 5
Accuracy of Logistic regression classifier on training set: 0.91
Accuracy of Logistic regression classifier on test set: 0.93
```

Accuracy on test set is not consistent. Therefore, we cannot select linear regression as a classifier.

Linear Discriminant Analysis:

By observing the co-relation coefficient of features, we found that Area, Perimeter, Length of kernel, width of kernel is highly co-related. To overcome the multi-co-linearity, we will retain the area and drop the rest of the variable specified above. Now we have 4 remaining variables (Area, compactness, length of groove and asymmetric coefficient. We ran our algorithm of these variables after applying 0-1 rescaling on it and our results are mentioned below:

```
cross validation set 1
Accuracy of LDA classifier on training set: 0.95
Accuracy of LDA classifier on test set: 0.95
cross validation set 2
Accuracy of LDA classifier on training set: 0.97
Accuracy of LDA classifier on test set: 0.95
cross validation set 3
Accuracy of LDA classifier on training set: 0.96
Accuracy of LDA classifier on test set: 0.93
cross validation set 4
Accuracy of LDA classifier on training set: 0.96
Accuracy of LDA classifier on test set: 0.95
cross validation set 5
Accuracy of LDA classifier on training set: 0.96
Accuracy of LDA classifier on test set: 0.98
```

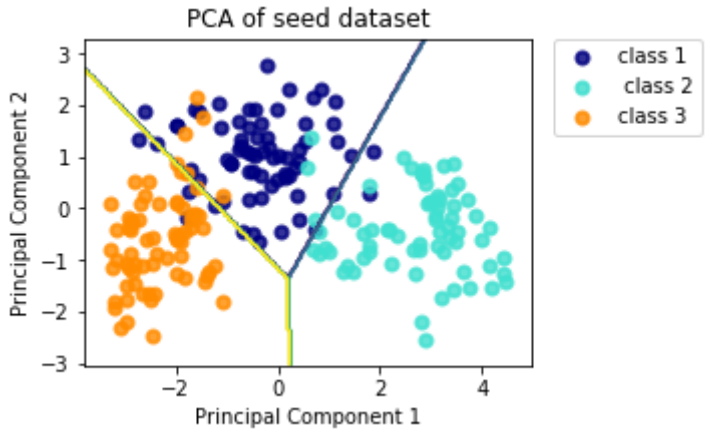
The average accuracy of LDA classifier on test set: **0.95.2.** The result are shown in confusion matrix mentioned below:

Confusion matrix:

Class	1	2	3
1	14	1	0
2	0	13	0
3	1	0	13

We can visualize our results by first 2 principal components, which is covering 89% variance (information of dataset) mentioned below:

LDA Decision boundaries visualization by 2 principal components:



Quadratic Discriminant Analysis:

We normalized our feature variables than applied the quadratic discriminant analysis algorithm. Results are mentioned below:

```
cross validation set
1
Accuracy of classifier on training set: 0.96
Accuracy of QDA classifier on test set: 0.95
cross validation set
2
Accuracy of classifier on training set: 0.98
Accuracy of QDA classifier on test set: 0.90
cross validation set
3
Accuracy of classifier on training set: 0.96
Accuracy of QDA classifier on test set: 1.00
cross validation set
```

```
4
Accuracy of classifier on training set: 0.98
Accuracy of QDA classifier on test set: 0.93
cross validation set
5
Accuracy of classifier on training set: 0.95
Accuracy of QDA classifier on test set: 1.00
```

Accuracy on the test set is not **consistent**.

Decision Tree:

Depth 2:

```
cross validation set 1
Accuracy of Decision tree classifier on training set: 0.93
Accuracy of Decision tree classifier on test set: 0.88
 cross validation set 2
Accuracy of Decision tree classifier on training set: 0.91
Accuracy of Decision tree classifier on test set: 0.95
 cross validation set 3
Accuracy of LDA classifier on training set: 0.93
Accuracy of Decision tree classifier on test set: 0.86
 cross validation set 4
Accuracy of Decision tree classifier on training set: 0.92
Accuracy of Decision tree classifier on test set: 0.93
cross validation set 5
 Accuracy of Decision tree classifier on training set: 0.94
Accuracy of Decision tree classifier on test set: 0.83
Accuracy of Decision Tree classifier on training set: 0.92
```

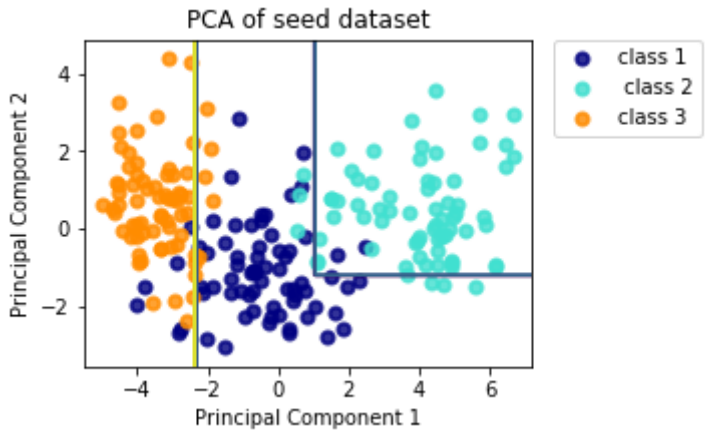
Accuracy of Decision Tree classifier on test set: 0.92. Confusion matrix is also mentioned below:

Confusion matrix:

Class	**1**	**2**	**3**
1	21	1	5
2	0	33	0
3	0	0	23

Average of true classification: $(21 + 33 + 23)/(27 + 34 + 23) = .916$

We can visualize our results by first 2 principal components, which is covering 89% variance (information of dataset).



Summary of classification:

Among the different classification methods, logistic regression and quadratic discriminant analysis gave inconsistent results while doing 5-cross validation. However, between decision tree and linear discriminant analysis, linear discriminant analysis gave 95% of correct classification. Therefore, we have selected the linear discriminant analysis as a classifier for our dataset.

Conclusion:

We have successfully analyzed different methods of clustering and classification on three different variety of wheat seeds. We have found that if class 3 gives us more financial revenue than hierarchical clustering ward linkage would be a good choice for agricultural analyst, else k-means clustering without outliers would be a better choice. We can also apply soft k-mean clustering and compare our result with that.

We have applied different methods of classification and found linear discriminant analysis is the best among all. Therefore, whenever any new seed of this group will come, agricultural analyst should choose linear discriminant analysis to find the true class of the seed.

Future work

Dataset used in this study is small and possibly may not be a true representative of the population. Therefore, a bigger set may give us better results. We can also apply bootstrap method to verify the accuracy of our results. On the big dataset, we should avoid the hierarchical clustering because of its computational complexity. Moreover, we can also try random forest or bootstrap aggregation methods of decision tree and compare all the methods.

References:

1. UCI Machine Learning Repository (2012). Seeds Data Set. http://archive.ics.uci.edu/ml/datasets/seeds.

2. M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak, 'A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010, pp. 15-24