Barbara Hammer, Philip Kenneweg Universität Bielefeld

Modern Data Analysis (WiSe 2023/2024)
5. Sheet

Start: Wednesday, 06.12.2023.
End: The worksheets should be solved using Python, in groups of 2-3 people and will be presented in the Tutorials.
Discussion: Monday, 18.12.2023 in den Tutorien.

Information

The worksheets and necessary toolboxes will be made available in the Lernraum "392246 Modern Data Analysis (V)". Worksheets will usually be released every two weeks on Wednesday, and discussed during the exercises on Monday two weeks later. In order to successfully finish the course, 60% of the available points have to be obtained and each participant has to present his/her results at least once. The Monday in between the release and discussion of the sheet will be used to discuss the implementation of the various algorithms presented in the lecture.

Exercise 1: The Core Set Algorithm

(*10 Points*)

This exercise concerns the Minimal Enclosing Ball (MEB) for the case of vectorial data, without a kernel. Reminder: In the lecture, the Core Set algorithm was introduced, which provides an approximative solution for the MEB of the data X by determining a Core Set with a chosen approximation factor ϵ. For the Core Set, which is usually a very small subset of X, the MEB (with center c and radius R) can be determined very quickly. This MEB is an approximation for the exact MEB of the total data set: all data points are contained within its ϵ-extension with center c and radius $R(1 + \epsilon)$.

(a) *(5 Pts.)* The calculation of the MEB for the current set S is a part of the Core Set algorithm. Write a function that solves the MEB problem (see Theorem 15 in the lecture notes) using an optimization method of your choice. We recommend the optimization package CVXPY. For example code on how to solve a quadratic program, see www.cvxpy.org/examples/basic/quadratic_program.html. Use the Lagrange multipliers α to calculate c and R of the MEB, as shown in the lecture. Test your MEB function on a small set of 10 random vectors in $\mathbb{R}^2$.

(b) *(5 Pts.)* Implement the Core Set method from the lecture notes using your MEB function. Test your algorithm with artificial data composed of 1000 Gaussian random vectors in $\mathbb{R}^2$. How does the size of the Core Set change if the parameter ϵ is chosen to be very small or large? Visualize some intermediate steps: create scatter plots that show the complete data set, the current core set, center and radius of the MEB, and the radius extended by ϵ.