Worksheet 5: Clustering Part 1

Rin Meng 51940633
Kevin Zhang 10811057
Mika Panagsagan 29679552
Priyansh Mathur 84491356

February 4, 2025

You're working on software to organize people's photos. Your algorithm receives as input:

1. A bunch of uncategorized photos.
2. A similarity measure for each pair of photos, where a 0 similarity indicates two photos are nothing like each other; a 1 indicates two photos are exactly the same. All other similarities are in between.
3. The number of categories to group them into.

Your algorithm should create a categorization: the requested number of categories, where a category is a non-empty set of photos. Every photo belongs to some category, and no photo belongs to more than one category. So, a categorization is a partition. We'd like similar photos to be in the same category.

3 Identify similar problems. What are the similarities?

1. Clustering: Clustering is the process of grouping similar items together. In this case, we are trying to group similar photos together.

2. Image Segmentation: Image segmentation is the process of partitioning an image into multiple segments. In this case, we are trying to partition the photos into categories.

3. Hierarchical Clustering: Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. In this case, we are trying to build a hierarchy of categories.

4. K-Means Clustering: K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. In this case, we are trying to partition the photos into k categories.

4 Evaluate brute force

1. The number of potential solutions is $2^{n-1} - 1$ which is exponentially.

2. A good potential solution is when we are trying to optimize using the greedy algorithms, by selecting the best possible solution at each step. Since we are using the greedy algorithm, asympotically, the time complexity is $O(n^2)$, since we are comparing each photo to every other photo.

5 Design a better algorithm.