University of Augsburg, Institute of Computer Science
Prof. Dr. P. Fischer
J. Kastner, L. Rudenko

SS 2019
07. June 2019
Sheet 6

Analyzing Massive Data Sets

Exercise 1: Expressing Similarity (homework)

The following documents are given:

- D_1 : "the red cat lies on the red chair next to the black cat on the black chair"
- D_2 : "the red chair is between the black chair and the blue chair"
- D_3 : "the black cat is between the red cat and the black chair"

Furthermore the following Query Q is given, specified also as a document:

- Q : "the black cat lies on the red chair between the black chair and the blue chair"

- First of all determine the **word frequencies**. After that calculate the **similarities** of **each document to the Query Q** by using
  a) the **Manhattan** distance.
  b) the **Canberra** distance.

Exercise 2: Locality Sensitive Hashing (homework)

Evaluate the S-curve $1 - (1 - s^r)^b$ for $s = 0.1, 0.2, ..., 0.9$, for the following values of r and b:

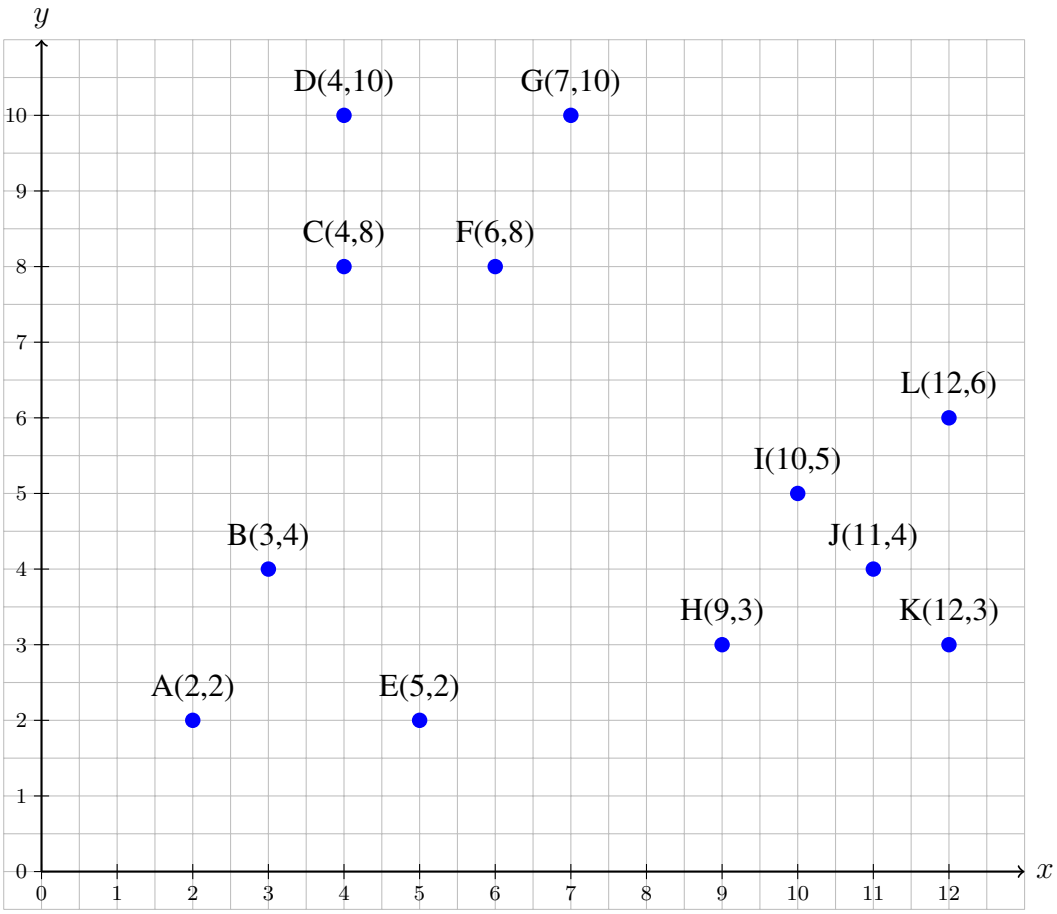
a) $r = 3, b = 10$

b) $r = 6, b = 20$

c) $r = 5, b = 50$

Exercise 3: Hierarchical Clustering (live)

The following points are given in a two-dimensional space:



Initially, each point is in a cluster by itself. Which clusters do you get after performing of an **agglomerative hierarchical** clustering using **Euclidean distance**. The clustering process can be stopped once we found **3 clusters**. Perform the clustering using

a) **centroid**: single "mid"-point

b) **single linkage**: pair with minimum distance

c) **complete linkage**: pair with maximum distance

d) **average distance** among all pairs of nodes in each cluster

to represent a cluster of many points. Are the clusters for each representation the same? Is the order in which the elements are added to the clusters the same?