Exercises for
Pattern Analysis Programming
Lina Felsner, Mathias Seuret, Dalia Rodriguez-Salas
Task 6: May 27 – June 3

Pattern Recognition Lab

K-Means and the Gap-Statistics

In this exercise, we will play with the gap statistics for model selection of k-means clusters. If you like to look at the literature, it can be found in Sec. 14.3.11 of Hastie/Tibshirani/Friedman. If you like to read the original source, you can find the original work by Tibshirani in the literature section of the studon class.

Exercise 1 Implement Tibshirani's gap statistic to automatically select the correct number of components of the clustering. In case that you compare Sec. 14.3.11 of Hastie/Tibshirani/Friedman with the original paper, you will notice that the original paper proposes to optionally align the sampling of the reference density with the principal components of the original data. Although this is certainly a smart move for cases of ill-shaped clusters, you may leave it out for this exercise and just follow Sec. 14.3.11. For each experiment, please produce four plots as shown in Fig. 1. These plots are:

(a) the input clusters, color-coded by their membership to the multivariate Gaussians

(b) the k-means result after selecting k with the gap statistics, color-coded by their membership to the k clusters.

(c) the gap statistics, including the standard deviation of the reference clusters (i.e., the right part of Fig. 14.11 in Hastie/Tibshirani/Friedman)

(d) the decrease of the within-cluster distance of the original data (i.e., the left part of Fig. 14.11 in Hastie/Tibshirani/Friedman)

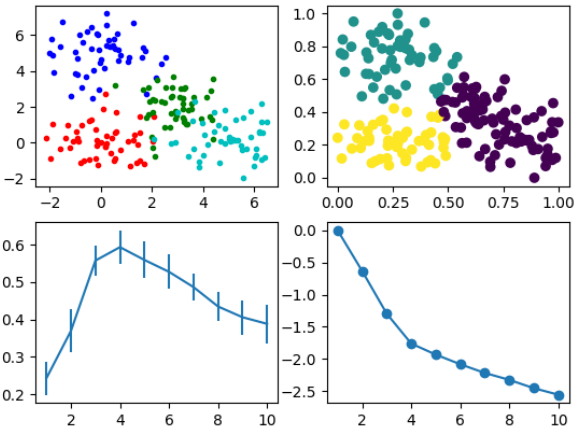

Figure 1: Top left: color-coded input clusters. Top right: color-coded k-means result after selecting k with the gap statistics. Bottom left: the gap statistics, including the standard deviation of the reference clusters. Bottom right: the decrease of the within-cluster distance.

Play with your new method! In which cases is the gap statistic likely to underestimate the number of clusters k? One aspect that complicates the analysis is that both our data generation and also the k-means clustering are randomized. What might be a reasonable approach to nevertheless measure the performance of the gap statistics? What are the issues with our raccoon density? Extra brain teaser: can you think of a case where the gap statistic overestimates the number of clusters? Extra brain teaser: assume that all we want from our raccoon is to cluster some of the bright parts of the original image – what might be a reasonable approach to make this happen?

Please select a density of your choice and post a figure similar to the one in Fig. 1 to the forum and add a short text about it.

Comments:
We ask for only one figure per group. Please also state your group number. Bring your code to the joint meeting on June 3 or June 4 for a potential little extra experiment.