CSE 6331 Cloud Computing
Summer 2018, © DL, UTA, 2018

Programming Assignment 4
Machine Learning (K-means clustering)
Due: In Blackboard

Task: You will get a data set (for example titanic) and use a k-means clustering tool to "better" understand your data.

The Titanic data set:
biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.xls

lists the passengers and information about them on the Titanic (when it sank).

There are many k-means clustering implementations:
Python, R, Weka, etc.

Your assignment is to do k-means clustering on the titanic (or similar) data set, on various (different) attributes (columns), specify number of clusters and show results in a table (number of points in a cluster, distances between centroids, how tightly "packed" clusters are…)

For example:
For 5 clusters:
Fare price and Age
Surviving and Fare Price
Cabin and Fare Price
Repeat for 20 clusters.
Repeat for 100 clusters.

Show results (total number of points, number of clusters centroid locations, distances, etc.) on a web page.

Interpret results:
How many clusters is "appropriate" (5, 10, 100?) and why?

Please, submit in Blackboard. Work must be individualized, but may be done in a group.

You must submit this lab, working (or partially) by the due date.
Your program should be well commented and documented, make sure the first few lines of your program contain your name, this course number, and the lab name and number.
Your comments should reflect your design and issues in your implementation.
Your design and implementation should address error conditions.