CPSC 320 Notes, Clustering Completed

AS BEFORE: We're given a complete, weighted, undirected graph $G = (V, E)$ represented as an adjacency list, where the weights are all between 0 and 1 and represent similarities—the higher the more similar—and a desired number $1 \leq k \leq |V|$ of categories.

We define the similarity between two categories C_1 and C_2 to be the maximum similarity between any pair of nodes $p_1 \in C_1$ and $p_2 \in C_2$. We must produce the categorization—partition into k (non-empty) sets—that minimizes the maximum similarity between categories.

Now, we'll prove this greedy approach optimal.

1. Sort a list of the edges E in decreasing order by similarity.
2. Initialize each node as its own category.
3. Initialize the category count to $|V|$.
4. While we have more than k categories:
   (a) Remove the highest similarity edge (u, v) from the list.
   (b) If u and v are not in the same category: Merge u's and v's categories, and reduce the category count by 1.

1 Greedy is at least as good as Optimal

We'll start by noting that any solution to this problem partitions the edges into the "intra-category" edges (those that connect nodes within a category) and the "inter-category" edges (those that cross categories).

1. **Getting to know the terminology:** Imagine we're looking at a categorization produced by our algorithm in which the inter-category edge with maximum similarity is e.

   Can our greedy algorithm's solution have an intra-category edge with **lower** weight than e? Either draw an example in which this can happen, or sketch a proof that it cannot.

2. Give a bound—indicating whether it's an upper- or lower-bound—on the maximum similarity of an arbitrary categorization $\mathcal{C}$ in terms of any one of its inter-category edge weights. That is, I tell you that $\mathcal{C}$ has an inter-category edge with weight s. How much can you tell me so far about $\text{Cost}(\mathcal{C})$?

3. Let $\mathcal{G}$ be the categorization produced by our greedy algorithm, and let $\mathcal{O}$ be an optimal categorization on that instance. Let E' be the set of edges removed from the list during iterations of the While loop. With respect to the greedy solution $\mathcal{G}$, are the edges in E' inter-category? Or intra-category? Or could both types of edges be in E'?

4. Suppose that some edge $e = (p, p', s)$ of E' is inter-category in the optimal solution $\mathcal{O}$. What can we say about $\text{Cost}(\mathcal{G})$ versus $\text{Cost}(\mathcal{O})$?

5. Suppose that all edges of E' are intra-category not only in $\mathcal{G}$, but also in the optimal solution $\mathcal{O}$. Can there be any edges that are inter-category in $\mathcal{G}$ but intra-category in $\mathcal{O}$? (Hint: imagine you have a solution produced by the greedy algorithm. Can you convert any of its inter-category edges to intra-category edges without either making some edges in E' inter-category or making your solution invalid?)

6. Apply the progress made in parts 3 to 5 to conclude that $\mathcal{G}$ must be an optimal solution.

2 Challenge

1. We can also give an "exchange" argument starting: "Consider a greedy solution $\mathcal{G}$ and an optimal solution $\mathcal{O}$. If they're different, then some edge (p, p', s) must be inter-category in $\mathcal{O}$ but intra-category in $\mathcal{G}$. In that case, we can make $\mathcal{O}$ more similar to $\mathcal{G}$ without decreasing $\text{Cost}(\mathcal{G})$ by..."

   Finish the proof. (Warning: even if you select (p, p', s) more carefully, you may not **just** want to move p into p''s category or vice versa.)