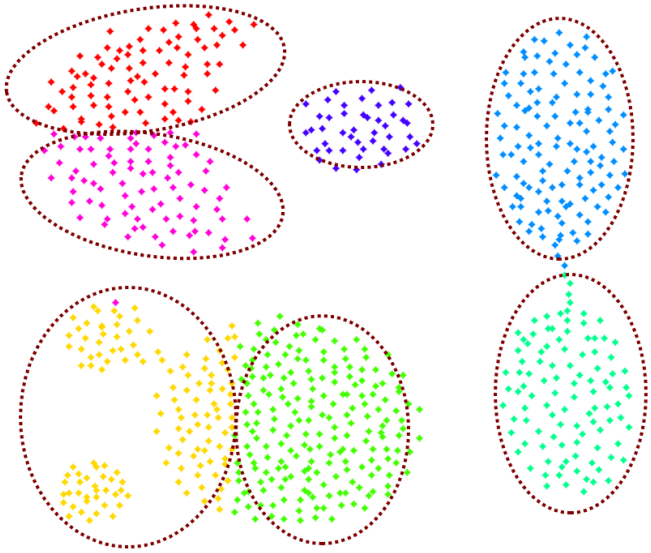
Clustering Methods

Exercises 5/7, 22.2.2022

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

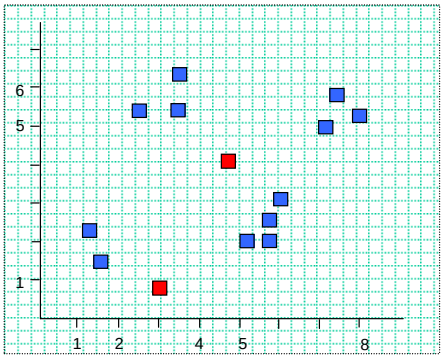
1. Below is one clustering result. Calculate (or just estimate) the number of shared points between the detected and real clusters between all those clusters where the value is >0.



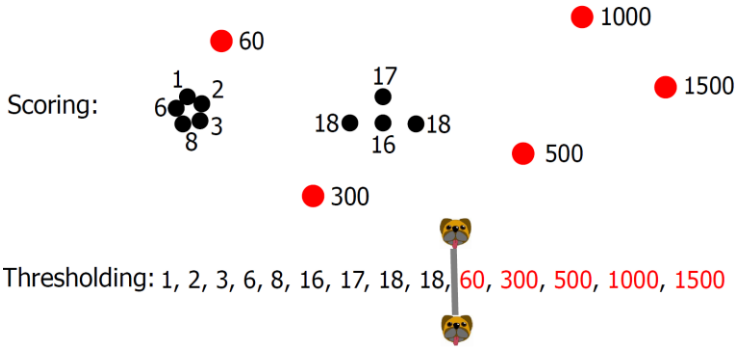
2. Based on the calculations in Task 1, calculate the following. (a) How many shared points there are relative to total number of points? (b) What are the corresponding Jaccard index values?

3. Calculate Centroid Index value for this clustering result.

4. Using k=2, are the following methods are able to detect the red points as outliers: (a) KDIST, (b) ODIN, (c) Mean-shift, (d) Medoid-shift.



5. The following data has 5 outliers that are easy to detect by many methods. The numbers shown are scores from one imaginary outlier detection methods. But how do we know how many outliers there are that we should detect? Threshold the scores by threshold value anywhere between 19 and 59 we would succeed. But how the computer would determine this value? Invent at least one idea for detecting either the threshold or guessing the number of outliers.



Bonus tasks:

6. Implement mean-shift outlier detection function. Apply it as a preprocessing before your k-means algorithm so that all outliers ignored in the centroid calculations step. Do you have any dataset where the result would be different?

7. Implement centroid index.