Due date: October 7th, by 11:59pm.

Task 1 (Programming):

Write a program in C++ (g++ >=4.2.1) / Java (>=1.7) / Python (2.7) implementing k-means clustering in n-dimensional (n varies with input data) space with Euclidean distance measure. You should not use any third party libraries that provide k-means.

Command line arguments:
+ 1st argument: number of clusters - k
+ 2nd argument: initialization method
  - "rand": random
  - "first": select first k points from input file as initialized centroids
+ 3rd argument: convergence threshold (if change between 2 iterations is smaller than this threshold, converged)
+ 4th argument: maximum number of iterations (stops after max number of trials even if convergence threshold not met)
+ 5th argument: input file name

Input:
Each line in the input file represents a data point; attribute values (all numeric) are separated by comma ",". Your code should dynamically detect the dimensionality of the input data. All of data points will have the same number of dimensions.

Output:
A file in the same folder with input file, has the name = input_file_name + ".output", and contains (k+N) lines: (N is number of data points)
- 1st to k-th line: the i-th line represent centroid of i-th cluster (comma-separated)
- (k + 1)-th to (k+N)-th lines: the (k+j)-th line represents cluster index of (j)-th point (cluster index is 0-based)

Example: for test1 file, run command line with parameters: "2 first 50 1e-9 test1", and output file is "test1.output" (see attachment)

Deliverables:
Source code and compiling instructions. Note that source files which fail to compile will receive 0 credit.

Task 2:

Learn the coefficients of the following linear model using Least Squares method given the car price dataset (see attachment):

$$\hat{y} = \beta_0 + \beta_1 x$$

where the independent variable is the odometer and the dependent variable is the price. Include brief intermediate steps of your calculation and explain the learned model in your own words.

Task 3:

Compute the Pearson linear correlation, the coefficient of determination (R^2), SSE of the lease squares linear regression, given training data, attached, with x and y variables (generated with logistic function $y = \frac{e^{8-0.3X}}{1+e^{8-0.3X}}$). Plot the linear regression model in comparison to the training data, and explain your findings.