University of Texas at Austin

Homework Assignment 5

$K-$Means clustering.

Please, provide your **complete solutions** to the following problems. Final answers only, even if correct will earn zero points for those problems.

Problem 5.1. (10 points) Provide an example of when **clustering** would be useful in **actuarial practice**.

Problem 5.2. (20 points) As you know from class, in $K-$means clustering, our objective is to minimize

$$\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$$

The above is a difficult formula to compute with, but there is an alternative called the *centroid formula*:

$$2\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

where

$$\bar{x}_{kj}=\frac{1}{|C_k|}\sum_{i\in C_k}x_{ij}$$

is the j^{th} component of the centroid $\bar{x}_k$ of the k^{th} cluster.

Prove that the above formula is correct.

Problem 5.3. (5 points) A $K-$means clustering algorithm based on squared Euclidean distance with $K=2$ produced these clusters:

$$I:(0,1),(1,2),(2,1),(3,2)$$
$$II:(0,3),(1,6),(2,6)$$

What is the value of the objective function, i.e., the function minimized by the clustering algorithm?

Problem 5.4. (15 points) *Source: MAS-II, Spring 2019.*
You have decided to perform $K-$means clustering with $K=2$ on the following data set and have already randomly assigned the clusters as follows:

Observation	x_1	x_2	Initial Cluster
1	5	5	2
2	4	6	2
3	3	0	1
4	5	3	1
5	5	1	2
6	3	6	1
7	2	5	2

- The centroid of the initial cluster 1 is $(3.667, 3)$.
- The centroid of the initial cluster 2 is $(4, 4.25)$.

Calculate the Euclidean distance of Observation 5 from the final centroid of Cluster 2.