Peter Subacz – CS534: Artificial Intelligence – Homework 2 – 9/26/19

Contents

[Instructure & TA Clarifications: 2](#_Toc23441633)

[Problem 1: K-Means Implementation 3](#_Toc23441634)

[A. Python implementation 3](#_Toc23441635)

[B. Python implementation 3](#_Toc23441636)

[Problem 2: Logistical Regression with Regularization 4](#_Toc23441637)

[A. Python implementation 4](#_Toc23441638)

[B. Logistical Regression with Regularization Performance 4](#_Toc23441639)

[C. Logistical Regression with Regularization and Standardization 4](#_Toc23441640)

[Problem 3: Clustering Techniques 5](#_Toc23441641)

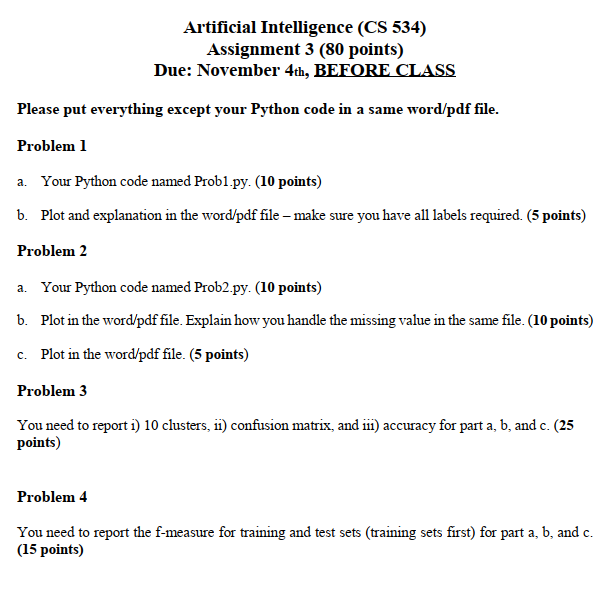
[A. K-Means on Digits Dataset 5](#_Toc23441642)

[B. Agglomerative Clustering on Digits Dataset 6](#_Toc23441643)

[C. Affinity Propagation on Digits Dataset 6](#_Toc23441644)

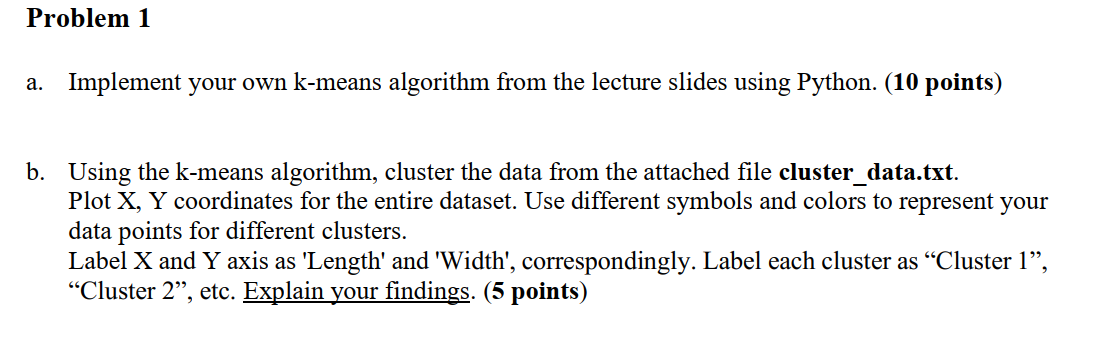
[Problem 4: Support Vector Machine 7](#_Toc23441645)

# Instructure & TA Clarifications:



Announcement from Oct. 30th: Problem 1a, you need to make sure your k-means algorithm is universal, which means it can take other datasets other than the provided one to perform clustering. The reason is that 1a is graded separately from 1b and based on your ability to implement k-means without dataset. 1b is graded based on your clusters. For problem 3a, you could either use the k-means from problem 1 or k-means from scikit. It will be graded based on the clusters, confusion matrix, and accuracy.

# Problem 1: K-Means Implementation



## Python implementation

See Python code attached titled ‘Prob1.py’. This code was developed to be a universal N dimensional clustering machine as clarified above. This K-Means will run assuming whatever method used to ingest the data can provide in list containing tuple of parameters. The parser implemented will run on any numerical data structured {index, value1, value2, value3…}. The plotting function assumes 2D data due to matplotlib’s implementation of scatter plots. Higher dimensional data turns off plotting.

## K-Means Clustering Analysis

The K-Means algorithm implemented here is designed to cluster n dimensional data with m number of features. If seed clusters positions are not given, then the algorithm will randomly seed initial cluster values based on minimum and maximum values for each feature. Figure 1 below shows Two clusters moving to central positions based on the derivative of newly calculated centroids.

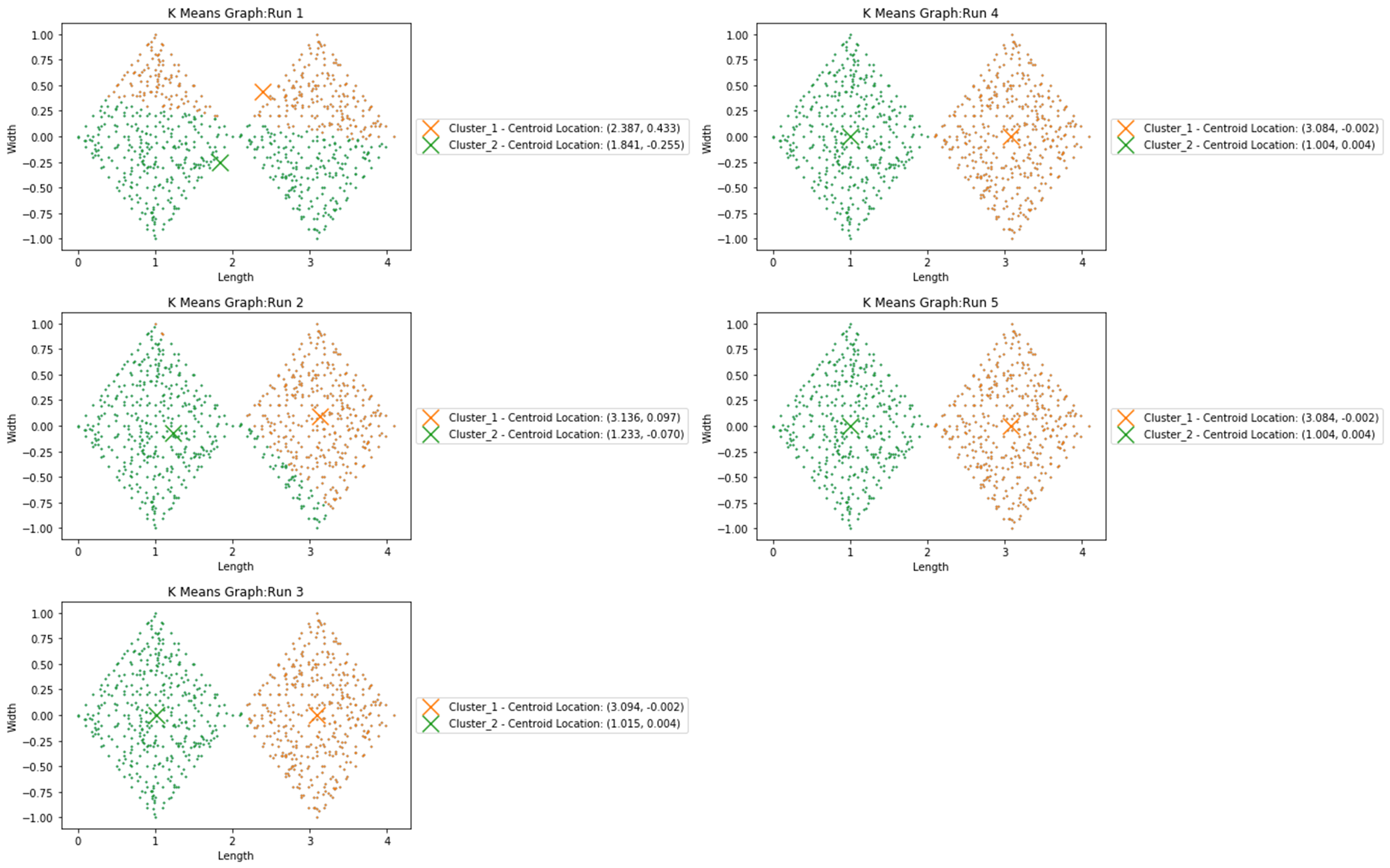


Figure - Two Clusters

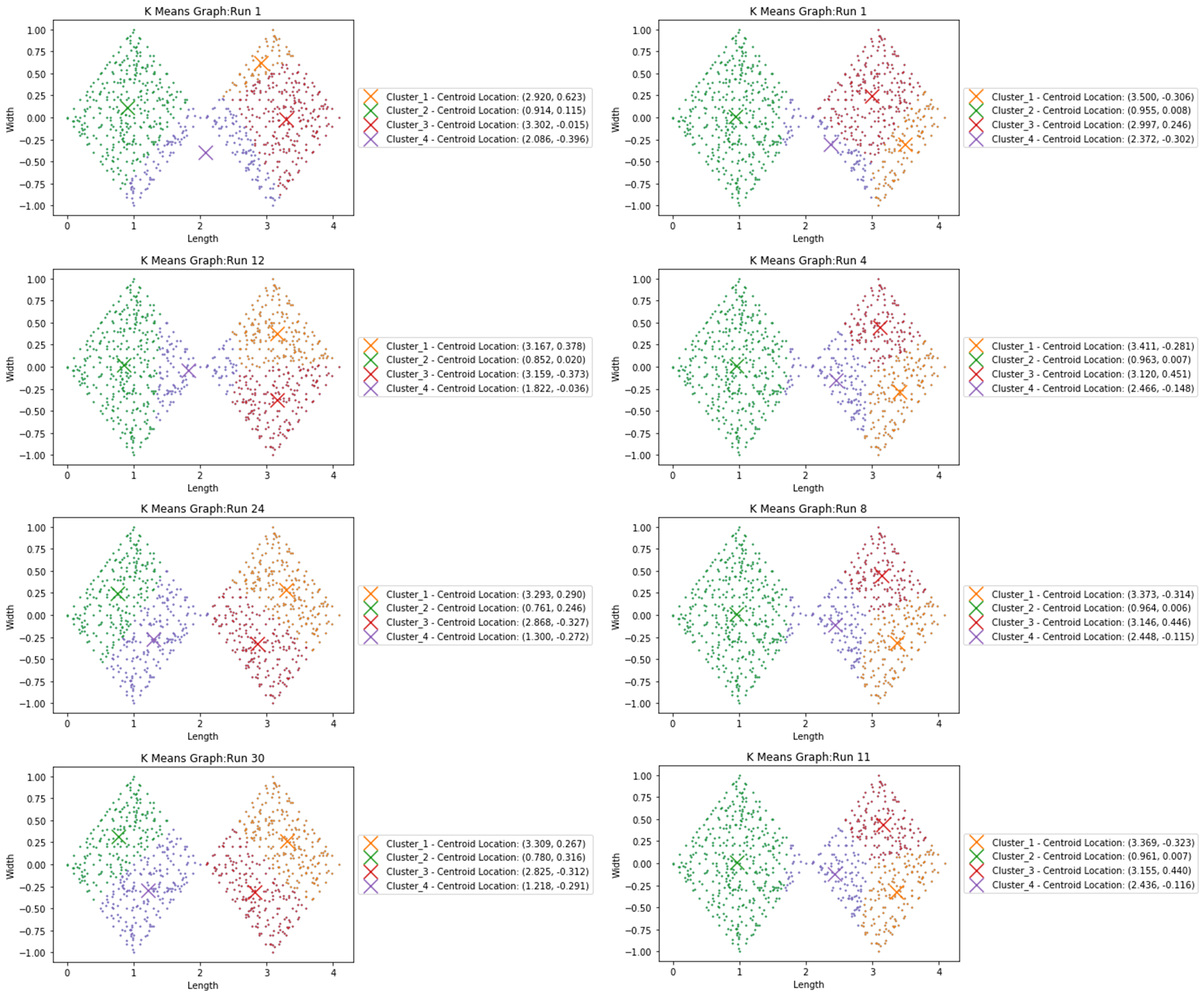
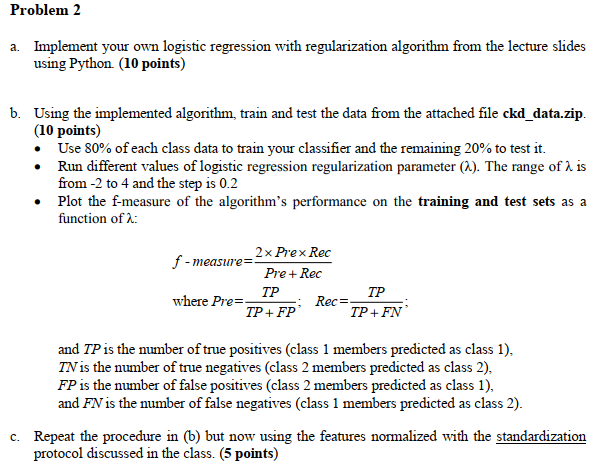
Different starting points can have huge consequences for K-Means as the direction each cluster moves is influenced by the initial position. On the left K-Means settles on 4 even distributed clusters while on the right there is 1 major cluster and 3 smaller clusters. 

Figure - Four Clusters

The threshold parameter is important in K- means as it defines the iteration stopping point. Higher thresholds means clusters will roughly have less iterations with sub-optimal cluster positions while smaller thresholds have more iterations but with more optimal cluster positions

# Problem 2: Logistical Regression with Regularization

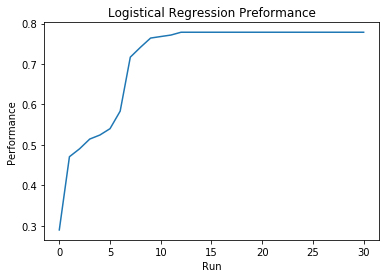


## Python implementation

See Python code attached titled ‘Prob2.py’

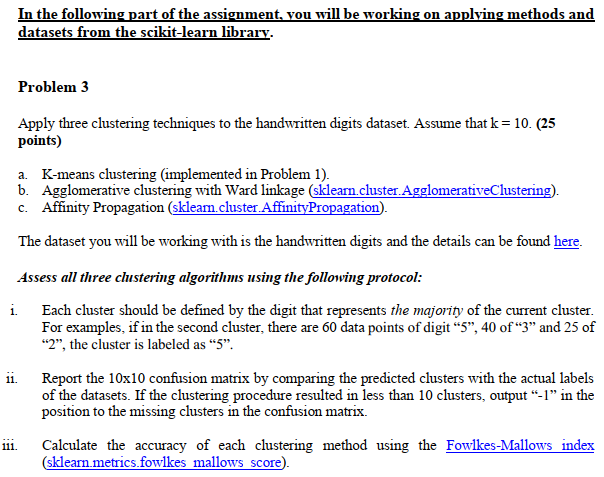
## Logistical Regression with Regularization Performance

This implementation dealt with missing datum by calculating the mean of available features and substituting them into as recommended in lecture 4. If Boolean values if the average was greater than 0.5 the value was chosen to be 1 and if less than 0.5 the value was 0.



## Logistical Regression with Regularization and Standardization

# Problem 3: Clustering Techniques



## K-Means on Digits Dataset

This K-Means implementation loops through the digits dataset and attempts to classify each digit with according to the Euclidian distance of the 64-dimensional pixel data. The distance to each cluster representative is calculated and assigned based on the minimum distance to a cluster. The classes are ordered 0- 9 based on the class labels. As seen in the output below, the algorithm took 14 iterations to converge but there is an issue with the 9th cluster as there are no values assigned to it. The confusion matrix shows that cluster 4 has a high number of 9 assigned to them. This is most likely due to distance function only using pixel brightness and that the digit 9 and 4 share a high percentage of similarities.

K-Means Run 14

Number of Clusters: 10

Cluster Confusion Matrix:

[[176. 0. 1. 0. 0. 0. 1. 0. 0. 0.]

[ 0. 153. 4. 0. 10. 0. 3. 2. 16. 20.]

[ 0. 25. 147. 2. 0. 0. 0. 0. 3. 0.]

[ 0. 0. 12. 160. 0. 39. 0. 0. 13. 145.]

[ 1. 0. 0. 0. 160. 1. 0. 0. 0. 0.]

[ 0. 1. 0. 1. 0. 139. 0. 4. 3. 5.]

[ 0. 2. 0. 0. 0. 2. 175. 0. 1. 0.]

[ 0. 0. 2. 8. 8. 0. 0. 169. 3. 8.]

[ 0. 0. 10. 11. 2. 0. 1. 3. 134. 2.]

[ 0. 0. 0. 0. 0. 0. 0. 0. 0. -1.]]

Fowlkes-Mallows Scores

Accuracy Score: 0.611097242866993

## Agglomerative Clustering on Digits Dataset

Number of Clusters: 10

Cluster Confusion Matrix:

[[177. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

[ 0. 154. 0. 0. 0. 0. 0. 0. 3. 20.]

[ 0. 27. 165. 0. 0. 0. 0. 0. 4. 0.]

[ 0. 0. 0. 168. 0. 2. 0. 0. 1. 145.]

[ 0. 0. 0. 0. 177. 0. 0. 0. 0. 0.]

[ 0. 0. 0. 0. 0. 178. 0. 0. 0. 2.]

[ 0. 0. 0. 0. 0. 1. 179. 0. 0. 0.]

[ 0. 0. 1. 1. 3. 0. 0. 178. 1. 11.]

[ 0. 0. 10. 13. 0. 0. 1. 0. 164. 2.]

[ -0. -0. -0. -0. -0. -0. -0. -0. -0. -1.]]

Fowlkes-Mallows Scores

Accuracy Score:80.132806

## Affinity Propagation on Digits Dataset

Estimated number of clusters: 10

Cluster Confusion Matrix:

[[103. 1. 4. 7. 3. 2. 1. 0. 1. 59.]

[ -0. -1. -0. -0. -0. -0. -0. -0. -0. -0.]

[ 5. 0. 138. 0. 1. 6. 0. 2. 0. 30.]

[ 1. 0. 0. 171. 1. 0. 7. 0. 0. 0.]

[ 0. 0. 0. 1. 175. 0. 0. 0. 1. 0.]

[ 0. 0. 0. 0. 0. 169. 9. 0. 0. 0.]

[ 24. 58. 29. 0. 7. 5. 192. 25. 1. 14.]

[ 0. 0. 10. 0. 2. 9. 8. 139. 0. 8.]

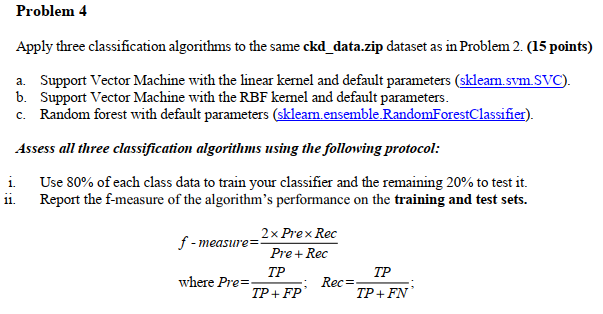
[ 2. 9. 0. 0. 0. 6. 10. 0. 153. 0.]

[ 4. 17. 7. 0. 9. 8. 0. 0. 0. 134.]]

Fowlkes-Mallows Scores

Accuracy Score:55.561656

# Problem 4: Support Vector Machine



1. **Support Vector Machine with the linear kernel and default parameters**

Train F-Measure: 1.000000

Test F-Measure: 0.730159

1. **Support Vector Machine with the RBF kernel and default parameters**

Train F-Measure: 1.000000

Test F-Measure: 0.814815

1. **Random forest with default parameters**

Train F-Measure: 0.987080

Test F-Measure: 0.973451