TU Dortmund
Artificial Intelligence / Data Science and Data Engineering
Prof. Dr. Erich Schubert
Prof. Dr. Emmanuel Müller
Daniel Wilmes

1.7.2022
Machine Learning for Complex Data

Summer 2022

Assignment 10: SCHISM and DUSC

Due: Thursday, 7.7.2022

Problem 10-1 SCHISM - Threshold Function 6

(a) Assuming the dimensions of a d-dimensional space are independent and uniformly distributed and discretized into $\xi = 10$ intervals.
Given the threshold function $thresh_{SCHISM}(p)$ and the following values

$$n = 1000, \tau = 0.5, f(p) = p, u = 0.05,$$

find the threshshold for $p = 7$ and $p = 2$.

(b) Derive the variable density threshold of SCHISM:

$$thresh(p) = \frac{E[X_p]}{n} + \sqrt{\frac{1}{2n} ln \frac{1}{\tau}}$$

Hint: A cell contains a cluster if the probability $Pr[X_p \geq n_p]$ is small ($Pr[X_p \geq n_p] \leq \tau$, i.e. the event that a cell contains more than n_p objects is unlikely). SCHISM uses the Chernoff-Hoeffding bound to upper bound this probability. Chernoff-Hoeffding bound:

$$Pr[Y \geq E[Y] + nt] \leq e^{-2nt^2}$$

Problem 10-2 DUSC 8

(a) Compare the density measures of SUBCLU and DUSC:

(i) What is the difference in their density threshold definition?

(ii) Explain the advantage of an unbiased density threshold for subspace clustering.

(iii) Are there also disadvantages? Explain how they possibly affect the result.

(b) Compute which of the clusters detected in 8-2 are redundant according to the DUSC redundancy definition using r=0.5.