**Lab 3: Clustering**

**Evgueni N. Smirnov, Gerasimos (Jerry) Spanakis**

This lab requires an extended version of Weka, which you can download from this [link](file:///This%20lab%20requires%20an%20extended%20version%20of%20Weka,%20which%20you%20can%20download%20from%20here%20%20%20.%20%20Please%20use%20only%20the%20supplied%20jar%20file%20since%20you%20will%20not%20be%20able%20to%20answer%20the%20questions%20using%20standard%20versions%20of%20Weka.%20%20You%20can%20start%20Weka%20using%20the%20following%20command):

You can start Weka using the following command (in Mac you can just click the downloaded .jar file)

java -jar weka.jar

If you don’t have Java installed, work with the regular version, you just won’t be able to see the entropy.

Also, wherever running clustering algorithms, use the default value for the 'Seed' parameter.

|  |  |
| --- | --- |
| **Clustering Algorithm** | **Weka Name** |
| K-means | SimpleKMeans |
| K-centered hierarchical clustering | FarthestFirst |

 Download the following data sets:

[Iris](https://eleum.maastrichtuniversity.nl/bbcswebdav/pid-1076583-dt-content-rid-5132224_1/xid-5132224_1)

[2d1](https://eleum.maastrichtuniversity.nl/bbcswebdav/pid-1076583-dt-content-rid-5132225_1/xid-5132225_1)

[2d2](https://eleum.maastrichtuniversity.nl/bbcswebdav/pid-1076583-dt-content-rid-5132226_1/xid-5132226_1)

[2d3](https://eleum.maastrichtuniversity.nl/bbcswebdav/pid-1076583-dt-content-rid-5132227_1/xid-5132227_1)

[food](https://eleum.maastrichtuniversity.nl/bbcswebdav/pid-1146442-dt-content-rid-5585275_1/xid-5585275_1)

Problem 1:

The goal is to become familiar with various clustering techniques. Load the **Iris** data set into the Weka Explorer. For this case, set the "Cluster Mode" to "Classes to clusters evaluation" (notice that this is one of the cases that we do have external information about the classes).

* Apply K-means clustering on this data set to find three clusters (you expect that right?).
* Report the confusion matrix, and the entropy of the resulting clusters. Now answer the following questions:
  1. Comment on the quality of this clustering based on your analysis of the confusion matrix and entropy. Which two classes are confused?
  2. Increase the number of clusters until you get entropy less than 0.2 (or just try to reduce SSE). For what number of clusters does that happen? Comment on how the classes are distributed among clusters and the purity of these clusters.
  3. Plot the SSE as the number of clusters varies from 2 to 10.  Describe the behavior of this curve. Does it make sense?

Problem 2:

Load the **2d1**, **2d2** and **2d3** data sets into the Weka Explorer and visualize them. In all these data sets, there are 6 underlying classes of points, two of which represent noise.

1. Apply **K-means** clustering and **FarthestFirst** clustering on the **2d1** set for **numClusters** equal to 6. Visualize and compare the clustering solutions using the centers of clusters.
2. Use **FarthestFirst** clustering to cluster the points in all the three data sets from 2 to 10 clusters in steps of 2 (2,4,...,10). Plot the trend followed by entropy for each of the three data sets at these number of clusters (plot all the three curves in the same figure). Comment on how entropy varies as the number of clusters increases, and also how the entropy compares between the three data sets as the number of clusters is varied. What characteristic difference(s) between the data sets lead to this variation in entropy? Explain in detail and if it is necessary visualize  the clustering solutions for the three data sets for numClusters equal to 6.
3. What are the general problems with these data sets that may lead to poor clustering results? Is there anything that can be done to improve the clustering results obtained with the FarthestFirst technique?  Which other clustering techniques could potentially do better on these data sets?

Problem 3:

Now it’s your turn! Use the **food** dataset which contains 6 features and try to create a meaningful clustering. Try different algorithms, different settings and come up with useful conclusions!  
Use the following steps as basic guidelines:

1. Choose a set of attributes for clustering and give a motivation. (**Hint**: always ignore attribute "name". Why does the name attribute need to be ignored?)
2. Experiment with different number of clusters
3. Do you think the clusters are "good" clusters? (Are all of its members "similar" to each other? Are members from different clusters dissimilar?)
4. What does each cluster represent? Choose one of the results. Make up labels (words or phrases in English) which characterize each cluster.