Score to Rank Columns for Clustering:

A typical problem in clustering very large datasets, is which columns to use for clustering. Think of a dataset with 1 Million rows (N) and 5000 (P) numeric columns which need to be clustered into a predetermined number of clusters (K). It is useless trying to cluster the data using all the columns – typically a smaller subset of columns (S) is enough to build decent clusters.

Can you come up with a numeric score that can be used to rank numeric columns in the data, by their efficiency to cluster the data? The idea is to then pick the top few say 25 (S) columns and cluster the data.

As a part of the function you are supposed to write two functions:

1. Simulating the data:

A function which takes as inputs N, P, S and K and build a dataset which has K clusters- each cluster might have the same for different datapoints. The K clusters are based on first S column in the data and remaining P-S column are just junk columns.

Use to function to build a dataset with N = 1000, P = 10, S = 5, K = 3.

1. Scoring every column in the data

A function which takes as input a dataset and return a numeric score for every column in the dataset. The higher the score, the better it is in building clusters from the dataset. Use the dataset created in (a) and output the score for each column in the dataset. If the score you have devised is good, then the scores of the first S columns in the data will be higher than the remaining P-S columns.