CSE 6331 Cloud Computing
Summer 2019, © DL, UTA, 2019

Programming Assignment 5
Machine Learning (K-means clustering)
Due: In Canvas

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).

If the Titanic worries you, there are also data sets for number of smokers (percent by year and country) and a table for average price of a pack of cigarettes.
Are they related, or do they cluster?

There are many k-means clustering implementations:
Python, R, Weka, etc.
You SHOULD (but not required) to use an already existing k-means implementation.

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?