Clustering Methods

Exercises 6/7, 29.2.2022

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

1. Apply k-means to any data of your choice using different *k*-values. For example, if there are 3 clusters, apply k-means with *k*=1 to 10. Plot the SSE-values of the results. Can you detect any knee/elbow point from the data?

2. Calculate the 1st and 2nd derivatives of the subsequent SSE-values in the previous task. Can they help to detect the number of clusters? Can you invent data with worst case behavior in this?

3. Implement sub-sampling process in your clustering algorithm (k-means or random swap). You can process the data simply by removing pre-defined number of points, or keeping, 20%, 10%, 5%, and so on. Instead of removing, you could just mark the other points as "obsolete" and exclude from the clustering process. How does the sampling affect the clustering result? How small sample you can have before clustering start to become unstable. What if you add one more cluster (*k*+1)? Does it behave differently?

4. Grid-based clustering can be applied to any data (others than location data). Draft an algorithm how to make simple grid-based clustering where each attribute is divided into 100 grid cells. How many cells there will be in total? How would you decide the number of cells? Does it work if you have 10 attributes? What about 100 attributes?

5. Based on the previous draft, implement grid-based algorithm without any overlap or merging steps, and use it as an initial solution for k-means. Compare the clustering result of the method to that of k-means with (a) random initialization, (b) grid-based initialization.