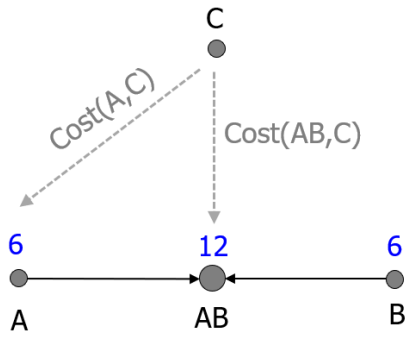
Clustering Methods

Exercises 2/7

1. Suppose we have three clusters in one-dimensional line as shown below. Does the branch-and-bound algorithm consider merge of A and C? If yes, then answer (a) why it happens, (b) does it make any sense, (c) could we avoid it somehow? If it does not happen, explain how the algorithm can avoid it.

A B C

2. Cluster centroid of merged cluster AB is closer to C than the centroid of cluster A before the merge: dist(C, AB) < dist(C, A). However, the merge cost does not decrease: Cost(C, AB) ≥ Cost(C, A). Explain how this is possible.



3. Assume agglomerative clustering variant that takes O(N^2) time. We can speed-up the algorithm by creating initial solution by k-means. Suppose we use $k_0 = 2 \cdot \mathrm{SQRT}(N)$ clusters in k-means and then start the merging process. What is the time complexity of this combined algorithm? What is the input data fed to the agglomerative algorithm?

4. Mumford-Shah -model minimizes both SSE and boundary length as follows. Their importance is weighted by λ parameter. In agglomerative clustering the following merge cost function is used. Can the merge cost be 0? If yes, show example. If not, explain why not.

Increase of SSE

Decrease of boundary length

$$cost_{i,nb} = \frac{n_i n_{nb}}{n_i + n_{nb}} \|c_i - c_{nb}\|^2 - \lambda \cdot 2 \cdot common_{i,nb}$$