Asmt 4: Clustering [1]

Turn in (**a pdf**) through Canvas by March 3, 2021

Student Name: - - - - - - -
Student UID: - - - - - - -

Overview

In this assignment you will explore clustering: hierarchical and point-assignment. You will also experiment with high dimensional data.

You will use three data sets for this assignment: You will use four text documents for this assignment:

- Canvas –> Files –> Assignments –> Clustering–> C1.txt

- Canvas –> Files –> Assignments –> Clustering–> C2.txt

- Canvas –> Files –> Assignments –> Clustering–> C3.txt

These data sets all have the following format. Each line is a data point. The lines have either 2 or 5 space separated items. The 2 (or 5 for C3) are the coordinates of the data point. C1 and C2 are in 2 dimensions, and C3 is in 5 dimensions. C1 should have n=20 points, C2 should have n=1221 points, and C3 should have n=900 points. We will always measure distance with Euclidean distance.

It is recommended that you use LaTeX for this assignment (or other option that can properly digitally render math). If you do not, you may lose points if your assignment is difficult to read or hard to follow. Find a sample form in: Canvas –> Files –> Assignments –> Assignment_Latex_Template.zip.

1   Hierarchical Clustering (35 points)

There are many variants of hierarchical clustering; here we explore 3. The key difference is how you measure the distance $d(S_1, S_2)$ between two clusters S_1 and S_2.

Single-Link: measures the shortest link $d(S_1, S_2) = \min\limits_{(s_1,s_2) \in S_1 \times S_2} \|s_1 - s_2\|_2$.

Complete-Link: measures the longest link $d(S_1, S_2) = \max\limits_{(s_1,s_2) \in S_1 \times S_2} \|s_1 - s_2\|_2$.

Mean-Link: measures the distances to the means. First compute $a_1 = \frac{1}{|S_1|} \sum_{s \in S_1} s$ and $a_2 = \frac{1}{|S_2|} \sum_{s \in S_2} s$ then $d(S_1, S_2) = \|a_1 - a_2\|_2$.

A (30 points):   Run all hierarchical clustering variants on data set C1.txt until there are $k = 5$ clusters, and report the results as sets. It may be useful to do this pictorially.

B (5 points):   Which variant did the best job, and which was the easiest to compute (think if the data was much larger)? Explain your answers.

2   Assignment-Based Clustering (65 points)

Assignment-based clustering works by assigning every point $x \in X$ to the closest cluster centers C. Let $\phi_C : X \to C$ be this assignment map so that $\phi_C(x) = \arg\min_{c \in C} \mathbf{d}(x, c)$. All points that map to the same cluster center are in the same cluster.

Two good heuristics for this type of clustering are the Gonzalez (Algorithm 8.2.1 in M4D book) and k-Means++ (Algorithm 8.3.2) algorithms.

A: (15 points)   Run Gonzalez and k-Means++ on data set `C2.txt` for $k = 4$. To avoid too much variation in the results, choose the point in the first line as c_1.

Report the centers and the subsets (as pictures) for Gonzalez. Report:

- the 4-center cost $\max_{x \in X} \mathbf{d}(x, \phi_C(x))$ and
- the 4-means cost $\sqrt{\frac{1}{|X|} \sum_{x \in X} (\mathbf{d}(x, \phi_C(x)))^2}$
  (Note this has been normalized so easy to compare to 4-center cost)

B: (20 points)   For k-Means++, the algorithm is randomized, so you will need to report the variation in this algorithm. Run it several trials (at least 20) and plot the *cumulative density function* of the 4-means cost.

C: (30 points)   Recall that Lloyd's algorithm for k-means clustering starts with a set of k centers C and runs as described in Algorithm 8.3.1 (in M4D).

1: Run Lloyds Algorithm with C initially with points in the $1st, 2nd, 3rd, 4th$ lines (the first 4 points). Report the final subset and the 4-means cost.

2: Run Lloyds Algorithm with C initially as the output of Gonzalez above. Report the final subset and the 4-means cost.

3: Run Lloyds Algorithm with C initially as the output of each run of k-Means++ above. Plot a *cumulative density function* of the 4-means cost. Also report the fraction of the trials that the subsets are the same as the input (where the input is the result of k-Means++).

3   BONUS k-Median Clustering (5 points)

The k-median clustering problem on a data set P is to find a set of k-centers $C = \{c_1, c_2, \ldots, c_k\}$ to minimize $\mathsf{Cost}_1(P, C) = \frac{1}{|P|} \sum_{p \in P} \mathbf{d}(p, \phi_C(p))$. We did not explicitly talk much about this formulation in class, but the techniques to solve it are all typically extensions of approaches we did talk about. This problem will be more open-ended, and will ask you to try various approaches to solve this problem. We will use data set `C3.txt`.

Find a set of 4 centers $C = \{c_1, c_2, c_3, c_4\}$ for the 4-medians problem on dataset `C3.txt`. Report the set of centers, as well as $\mathsf{Cost}_1(P, C)$. The centers should be in the write-up you turn in, but also in a file formatted the same was as the input so we can verify the cost you found. That is each line has 1 center with 5 space separated numbers. 5 is the 5-dimensional coordinates of that center.

Very briefly describe how you found the centers.