Ruolin Li
3176416

CPSC 320: Clustering Completed *

Step 5 Continued: Correctness of Greedy-Clustering

Our goal is to show that our greedy algorithm below for our photo clustering problem produces a categorization that minimizes Cost($\mathcal{C}$). Recall that an instance of the problem is

- n, the number of photos (numbered from 1 to n);
- E, a set of weighted edges, one for each pair of photos, where the weight is a similarity in the range between 0 and 1 (the higher the weight, the more similar the photos); and
- c the desired number of categories, where $1 \leq c \leq n$.

A *categorization* $\mathcal{C}$ is a partition of the photos into c (non-empty) sets, or categories. If $\mathcal{C}$ has more than one category, the *inter-category* similarity between two of its categories C_1 and C_2 is the maximum similarity between any pair of photos $p_1 \in C_1$ and $p_2 \in C_2$. Edges between photos in the same category are called *intra-category* edges. The *cost* of the categorization is the maximum inter-category similarity, taken over all pairs of categories. We'll denote the cost by Cost($\mathcal{C}$). The lower the cost, the better the categorization, so we are trying to find the categorization with minimum cost.

```
function CLUSTERING-GREEDY(n, E, c)
▷ n ≥ 1 is the number of photos
▷ E is a set of edges of the form (p, p', s), where s is the similarity of p and p'
▷ c is the number of categories, 1 ≤ c ≤ n
    create a list of the edges of E, in decreasing order by similarity
    let C be the categorization with each photo in its own category
    Num-C ← n               ▷ Initial number of categories
    while Num-C > c do
        remove the highest-similarity edge (p, p', s) from the list
        if p and p' are in different categories of C then
            "merge" the categories containing p and p'
            Num-C ← Num-C − 1
    return C
```

1. We'll start by getting to know the terminology. Imagine that we're looking at a categorization produced by our algorithm in which e is the inter-category edge with highest similarity. **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.

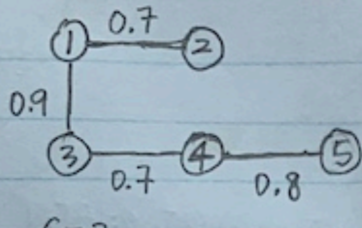
2. **Suppose that I tell you that $\mathcal{C}$ has an inter-category edge e with weight s. Can you find an upper bound or lower bound on Cost($\mathcal{C}$) in terms of s?**

3. On to proof of correctness of our greedy algorithm. Fix an instance of the problem. In what follows, let $\mathcal{C}^G$ be the categorization produced by our greedy algorithm, and let $\mathcal{C}^*$ 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 $\mathcal{C}^G$, 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 $\mathcal{C}^*$. What can we say about Cost($\mathcal{C}^G$) versus Cost($\mathcal{C}^*$)?**

5. **Suppose that all edges of E' are intra-category not only in $\mathcal{C}^G$, but also in the optimal solution $\mathcal{C}^*$. Can $\mathcal{C}^G$ and $\mathcal{C}^*$ be different?**

6. **Apply the progress made in parts 4 to 5 to conclude that $\mathcal{C}^G$ must be an optimal solution.**

4. 

$C = 2$

Greedy Rounds
0 $\{1\}$ $\{2\}$ $\{3\}$ $\{4\}$ $\{5\}$
1 $\{1,3\}$ $\{2\}$ $\{4\}$ $\{5\}$
2 $\{1,3\}$ $\{2\}$ $\{4,5\}$
$C^G \rightarrow$ 3 $\{1,2,3\}$ $\{4,5\}$.

$C^* = \{1,3,4,5\}$ $\{2\}$.

Edge $(1,2,0.7)$ is inter-category in C^*, intra-category in C^G.

$Cost(C^*) \geq S$

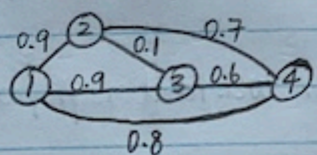
Also $Cost(C^G) \leq S$

Since C^* is the optimal solution, Then $Cost(C^G) = Cost(C^*)$.

5. No. Since greedy makes edges intra-category in a manner completely consistent the optimal C^*, it must eventually produce C^*.

6. When greedy completes, either the scenario described in part 4 or described in part 5, must occur. Either way $cost(C^G) = cost(C^*)$ and so C^G is also optimal.

Step 5 Correctness of Greedy-Clustering
(Con't) 1. inter-category: similarity between two of its categpries.
intra-category: edge between photos in the same category



{1, 2, 3} and {4}. (2, 3) is an intra-category

2. maximum similarity = maximum similarity of any inter-category edge.
maximum similarity cannot be any smaller than w.
w is the lower bound, no upper bound.

3. All edges in E' (p, p', s) will be intra-category, because when removed from the list,
either (i) p, p' are already in the same category
(ii) the algorithm merges the categories containing p and p', making (p, p', s) a intra-category similarity.