Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Yifeng Lu

Knowledge Discovery in Databases II

SS 2019

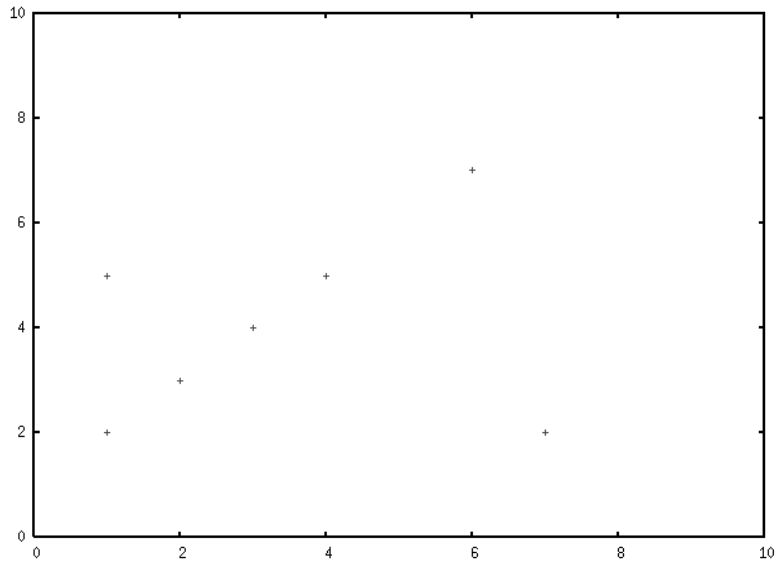
Exercise 6: Correlation Clustering and Stream 1

Exercise 6-1 CASH: Hough-Transform

Consider the data set "`cashDaten.txt`".

(To visualize the data space, use the following gnuplot command:

```
plot [0:10][0:10] ``cashDaten.txt'' title '' )
```



Determine the parameter space associated with this data space, i.e. for each point a parameter function of the following form:

$$f_p(\alpha_1, \ldots, \alpha_{d-1}) \quad = \quad \sum_{i=1}^{d} p_i \cdot \left(\prod_{j=1}^{i-1} \sin(\alpha_j) \right) \cdot \cos(\alpha_i)$$

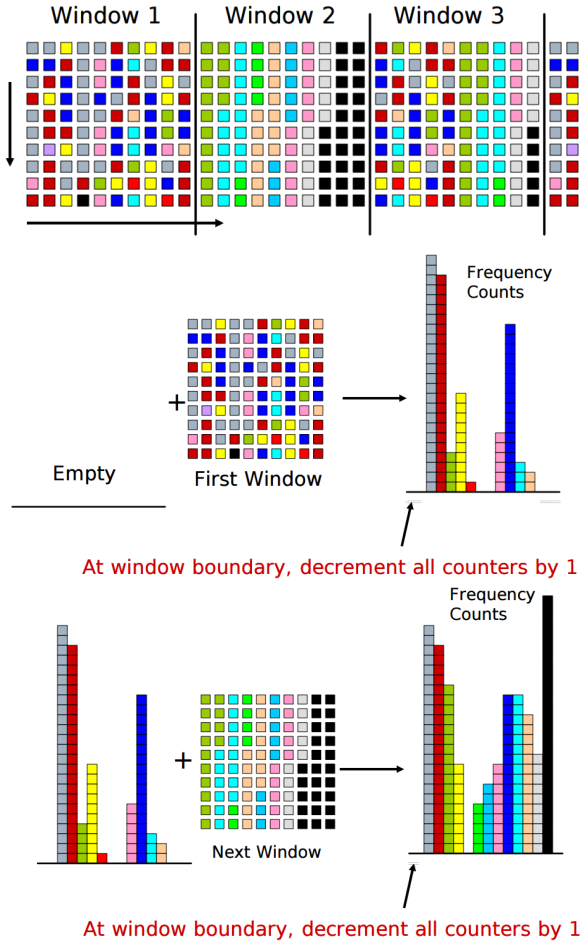
(Note: $\alpha_d = 0$).

Visualize the parameter functions. Where are dense regions located?

Exercise 6-2 Lossy Counting

Before stream clustering, let's take a look at a more fundamental task in stream: count the occurrence of objects in a stream and output objects with a count larger or equal to some given threshold: $minSup \times L$, where L is the length of the stream up to now and $minSup$ is the given threshold (minimum support).

Lossy Counting is one of the basic algorithms that solve this problem. Given the windows size as $w = \frac{1}{\epsilon}$, the lossy counting algorithm works as the follows: cut stream into windows, process one window a time and prune histogram entries with 0 counts at each window boundary. The illustration is given below:



Please prove that

(a) the maximum count error (the maximum difference between the real count and the estimated count) of the lossy counting algorithm is ϵL

(b) the memory consumption, i.e., the number of entries stored in the histogram, is $O(\frac{1}{\epsilon}\log(\epsilon L))$. (Optional)