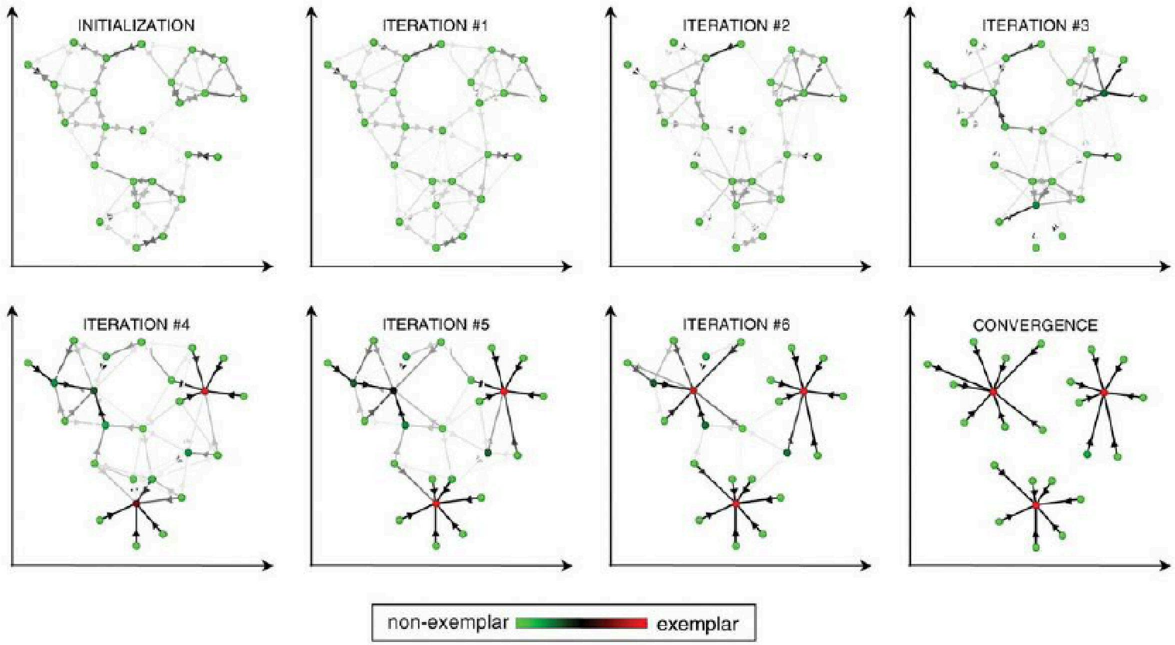
Project # 5

Assignment

The goal of the project is the implementation of the Affinity Propagation clustering algorithm.

Problem description. Affinity Propagation (AP) is a clustering algorithm that belongs to the category of representative-based clustering methods. Unlike some clustering algorithms that require a predetermined number of clusters, AP identifies the centers of the clusters, called representatives, and assigns data to these defined clusters iteratively.

Fig 1: AP example



The AP principle can be described in several steps:

1. AP is based on the similarity matrix, S, which represents the similarity between pairs of points, x_i, x_j. Similarity can be calculated as the negative of the square of the distance, $S(i,j) = -||x_i - x_j||^2$. In other words, for d-dimensional points the calculation is:

$S(i,j) = -\sum_{k=1}^{d}(x_i[k] - x_j[k])^2$.
The values on the diagonal of S have a special role. We should set tem up to a particular value, e.g., the mean of values in S (some strategy see, e.g., here).

2. The AP itself uses two matrices: the responsibility matrix R and the availability matrix A.
   - Responsibility (R): $R(i,k)$ represents the "responsibility" (appropriateness) of point i to represent point k, when compared to other possible representatives of k.
   - Availability (A): $A(i,k)$ represents the "availability" of point k to select point i as its representative. it shows how much point k prefers point i as its representative.

   Both matrices are initialized by all zeros
3. The algorithm iteratively updates the responsibility and availability matrices based on the following rules:
   - $R(i,k) = S(i,k) - \max_{k' \neq k}\{A(i,k') + S(i,k')\}$
   - $A(i,k) = \min\{0, R(k,k) + \sum_{i' \neq i} \max\{0, R(i',k)\}\}$
     $A(k,k) = \sum_{i' \neq k} \max\{0, R(i',k)\}\}$
4. Representatives and cluster assignments are determined based on values in the responsibility and availability matrices. Points with high values in both matrices have a probability of becoming representatives. Consider the criterion matrix, $C = R + A$. In such a matrix, the representative of each row is the point with the largest value in the column.
   For example, in this matrix $C = \begin{bmatrix} 5 & -16 & -15 & -11 & -21 \\ 5 & -15 & -25 & -15 & -25 \\ 5 & -26 & -15 & -17 & -25 \\ -9 & -29 & -30 & -5 & -10 \\ -14 & -34 & -33 & -5 & -10 \end{bmatrix}$
   is the point x_1 (first line) represented by itself (x_1), because the highest value in the first row is in the first column. Point x_2 is also represented by x_1 (the highest value in the second row is again in the first column), the same for x_3. By the same logic, x_4 is a represented by x_4 and x_5 by x_4. So there are two clusters in the data, $\{x_1, x_2, x_3\}$ a $\{x_4, x_5\}$
5. Points 3 - 4 are repeated until the clusters stabilize or after a predetermined number of iterations.

Based on the description of Affinity Propagation, we can formulate the following assignment.

Assignment. Implement an Affinity Propagation solution (e.g., on a part of the dataset MNIST). You can find the description of the dataset here (original) or here (simpler format). The goal of the task is not to learn how to cluster MNIST by real classes, but to test a parallel implementation. So we can ignore the class label for clustering purposes. It can be used to verify how many clusters the algorithm finds and how correctly it assigns objects.

References

1. Brendan J. Frey; Delbert Dueck (2007). "Clustering by passing messages between data points". Science. 315 (5814): 972-976. Bibcode:2007Sci...315..972F. CiteSeerX 10.1.1.121.3145. doi:10.1126/science.1136800. PMID 17218491. S2CID 6502291
2. Thavikulwat, Precha. “Affinity Propagation: A Clustering Algorithm for Computer-Assisted Business Simulations and Experiential Exercises.” Developments in Business Simulation and Experiential Learning 35 (2014): n. pag.
3. https://www.geeksforgeeks.org/affinity-propagation-in-ml-to-find-the-number-of-clusters/