MSU CSC 325 Spring 2016
Project 1. K-Means clustering algorithm
Mar. 14, 2016
Due Wed. Mar 30, 11:59pm

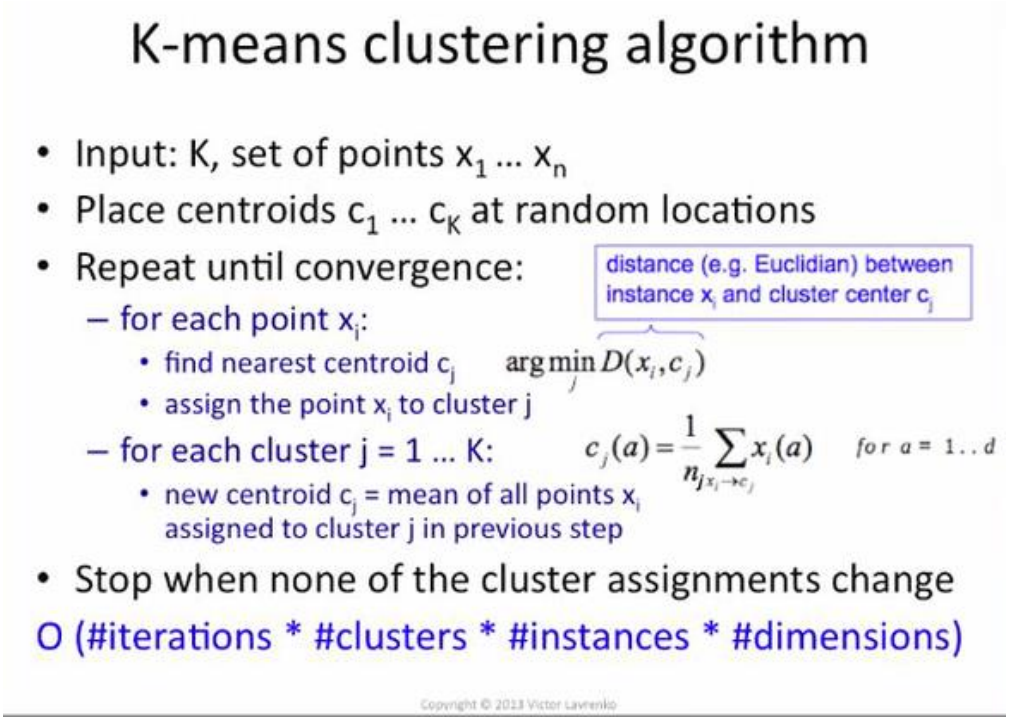
Write a program in either C++ or Python that implements the K-Means clustering algorithm shown below. **Clustering** is the act of partitioning a population into sub-groups (clusters) using some characteristic.

The K-means clustering algorithm starts by placing K points (centroids) at random locations in space. We then perform the following steps iteratively:

(1) for each instance, we assign it to a cluster with the nearest centroid, and

(2) we move each centroid to the mean of the instances assigned to it.

The algorithm continues until no instances change cluster membership.



Source -- https://people.sc.fsu.edu/~jburkardt/classes/isc_2009/clustering_kmeans.pdf

https://www.youtube.com/watch?v=IuRb3y8qKX4 0:00-3:00 – skip. 3:00-- example

https://www.youtube.com/watch?v=_aWzGGNrcic
2:00-2:55 – skip 3:20 – "K-means is blazingly fast." 4:20 -- example

https://www.youtube.com/watch?v=BVFG7fd1H30 The data of this animation is uninterestingly uniform. The clusters form a Voronoi Diagram, http://www.ams.org/samplings/feature-column/fcarc-voronoi

In Project 1, the "population" is US states, for which two characteristics are chosen to define a two-dimensional "coordinate":

- the x coordinate is the percentage of unemployment in that state in Dec. 2015
- the y coordinate is the percentage of change in that state's 2016 funding for higher education

The clustering characteristic to be used is "difference" measured by distance between (x, y) coordinates of two different points.

Project steps:

1. Get the data from websites below. You'll need to cut-and-paste, edit and format into Excel and perhaps your program. Plot the data in Excel using a scatter plot.
   - US state funding for higher education *(NOTE: two states are missing and have no data)*
   https://www.insidehighered.com/news/2016/01/25/state-support-higher-education-rises-41-percent-2016
   - US State Unemployment Rates, December 2015
   http://www.ncsl.org/research/labor-and-employment/state-unemployment-update.aspx

2. Write the K-means clustering program that creates BearPlot output with your choice of number of clusters.
   - Your program does not need to read data from a file. You could hardcode the data into a structure.
   - The BearPlot output does not need to be a Voronoi Diagram. Use color to graph the clusters as in the example below.
   - The BearPlot output does not need to be an *animated* graph output.

3. Put a screenshot of the BearPlot output into your report document.

4. Make a one-sentence observation about the clustering of the data. *(Does the clustering show any patterns? Is the clustering surprising, unhelpful, inconclusive?)*

You will turn in **two files** in your eccentric folder (each member of group turns them in):

- a program named for the language you chose -- **Proj1Cluster.cpp** or **Proj1Cluster.py**
- a one-page MS Word report named **Proj1Cluster** that includes:
  - The plot of the data made in Excel and cut-and-pasted into this Word doc
  - A screenshot of your program's output
  - A one-sentence observation about the clustering of the data