Asmt 4: Clustering

Turn in through Canvas by 1:00pm:
Wednesday, February 26
100 points

Overview

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

You will use four data sets for this assignment:

- `http://www.cs.utah.edu/~jeffp/teaching/DM/A4/C1.txt`
- `http://www.cs.utah.edu/~jeffp/teaching/DM/A4/C2.txt`
- `http://www.cs.utah.edu/~jeffp/teaching/DM/A4/C3.txt`
- `http://www.cs.utah.edu/~jeffp/teaching/DM/A4/C4.txt`

Below is the information about data set formats:

- C1/C2.txt: First integer describes the index, the next 2 numbers represent the 'x' and 'y' coordinates of a point respectively.
- C3.txt: Each row in the dataset represents a word's embedding vector.
- C4.txt: First integer describes the index, the next 5 are the coordinates of the data point.

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 this directory: `http://www.cs.utah.edu/~jeffp/teaching/latex/`

Click here for an example template specifically created for this assignment.

1 Hierarchical Clustering (25 points)

There are many variants of hierarchical clustering; here we explore 2. 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 $d(S_1, S_2) = \min_{(s_1,s_2)\in S_1\times S_2} \|s_1 - s_2\|_2$.

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

A (20 points): Run the given two hierarchical clustering variants on data set `C1.txt` until there are $k = 4$ clusters, and report the results as sets. It may be useful to do this pictorially.

Hint: *You can create a scatter plot where data points are colored according to their assigned cluster groups.*

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 on data set `C2.txt` for $k = 3$. To avoid too much variation in the results, choose c_1 as the point with index `1`.

Report:

i For Gonzalez, report the centroids and clusters (make a figure using scatter in matplotlib).

ii the 3-center cost $\max_{x \in X} \mathbf{d}(x, \phi_C(x))$ and

iii the 3-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 3-center cost)

B: (20 points) Now run and k-Means++ on data set `C2.txt` for $k = 3$. Also use c_1 as the point with index `1`. This algorithm is randomized, so you will need to report the variation in this algorithm.

i Run it several trials (at least 20) and plot the *cumulative density function* of the 3-means cost.

ii Report what fraction of the time the subsets are the same as the result from Gonzalez.

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 indexed $\{$`1,2,3`$\}$. Report the final subset and the 3-means cost.

2: Run Lloyds Algorithm with C initially as the output of Gonzalez above. Report the final subset and the 3-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 3-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 Number of Clusters (10 points)

For data sets `C1.txt`, `C2.txt`, and `C3.txt` run any clustering method you want, and estimate how many clusters you think there should be. Explain your reasoning for each.

4 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 $\text{Cost}_1(P, C) = \frac{1}{|P|} \sum_{p \in P} \mathbf{d}(p, \phi_C(p))$. 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 `C4.txt`.

Find a set of 4 centers $C = \{c_1, c_2, c_3, c_4\}$ for the 4-medians problem on dataset `C4.txt`. Report the set of centers, as well as $\text{Cost}_1(P, C)$. The centers should be in the write-up you turn in, but also include a text block in the assignment pdf formatted the same as the input file so we can verify the cost you found; ideally we should be able to use copy+paste from the single pdf you turn in. That is each line has 1 center with 6 tab separated numbers. The first being the index (e.g., 1, 2, 3 or 4), and the next 5 being the 5-dimensional coordinates of that center.

Your score will be based on how small a $\textsf{Cost}_1(P, C)$ you can find. You can get 2 points for reasonable solution. The smallest found score in the class will get all 5 points. Other scores will obtain points in between.

Very briefly describe how you found the centers.