Homework 4

CAP 4453

Fall 2021

Assume you have the following data:

Data:

Name	Gender	Age	Height
Carlos	Male	24	6.2
Jhon	Male	2	2.4
Jessica	Female	41	5.1
Keneddy	Female	12	4.4
Peter	Male	19	6.0
Lathia	Female	17	5.3
Nancy	Female	13	4.8
Ana	Female	22	5.8
Jared	Male	27	6.1
Lebron	Male	35	6.6
David	Male	16	5.6
Henry	Male	8	4.6
Claire	Female	8	4.5
Jude	Female	10	4.9
Grace	Female	12	5.0
Mason	Male	14	5.5
Blake	Male	5	4.3

You are going to cluster the data base in two features simultaneously. They are: Age and Height.

Given that the units are not comparable, you are going to use a value to scale one of the axis when you are computing distances.

Define the distance between any two persons I and j as:

$$d_{ij} = \left|Age_i - Age_j\right| + m|Height_i - Height_j|$$

Note that m weights which component (age or height) is more important when you are computing distances.

Assuming a m=10,

1. Plot your data in a cartesian plane.
    a. Put Age in the x-axis, Height in the Y-axis (you can use the factor of m=10 in this axis) to visualize how close they are using the defined distance

2. Define the number of clusters K=4, with seeds on Blake, Lebron, Peter and Grace, and compute:
    a) Assign a group for each of the persons of the table
    b) Compute new center centers

Shows your results in tables

3. Compute and show results in tables:
    a) Assign a group for each of the persons of the table
    b) Compute a new center

4. Compute and show results in tables:
    a) Assign a group for each of the persons of the table
    b) Compute a new center

5. Compute and show results in tables:
    a) Assign a group for each of the persons of the table
    b) Compute a new center

6. Propose a good criterion to stop the algorithm.
7. Could you tell something about the clusters created. Do they make sense?