Clustering Methods

Exercises 3/7, 6.2.2024

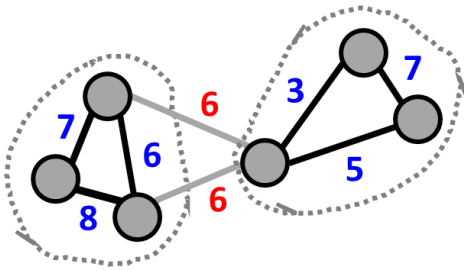
Five tasks required. If done more, each earn one bonus point.

1. Implement random swap algorithm. Last week task #5 compared k-means and random swap theoretically. Perform similar comparison but now based on empirical tests. How many trial swaps can we perform using the same as k-means algorithms requires?

2. Assume our data has only similarity function but the algorithm needs a distance function. How would you solve this issue? Hint: convert the similarities to distances. How would it affect the clustering in practice (if at all)?

3. Consider a black-and-white image of size which is all-white, except a small 10×10 pixel rectangular in the middle which is black. Assume perfect segmentation with two segments: one consisting of the white pixels and one consisting of the black pixels. Calculate *Mumford-Shah cost function* for this segmentation. Is this segmentation optimal?

4. Calculate mean internal weight (MIW) for the following clustering. Can you improve it?



5. Can we use standard k-means for the above data example? If yes, how? If not, why not?

Bonus task:

6. Considering the random swap is quite … random. Can you invent some better way to select *where* the new centroid should be added (instead of random point)? Test it and compare against the standard random swap algorithm.

7. Implement semi-random variant of the deterministic swap by selecting two random clusters for removal, and then actually remove the one with higher larger MSE. Can you implement similar semi-random heuristic for addition?

8. Outline some reasonable stopping criterion for the *Random Swap* algorithm and demonstrate it by an example (artificial or real).