Asmt 3: Clustering

Turn in through Canvas by 1:00pm:
Wednesday, October 1
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/A3/C1.csv`
- `http://www.cs.utah.edu/~jeffp/teaching/DM/A3/C2.csv`
- `http://www.cs.utah.edu/~jeffp/teaching/DM/A3/C3.csv`
- `http://www.cs.utah.edu/~jeffp/teaching/DM/A3/C4.csv`

Below is the information about data set formats all csv files:

- C1/C2.csv: are in 2 dimensions
- C3.csv: Each row in the dataset represents a word's embedding vector in 50 dimensions.
- C4.csv: this is in 5 dimensions

We will always measure the base point-wise distance with Euclidean distance $\mathbf{d}(a, b) = \|a - b\|_2$.

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 $\mathbf{d}(S_1, S_2)$ between two *clusters* S_1 and S_2.

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

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

A (5 points): Plot data in `C1.txt` in a scatter plot (e.g., using `scatter` in `matplotlib`). You may decide there may or may not be useful clusters; we will run some algorithms anyway to see what happens.

B (15 points): Run the above 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 using the drawing; or by writing code to do it.

Report the clusters by encoding the assignment of each point to its cluster via a different shape/color of its glyph.

C (10 points): Which variant in your opinion did the best job? Explain your answer.

2 Assignment-Based Clustering (65 points)

Assignment-based clustering works by assigning every point $x \in X$ to the closest cluster centers S. Let $\phi_S : X \to C$ be this assignment map so that $\phi_S(x) = \arg\min_{s \in S} \mathbf{d}(x, s)$. All points that map to the same cluster center are in the same cluster. For this section **2** we will use data in `C2.txt`

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. Then we will optimize with Lloyd's algorithm (Algorithm 8.3.1 in M4D book).

A: (5 points) Plot the data.

A: (15 points) Run Gonzalez for $k = 3$. To avoid too much variation in the results, choose c_1 to be the first point in the file.

Report:

i For Gonzalez, report the centroids and clusters (use the plot to report so we can grade visually).

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

iii the 3-means cost $\sqrt{\frac{1}{|X|} \sum_{x \in X} (\mathbf{d}(x, \phi_S(x)))^2}$
(Note this has been normalized so easy to compare to 3-center cost)

B: (20 points) Now run and k-Means++ for $k = 3$. Also use s_1 as the first point in the file. This algorithm is randomized, so you will need to report the variation in this algorithm.

i Run it several trials (at least 20; more if you think the plot does not capture the result well) 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: (25 points) Recall that Lloyd's algorithm for k-means clustering starts with a set of k centers S and runs as described in Algorithm 8.3.1 (in M4D).

1: Run Lloyds Algorithm with S initially with the first 3 points in the file. Report the final subset (as a scatter plot) and the 3-means cost.

2: Run Lloyds Algorithm with S initially as the output of Gonzalez above. Report the final subset (as a scatter plot) and the 3-means cost.

3: Run Lloyds Algorithm with S 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. There may or may not be right/wrong answers.
• Explain your reasoning for each dataset.

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 $S = \{s_1, s_2, \ldots, s_k\}$ to minimize $\mathsf{Cost}_1(P, S) = \frac{1}{|P|}\sum_{p \in P} \mathsf{d}(p, \phi_S(p))$. 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 $S = \{s_1, s_2, s_3, s_4\}$ for the 4-medians problem on dataset `C4.txt`. Report the set of centers, as well as $\mathsf{Cost}_1(P, S)$. 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 5 comma separated numbers as the 5-dimensional coordinates of that center.

Your score will be based on how small a $\mathsf{Cost}_1(P, S)$ 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.