Asmt 3: Clustering

Turn in through Canvas by 2:45pm:
Wednesday, March 02
due to a mistake, and grand generosity, I am extending the deadline to
Saturday, March 05 @ 11:59pm
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 three data sets for this assignment:

- `http://www.cs.utah.edu/~jeffp/teaching/cs5140/A3/C1.txt`
- `http://www.cs.utah.edu/~jeffp/teaching/cs5140/A3/C2.txt`
- `http://www.cs.utah.edu/~jeffp/teaching/cs5140/A3/C3.txt`

These data sets all have the following format. Each line is a data point. The lines have either 3 or 6 tab separated items. The first one is an integer describing the index of the points. The next 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=1004 points, and `C3` should have n=1000 points. We will always measure distance with Euclidean distance.

As usual, it is highly recommended that you use LaTeX for this assignment. 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/`

1 Hierarchical Clustering (25 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 $d(S_1, S_2) = \min_{(s_1,s_2)\in S_1\times S_2} \|s_1 - s_2\|_2$.

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

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 (25 points): Run all 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.

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 (50 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 these types of cluster are the Gonzalez (Algorithm 9.4.1) and k-Means++ (Algorithm 10.1.2) algorithms.

A: (20 points) Run Gonzalez and k-Means++ 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 the centers and the subsets (as pictures) for Gonzalez. Report:

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

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 3-means cost. Also report what fraction of the time the subsets are the same as the result from Gonzalez.

B: (20 points) Recall that Lloyd's algorithm for k-means clustering starts with a set of k centers C and runs as described in Algorithm 10.1.1.

- Run Lloyds Algorithm with C initially with points indexed $\{1, 2, 3\}$. Report the final subset and the 3-means cost.
- Run Lloyds Algorithm with C initially as the output of Gonzalez above. Report the final subset and the 3-means cost.
- 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.

C: (10 points) Consider a set of points $S \subset \mathbb{R}^d$ and $\mathbf{d}$ the Euclidean distance. Prove that

$$\arg\min_{p \in \mathbb{R}^d} \sum_{x \in S} (\mathbf{d}(x, p))^2 = \frac{1}{|S|} \sum_{x \in S} x.$$

Here are some suggested steps to follow towards the proof (note there are also other valid ways to prove this, but, for instance, achieving some of these steps will get you partial credit):

1. *First prove the same results for $S \in \mathbb{R}^1$.*
   (a) Expand each term $(\mathbf{d}(x, p))^2 = (x - p)^2 = x^2 + p^2 - 2xp$.
   (b) Take the derivative of each term.
   (c) Add the above terms back together and find where the total derivative is 0.
2. *Show the results for each dimension can be solved independently (use properties of edge lengths in a right triangle – you may want to just redo the above steps using vector calculus).*

3 k-Median Clustering (25 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`.

A: (20 points) 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 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 $\mathsf{Cost}_1(P, C)$ you can find. You can get 15 points for reasonable solution. The smallest found score in the class will get all 20 points. Other scores will obtain points in between.

Very briefly describe how you found the centers.

B: (5 points) Run your algorithm again for the 5-medians problem on dataset `C3.txt`. Report the set of 5 centers and the $\mathsf{Cost}_1(P, C)$. You do not need to turn in a file for these, just write it in your report.

4 BONUS (2 points)

Recall that the k-center problem is to find a set of k centers C to minimize

$$\mathsf{Cost}_0(P, C) = \max_{p \in P} \min_{c \in C} \mathbf{d}(p, c).$$

Let C^* be the optimal choice of k centers for the k-center problem, and let $V^* = \mathsf{Cost}_0(P, C^*)$.

Prove that the Gonzalez algorithm always finds a set of k centers C such that

$$\mathsf{Cost}_0(P, C) \leq 2V^*.$$