CPSC 320 Sample Solution, Clustering Completed

February 8, 2017

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).

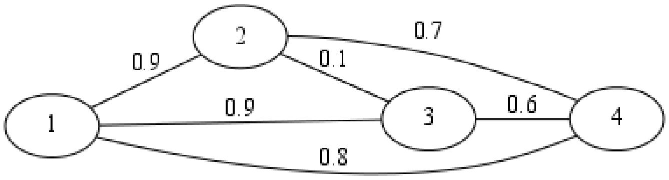
1. **Getting to know the terminology:** Imagine we're looking at a categorization produced by our algorithm in which the inter-category edge with maximum similarity is e.

   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.

   SOLUTION: Can an edge between two nodes in the same category have a similarity lower than the largest-similarity edge that goes across categories?

   Why would we think this could **not** happen? Because we created the categories by merging on edges in order from highest-similarity down. However, if you've tried a few problems, you've noticed that some of the intra-category edges were never merged on. They're intra-category because a series of **other** edges leading between their endpoints all got merged.

   Let's build the smallest instance we can where there's an intra-category edge that was never merged on and then make that edge's weight low. We can get that with 2 desired categories and the graph:



(1, 3) and (1, 2) have the highest similarities and will both be merged on in 4(b). Now, we have two clusters: $\{1, 2, 3\}$ and $\{4\}$. Note that (2, 3) is intra-category, even though its weight is much lower than every inter-category edge, not just the highest-similarity one (which is (1, 4) at 0.8).

2. Give a bound—indicating whether it's an upper- or lower-bound—on the maximum similarity of an arbitrary solution in terms of any one of its inter-category edge weights. (That is, I tell you that the solution has an inter-category edge with weight w. How much can you tell me so far about the solution's overall "goodness"?)

   SOLUTION: The maximum similarity of an arbitrary solution is the maximum similarity of any pair of its categories, which in turn is the maximum similarity of any inter-category edge. Nothing here says that the inter-category edge we're looking at has the **maximum** similarity among all inter-category edges, however.

   So, w is not necessarily actually the maximum similarity because some other edge's weight may be larger. Even if every other inter-category edge has lower weight than w, however, the maximum similarity cannot be any **smaller** than w.

   Therefore the weight of any inter-category edge gives a **lower** bound on the maximum similarity. (I.e., max similarity $\geq w$.)

   (Neither lower- nor upper-bounds need be asymptotic bounds. For example, if you know you passed a class but you don't know the specific grade you got, you have a lower-bound on your grade of 50.)

3. Give a bound on the maximum similarity of a solution produced by the greedy algorithm in terms of the weight of any one of the edges it merged on (in step 4(b)).

   SOLUTION: Since the algorithm inspects edges in order of decreasing similarity and ensures every one is intra-category (either because it's merged on in 4(b) or because it's already intra-category), no inter-category edge can have a higher weight than any edge merged on in 4(b).

   Thus, the weight of any edge merged on in 4(b)—or even just any edge considered in step 4—forms an **upper** bound on the maximum similarity.

4. Consider an optimal solution $\mathcal{O}$ to an instance of the problem. Prove that its intra-category edges cannot be a proper superset of greedy's intra-category edges (i.e., cannot be the same plus at least one more intra-category edge).

   You should **assume** that both $\mathcal{O}$ and the greedy algorithm produce valid solutions, i.e., partitions of V into exactly k subsets. (That's clearly true for $\mathcal{O}$ since it's an optimal solution and not too hard to prove for the greedy approach.)

   SOLUTION: Let's try a proof by contradiction. Imagine $\mathcal{O}$'s set of intra-category edges **is** a proper superset of our algorithm's. Then, everything that was intra-category before still is. (So, none of the existing categories have been "broken up" in any way.) Furthermore, because it's a *proper* superset, at least one of the existing inter-category edges must now be intra-category. That means the two otherwise intact categories on either end of (at least) one inter-category edge have now been merged into a single category. (Note that "ripping out" part of one category and moving it to the other is not an option because of our argument above that no existing categories have been "broken up".)

   Thus, $\mathcal{O}$ has one fewer categories than the greedy solution. But... the greedy solution had the correct number of categories, which means $\mathcal{O}$ has the wrong number, but that's a contradiction with $\mathcal{O}$ being a solution at all.

Therefore our assumption was wrong, and $\mathcal{O}$'s set of intra-category edges is **not** a superset of the greedy solution's.

5. If $\mathcal{O}$'s set of intra-category edges is not a superset of greedy's and it's not the same solution (i.e., the edge sets of the two are not the same), prove that at least one edge that greedy merged on in step 4(b) is an inter-category edge in $\mathcal{O}$.

   SOLUTION: First, if $\mathcal{O}$'s set is neither a superset of the greedy solution's nor equal (because then it would be the same solution), then $\mathcal{O}$'s set has at least one edge that greedy's doesn't and lacks at least one that greedy's has.

   That doesn't yet show that one of the missing (not intra-category, and therefore inter-category) edges is an edge that was merged on in step 4(b), however.

   So, let's imagine (for contradiction) that all the "4(b)" intra-category edges in the greedy solution are also intra-category in $\mathcal{O}$. The trouble here is that these are the only edges that the greedy solution "forces" to be intra-category. Every other intra-category edge in the greedy solution is intra-category **because** the 4(b) edges are. Thus, making all the 4(b) edges intra-category in $\mathcal{O}$ causes the set of intra-category edges in $\mathcal{O}$ to also include every intra-category edge from the greedy solution, which is a contradiction.

   Therefore, $\mathcal{O}$'s intra-category set is missing at least one "4(b)" edge from greedy's.

6. Prove that if $\mathcal{O}$'s set of intra-category edges is neither equal to nor a superset of greedy's, then greedy's solution is optimal. (Remember: optimal doesn't mean "better than all other solutions", just "at least as good as all other solutions".)

   SOLUTION: It's a common but incorrect conclusion at this point to think that we've reached a contradiction that "$\mathcal{O}$ is not optimal". That's similar to what has happened, but isn't quite right.

   We showed that a "4(b)" edge from greedy is inter-category in $\mathcal{O}$. Let's say that edge weighs w. By our reasoning above w is an upper-bound on the maximum similarity of greedy's solution (a 4(b) edge) and a lower-bound on the maximum similarity of $\mathcal{O}$. To make this concrete, I'll call the greedy solution $\mathcal{G}$ and use $M(S)$ to refer to the maximum similarity of a solution S. Then, we've established that $w \geq M(\mathcal{G})$ and $w \leq M(\mathcal{O})$. So, $M(\mathcal{O}) \geq M(\mathcal{G})$.

   But wait! $\mathcal{O}$ is optimal, which means for any solution S, $M(\mathcal{O}) \leq M(S)$. That's true for all solutions; so, it's true for $\mathcal{G}$: $M(\mathcal{O}) \leq M(\mathcal{G})$.

   Is there a contradiction in $M(\mathcal{O}) \geq M(\mathcal{G})$ and $M(\mathcal{O}) \leq M(\mathcal{G})$?

   No, it just means $M(\mathcal{O}) = M(\mathcal{G})$, and the greedy solution is **also** optimal.